UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS FÍSICAS E MATEMÁTICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA PURA E APLICADA

Rubén Alex Martínez Muñoz

Novas regras de aprendizagem supervisionada utilizando a estrutura dos números p-ádicos

Florianópolis
2023

Rubén Alex Martínez Muñoz

Novas regras de aprendizagem supervisionada utilizando a estrutura dos números p-ádicos

Tese submetida ao Programa de Pós-Graduação em Matemática Pura e Aplicada da Universidade Federal de Santa Catarina para a obtenção do título de Doutor em Matemática, com Área de Concentração em Análise.
Orientador: Prof. Vladimir G. Pestov, Dr.
Coorientador: Prof. Edson Cilos Vargas Júnior, Dr.

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Martínez Muñoz, Rubén Alex
Novas regras de aprendizagem supervisionada utilizando a estrutura dos números p-ádicos / Rubén Alex Martínez Muñoz ; orientador, Vladimir Pestov, coorientador, Edson Cilos Vargas Júnior, 2023.
189 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências Físicas e Matemáticas, Programa de Pós-Graduação em Matemática Pura e Aplicada, Florianópolis, 2023.

Inclui referências.

1. Matemática Pura e Aplicada. 2. Aprendizagem de máquina supervisionada não paramétrica. 3. Números p ádicos. 4. Redução de dimensionalidade boreliana. 5. Floresta p-ádica Aleatória. I. Pestov, Vladimir. II. Vargas Júnior, Edson Cilos. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Matemática Pura e Aplicada. IV. Título.

Rubén Alex Martínez Muñoz

Novas regras de aprendizagem supervisionada utilizando a estrutura dos números p-ádicos

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Douglas Soares Gonçalves, Dr.
Universidade Federal de Santa Catarina

Prof[a]. Florencia Graciela Leonardi, Dra.
Universidade de São Paulo

Prof. Nacib André Gurgel e Albuquerque, Dr.
Universidade Federal da Paraíba

Prof. Roberto Imbuzeiro Moraes Felinto de Oliveira, Dr.
IMPA

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutor em Matemática, com Área de Concentração em Análise.

ICP-Edu
Documento assinado digitalmente
Daniel Goncalves
Data: 28/12/2023 15:58:13-0300
CPF: ***.022.459-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

ICP-Edu
Documento assinado digitalmente
Vladimir Pestov
Data: 28/12/2023 10:53:20-0300
CPF: ***.231.711-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Vladimir G. Pestov, Dr.
Orientador

Florianópolis, 2023.

À memória do meu amado pai.

AGRADECIMENTOS

Agradeço aos meus pais, Juan Martínez e Elma Muñoz, e à minha família por seu apoio incondicional ao longo da minha vida. À minha filha, Violeta, que com sua existência e seu carinho foi um pilar fundamental nos momentos mais difícies desse processo, e também a todos os meus amigos por ajudarem a tornar esse caminho mais ameno.

Agradeço especialmente ao meu orientador, Prof. Dr. Vladimir Pestov, por me guiar com infinita paciência e compreensão durante o doutorado e compartilhar comigo seus valiosos conhecimentos tanto matemáticos como da vida mesma. Ao Prof. Dr. Edson Cilos Vargas Júnior, agradeço pelas valiosas contribuições no âmbito computacional e no estágio da docência, sua ajuda foi fundamental. Aos demais membros da banca, Prof. Dr. Douglas Soares Gonçalves, Prof[a]. Dra. Florencia Graciela Leonardi, Prof. Dr. Nacib André Gurgel e Albuquerque e o Prof. Dr. Roberto Imbuzeiro Moraes Felinto de Oliveira, por aceitar o convite e pelas valiosas contribuições ao trabalho.

A todos os meus professores do Departamento de Matemática da UFSC, por todas as lições e valiosos ensinamentos oferecidos durante toda a minha estadia no doutorado.

Agradeço aos funcionários do Programa de Pós-Graduação em Matemática Pura e Aplicada da UFSC pela dedicação e apoio durante o período de doutorado.

Finalmente, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo vital suporte financeiro durante os quatro anos do doutorado e especialmente ao povo brasileiro, quem faz possível o importantíssimo trabalho realizado pela CAPES; e pelo qual fui acolhido com muita hospitalidade e carinho.

RESUMO

No presente trabalho, desenvolvemos novos algoritmos de aprendizagem supervisionada (não paramétrica) num domínio boreliano padrão, onde o nosso principal exemplo será o espaço euclidiano de dimensão finita; utilizando para tomar a decisão a *estrutura métrica* dos números p-ádicos em lugar da estrutura métrica dos números reais. Analisando de perto o modelo teórico da aprendizagem estatística, nota-se que a noção de *consistência universal* de um algoritmo depende apenas da estrutura boreliana do domínio e portanto ela é insensível à estrutura métrica ou mesmo topológica no domínio, desde que a estrutura boreliana permaneça intacta. Isto permite, através de uma *injecão boreliana*, reduzir os dados para um outro domínio onde os algoritmos sejam universalmente consistentes e mais eficientes, após o qual o algoritmo de aprendizagem composto com a redução boreliana continua a ser universalmente consistente. Esta ideia foi sugerida no artigo de um dos orientadores [53]. A ideia principal da nossa abordagem é de construir uma redução do domínio original (tipicamente o espaço euclidiano) para um espaço vetorial sobre o corpo $\mathbb{Q}_p$ de números p-ádicos, e experimentar com as novas possibilidades que a estrutura p-ádica fornece. A diferença principal com o espaço euclidiano é que o espaço p-ádico é *não-arquimediano*, o que implica uma propriedade muito particular: duas bolas ou são disjuntas, ou uma está contida na outra. Como consequência, as árvores de busca são muito mais eficientes. Ao mesmo tempo, o espaço p-ádico possui uma estrutura linear rica, permitindo usar as ferramentas da teoria de matrizes, análise funcional, etc. O algoritmo principal, que é universalmente consistente, é definido nos números p-ádicos e combina as caraterísticas de algoritmos clássicos no espaço euclidiano, tais como o classificador k-NN de *k vizinhos mais próximos*, *regra do histograma* e *árvores de decisão*. Enquanto no espaço euclidiano todos esses algoritmos são bastante distintos, a geometria específica do espaço p-ádico permite *combinar* as vantagens deles em um algoritmo de aprendizagem só, e fazendo a composição do algoritmo p-ádico com uma injeção boreliana, obtemos um algoritmo universalmente consistente no espaço euclidiano que de certa maneira *transporta* também esse tipo de comportamento *combinado*. Finalizamos o plano teórico utilizando as novas regras desenvolvidas para construir algoritmos do tipo *ensemble* que preservam a consistência universal. Simples experimentos numéricos sobre *alguns* conjuntos de dados foram realizados, e mesmo não sendo extensivos nem aprofundados, eles mostram que para alguns tipos de dados o desempenho do novo algoritmo foi melhor que o de alguns classificadores tradicionais. Parece que a ideia de pesquisar algoritmos de aprendizagem no espaço p-ádico foi apenas usada antes na área de aprendizagem *não supervisionada*, especificamente na classificação via clusters [11, 12, 25], e na aprendizagem de máquina supervisionada *paramétrica*, especificamente em redes neurais p-ádicas [38]; ideias que são bem diferentes da nossa, pois trabalhamos no contexto da aprendizagem estatística supervisionada *não paramétrica*.

Palavras-chave: Aprendizagem de máquina supervisionada não paramétrica, Números p-ádicos, Redução de dimensionalidade boreliana, Floresta p-ádica Aleatória.

ABSTRACT

In the present work, we develop certain new (nonparametric) supervised learning algorithms in a domain which is a standard Borel space, with the finite-dimensional Euclidean space as the main example. Our algorithms are using the metric structure of p-adic numbers instead of the metric structure of real numbers. Taking a closer look at the theoretical model of statistical learning, it can be noted that the notion of universal consistency of an algorithm depends only on the Borel structure of the domain and therefore it is insensitive to the metric or even topological structure in the domain, as long as the Borel structure remains intact. This allows, through a Borel measurable injection, to reduce the data to another domain where the algorithms are universally consistent yet more efficient, after which the learning algorithm composed with the Borel dimensionality reduction continues to be universally consistent. This idea was suggested in an article by one of the advisors [53]. The main idea of our approach is to construct a reduction of the original domain (typically a Euclidean space) to a vector space over the field $\mathbb{Q}_p$ of p-adic numbers, and to experiment with the new possibilities that the p-adic structure provides. The main difference with Euclidean space is that p-adic space is non-archimedean, which implies a very particular geometric property: two balls are either disjoint, or one is contained in the other. As a consequence, search trees are much more efficient. At the same time, p-adic space has a rich linear structure, allowing to use the tools of matrix theory, functional analysis, etc. Our main learning algorithm, which is universally consistent, is defined in p-adic numbers and combines the features of classical algorithms in Euclidean space, such as the k-NN classifier, histogram rule and decision trees. While in Euclidean space all these algorithms are quite distinct, the specific geometry of the p-adic space allows us to combine their advantages in a single learning algorithm, and by composing the p-adic algorithm with a Borel injection, we obtain a universally consistent algorithm in Euclidean space that in a certain way also transports this type of combined behavior. We finalize the theoretical part by using the new rules to build ensemble-type algorithms that preserve universal consistency. Simple numerical experiments on some data sets were carried out, and although they were not extensive or in-depth, they showed that for some types of data the performance of the new algorithm was better than that of some traditional classifiers. It appears that the idea of constructing learning algorithms in p-adic space has only been used before in the area of unsupervised learning, specifically in classification via clusters [11, 12, 25], and in parametric supervised machine learning, specifically in the so-called p-adic neural networks [38]. Those ideas are very different from ours, as we work in the context of nonparametric supervised statistical learning.

Keywords: Nonparametric supervised machine learning, p-adic numbers, Borel dimensionality reduction, p-adic Random Forest.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

A *aprendizagem de máquina*, também conhecida como *aprendizado de máquina* ou *aprendizado automático*, é uma área do conhecimento interdisciplinar que se encontra na interseção da *matemática*, da *estatística* e das *ciências da computação*, cujo objetivo é desenvolver técnicas que deem às *máquinas* a capacidade de *aprender*. A aprendizagem de máquina, principalmente pode ser dividida nos seguintes quatro grupos.

- **Aprendizagem supervisionada**:

  Formaliza a noção algorítmica de aprendizagem e construção de previsões a partir de *dados rotulados*. Matematicamente, um conjunto de dados rotulados de tamanho $n \in \mathbb{N}$, também chamados de amostras rotuladas, é denotado por:

  $$d_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\Omega \times \mathcal{Y})^n,$$

  onde no contexto mais geral, Ω é um espaço métrico ou um espaço boreliano padrão e $\mathcal{Y}$ é o conjunto de *respostas* ou *rótulos*. A aprendizagem de máquina supervisionada, procura extrair informações e fazer previsões para pontos ainda *não rotulados* a partir de amostras rotuladas.

  Alguns dos principais exemplos de regras de aprendizagem de máquina supervisionada, são: o algoritmo de k-vizinhos mais próximos, que denotamos por k-NN; a árvore de decisão, a floresta aleatória, as máquinas de vetores de suporte (Support Vector Machines) e as redes neurais. Detalhes podem ser encontrados em [24].

- **Aprendizagem não supervisionada**:

  Diferente da aprendizagem supervisionada, a aprendizagem não supervisionada procura extrair informações e padrões de dados *não rotulados*, isto é, que não possuem rótulos de saída predefinidos. O objetivo é descobrir padrões, relacionamentos ou estruturas nos dados. Uma amostra de dados não rotulados é da forma:

  $$\varsigma_n = (x_1, x_2, \ldots, x_n) \in \Omega^n,$$

  onde Ω é como antes.

  Alguns dos principais exemplos de regras de aprendizagem de máquina *não* supervisionada, incluem algoritmos de agrupamento como: k-means, agrupamento hierárquico e o algoritmo DBSCAN. Detalhes podem ser consultados em [45].

- **Aprendizagem semi-supervisionada**:

  Esse tipo de aprendizagem de máquina, busca extrair informações a partir de amostras com *instâncias* tanto rotuladas como não rotuladas. Os algoritmos que implementam esse tipo de aprendizagem, aproveitam os dados rotulados limitados e um conjunto maior de dados não rotulados para melhorar o processo de aprendizagem.

Esta abordagem é particularmente útil quando a aquisição de dados rotulados é cara ou demorada. Técnicas de aprendizagem semi-supervisionada podem ser aplicadas a tarefas de classificação e regressão, permitindo que os modelos façam previsões mais precisas e generalizem melhor em cenários do mundo real. Mais detalhes podem ser encontrados em [1].

- **Aprendizagem por reforço**:

  A aprendizagem por reforço, é um tipo de aprendizagem de máquina inspirado em como os seres humanos aprendem principalmente: por *tentativa* e *erro*. Nesse tipo de aprendizagem, um *agente* interage com um *ambiente* e aprende a tomar decisões *ótimas* que procuram maximizar recompensas cumulativas. O agente recebe um *feedback* por meio de recompensas ou penalidades com base em suas ações e aprende a tomar decisões que levem aos resultados mais favoráveis ao longo do tempo. É comumente usada em robótica, jogos e sistemas autônomos. Detalhes podem ser encontrados em [35].

Independente do tipo de aprendizagem de máquina considerado, na prática quase a totalidade dos algoritmos ou modelos, são desenvolvidos para funcionar com amostras onde $\Omega = \mathbb{R}^d$, $d \in \mathbb{N}$. Dos cursos básicos de análise na reta, é sabido que o *corpo com valor absoluto* $(\mathbb{R}, |\cdot|)$, onde $|\cdot|$ é o valor absoluto usual, é o completamento (como espaço métrico com a métrica induzida pelo valor absoluto) do corpo com valor absoluto $(\mathbb{Q}, |\cdot|)$. Um resultado importante e não tão conhecido, o *Teorema de Ostrowski*, diz que se o que se procura é dotar o corpo $(\mathbb{Q}, +, \cdot)$ com a topologia induzida por um valor absoluto *não trivial*, apenas existem duas possibilidades: que a topologia seja *equivalente* com a induzida pelo valor absoluto usual, ou que seja equivalente com a topologia induzida por algum valor absoluto do tipo p-ádico, $|\cdot|_p$, onde $p \in \mathbb{N}$ é um número *primo* [29]. Do mesmo modo como acontece com o valor absoluto usual, $\mathbb{Q}$ munido de qualquer valor absoluto p-ádico não é completo e o completamento do corpo $(\mathbb{Q}, |\cdot|_p)$ gera o corpo com valor absoluto dos números p-ádicos $(\mathbb{Q}_p, |\cdot|_p)$. Esses números, mesmo representando os únicos completamentos possíveis dos números racionais via valor absoluto (além do $\mathbb{R}$), *praticamente* não são utilizados nas ciências da computação e na aprendizagem de máquina. Alguns trabalhos de aprendizagem de máquina que utilizam números p-ádicos, são na área da aprendizagem *não supervisionada*, especificamente na classificação via clusters [11, 12, 25], e na área da aprendizagem de máquina supervisionada *paramétrica*, especificamente em redes neurais p-ádicas [38].

Uma propriedade do valor absoluto p-ádico, é que ele satisfaz a desigualdade triangular forte:

$$|x + y|_p \leq \max\{|x|_p, |y|_p\}, \ \forall x, y \in \mathbb{Q}_p.$$

Os valores absolutos satisfazendo a desigualdade acima, são chamados de valores absolutos *não-arquimedianos*, e possuem uma propriedade importante e muito particular: duas bolas

nessa topologia, ou são disjuntas ou uma bola está contida na outra; propriedade que claramente não vale no caso euclidiano e que pode ser útil na hora de definir árvores de busca em tais espaços. Além do anterior, o $\mathbb{Q}_p$-espaço vetorial *normado* d-dimensional $(\mathbb{Q}_p^d, \|\cdot\|_p)$, possui uma estrutura linear rica, permitindo usar as ferramentas da teoria de matrizes, análise funcional, etc. de modo similar ao caso real.

1.1 OBJETIVOS E ESTÁGIOS DA PESQUISA

O objetivo principal da pesquisa, é desenvolver novos algoritmos de aprendizagem supervisionada *não paramétrica*[1] no *espaço euclidiano* de dimensão finita, utilizando para tomar a decisão, a particular estrutura métrica dos números p-ádicos em lugar da estrutura métrica dos números reais. Como é uma área de investigação aparentemente inexplorada, a pesquisa começou praticamente desde *zero*, contando apenas com as ferramentas da matemática e da aprendizagem de máquina supervisionada em espaços métricos. Analisando de perto o modelo teórico da aprendizagem estatística, nota-se que a noção de *consistência universal* de um algoritmo depende apenas da estrutura boreliana do domínio e portanto ela é insensível à estrutura métrica ou mesmo topológica no domínio, enquanto a estrutura boreliana permaneça intacta. Isto permite, através de uma *injecão boreliana*, reduzir os dados para um outro domínio onde os algoritmos sejam *universalmente consistentes* e mais eficientes, após a qual o algoritmo de aprendizagem composto com a redução boreliana continua a ser universalmente consistente. Essa ideia foi sugerida no artigo de um dos orientadores no ano 2013 [53]. Com essa ideia em mente, a estratégia utilizada será construir uma redução do domínio original, tipicamente o domínio euclidiano, para um espaço vetorial sobre o corpo dos números p-ádicos $\mathbb{Q}_p$ e experimentar com as novas possibilidades que a estrutura p-ádica fornece.

Nessas condições, os estágios da pesquisa são os seguintes. Primeiro, uma análise do modelo teórico da aprendizagem estatística é feita, incluindo conceitos chaves como o de *consistência universal* e o de *redução boreliana de dimensionalidade*.

Também é feito um estudo do corpo dos números p-ádicos junto a suas principais propriedades topológicas e algébricas, incluindo resultados fundamentais como o já mencionado *Teorema de Ostrowski*. Como são raros os pesquisadores que dominam simultaneamente os princípios da aprendizagem estatística e os conceitos da análise p-ádica, no presente trabalho esses tópicos são analisados em detalhe.

Uma vez adquiridos os conhecimentos necessários, o seguinte passo na investigação foi explorar alternativas de algoritmos de aprendizagem de máquina supervisionada não paramétrica no corpo dos números p-ádicos, conseguindo desenvolver um algoritmo que

[1] Um algoritmo de aprendizagem supervisionada *não paramétrico*, é um tipo de algoritmo que constroi a função que faz as previsões usando apenas a informação contida numa amostra rotulada, sem estimar ou calcular nenhum tipo de parâmetro que defina a função. Um exemplo de algoritmo não paramétrico é o k-NN, e uma *rede neural* é um exemplo de algoritmo *paramétrico*.

pode ser implementado computacionalmente usando árvores de prefixos. Especificamente, é desenvolvido um simples algoritmo de classificação binária que trabalha com amostras rotuladas $d_n \in (\mathbb{Z}_p \times \{0,1\})^n$, onde $\mathbb{Z}_p \subset \mathbb{Q}_p$ é o conjunto de *inteiros p-ádicos*, e que possui as vantagens da simplicidade das árvores de prefixos. Uma vez feito isso, o algoritmo unidimensional é extendido de maneira natural ao espaço métrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$, obtendo a regra de classificaçao p-ádica d-dimensional que é denotada por $\mathcal{L}_{(k,p)}$, onde $1 \leq k \leq n$ é um hiperparâmetro que representa o número de *vizinhos próximos* do ponto $x \in \mathbb{Z}_p$ que desejamos classificar.

Para *transportar* as qualidades do classificador p-ádico ao espaço euclidiano, o seguinte passo foi construir uma função injetora e boreliana $\Phi_p : \mathbb{R}_+^d \to \mathbb{Q}_p^d$ e assim obter, mediante a composição da regra $\mathcal{L}_{(k,p)}$ com a função Φ_p, uma regra de classificação binária no espaço euclidiano, denotada por $\mathcal{L}_k^{\Phi_p} := \mathcal{L}_{(k,p)}^{\Phi_p}$, que usufrui das boas propriedades do algoritmo nos inteiros p-ádicos. Além do anterior e usando novamente o método de redução boreliana de dimensionalidade, considerando a composição do classificador p-ádico com aplicações lineares e bijetoras $T : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$, são obtidas novas regras de classificação no espaço p-ádico, $\mathcal{L}_{(k,p)}^T$; e no espaço euclidiano $\mathcal{L}_k^{T\circ\Phi_p} := \mathcal{L}_{(k,p)}^{T\circ\Phi_p}$.

O seguinte estágio da investigação, foi estudar a consistência do novo classificador desenvolvido em $\mathbb{Z}_p^d$, e para esse fim, foi necessário obter uma *expressão matemática* do algoritmo que possa ser utilizada para analisar a consistência, obtendo como resultado uma *regra de aprendizagem supervisionada* que pode ser aplicada, ao menos teoricamente e independentemente do seu desempenho, em qualquer espaço métrico e que é denotada por ${}^+k$-NN. O resultado do estudo de consistência, revela que a nova regra de aprendizagem é universalmente consistente em certa classe de espaços métricos: os espaços métricos de dimensão σ-finita no sentido de *Nagata* [48], [54].

Junto com o estudo de consistência da nova regra, é feito um estudo de consistência sobre a *aprendizagem ensemble*, especificamente da regra de classificação binária dada pelo voto majoritário de regras de classificação consistentes, resultado que permite definir novas regras de aprendizagem universalmente consistentes baseadas nas regras $\mathcal{L}_{(k,p)}^T$ para o caso p-ádico, e nas regras $\mathcal{L}_k^{T\circ\Phi_p}$ para o caso euclidiano.

Finalmente, experimentos numéricos são realizados para ver em ação a nova regra de aprendizagem $\mathcal{L}_k^{\Phi_p}$, onde é comparado o desempenho da nova regra com o desempenho dos membros de uma família de algoritmos clássicos de classificação binária que são largamente utilizados na prática.

1.2 CONTRIBUIÇÕES DA TESE

As contribuições da tese são no âmbito teórico e prático. A principal contribuição teórica, é a descoberta e subsequente prova de consistência, da nova regra de classificação binária que é denotada por ${}^+k$-NN, a qual é universalmente consistente em espaços métricos

de dimensão σ-finita no sentido de Nagata. O espaço métrico não-arquimediano $(\mathbb{Z}_p^d, \|\cdot\|_p)$ pertence à essa categoria de espaços métricos, e a regra ${}^+k$-NN aplicada em $\mathbb{Z}_p^d$ combinada com a técnica de redução boreliana de dimensionalidade, gera uma família de regras universalmente consistentes em $\mathbb{Z}_p^d$ e em $[0,1]^d$, com $d \in \mathbb{N}$, respectivamente. A segunda contribuição teórica mais importante, é a prova da consistência da regra de classificação binária dada pelo voto majoritário de regras de classificação consistentes; resultado do qual se desprendem novos resultados de consistência de classificadores do tipo ensemble usando regras p-ádicas do tipo $\mathcal{L}_{(k,p)}^T$, ou regras euclidianas do tipo $\mathcal{L}_k^{T\circ\Phi_p}$, com $T : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$, aplicação linear bijetora.

Como contribuição no âmbito prático, o algoritmo de classificação no espaço euclidiano, $\mathcal{L}_k^{\Phi_p}$, resultou ser um algoritmo de fácil implementação computacional e que na maioria dos testes numéricos realizados mostra ter um desempenho similar ao desempenho dos membros da família de classificadores clássicos utilizados na comparação, mostrando ainda, um desempenho nas melhores posições do ranking em algumas situações.

1.3 ORGANIZAÇÃO DA TESE

A organização da tese é a seguinte. No Capítulo 2, de preliminares, primeiro é feita uma revisão dos conceitos fundamentais da teoria da aprendizagem de máquina supervisionada. Em particular, serão vistos o conceito de *consistência universal* e a prova com todos os detalhes de um resultado sobre *redução boreliana de dimensionalidade* de autoria de uns dos orientadores (introduzido no artigo [53]), que será vital nos capítulos subsequentes. Também, será abordado o conceito de *dimensão de uma métrica*, em particular o conceito de *dimensão de Nagata* e a conexão desse conceito, mediante o resultado de Preiss [54], com o *Teorema da diferenciação de Lebesgue-Besicovitch*.

No Capítulo 3, será feito um estudo, com o grau de profundidade necessário para abordar os capítulos subsequentes, do corpo com valor absoluto dos *números p-ádicos*, $(\mathbb{Q}_p, |\cdot|_p)$. Em particular, serão expostos resultados fundamentais como o *Teorema de Ostrowski* e será feita uma construção de $(\mathbb{Q}_p, |\cdot|_p)$ como o completamento de um espaço métrico. Também serão vistas suas principais propriedades algébricas e topológicas, assim como as propriedades topológicas do espaço vetorial p-ádico $d \in \mathbb{N}$ dimensional, $(\mathbb{Q}_p^d, \|\cdot\|_p)$, propriedades que serão vitais na hora de construir o algoritmo de classificação binária do seguinte capítulo.

No Capítulo 4 é feita a construção do algoritmo de classificação binária no espaço métrico $d \in \mathbb{N}$ dimensional $(\mathbb{Z}_p^d, \|\cdot\|_p)$, onde $\mathbb{Z}_p \subset \mathbb{Q}_p$ é o conjunto de *inteiros p-ádicos*. Primeiro é definida uma representação do espaço métrico $(\mathbb{Z}_p, |\cdot|_p)$ como *árvore de prefixos* para logo utilizar essa estrutura para construir um algoritmo de classificação *unidimensional* que depois será estendido ao espaço d-dimensional, com $d \geq 2$. Além do anterior, é feita a construção de uma função *injetora* e *boreliana* entre os espaços métricos $(\mathbb{R}_+^d, \|\cdot\|)$ e $(\mathbb{Q}_p^d, \|\cdot\|_p)$, denotada por Φ_p, função que combinada com o conceito

de *redução boreliana de dimensionalidade*, define um algoritmo de classificação no espaço métrico $([0,1]^d, \|\cdot\|)$, com $[0,1] \subset \mathbb{R}_+$. Seguindo a mesma senda, são definidas famílias de classificadores nos espaços métricos $(\mathbb{Z}_p^d, \|\cdot\|_p)$ e $([0,1]^d, \|\cdot\|)$, como a *composição* de uma transformação linear bijetora sobre $\mathbb{Z}_p^d$ com o algoritmo de classificação em $(\mathbb{Z}_p^d, \|\cdot\|_p)$, e como a composição dessas últimas regras com a função Φ_p, respectivamente.

O Capítulo 5 é dedicado à consistência dos algoritmos desenvolvidos no Capítulo 4. Especificamente, primeiro é obtida uma expressão matemática para a regra de aprendizagem que define o algoritmo em $\mathbb{Z}_p^d$, regra denotada por ${}^+k$-NN, para logo provar um resultado de consistência que implica, em particular, que a regra ${}^+k$-NN aplicada no espaço $(\mathbb{Z}_p^d, \|\cdot\|_p)$ é *universalmente consistente*. Também, são provados resultados sobre a consistência do classificador binário dado pelo voto majoritário de classificadores consistentes, que garantem a consistência de certos modelos do tipo *ensemble*, como por exemplo, o ensemble de classificadores definidos como a composição de uma aplicação linear bijetora com a regra ${}^+k$-NN em $\mathbb{Z}_p^d$, e também os resultantes da composição dessas últimas regras com a função Φ_p.

No Capítulo 6, são realizados experimentos numéricos com o novo classificador $\mathcal{L}_k^{\Phi_p}$. Como a qualidade de um classificador depende do conjunto de dados sobre o qual aplica e da *métrica de avaliação* utilizada, o novo classificador será visto em ação mediante a comparação do seu desempenho com o dos membros de uma família de *modelos clássicos* e largamente utilizados na prática, ao longo de 8 conjuntos de dados disponíveis gratuitamente na web. Especificamente, na seção de preliminares, será visto todo o necessário para realizar a comparação entre modelos: estratégia de comparação, métricas de avaliação e os modelos clássicos que serão os concorrentes, e seguido disso, é aplicada a estratégia de comparação para cada um dos conjuntos de dados considerados.

Finalmente, no Capítulo 7, são feitas as conclusões finais e também são abordadas as limitações do presente trabalho, finalizando com uma lista de possíveis trabalhos futuros.

Como último comentário, é importante apontar que o presente trabalho foi escrito pensando no seu público alvo, composto principalmente por *matemáticos* e *cientistas de dados*, por isso, um pré-requisito para um bom aproveitamento da leitura é que o leitor possua conhecimentos básicos em teoria de probabilidade e análise real.

2 PRELIMINARES

Neste capítulo, vamos introduzir alguns conceitos fundamentais da aprendizagem estatística supervisionada. Começamos com o problema de classificação binária para logo discutir sobre erro de classificação, regras de aprendizagem, consistência e veremos os conceitos anteriores no caso de classificação não binária. Também veremos o conceito de redução de dimensionalidade boreliana, que é uma forma de compor uma regra de aprendizagem com uma função para gerar uma nova regra de apendizagem em um outro espaço, preservando importantes propriedades e por último, vamos lembrar alguns resultados úteis sobre *dimensão de Nagata*. Como comentário final, para acompanhar o conteúdo do texto, o leitor deve ter conhecimentos básicos de *ánalise funcional*, *teoria da medida* e *teoria de probabilidades*. Uma boa e prática revisão desses tópicos pode ser encontrada nos apêndices do livro de um dos orientadores [52].

2.1 CLASSIFICAÇÃO BINÁRIA E ERRO DE CLASSIFICAÇÃO

O objetivo da aprendizagem estatística supervisionada, consiste basicamente em construir uma função entre um conjunto não-vazio Ω, de *observações* e um conjunto $\mathcal{Y}$ de *respostas*, usando a informação dada por uma *amostra rotulada*, d_n de $n \in \mathbb{N}$ observações e suas respectivas respostas

$$d_n = ((x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)) \in (\Omega \times \mathcal{Y})^n,$$

que procure predizer as respostas de novas observações.

O conjunto não-vazio Ω, é chamado de *domínio*. Quando $\mathcal{Y}$ é um conjunto com a *cardinalidade do contínuo*, como por exemplo $\mathcal{Y} = \mathbb{R}$, então o problema de aprendizagem é chamado de *problema de regressão*. Quando $\mathcal{Y}$ é um conjunto *enumerável* ou *discreto*, ele é chamado de *conjunto de rótulos* e o problema de aprendizagem é chamado *problema de classificação* e em particular, se $\mathcal{Y} = \{0, 1\}$, o problema é chamado de *problema de classificação binaria*. Para simplificar a notação ao longo do texto, vamos denotar o conjunto dos primeiros $n \in \mathbb{N}$ números naturais por $[n] := \{1, 2, \ldots, n\}$. Mais detalhes sobre esses tópicos estão nas referências [24, 32, 45, 52, 57].

Exemplo 2.1.1. Considere o problema de determinar se um paciente tem uma certa doença. Suponha que associado a cada paciente, temos um vetor $x \in \Omega = \mathbb{R}^d$, com $d \in \mathbb{N}$, cujas componentes são valores de métricas como: idade, peso, altura, etc. Como rótulos, temos o conjunto $\mathcal{Y} = \{0, 1\}$, onde $y = 1$ é o rótulo para um paciente com a doença e $y = 0$ é o rótulo para um paciente sem a doença. Suponha que dispomos de uma amostra de tamanho $n \in \mathbb{N}$, d_n, de registros de pacientes x_i junto com seu diagnóstico y_i:

$$(x_i, y_i) \in \mathbb{R}^d \times \{0, 1\},\ i \in [n].$$

Então, aqui o problema de classificação binária, é predizer o diagnóstico $y \in \{0, 1\}$ associado ao registro de um novo paciente $x \in \mathbb{R}^d$ usando a informação contida na amostra d_n.

Um trabalho que possui exemplos e detalhes sobre esse tipo de dados, pode ser consultado em [26].

Com relação ao domínio Ω, lembrando que domínios separáveis são espaços topológicos que contém um subconjunto denso enumerável, no paradigma atual da aprendizagem estatística supervisionada não se sabe se a aprendizagem consistente é possível em domínios *não separáveis*. Uma referência sobre essa discussão é o artigo [51]. Também, nessa situação, no melhor dos casos o problema de aprendizagem pode ser resolvido em um subdomínio separável. Além disso, uma outra restrição natural para Ω, é que o espaço seja metrizável com uma métrica completa para poder usar as ferramentas da Análise. Assim, para definir formalmente o problema de classificação binária, vamos exigir que o conjunto Ω possua pelo menos as duas características mencionadas. Antes de definir o tipo de domínio que utilizaremos, o espaço boreliano padrão, vamos lembrar o conceito de σ-álgebra.

Definição 2.1.1 (σ-Álgebra e Espaços Mensuráveis)**.** Seja Ω um conjunto não vazio. Uma *σ-álgebra* (lê-se *sigma álgebra*) de conjuntos de Ω, é uma família $\mathscr{A} \subset 2^{\Omega}$, tal que

(i) $\mathscr{A} \neq \emptyset$,

(ii) se $E \in \mathscr{A}$, então $E^c = \Omega \setminus E \in \mathscr{A}$,

(iii) se $\{E_n\}_{n\in\mathbb{N}} \subset \mathscr{A}$, então $\cup_{n\in\mathbb{N}} E_n \in \mathscr{A}$.

Dada $\mathcal{C} \subset 2^{\Omega}$, denotamos por $\sigma(\mathcal{C})$, a menor σ-álgebra (com relação à ordem dada pela inclusão) de subconjuntos de Ω que contém a família $\mathcal{C}$. Finalmente, o par $(\Omega, \mathscr{A})$, onde $\mathscr{A}$ é uma σ-álgebra de subconjutos de Ω, é chamado de *espaço mensurável*. No caso em que (Ω, τ) é um espaço topológico, denotamos a menor σ-álgebra de subconjuntos de Ω que contém τ, por $\mathscr{B}_{\Omega} = \sigma(\tau)$, que é chamada de *$\sigma$-álgebra de Borel* de Ω.

Observação 2.1.1. Comumente, a condição **(i)** da Definição 2.1.1 é dada como **(i')** $\Omega \in \mathscr{A}$, ou **(i")** $\emptyset \in \mathscr{A}$; mas aqui observamos que é *suficiente* com que a família $\mathscr{A}$ seja não vazia, pois se existe algum $E \in \mathscr{A}$, então por **(ii)** temos $E^c \in \mathscr{A}$ e assim de **(iii)**, obtemos $\Omega = E \cup E^c \in \mathscr{A}$. De forma similar inferimos **(i")**.

Definição 2.1.2 (Espaço boreliano padrão)**.** Um *espaço boreliano padrão*, é um par $(\Omega, \mathscr{B})$, formado por um conjunto não-vazio Ω e uma σ-álgebra de Borel $\mathscr{B}$, que é gerada por uma topologia em Ω que é metrizável, completa e separável.

Pelo exposto, ao longo do texto Ω será um espaço boreliano padrão. Outro conceito que vamos lembrar antes de definir o problema de classificação binária é o de função mensurável.

Definição 2.1.3 (Função Mensurável)**.** Sejam $(\Omega, \mathscr{A}_\Omega)$ e $(\Gamma, \mathscr{A}_\Gamma)$ espaços mensuráveis, e $f : \Omega \to \Gamma$ uma função. Chamamos f de função *$(\mathscr{A}_\Omega, \mathscr{A}_\Gamma)$-mensurável*, se

$$f^{-1}(E) \in \mathscr{A}_\Omega, \ \forall E \in \mathscr{A}_\Gamma.$$

Se f for $(\mathscr{B}_\Omega, \mathscr{B}_\Gamma)$-mensurável, ela é chamada de *Borel mensurável*.

Definição 2.1.4 (O problema de classificação binária)**.** Sejam Ω um espaço boreliano padrão, e $d_n \in (\Omega \times \{0,1\})^n$ uma amostra rotulada. O *problema de classificação binária*, consiste em construir uma função Borel mensurável, $g : \Omega \to \{0,1\}$, chamada de *classificador boreliano*, utilizando a informação da amostra d_n, tal que $y = g(x)$ representa a predição $y \in \{0,1\}$ do rótulo do elemento $x \in \Omega$.

Na situação do Exemplo 2.1.1, podemos construir um classificador calculando, por exemplo, o peso médio dos pacientes na amostra d_n, P_m, que é a média aritmética da variável peso dos registros em d_n, e associar ao registro $x \in \mathbb{R}^d$ de um novo paciente o rótulo 1 se a variável peso em x é maior que P_m e 0 no caso contrário. Como as causas de uma doença podem ser múltiplas, o classificador assim construído pode errar na hora de predizer o rótulo de alguns pacientes, mais ainda se a doença em estudo não tem relação conhecida com o peso do paciente.

Assim, no problema de classificação, devemos preferir os classificadores que apresentam menos erros na hora de predizer, tendo portanto uma melhor *acurácia*. Então, a pergunta que fica é: como medir a capacidade para predizer de um classificador? Como estamos trabalhando com incertezas, o enfoque probabilístico é o mais indicado.

Seja Ω um espaço boreliano padrão. O ponto rotulado $(x, y) \in \Omega \times \{0,1\}$ será modelado por uma variável aleatória $(X, Y) \in \Omega \times \{0,1\}$, onde $X \in \Omega$ representa um ponto do domínio e $Y \in \{0,1\}$ representa o rótulo marcando o ponto. Lembrando que *uma realização*[1] de uma variável aleatória $X \in \Omega$ sobre o espaço de probabilidade (Γ, ν) é qualquer função mensurável $f : \Gamma \to \Omega$ satisfazendo $P[X \in A] = (\nu \circ f^{-1})(A)$, onde $P[X \in A]$ é a probabilidade de X pertencer a A; temos que um ponto $x \in \Omega$ representará uma *instância* da variável aleatória X, se x pertence à imagem de alguma realização de X, isto é, se existem (Γ, ν) e f com f realização de X sobre (Γ, ν), tal que $x = f(\gamma)$, para algum $\gamma \in \Gamma$. Em palavras simples, uma instância é um dos valores possíveis que a variável aleatória pode tomar. Similarmente, $y \in \{0,1\}$ representará uma instância da variável aleatória Y e assim, o ponto (x, y) denotará uma instância da variável aleatória (X, Y). Também, denotaremos por $\widetilde{\mu}$ a medida de probabilidade boreliana em $\Omega \times \{0,1\}$ que é a *lei de probabilidade conjunta* da variável aleatória (X, Y), isto é, $\widetilde{\mu}$ satisfaz

$$P[(X, Y) \in B] = \widetilde{\mu}(B), \ \forall B \in \mathscr{B}_{\Omega \times \{0,1\}},$$

[1] Uma variável aleatória pode ter várias realizações, onde todas elas diferem apenas em um conjunto de medida nula.

onde $\mathscr{B}_{\Omega\times\{0,1\}}$ é a σ-álgebra de Borel no espaço $\Omega \times \{0,1\}$ e abreviamos que $\widetilde{\mu}$ seja a lei de probabilidade de (X,Y) por: $(X,Y) \sim \widetilde{\mu}$.

Definição 2.1.5 (Erro de classificação)**.** Sejam $\widetilde{\mu}$ uma medida de probabilidade boreliana no espaço $\Omega \times \{0,1\}$ que é a lei de probabilidade dos dados rotulados $(X,Y) \in \Omega \times \{0,1\}$ e $g : \Omega \to \{0,1\}$ um classificador boreliano. Definimos o *erro de classificacão* de g com relação a $\widetilde{\mu}$, $\text{erro}_{\widetilde{\mu}}(g)$, como a probabilidade de g errar na predição dos rótulos

$$\begin{aligned}\text{erro}_{\widetilde{\mu}}(g) :&= P[g(X) \neq Y] \\ &= P[(X,Y) \in \{(x,y) \in \Omega \times \{0,1\} : g(x) \neq y\}] \\ &= \widetilde{\mu}(\{(x,y) \in \Omega \times \{0,1\} : g(x) \neq y\}).\end{aligned}$$

O objetivo principal de um classificador é predizer o rótulo para um novo dado $x \in \Omega$ e o *erro de classificacão* dá a probabilidade do classificador errar. Assim, no problema de classificação, vamos preferir os classificadores que tenham um erro de classificação o mais baixo possível.

Dado $(X,Y) \sim \widetilde{\mu}$, podemos esperar encontrar um classificador boreliano $g : \Omega \to \{0,1\}$, tal que $\text{erro}_{\widetilde{\mu}}(g) = 0$?. Em geral, como vamos ver, não. Mas, como para qualquer classificador boreliano g, definido em Ω, temos $0 \leq \text{erro}_{\widetilde{\mu}}(g) \leq 1$ e como o conjunto de classificadores borelianos em Ω é claramente não vazio (contém pelo menos o classificador $g_0 : \Omega \to \{0,1\}$ definido por $x \mapsto 0, \forall x \in \Omega$), concluímos que o ínfimo do erro de classificação sobre todos os classificadores borelianos em Ω, $\inf_g \text{erro}_{\widetilde{\mu}}(g)$, existe. Chamaremos este ínfimo de *erro de Bayes*.

Definição 2.1.6 (Erro de Bayes)**.** Definimos o *erro de Bayes* com relação à medida de probabilidade boreliana $\widetilde{\mu}$ sobre $\Omega \times \{0,1\}$, que denotamos por $\ell^*(\widetilde{\mu})$; como o ínfimo dos erros de classificação sobre todos os classificadores borelianos possíveis sobre Ω, ou seja,

$$\ell^*(\widetilde{\mu}) := \inf_{g \in \mathfrak{B}(\Omega,\{0,1\})} \text{erro}_{\widetilde{\mu}}(g),$$

onde $\mathfrak{B}(\Omega,\{0,1\})$ é o conjunto de funções borelianas de Ω em $\{0,1\}$. Em outras palavras, o erro de Bayes, $\ell^*(\widetilde{\mu})$, é o menor erro que podemos esperar atingir ao tentar predizer os rótulos dos elementos de Ω seguindo a distribuição $\widetilde{\mu}$.

O ínfimo na definição acima, é na verdade um mínimo, mas antes de constatar isso veremos outra forma de descrever a distribuição de probabilidade de (X,Y) mediante o par (μ,η), onde μ é a distribuição dos dados *não rotulados* e η é chamada de *função de regressão* de $\widetilde{\mu}$. Para obter essa representação, precisamos do teorema de Radon-Nikodým. Começamos lembrando as seguintes definições.

Definição 2.1.7. Sejam $\mu : \mathscr{A} \to [0,+\infty]$ e $\nu : \mathscr{A} \to [0,+\infty]$, medidas definidas na mesma σ-álgebra $\mathscr{A}$. Dizemos que ν é *absolutamente contínua* com relação a μ, conceito que denotaremos por $\nu \ll \mu$, se para todo $E \in \mathscr{A}$ tal que $\mu(E) = 0$, temos $\nu(E) = 0$.

Definição 2.1.8. (Noção q.t.p) Dizemos que uma propriedade $\mathscr{P}(x)$, que depende do elemento $x \in \Omega$, onde $(\Omega, \mathscr{A}, \mu)$ é um espaço de medida, é verdadeira em **q**uase **t**oda **p**arte segundo a medida μ, ou para μ **q**uase **t**odo $x \in \Omega$, se existe um conjunto mensurável $N \in \mathscr{A}$, tal que $\mu(N) = 0$ e $\mathscr{P}(x)$ seja verdadeira para todo $x \in \Omega \setminus N$. Neste caso, escrevemos

$$\mathscr{P}\ ,\ \mu\text{-q.t.p}$$

ou

$$\mathscr{P}(x),\ \text{ para }\ \mu\text{-q.t }\ x \in \Omega.$$

O teorema de Radon-Nikodým é um resultado que permite escrever certas medidas como uma integral e uma das hipóteses do teorema considera duas medidas que satisfazem a propriedade $\nu \ll \mu$. Por outro lado, se duas medidas ν e μ satisfazem

$$\nu(E) \leq \mu(E),$$

para todo E conjunto mensurável, então elas necessariamente satisfazem $\nu \ll \mu$. Utilizando a hipótese acima em lugar de $\nu \ll \mu$, a prova do teorema de Radon-Nikodým fica consideravelmente mais simples e por isso aqui usaremos o caso especial do teorema. Uma prova do caso especial do resultado, pode ser encontrada no apêndice H de [52] e detalhes sobre o resultado original podem ser consultados em [9, 28, 31, 55].

Teorema 2.1.1 (Radon-Nikodým (caso especial))**.** *Sejam μ e ν duas medidas finitas sobre um espaço boreliano padrão, $(\Omega, \mathscr{B}_\Omega)$, tais que*

$$\nu(B) \leq \mu(B),\ \ \forall B \in \mathscr{B}_\Omega.$$

Então, existe uma função boreliana $f : \Omega \to [0,1]$ tal que

$$\nu(B) = \int_B f(x)\, d\mu(x),\ \ \forall B \in \mathscr{B}_\Omega. \tag{1}$$

A função f é chamada de derivada de Radon-Nikodým de ν com relação a μ e escrevemos

$$f = \frac{d\nu}{d\mu}.$$

A derivada de Radon-Nikodým é única μ-q.t.p, isto é, se $f_1, f_2 : \Omega \to [0,1]$ são duas funções satisfazendo (1), então $f_1 = f_2$, μ-q.t.p.

Agora, voltando sobre o assunto do par (μ, η), que foi apontado depois da Definição 2.1.6, primeiro observamos que a lei de probabilidade dos dados não rotulados, μ, é dada pela imagem direta da lei $\widetilde{\mu}$ por π_Ω, onde π_Ω é a projeção canônica de $\Omega \times \{0,1\}$ sobre Ω, isto é, $\pi_\Omega(x,y) = x, \forall (x,y) \in \Omega \times \{0,1\}$. Logo, para cada $B \in \mathscr{B}_\Omega$, temos

$$\begin{aligned}
\mu(B) &= \widetilde{\mu}(\pi_\Omega^{-1}(B)) \\
&= \widetilde{\mu}(B \times \{0,1\}) \\
&= \widetilde{\mu}(B \times \{0\}) + \widetilde{\mu}(B \times \{1\}) \\
&= \mu_0(B) + \mu_1(B),
\end{aligned}$$

onde μ_0 e μ_1 são medidas finitas em Ω, não necessariamente de probabilidade, definidas para cada $B \in \mathscr{B}_\Omega$ por $\mu_0(B) = \widetilde{\mu}(B \times \{0\})$ e $\mu_1(B) = \widetilde{\mu}(B \times \{1\})$. Como $\mu_1(B) \leq \mu(B)$, $\forall B \in \mathscr{B}_\Omega$, pelo teorema de Radon-Nikodým 2.1.1, existe uma derivada (de Radon–Nikodým) de μ_1 com relação a μ, isto é, existe uma função boreliana $\eta : \Omega \to [0,1]$, definida para cada $x \in \Omega$, por

$$\eta(x) = \frac{d\mu_1}{d\mu}(x),$$

tal que para todo $B \in \mathscr{B}_\Omega$, satisfaz

$$\mu_1(B) = \int_B \eta(x)\, d\mu(x),$$

de onde inferimos que

$$\mu_0(B) = \int_B (1 - \eta(x))\, d\mu(x).$$

Na linguagem probabilística, $\eta(x)$ é a *probabilidade condicional* de ser rotulado 1 dado que $X = x$:

$$\eta(x) = P[Y = 1 | X = x].$$

Por outro lado, dado $D \in \mathscr{B}_{\Omega \times \{0,1\}}$ qualquer, podemos escrever

$$D = (D_0 \times \{0\}) \cup (D_1 \times \{1\}),$$

onde $D_i = \{x \in \Omega : (x, i) \in D\}$, $i = 0, 1$. Portanto

$$\begin{aligned} P[(X,Y) \in D] &= P[(X,Y) \in D_0 \times \{0\}] + P[(X,Y) \in D_1 \times \{1\}] \\ &= \widetilde{\mu}(D_0 \times \{0\}) + \widetilde{\mu}(D_1 \times \{1\}) \\ &= \mu_0(D_0) + \mu_1(D_1) \\ &= \int_{D_0} (1 - \eta(x))\, d\mu(x) + \int_{D_1} \eta(x)\, d\mu(x), \end{aligned}$$

recuperando assim a distribuição de (X, Y) a partir do par (μ, η) e portanto, indistintamente utilizaremos $\widetilde{\mu}$ ou (μ, η) para denotar a distribuição de (X, Y).

Agora, com a ajuda da função de regressão, veremos algumas propriedades do erro de classificação. Os resultados a seguir, podem ser consultados em [52].

Proposição 2.1.1. *Para $\widetilde{\mu}$ descrita pelo par (μ, η), e $g : \Omega \to \{0,1\}$ um classificador boreliano qualquer, temos*

$$erro_{\widetilde{\mu}}(g) = \|g - \eta\|_{L^1(\mu)} = \int_\Omega |g(x) - \eta(x)|\, d\mu(x).$$

Demonstração.

$$\begin{aligned}
\text{erro}_{\widetilde{\mu}}(g) &= P[g(X) \neq Y] \\
&= \int_{\Omega} (P[g(X)=1, Y=0|X=x] + P[g(X)=0, Y=1|X=x])\, d\mu(x) \\
&= \int_{\Omega} \left(\mathbb{I}_{\{g(X)=1\}}(x) P[Y=0|X=x] + \mathbb{I}_{\{g(X)=0\}}(x) P[Y=1|X=x] \right) d\mu(x) \\
&= \int_{\{g(X)=1\}} P[Y=0|X=x]\, d\mu(x) + \int_{\{g(X)=0\}} P[Y=1|X=x]\, d\mu(x) \\
&= \int_{\{g(X)=1\}} (1-\eta(x))\, d\mu(x) + \int_{\{g(X)=0\}} \eta(x)\, d\mu(x) \\
&= \int_{\{g(X)=1\}} |g(x)-\eta(x)|\, d\mu(x) + \int_{\{g(X)=0\}} |g(x)-\eta(x)|\, d\mu(x) \\
&= \int_{\Omega} |g(x)-\eta(x)|\, d\mu(x).
\end{aligned}$$

Aqui utilizamos a notação simplificada usual: $\{g(X)=y\} = \{x \in \Omega : g(x)=y\}$ para $y \in \{0,1\}$ e $\mathbb{I}_{\{g(X)=y\}}$ é a função indicadora do conjunto $\{g(X)=y\}$. ■

Corolário 2.1.1.1. *Nas condições da Proposição 2.1.1, seja* $\eta_{1/2} = \{\eta(X) = 1/2\}$. *Então*

$$erro_{\widetilde{\mu}}(g) = \int_{\Omega \setminus \eta_{1/2}} |g(x)-\eta(x)|\, d\mu(x) + \frac{1}{2}\mu(\eta_{1/2}).$$

Demonstração. É suficiente notar que para qualquer classificador g, a função $c(x) := |g(x)-\eta(x)|$ é constante no conjunto $\eta_{1/2}$:

$$c|_{\eta_{1/2}} = |g-\eta||_{\eta_{1/2}} \equiv \frac{1}{2}.$$

■

Graças ao corolário acima, obtemos de forma imediata o seguinte resultado.

Corolário 2.1.1.2. *Sejam* g *e* g' *dois classificadores borelianos em* Ω, *satisfazendo*

$$\forall x \in \Omega : \eta(x) \neq \frac{1}{2} \Rightarrow g(x) = g'(x).$$

Então,

$$erro_{\widetilde{\mu}}(g) = erro_{\widetilde{\mu}}(g').$$

Agora, definimos o conceito de classificador de Bayes.

Definição 2.1.9 (Classificador de Bayes)**.** Suponha que $(X,Y) \sim \widetilde{\mu}$, onde $\widetilde{\mu}$ é uma medida de probabilidade boreliana em $\Omega \times \{0,1\}$ descrita pelo par (μ, η). O classificador $g^*_{\widetilde{\mu}}$ será *um classificador de Bayes* para a lei $\widetilde{\mu}$, se satisfaz:

$$g^*_{\widetilde{\mu}}(x) = \begin{cases} 1, & \eta(x) > \frac{1}{2} \\ 0, & \eta(x) < \frac{1}{2}. \end{cases}$$

Quando estamos trabalhando apenas com uma lei de probabilidade de dados rotulados e portanto o contexto é claro, denotaremos um classificador de Bayes simplesmente por g^*.

Observação 2.1.2. Um classificador de Bayes, $g^*_{\widetilde{\mu}}$, depende da lei $\widetilde{\mu}$ de (X, Y), que é desconhecida, assim $g^*_{\widetilde{\mu}}$ também é desconhecido, porém, assumimos a existência da medida $\widetilde{\mu}$ para fazer possível o estudo teórico. Por outro lado, do Corolário 2.1.1.1, vemos que para calcular o erro de classificação utilizamos apenas os pontos do conjunto $\Omega \setminus \eta_{1/2}$, assim, para os "olhos" do erro de classificação o comportamento de um classificador no conjunto $\eta_{1/2}$ é irrelevante. Em particular, um classificador de Bayes, $g^*_{\widetilde{\mu}}$, pode tomar qualquer valor, 0 ou 1 nos pontos do conjunto $\eta_{1/2}$, o importante é que $g^*_{\widetilde{\mu}}$ tome o valor 1 no conjunto $\{\eta(X) > 1/2\}$ e tome o valor 0 em $\{\eta(X) < 1/2\}$, por isso, preferimos falar de *um classificador de Bayes* em lugar de *o classificador de Bayes*; e aqui faremos dito tratamento para esse importante conceito, por ser mais exato do que o tratamento feito na maioria dos textos.

Agora podemos mostrar que $\inf_g \text{erro}_{\widetilde{\mu}}(g) = \min_g \text{erro}_{\widetilde{\mu}}(g)$, e que esse mínimo é atingido por qualquer classificador de Bayes para a lei $\widetilde{\mu}$.

Teorema 2.1.2 (Teorema 2.1 de [24])**.** *Sejam g um classificador boreliano definido em Ω e g^* um classificador de Bayes para a lei dos dados rotulados $\widetilde{\mu}$, então*

$$P[g(X) \neq Y] \geq P[g^*(X) \neq Y],$$

ou

$$erro_{\widetilde{\mu}}(g) \geq erro_{\widetilde{\mu}}(g^*),$$

isto é, $erro_{\widetilde{\mu}}(g^*) = \ell^*(\widetilde{\mu})$.

Demonstração. Para $x \in \Omega$ e lembrando que $\eta(x) = P[Y = 1 | X = x]$, dado um classificador boreliano g sobre Ω, qualquer, podemos escrever:

$$\begin{aligned}
&P[g(X) = Y | X = x] \\
&\quad = P[g(X) = 0, Y = 0 | X = x] + P[g(X) = 1, Y = 1 | X = x] \\
&\quad = \mathbb{I}_{\{g(X)=0\}}(x) P[Y = 0 | X = x] + \mathbb{I}_{\{g(X)=1\}}(x) P[Y = 1 | X = x] \\
&\quad = \mathbb{I}_{\{g(X)=0\}}(x)(1 - \eta(x)) + \mathbb{I}_{\{g(X)=1\}}(x)\eta(x).
\end{aligned}$$

Da igualdade acima, observamos que se $x \in \eta_{1/2}$, então $P[g(X) = Y | X = x] = \frac{1}{2}$, e em particular, para qualquer classificador de Bayes g^*, temos $P[g^*(X) = Y | X = x] = \frac{1}{2}$. Assim, $P[g(X) \neq Y | X = x] - P[g^*(X) \neq Y | X = x] = 0$ nos pontos $x \in \eta_{1/2}$. Procedendo como antes, para g^* podemos escrever:

$$P[g^*(X) = Y | X = x] = \mathbb{I}_{\{g^*(X)=0\}}(x)(1 - \eta(x)) + \mathbb{I}_{\{g^*(X)=1\}}(x)\eta(x).$$

Logo, para todo $x \in \Omega \setminus \eta_{1/2}$, temos

$$\begin{aligned}
&P[g(X) \neq Y \,|\, X = x] - P[g^*(X) \neq Y \,|\, X = x] \\
&\quad = \eta(x) \left(\mathbb{I}_{\{g^*(X)=1\}}(x) - \mathbb{I}_{\{g(X)=1\}}(x) \right) + (1 - \eta(x)) \left(\mathbb{I}_{\{g^*(X)=0\}}(x) - \mathbb{I}_{\{g(X)=0\}}(x) \right) \\
&\quad = (2\eta(x) - 1) \left(\mathbb{I}_{\{g^*(X)=1\}}(x) - \mathbb{I}_{\{g(X)=1\}}(x) \right) \\
&\quad = |2\eta(x) - 1| \mathbb{I}_{\{g(X) \neq g^*(X)\}}(x) \qquad (2) \\
&\quad = |2\eta(x) - 1| |g(x) - g^*(x)| \\
&\quad \geq 0,
\end{aligned}$$

pois na penúltima igualdade, se $g^*(x) = g(x)$, então

$$\mathbb{I}_{\{g^*(X)=1\}}(x) - \mathbb{I}_{\{g(X)=1\}}(x) = \mathbb{I}_{\{g(X) \neq g^*(X)\}}(x) = 0$$

e se $g^*(x) \neq g(x)$ e $\eta(x) > 1/2$, temos

$$\mathbb{I}_{\{g^*(X)=1\}}(x) - \mathbb{I}_{\{g(X)=1\}}(x) = \mathbb{I}_{\{g(X) \neq g^*(X)\}}(x) = 1,$$

portanto o produto fica

$$\begin{aligned}
|2\eta(x) - 1| \mathbb{I}_{\{g(X) \neq g^*(X)\}}(x) &= 2\eta(x) - 1 \\
&= (2\eta(x) - 1) \left(\mathbb{I}_{\{g^*(X)=1\}}(x) - \mathbb{I}_{\{g(X)=1\}}(x) \right).
\end{aligned}$$

Por último, se $g^*(x) \neq g(x)$ e $\eta(x) < 1/2$, então $\mathbb{I}_{\{g^*(X)=1\}}(x) - \mathbb{I}_{\{g(X)=1\}}(x) = -1$, $\mathbb{I}_{\{g(X) \neq g^*(X)\}}(x) = 1$ e portanto o produto fica

$$\begin{aligned}
|2\eta(x) - 1| \mathbb{I}_{\{g(X) \neq g^*(X)\}}(x) &= 1 - 2\eta(x) \\
&= (2\eta(x) - 1) \left(\mathbb{I}_{\{g^*(X)=1\}}(x) - \mathbb{I}_{\{g(X)=1\}}(x) \right).
\end{aligned}$$

Finalmente

$$\begin{aligned}
P[g(X) \neq Y] - P[g^*(X) \neq Y] &= \int_\Omega \left(P[g(X) \neq Y \,|\, X = x] - P[g^*(X) \neq Y \,|\, X = x] \right) d\mu(x) \\
&= \int_{\Omega \setminus \eta_{1/2}} |2\eta(x) - 1| |g(x) - g^*(x)| \, d\mu(x) \\
&\geq 0,
\end{aligned}$$

e assim $P[g^*(X) \neq Y] \leq P[g(X) \neq Y]$ para qualquer classificador boreliano g e portanto $\text{erro}_{\widetilde{\mu}}(g^*) = P[g^*(X) \neq Y] = \ell^*(\widetilde{\mu})$. ■

2.2 REGRAS DE APRENDIZAGEM E CONSISTÊNCIA

No problema de aprendizagem, dispomos de amostras rotuladas para construir a função boreliana que tentará predizer as respostas das novas observações. Segundo o Teorema 2.1.2, o melhor que podemos esperar de um classificador boreliano é que ele atinja o erro de Bayes. No problema de classificação binária da Definição 2.1.4, buscamos uma função boreliana da forma $g_n(d_n) : \Omega \to \{0,1\}$ que tente predizer os rótulos de novas observações, por isso, agora vamos considerar famílias de classificadores indexadas pelo tamanho da amostra rotulada que aceitem como argumento, chamadas regras de aprendizagem.

Definição 2.2.1 (Regra de aprendizagem)**.** Uma regra de aprendizagem sobre Ω, espaço boreliano padrão, é uma família $\mathcal{L} = (g_n)_{n=1}^{\infty}$, de funções

$$g_n : (\Omega \times \{0,1\})^n \to \mathfrak{B}(\Omega, \{0,1\}), n \in \mathbb{N}$$

com $\mathfrak{B}(\Omega, \{0,1\})$ o espaço de classificadores borelianos sobre Ω, tais que as aplicações:

$$(\Omega \times \{0,1\})^n \times \Omega \ni (d_n, x) \mapsto g_n(d_n)(x) \in \{0,1\}$$

sejam borelianas no espaço produto.

Em outras palavras, uma regra de aprendizagem, para $n \in \mathbb{N}$, toma uma amostra rotulada d_n e devolve um classificador $g_n(d_n)$. Esse classificador, para $x \in \Omega$, retorna o rótulo $g_n(d_n)(x) \in \{0,1\}$.

Para analisar a acurácia de uma regra de aprendizagem, mesmo sendo ela uma família de funções *determinísticas*, como as amostras rotuladas são instâncias de variáveis aleatórias, vamos precisar de um enfoque probabilístico para modelar as amostras rotuladas. No problema de classificação, vamos supor que os elementos (x_i, y_i), $i \in [n]$ da amostra d_n, satisfazem:

(i) Cada (x_i, y_i), $i \in [n]$ é uma instância de uma variável aleatória $(X_i, Y_i) \sim \widetilde{\mu}$.

(ii) As amostras rotuladas (x_i, y_i), $i \in [n]$ são obtidas de forma *independente*.

Resumimos (i) e (ii) em $(X_i, Y_i) \overset{i.i.d}{\sim} \widetilde{\mu}$, onde $i.i.d$ lê-se: "...são **i**ndependentes e **i**dênticamente **d**istribuídas com medida...". Escrevemos tudo isto formalmente na seguinte definição.

Definição 2.2.2 (Variáveis aleatórias $i.i.d$)**.** Seja $\mathcal{I}$ um conjunto de índices. Dizemos que a família de variáveis aleatórias, $\{W_i\}_{i\in\mathcal{I}}$, tomando valores no espaço de probabilidade boreliano $(E, \mathscr{B}_E, \nu)$, é **i**ndependente e **i**dênticamente **d**istribuida com medida ν, conceito que denotamos por $W_i \overset{i.i.d}{\sim} \nu$, se para todo $i \in \mathcal{I}$, $W_i \sim \nu$ e para cada subconjunto finito

de índices $\mathcal{J} \subset \mathcal{I}$, $\mathcal{J} = \{i_1, i_2, \ldots, i_m\}$, temos

$$\begin{aligned} P\left[W \in \prod_{j=1}^{m} B_j\right] &= \prod_{j=1}^{m} P[W_{i_j} \in B_j] \\ &= \prod_{j=1}^{m} \nu(B_j) \\ &= \left(\bigotimes_{j=1}^{m} \nu\right)\left(\prod_{j=1}^{m} B_j\right) \\ &=: \nu^m \left(\prod_{j=1}^{m} B_j\right), \end{aligned}$$

onde $W = (W_{i_1}, W_{i_2}, \ldots, W_{i_m}) \in \prod_{j=1}^{m} E =: E^m$, os conjuntos $B_j \in \mathscr{B}_E$, $\forall j \in [m]$, são arbitrários e $\nu^m := \bigotimes_{j=1}^{m} \nu$ é a medida produto no espaço E^m.

Observação 2.2.1. Considere o conjunto $\mathcal{I} = [n]$ com $n \in \mathbb{N}$ e as variáveis aleatórias $\{W_i\}_{i\in\mathcal{I}} \subset E$ tais que $W_i \overset{i.i.d}{\sim} \nu$, onde $(E, \mathscr{B}_E, \nu)$ é um espaço de probabilidade boreliano. Se μ_W é a lei de probabilidade da variável aleatória $W = (W_1, W_2, ..., W_n) \in E^n$, então para todo $B \in \mathscr{B}_{E^n}$ temos $\mu_W(B) = P[W \in B]$, mas como E^n é *produto de espaços de probabilidade*, temos que *existe uma única medida de probabilidade* sobre E^n, a medida produto $\otimes_{i=1}^{n}\nu =: \nu^n$, que satisfaz $\nu^n\left(\prod_{i=1}^{n} B_i\right) = \prod_{i=1}^{n} \nu(B_i)$ para conjuntos $B_i \in \mathscr{B}_E$ arbitrários, logo $\mu_W = \nu^n$ e assim, para todo $B \in \mathscr{B}_{E^n}$ temos $\mu_W(B) = \nu^n(B)$ e escrevemos $W \sim \nu^n$. Similarmente, se $\mathcal{I} = \mathbb{N}$, no espaço $E^\infty := \prod_{i=1}^{\infty} E$ temos a medida produto $\otimes_{i=1}^{\infty}\nu =: \nu^\infty$, que é a única medida de probabilidade em E^∞ que satisfaz $\nu^\infty\left(\prod_{i=1}^{\infty} B_i\right) = \prod_{i=1}^{\infty} \nu(B_i)$ para conjuntos $B_i \in \mathscr{B}_E$ arbitrários e a lei μ_W da variável aleatória $W = (W_1, W_2, ...) \in E^\infty$ onde $W_i \overset{i.i.d}{\sim} \nu$, é $\mu_W = \nu^\infty$ e assim, para todo $B \in \mathscr{B}_{E^\infty}$ temos $\mu_W(B) = \nu^\infty(B)$ e escrevemos $W \sim \nu^\infty$. Para maiores detalhes, ver [9, Sec. 18] e [52, Sec. E.1.6].

Agora que temos o conceito de independência de variáveis aleatórias, podemos modelar uma amostra d_n usando a amostra aleatória

$$D_n := ((X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)) \in (\Omega \times \{0, 1\})^n$$

com $(X_i, Y_i) \overset{i.i.d}{\sim} \widetilde{\mu}$, logo, para $\prod_{i=1}^{n} B_i \in \mathscr{B}_{(\Omega\times\{0,1\})^n}$ com conjuntos $B_i \in \mathscr{B}_{\Omega\times\{0,1\}}$

arbitrários, temos

$$\begin{aligned} P\left[D_n \in \prod_{i=1}^{n} B_i\right] &= \prod_{i=1}^{n} P[(X_i, Y_i) \in B_i] \\ &= \prod_{i=1}^{n} \widetilde{\mu}(B_i) \\ &= \left(\bigotimes_{i=1}^{n} \widetilde{\mu}\right)\left(\prod_{i=1}^{n} B_i\right) \\ &=: \widetilde{\mu}^n \left(\prod_{i=1}^{n} B_i\right) \end{aligned}$$

e assim pelo mesmo argumento utilizado na Observação 2.2.1, $D_n \sim \widetilde{\mu}^n$.

Para medir o erro de uma regra de aprendizagem em Ω, primeiro observamos que para cada amostra rotulada $d_n \in (\Omega \times \{0,1\})^n$, o classificador $g_n(d_n)$ tem erro de classificação

$$\text{erro}_{\widetilde{\mu}}(g_n(d_n)) = P[g_n(D_n)(X) \neq Y \mid D_n = d_n] = h_n(d_n),$$

onde $h_n : (\Omega \times \{0,1\})^n \to [0,1]$, é uma função mensurável tal que

$$P[g_n(D_n)(X) \neq Y \mid D_n] = h_n \circ D_n.$$

A função $L(g_n) := h_n$, definida a seguir, é chamada de erro de generalização de $\mathcal{L}$.

Definição 2.2.3 (Erro de generalização)**.** Para uma regra de aprendizagem $\mathcal{L} = (g_n)_{n=1}^{\infty}$ e para cada $n \in \mathbb{N}$, defina o erro de generalização de g_n, como a função mensurável $L(g_n) : (\Omega \times \{0,1\})^n \to [0,1]$, definida para cada amostra aleatória $D_n \in (\Omega \times \{0,1\})^n$, pela *variável aleatória*

$$\begin{aligned} L(g_n)(D_n) &= P[g_n(D_n)(X) \neq Y \mid D_n] \\ &= \|g_n(D_n) - \eta\|_{L^1(\mu)} \\ &= \int_{\Omega \setminus \eta_{1/2}} |g_n(D_n)(x) - \eta(x)| \, d\mu(x) + \frac{1}{2}\mu(\eta_{1/2}), \end{aligned}$$

onde $\eta_{1/2} = \{\eta(X) = 1/2\}$ é como no Corolário 2.1.1.1. Também, definimos o erro esperado ao nível n para $\mathcal{L}$, pelo *número real*

$$\begin{aligned} P[g_n(D_n)(X) \neq Y] &= \mathbb{E}\left[L(g_n)(D_n)\right] \\ &= \int_{(\Omega \times \{0,1\})^n} L(g_n)(d_n) \, d\widetilde{\mu}^n(d_n). \end{aligned}$$

A acurácia de uma regra de aprendizagem é medida pela convergência do erro de generalização para o erro de Bayes. Assim, em uma regra de aprendizagem "boa", o que esperamos é que a medida que o valor de n cresce, o erro esperado $\mathbb{E}\left[L(g_n)(D_n)\right]$ tenda para o erro de Bayes.

Observação 2.2.2. Como para todo $n \in \mathbb{N}$ as variáveis aleatórias $L(g_n)(D_n)$ tomam valores no intervalo $[0,1]$, temos que *convergência em média* para o erro de Bayes de $L(g_n)$ é *equivalente* com *convergência em probabilidade* de $L(g_n)$ para o erro de Bayes $\ell^* = \ell^*(\widetilde{\mu})$, isto é, quando $n \to \infty$

$$\mathbb{E}[L(g_n)(D_n)] \to \ell^* \quad \Leftrightarrow \quad L(g_n)(D_n) \xrightarrow{P} \ell^*.$$

Demonstração. ($\Rightarrow$) Imediata, pois em um espaço de medida, convergência em média (no nosso caso em $L^1(\widetilde{\mu}^n)$) implica convergência em probabilidade.
($\Leftarrow$) Para $\epsilon > 0$ qualquer, defina $A_\epsilon = \{d_n \in (\Omega \times \{0,1\})^n : L(g_n)(d_n) - \ell^* > \epsilon\}$.
Então

$$\begin{aligned}
&\mathbb{E}[L(g_n)(D_n)] - \ell^* \\
&\quad = \int_{(\Omega\times\{0,1\})^n} (L(g_n)(d_n) - \ell^*)\, d\widetilde{\mu}^n(d_n) \\
&\quad = \int_{A_\epsilon} (L(g_n)(d_n) - \ell^*)\, d\widetilde{\mu}^n(d_n) + \int_{\Omega\setminus A_\epsilon} (L(g_n)(d_n) - \ell^*)\, d\widetilde{\mu}^n(d_n) \\
&\quad \leq \int_{A_\epsilon} (1 - \ell^*)\, d\widetilde{\mu}^n(d_n) + \int_{\Omega\setminus A_\epsilon} \epsilon\, d\widetilde{\mu}^n(d_n) \\
&\quad = (1 - \ell^*) \cdot \widetilde{\mu}^n(A_\epsilon) + \epsilon \cdot \widetilde{\mu}^n(\Omega \setminus A_\epsilon) \\
&\quad \leq (1 - \ell^*) \cdot P[L(g_n)(D_n) - \ell^* > \epsilon] + \epsilon,
\end{aligned}$$

logo, se $L(g_n)(D_n) \xrightarrow{P} \ell^*$, então $P[L(g_n)(D_n) - \ell^* > \epsilon] \to 0$, quando $n \to \infty$ e como $\epsilon > 0$ é arbitrário, obtemos $\mathbb{E}[L(g_n)(D_n)] \to \ell^*$, quando $n \to \infty$. ■

Quando temos convergência do erro de generalização para o erro de Bayes, dizemos que a regra é consistente. Mais precisamente.

Definição 2.2.4 (Regra de aprendizagem consistente)**.** Sejam Ω espaço boreliano padrão e $\widetilde{\mu}$ uma medida de probabilidade boreliana sobre $\Omega \times \{0,1\}$. Dizemos que a regra de aprendizagem $\mathcal{L} = (g_n)_{n=1}^{\infty}$ é (*fracamente*) consistente com a medida $\widetilde{\mu}$, se o erro de generalização converge ao erro de Bayes $\ell^* = \ell^*(\widetilde{\mu})$, em probabilidade, isto é, se

$$\forall \epsilon > 0,\ P[L(g_n)(D_n) > \ell^* + \epsilon] \overset{n\to\infty}{\longrightarrow} 0$$

ou na linguagem da teoria da medida

$$\forall \epsilon > 0,\ \widetilde{\mu}^n(\{d_n \in (\Omega \times \{0,1\})^n : L(g_n)(d_n) > \ell^* + \epsilon\}) \overset{n\to\infty}{\longrightarrow} 0.$$

Denotamos este tipo de convergência por $L(g_n)(D_n) \xrightarrow{P} \ell^*$.

Ao longo do texto vamos utilizar apenas a noção de consistência fraca e portanto, no que segue chamaremos a consistência fraca simplesmente de *consistência*.

Também dizemos que a regra de aprendizagem $\mathcal{L} = (g_n)_{n=1}^{\infty}$, é *fortemente* consistente com a medida $\widetilde{\mu}$, se a convergência para o erro de Bayes $\ell^* = \ell^*(\widetilde{\mu})$, é *quase certa*, isto é

$$P\left[\lim_{n\to\infty} L(g_n)(D_n) = \ell^*\right] = 1$$

ou na linguagem da teoria da medida

$$\widetilde{\mu}^{\infty}(\{d_\infty \in (\Omega \times \{0,1\})^{\infty} : \lim_{n\to\infty} L(g_n)(d_n) = \ell^*\}) = 1,$$

onde $d_\infty := ((x_1, y_1), (x_2, y_2), \cdots) \in (\Omega \times \{0,1\})^{\infty}$ e denotamos esse tipo de convergência por $L(g_n)(D_n) \overset{q.c}{\to} \ell^*$.

A noção de consistência (fraca), diz que se n crescer o suficiente, o erro esperado sobre todas as amostras rotuladas de tamanho n fica tão próximo do erro de Bayes quanto a gente quiser. Por outro lado, para a consistência forte, o erro de generalização converge para o erro de Bayes para $\widetilde{\mu}^{\infty}$-quase todos os *caminhos amostrais* $d_\infty \in (\Omega \times \{0,1\})^{\infty}$ ou equivalentemente, em $\widetilde{\mu}^{\infty}$ - q.t.p.

Observação 2.2.3 (Consistência Forte $\Rightarrow$ Consistência Fraca)**.** Para cada $n \in \mathbb{N}$, considere a função $\pi^n : (\Omega \times \{0,1\})^{\infty} \to (\Omega \times \{0,1\})^n$ definida por

$$\pi^n(d_\infty) = \pi^n(((x_1, y_i), (x_2, y_2), \ldots, (x_n, y_n), \ldots)) = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) = d_n.$$

A função π^n é Borel mensurável graças à Proposição A.1.3, pois $\pi_i^n(d_\infty) = \pi_i(d_\infty)$, onde as funções *borelianas* (contínuas) $\pi_i : (\Omega \times \{0,1\})^{\infty} \to \Omega \times \{0,1\}$, para $i \in \mathbb{N}$, são as projeções canônicas. De maneira similar de como foi feito na Observação 2.2.1, é fácil verificar que

$$\widetilde{\mu}^{\infty} \circ (\pi^n)^{-1} = \widetilde{\mu}^n$$

e assim, por exemplo, para $a \in \mathbb{R}$ qualquer, temos

$$\begin{aligned} P[L(g_n)(\pi^n(D_\infty)) > a] &= \widetilde{\mu}^{\infty}(\{d_\infty \in (\Omega \times \{0,1\})^{\infty} : L(g_n)(\pi^n(d_\infty)) > a\}) \\ &= \widetilde{\mu}^{\infty}((\pi^n)^{-1}(\{d_n \in (\Omega \times \{0,1\})^n : L(g_n)(d_n) > a\})) \\ &= \widetilde{\mu}^n(\{d_n \in (\Omega \times \{0,1\})^n : L(g_n)(d_n) > a\}) \\ &= P[L(g_n)(D_n) > a]. \end{aligned}$$

Suponha que a regra $\mathcal{L} = (g_n)_{n=1}^{\infty}$ é fortemente consistente com a medida dos dados rotulados $\widetilde{\mu}$ e considere $\ell^* = \ell^*(\widetilde{\mu})$ o erro de Bayes para $\widetilde{\mu}$ e o conjunto

$$A_1 = \{d_\infty \in (\Omega \times \{0,1\})^{\infty} : \lim_{n\to\infty} L(g_n)(\pi^n(d_\infty)) = \ell^*\}$$

que tem medida $\widetilde{\mu}^{\infty}(A_1) = 1$. Então, a regra $\mathcal{L}$ e fracamente consistente com $\widetilde{\mu}$, pois para $\epsilon > 0$ *arbitrário*, utilizando a *desigualdade de Markov* e tomando limite quando $n \to \infty$

junto com o Teorema da Convergência Dominada de Lebesgue, obtemos

$$\begin{aligned}\lim_{n\to\infty} P[L(g_n)(D_n) > \ell^* + \epsilon] &= \lim_{n\to\infty} P[L(g_n)(\pi^n(D_\infty)) - \ell^* > \epsilon] \\ &\leq \frac{1}{\epsilon} \lim_{n\to\infty} \mathbb{E}[L(g_n)(\pi^n(D_\infty)) - \ell^*] \\ &= \frac{1}{\epsilon} \int_{A_1} \lim_{n\to\infty} (L(g_n)(\pi^n(d_\infty)) - \ell^*)\, d\widetilde{\mu}^\infty(d_\infty) \\ &= 0.\end{aligned}$$

Na definição acima, temos a noção de consistência com relação a uma medida de probabilidade boreliana em *particular*, no nosso caso a medida $\widetilde{\mu}$, portanto perfeitamente podemos ter que uma regra de aprendizagem seja consistente com uma medida e não consistente com uma outra medida de probabilidade boreliana sobre o mesmo domínio. Como a lei $\widetilde{\mu}$ dos dados é desconhecida, vamos preferir aquelas regras de aprendizagem que sejam consistentes *para toda* medida de probabilidade boreliana sobre o domínio em estudo. Isto leva ao conceito de *consistência universal*.

Definição 2.2.5 (Regra de aprendizagem universalmente consistente)**.** Seja Ω um espaço boreliano padrão. Dizemos que a regra de aprendizagem em Ω, $\mathcal{L} = (g_n)_{n=1}^{\infty}$, é *universalmente (fracamente) consistente*, se $\mathcal{L}$ é (fracamente) consistente com *toda* medida de probabilidade boreliana sobre $\Omega \times \{0,1\}$. Similarmente, a regra de aprendizagem $\mathcal{L}$ é *universalmente fortemente consistente*, se $\mathcal{L}$ é fortemente consistente com *toda* medida de probabilidade boreliana sobre $\Omega \times \{0,1\}$.

Um exemplo de regra de aprendizagem universalmente consistente, é a regra dos k vizinhos mais próximos ou k-NN. Veremos a definição e alguns detalhes desta regra no Capítulo 6. No artigo [59] do ano 1977, Charles Stone mostra que a regra k-NN é universalmente (fracamente) consistente no espaço euclidiano, e no recente artigo [41], mostram que a regra k-NN é *universalmente fortemente consistente* em espaços métricos *ultramétricos*; classe de espaços métricos à qual pertence o corpo dos números p-ádicos que estudaremos no Capítulo 3. No Capítulo 5, mostramos outro exemplo de uma regra de aprendizagem universalmente consistente, uma regra que é um "parente" próximo do k-NN, a regra "^+k-NN". Desde o ponto de vista teórico vamos preferir regras de aprendizagem universalmente consistentes, mas na aplicações práticas, algumas regras cuja consistência ainda não foi estabelecida são usadas porque conseguem uma alta acurácia. Um exemplo é a Floresta Aleatória (Random Forest), regra amplamente usada nas aplicações, mas cuja consistência na versão original [14] é um problema em aberto. Detalhes sobre resultados teóricos e práticos da Floresta Aleatória, em [7] e [8].

A intuição nos diz que quanto mais dados usamos para construir o classificador melhor será o desempenho dele, mas o erro esperado ao nível n não é necessariamente maior que o erro no nível $n+1$. Uma regra de aprendizagem cujo erro esperado ao nível n é maior ou igual que o erro esperado no nível $n+1$, é chamada de regra de aprendizagem *inteligente*

(Smart) ver [24, Sec. 6.8]. Surpreendentemente, as regras consistentes mais comuns, como o k-NN, não são regras inteligentes, esse fato inspirou no ano 1996 a seguinte conjectura (ver [24, Ex 6.16, p. 109]): "*nenhuma regra de aprendizagem universalmente consistente é inteligente*". Mais tarde, no ano 2022, um dos orientadores provou que esta conjectura é *falsa*, construindo uma regra de aprendizagem universalmente consistente e inteligente. Detalhes podem ser encontrados em [50].

2.2.1 Regras de aprendizagem do tipo Plug-In

Como visto anteriormente, uma regra de aprendizagem em Ω, $\mathcal{L} = (g_n)_{n=1}^{\infty}$ é uma família de funções $g_n : (\Omega \times \{0,1\})^n \to \mathfrak{B}(\Omega, \{0,1\})$ onde para cada d_n, $g_n(d_n)$ é um classificador boreliano em Ω, tal que as aplicações

$$(\Omega \times \{0,1\})^n \times \Omega \ni (d_n, x) \mapsto g_n(d_n)(x) \in \{0,1\}$$

sejam borelianas no espaço produto. Uma forma simples de gerar classificadores é construindo uma função $\widetilde{\eta}$ que aproxime à função de regressão η, para logo usar a forma do classificador de Bayes na Definição 2.1.9.

Definição 2.2.6 (Classificador do tipo *plug-in*)**.** Seja Ω um espaço boreliano padrão e $\widetilde{\eta} : \Omega \to [0,1]$, uma função Borel mensurável, que é vista como uma *aproximação da função de regressão*, η. Defina o classificador $g : \Omega \to \{0,1\}$ mediante

$$g(x) = \begin{cases} 1, & \tilde{\eta}(x) \geq \frac{1}{2} \\ 0, & \text{caso contrário.} \end{cases}$$

Um classificador g construído assim, é chamado de classificador do tipo *plug-in*.

Agora, se para cada d_n geramos o classificador do tipo plug-in $g_n(d_n)$, então a regra resultante também leva o mesmo nome.

Definição 2.2.7 (Regra de aprendizagem do tipo *plug-in*)**.** Considere a família de funções Borel mensuráveis $\mathcal{F} = \{\eta_n\}_{n\in\mathbb{N}}$, tais que $\eta_n : (\Omega \times \{0,1\})^n \times \Omega \to [0,1]$. Para $n \in \mathbb{N}$ e para cada $d_n \in (\Omega \times \{0,1\})^n$, defina o classificador do tipo plug-in, $g_n(d_n) : \Omega \to \{0,1\}$ para $x \in \Omega$, por

$$g_n(d_n)(x) = \begin{cases} 1, & \eta_n(d_n, x) \geq \frac{1}{2} \\ 0, & \text{caso contrário.} \end{cases}$$

Uma regra de aprendizagem $\mathcal{L} = (g_n)_{n=1}^{\infty}$ construida desta maneira, é chamada de regra de aprendizagem do tipo plug-in.

Para os classificadores do tipo plug-in, temos o seguinte resultado, que como veremos depois do teorema, fornece uma condição suficiente para garantir a consistência fraca de uma regra de aprendizagem do tipo *plug-in*.

Teorema 2.2.1 (Teorema 2.2 de [24]). *O erro de classificação do classificador do tipo plug-in, g da Definição 2.2.6, satisfaz*

$$P[g(X) \neq Y] - \ell^* = 2\int_{\Omega\setminus\eta_{1/2}} |\eta(x) - 1/2|\, \mathbb{I}_{\{g(X)\neq g^*(X)\}}(x)\, d\mu(x),$$

e

$$P[g(X) \neq Y] - \ell^* \leq 2\int_{\Omega\setminus\eta_{1/2}} |\eta(x) - \tilde{\eta}(x)|\, d\mu(x) = 2\mathbb{E}[|\eta(X) - \tilde{\eta}(X)|\mathbb{I}_{\Omega\setminus\eta_{1/2}}(X)],$$

onde g^* e ℓ^*, são o classificador e o erro de Bayes respeito da lei dos dados rotulados $\widetilde{\mu}$, respectivamente, e como antes, $\eta_{1/2} = \{\eta(X) = 1/2\}$.

Demonstração. Usando a equação (2) da prova do Teorema 2.1.2, para cada $x \in \Omega \setminus \eta_{1/2}$, temos

$$P[g(X) \neq Y|\, X = x] - P[g^*(X) \neq Y|\, X = x] = |2\eta(x) - 1|\mathbb{I}_{\{g(X)\neq g^*(X)\}}(x).$$

Integrando em $\Omega \setminus \eta_{1/2}$, obtemos

$$P[g(X) \neq Y] - P[g^*(X) \neq Y] = 2\int_{\Omega\setminus\eta_{1/2}} |\eta(x) - 1/2|\mathbb{I}_{\{g(X)\neq g^*(X)\}}(x)\, d\mu(x).$$

Para provar a desigualdade, note que se $g(x) \neq g^*(x)$, temos $|\eta(x) - 1/2| \leq |\eta(x) - \tilde{\eta}(x)|$, e como para todo $x \in \Omega \setminus \eta_{1/2}$, $\mathbb{I}_{\{g(X)\neq g^*(X)\}}(x) \leq 1$, então

$$|\eta(x) - 1/2|\mathbb{I}_{\{g(X)\neq g^*(X)\}}(x) \leq |\eta(x) - \tilde{\eta}(x)|, \forall x \in \Omega \setminus \eta_{1/2},$$

logo, integrando em $\Omega \setminus \eta_{1/2}$ o resultado segue. ■

Do Teorema 2.2.1, temos o seguinte simples corolário.

Corolário 2.2.1.1. *Seja $\mathcal{L} = (g_n)_{n=1}^{\infty}$ uma regra de aprendizagem em Ω do tipo plug-in como na Definição 2.2.7. Então, para uma amostra aleatória $D_n \in (\Omega \times \{0,1\})^n$, temos*

$$P[g_n(D_n)(X) \neq Y|D_n] - \ell^*(\widetilde{\mu}) \leq 2\mathbb{E}[|\eta_n(D_n, X) - \eta(X)|\mathbb{I}_{\Omega\setminus\eta_{1/2}}(X)\,|D_n].$$

Finalmente, uma consequência imediata do corolário acima, que fornece uma condição suficiente para a consistência de uma regra de aprendizagem do tipo plug-in, é o seguinte resultado.

Corolário 2.2.1.2. *Nas condições do corolário anterior, se*

$$\mathbb{E}[|\eta_n(D_n, X) - \eta(X)|\mathbb{I}_{\Omega\setminus\eta_{1/2}}(X)] \overset{n\to\infty}{\longrightarrow} 0,$$

então a regra $\mathcal{L}$ é consistente com a medida dos dados rotulados $\widetilde{\mu}$.

2.3 COMPOSIÇÃO DE UMA REGRA DE APRENDIZAGEM COM UMA APLICAÇÃO

Nesta seção, vamos definir uma forma de compor uma regra de aprendizagem com uma aplicação, para poder transportar o problema de classificação/regressão para um outro domínio onde a classificação/regressão seja mais eficiente em algum sentido de interesse. Também veremos que se a aplicação que utilizamos para a composição é *injetora e Borel mensurável*, então conservamos a consistência universal da regra de aprendizagem, caso houver.

Observação 2.3.1. Considere agora, Ω e Γ espaços borelianos padrão, $\mathcal{Y}$ o espaço de rótulos e seja $f : \Omega \to \Gamma$ uma função injetiva e boreliana. Podemos estender f de forma canônica a uma função injetiva e boreliana entre os espaços $\Omega \times \mathcal{Y}$ e $\Gamma \times \mathcal{Y}$, mediante $F : \Omega \times \mathcal{Y} \to \Gamma \times \mathcal{Y}$, que é definida por $F(x, y) := (f(x), y)$.

Agora, se

$$d_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\Omega \times \mathcal{Y})^n,$$

defina a função F_n para cada d_n

$$\begin{aligned} F_n(d_n) :&= (F(x_1, y_1), F(x_2, y_2), \ldots, F(x_n, y_n)) \\ &= ((f(x_1), y_1), (f(x_2), y_2), \ldots, (f(x_n), y_n)) \in (\Gamma \times \mathcal{Y})^n. \end{aligned}$$

Definição 2.3.1 (Composição da regra $\mathcal{L}$ com a injeção boreliana f)**.** Sejam Ω e Γ espaços borelianos padrão, $\mathcal{L} = (g_n)_{n=1}^{\infty}$ uma regra de aprendizagem em Γ e $f : \Omega \to \Gamma$ uma função injetiva e boreliana. Definimos em Ω, a regra de aprendizagem $\mathcal{L}^f = \left(g_n^f\right)_{n=1}^{\infty}$, obtida da composição da regra $\mathcal{L}$ com a função f, mediante

$$g_n^f(d_n)(x) := g_n(F_n(d_n))(f(x)),$$

onde $x \in \Omega$, $d_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\Omega \times \{0, 1\})^n$ e F_n é como na Observação 2.3.1.

Nosso objetivo agora é mostrar que se a regra $\mathcal{L}$ na definição acima é universalmente consistente em Γ, então a regra $\mathcal{L}^f$ será uma regra universalmente consistente em Ω, mas primeiro veremos alguns resultados auxiliares necessários.

Na seguinte proposição, veremos que uma função boreliana entre dois espaços borelianos padrão preserva a propriedade "$i.i.d$".

Proposição 2.3.1. *Sejam $\{W_i\}_{i \in \mathbb{N}}$ uma sequência de variáveis aleatórias com valores no espaço de probabilidade boreliano $(E, \mathscr{B}_E, \mu)$, tais que $W_i \overset{i.i.d}{\sim} \mu$ e $F : E \to G$, uma função boreliana, com G espaço boreliano padrão, então,*

$$F(W_i) \overset{i.i.d}{\sim} \nu,$$

onde $\nu := F_{*\mu}$, isto é, ν é a imagem direta de μ mediante F, ou seja, é a medida boreliana em G tal que $\nu(B) = \mu(F^{-1}(B))$ para todo $B \in \mathscr{B}_G$.

Demonstração. Primeiro, vemos que para todo $i \in \mathbb{N}$ e todo $B \in \mathscr{B}_G$ temos

$$P[F(W_i) \in B] = P[W_i \in F^{-1}(B)] = \mu(F^{-1}(B)) = \nu(B),$$

logo para todo $i \in \mathbb{N}$, $F(W_i) \sim \nu$. Agora, se $B_i \in \mathscr{B}_G$ para $i \in \mathbb{N}$, são arbitrários e $W = (W_1, W_2, ...) \in E^\infty$, temos

$$\begin{aligned} P\left[(F(W_1), F(W_2), ...) \in \prod_{i=1}^{\infty} B_i\right] &= P\left[W \in \prod_{i=1}^{\infty} F^{-1}(B_i)\right] \\ &\overset{W_i \text{ i.i.d}}{=} \prod_{i=1}^{\infty} P[W_i \in F^{-1}(B_i)] \\ &= \prod_{i=1}^{\infty} P[F(W_i) \in B_i] \end{aligned}$$

de onde vemos que a sequência $\{F(W_i)\}_{i\in\mathbb{N}}$ satisfaz, $F(W_i) \overset{i.i.d}{\sim} \nu$. ■

Agora, vamos a enunciar sem prova, um resultado clássico da teoria da medida conhecido como teorema de mudança de variável.

Teorema 2.3.1 (Mudança de variável (Teorema 16.12 de [9]))**.** *Sejam $(\Omega, \mathscr{A}_\Omega, \mu)$ espaço de medida, $(\Gamma, \mathscr{A}_\Gamma)$ espaço mensurável, $f : \Omega \to \Gamma$ função $(\mathscr{A}_\Omega, \mathscr{A}_\Gamma)$-mensurável e considere em Γ a medida $\nu = \mu \circ f^{-1}$. Se $g : \Gamma \to \mathbb{R}$ é ν-integrável, então para todo $B \in \mathscr{A}_\Gamma$*

$$\int_{f^{-1}(B)} (g \circ f)(x)\, d\mu(x) = \int_B g(w)\, d\nu(w).$$

A seguir, apresentamos também sem prova, um resultado que é um corolário do *Teorema de Luzin-Souslin* [37, Teo. 15.1] o qual garante que a imagem direta de um subconjunto boreliano de um espaço boreliano padrão mediante uma função injetora e boreliana, é também um conjunto boreliano.

Teorema 2.3.2 (Corolário 15.2 de [37])**.** *Sejam Ω e Γ espaços borelianos padrão, B subconjunto boreliano de Ω e $f : \Omega \to \Gamma$ função boreliana. Se $f|_B$ é injetora, então $f(B)$ é boreliano em Γ e f induz um isomorfismo boreliano de B sobre $f(B)$.*

A seguir, vamos determinar a relação entre a função de regressão da medida $\widetilde{\mu}$ e a função de regressão da medida $\widetilde{\nu} = \widetilde{\mu} \circ F^{-1}$.

Proposição 2.3.2. *Sejam Ω e Γ espaços borelianos padrão, η a função de regressão da medida $\widetilde{\mu}$ sobre $\Omega \times \{0,1\}$ e ι a função de regressão da medida $\widetilde{\nu} = \widetilde{\mu} \circ F^{-1}$ sobre $\Gamma \times \{0,1\}$, com $F : \Omega \times \{0,1\} \to \Gamma \times \{0,1\}$, $F(x,y) := (f(x), y)$ e $f : \Omega \to \Gamma$ injecão boreliana. Então,*

$$\eta = \iota \circ f, \quad \mu\text{-q.t.p.}$$

Demonstração. Para $B \in \mathscr{B}_\Omega$ arbitrário, temos

$$\mu_1(B) = \widetilde{\mu}(B \times \{1\}) = \int_B \eta(x)\, d\mu(x)$$

e analogamente, para $E \in \mathscr{B}_\Gamma$ qualquer, temos

$$\nu_1(E) = \widetilde{\nu}(E \times \{1\}) = \int_E \iota(w)\, d\nu(w)$$

logo

$$\begin{aligned}
\nu_1(E) &= \widetilde{\mu}(F^{-1}(E \times \{1\})) \\
&= \widetilde{\mu}\left(\{(x,y) \in \Omega \times \{0,1\} : F(x,y) \in E \times \{1\}\}\right) \\
&= \widetilde{\mu}\left(\{(x,y) \in \Omega \times \{0,1\} : (f(x),y) \in E \times \{1\}\}\right) \\
&= \widetilde{\mu}\left(f^{-1}(E) \times \{1\}\right) \\
&= \mu_1\left(f^{-1}(E)\right)
\end{aligned}$$

e assim, para todo $E \in \mathscr{B}_\Gamma$

$$\nu_1(E) = \int_E \iota(w)\, d\nu(w) = \mu_1(f^{-1}(E)) = \int_{f^{-1}(E)} \eta(x)\, d\mu(x).$$

Agora, para todo $B \in \mathscr{B}_\Omega$, pelo Teorema 2.3.2, $f(B)$ é um subconjunto boreliano de Γ e

$$\begin{aligned}
\int_B \eta(x)\, d\mu(x) &= \mu_1(B) \\
&= \nu_1(f(B)) \\
&= \int_{f(B)} \iota(w)\, d\nu(w) \\
&\overset{(m.v)}{=} \int_B (\iota \circ f)(x)\, d\mu(x),
\end{aligned}$$

onde $(m.v)$ significa que foi aplicada uma mudança de variável. Assim, vemos que para todo $B \in \mathscr{B}_\Omega$,

$$\mu_1(B) = \int_B (\iota \circ f)(x)\, d\mu(x),$$

ou seja, $\iota \circ f$ é uma derivada de *Radon-Nikodým* de μ_1 com relação a μ, mas essa derivada é única μ - q.t.p, logo

$$\eta = \iota \circ f, \quad \mu \text{ - q.t.p.}$$

■

Um simples e importante corolário da proposição anterior, é o seguinte.

Corolário 2.3.2.1. *Nas condições da Proposição 2.3.2, temos*

$$g^*_{\widetilde{\mu}} = g^*_{\widetilde{\nu}} \circ f, \ \mu \text{ - } q.t.p$$

e

$$\ell^*(\widetilde{\nu}) = \ell^*(\widetilde{\mu}),$$

onde $g^*_{\widetilde{\mu}}$ e $g^*_{\widetilde{\nu}}$, são classificadores de Bayes para $\widetilde{\mu}$ e $\widetilde{\nu}$, respectivamente.

Demonstração. Da proposicão anterior, seja $A_\eta \subset \Omega$, com $\mu(A_\eta) = 1$ tal que $\eta = \iota \circ f$ em A_η. Então,

$$\begin{aligned} g^*_{\widetilde{\mu}}(x) &= \begin{cases} 1, & \eta(x) > \frac{1}{2} \\ 0, & \eta(x) < \frac{1}{2} \end{cases} \\ &= \begin{cases} 1, & \iota(f(x)) > \frac{1}{2} \\ 0, & \iota(f(x)) < \frac{1}{2} \end{cases} \\ &= g^*_{\widetilde{\nu}}(f(x)),\ \forall x \in A_\eta. \end{aligned}$$

Por outro lado, do fato anterior

$$\begin{aligned} \ell^*(\widetilde{\nu}) &= P[g^*_{\widetilde{\nu}}(W) \neq Y] \\ &= \widetilde{\nu}(\{(w,y) \in \Gamma \times \{0,1\} : g^*_{\widetilde{\nu}}(w) \neq y\}) \\ &= \widetilde{\mu}(F^{-1}(\{(w,y) \in \Gamma \times \{0,1\} : g^*_{\widetilde{\nu}}(w) \neq y\})) \\ &= \widetilde{\mu}(\{(x,y) \in \Omega \times \{0,1\} : F(x,y) \in \{(w,y) \in \Gamma \times \{0,1\} : g^*_{\widetilde{\nu}}(w) \neq y\}\}) \\ &= \widetilde{\mu}(\{(x,y) \in \Omega \times \{0,1\} : (f(x),y) \in \{(w,y) \in \Gamma \times \{0,1\} : g^*_{\widetilde{\nu}}(w) \neq y\}\}) \\ &= \widetilde{\mu}(\{(x,y) \in \Omega \times \{0,1\} : g^*_{\widetilde{\nu}}(f(x)) \neq y\}) \\ &= \widetilde{\mu}(\{(x,y) \in A_\eta \times \{0,1\} : g^*_{\widetilde{\nu}}(f(x)) \neq y\}) \\ &= \widetilde{\mu}(\{(x,y) \in A_\eta \times \{0,1\} : g^*_{\widetilde{\mu}}(x) \neq y\}) \\ &= \widetilde{\mu}(\{(x,y) \in \Omega \times \{0,1\} : g^*_{\widetilde{\mu}}(x) \neq y\}) \\ &= P[g^*_{\widetilde{\mu}}(X) \neq Y] \\ &= \ell^*(\widetilde{\mu}). \end{aligned}$$

■

Agora, para o erro de generalização da regra de aprendizagem composta com uma função injetora e boreliana, temos o seguinte simples resultado.

Proposição 2.3.3. *Seja $\mathcal{L}^f = (g^f_n)_{n=1}^\infty$, a composição da regra $\mathcal{L} = (g_n)_{n=1}^\infty$ com a função $f : \Omega \to \Gamma$, injeção boreliana. Para $d_n \in (\Omega \times \{0,1\})^n$ qualquer, temos*

$$erro_{\widetilde{\mu}}(g^f_n(d_n)) = erro_{\widetilde{\nu}}(g_n(F_n(d_n))).$$

Demonstração. Seja $d_n \in (\Omega \times \{0,1\})^n$ fixa, então

$$\begin{aligned}
\text{erro}_{\widetilde{\mu}}(g_n^f(d_n)) &= P[g_n^f(d_n)(X) \neq Y] \\
&= P[g_n(F_n(d_n))(f(X)) \neq Y] \\
&= P[(f(X),Y) \in \{(w,y) \in \Gamma \times \{0,1\} : g_n(F_n(d_n))(w) \neq y\}] \\
&= \widetilde{\nu}(\{(w,y) \in \Gamma \times \{0,1\} : g_n(F_n(d_n))(w) \neq y\}) \\
&= P[g_n(F_n(d_n))(W) \neq Y] \\
&= \text{erro}_{\widetilde{\nu}}(g_n(F_n(d_n))).
\end{aligned}$$

■

Finalmente estamos em condições de mostrar um resultado de consistência para a regra de aprendizagem composta com uma injeção boreliana, que permite criar novas regras de aprendizagem universalmente consistentes a partir de uma regra universalmente consistente em um espaço diferente. Mais detalhes sobre o seguinte resultado em [53] e [52]. A prova a seguir, conta com mais detalhes dos contidos na prova do Teorema 6.3.8 de [52].

Teorema 2.3.3 (Teorema 6.3.8 de [52])**.** *Sejam Ω e Γ espaços borelianos padrão, $f : \Omega \to \Gamma$ uma injeção boreliana e $\mathcal{L} = (g_n)_{n=1}^{\infty}$ uma regra de aprendizagem universalmente consistente (respectivamente universalmente fortemente consistente) em Γ. Então, a regra de aprendizagem obtida pela composição da regra $\mathcal{L}$ e a função f, $\mathcal{L}^f = \left(g_n^f\right)_{n=1}^{\infty}$, é universalmente consistente (respectivamente universalmente fortemente consistente) em Ω.*

Demonstração. Suponha que a regra de aprendizagem sobre Γ, $\mathcal{L} = (g_n)_{n=1}^{\infty}$ é universalmente consistente, e sejam $\epsilon > 0$ e $n \in \mathbb{N}$ arbitrários. Então

$$\begin{aligned}
&P[L(g_n^f)(D_n) > \ell^*(\widetilde{\mu}) + \epsilon] \\
&\quad = \widetilde{\mu}^n(\{d_n \in (\Omega \times \{0,1\})^n : \text{erro}_{\widetilde{\mu}}(g_n^f(d_n)) > \ell^*(\widetilde{\mu}) + \epsilon\}) \\
&\quad = \widetilde{\mu}^n(\{d_n \in (\Omega \times \{0,1\})^n : \text{erro}_{\widetilde{\nu}}(g_n(F_n(d_n))) > \ell^*(\widetilde{\nu}) + \epsilon\}) \\
&\quad = \widetilde{\mu}^n(F_n^{-1}(\{e_n \in (\Gamma \times \{0,1\})^n : \text{erro}_{\widetilde{\nu}}(g_n(e_n)) > \ell^*(\widetilde{\nu}) + \epsilon\})) \\
&\quad \overset{(*)}{=} (\widetilde{\mu} \circ F^{-1})^n(\{e_n \in (\Gamma \times \{0,1\})^n : \text{erro}_{\widetilde{\nu}}(g_n(e_n)) > \ell^*(\widetilde{\nu}) + \epsilon\}) \\
&\quad = \widetilde{\nu}^n(\{e_n \in (\Gamma \times \{0,1\})^n : \text{erro}_{\widetilde{\nu}}(g_n(e_n)) > \ell^*(\widetilde{\nu}) + \epsilon\}) \\
&\quad = P[L(g_n)(E_n) > \ell^*(\widetilde{\nu}) + \epsilon] \overset{n\to\infty}{\longrightarrow} 0
\end{aligned}$$

obtendo o resultado no caso da consistência.

Agora, suponha que a regra de aprendizagem no domínio Γ, $\mathcal{L} = (g_n)_{n=1}^{\infty}$ é universalmente *fortemente* consistente, então com $d_\infty = ((x_1,y_1),(x_2,y_2),...) \in (\Omega \times \{0,1\})^\infty$

e $e_\infty = ((w_1, y_1), (w_2, y_2), ...) \in (\Gamma \times \{0,1\})^\infty$, temos

$$\begin{aligned}
&P\left[\lim_{n\to\infty} L(g_n^f)(D_n) = \ell^*(\widetilde{\mu})\right] \\
=&\widetilde{\mu}^\infty\left(\left\{d_\infty \in (\Omega \times \{0,1\})^\infty : \lim_{n\to\infty} \text{erro}_{\widetilde{\mu}}(g_n^f(d_n)) = \ell^*(\widetilde{\mu})\right\}\right) \\
=&\widetilde{\mu}^\infty\left(\left\{d_\infty \in (\Omega \times \{0,1\})^\infty : \lim_{n\to\infty} \text{erro}_{\widetilde{\nu}}(g_n(F_n(d_n))) = \ell^*(\widetilde{\nu})\right\}\right) \\
=&\widetilde{\mu}^\infty\left(F_\infty^{-1}\left(\left\{e_\infty \in (\Gamma \times \{0,1\})^\infty : \lim_{n\to\infty} \text{erro}_{\widetilde{\nu}}(g_n(e_n)) = \ell^*(\widetilde{\nu})\right\}\right)\right) \\
\overset{(*)}{=}&(\widetilde{\mu}\circ F^{-1})^\infty\left(\left\{e_\infty \in (\Gamma \times \{0,1\})^\infty : \lim_{n\to\infty} \text{erro}_{\widetilde{\nu}}(g_n(e_n)) = \ell^*(\widetilde{\nu})\right\}\right) \\
=&\widetilde{\nu}^\infty\left(\left\{e_\infty \in (\Gamma \times \{0,1\})^\infty : \lim_{n\to\infty} \text{erro}_{\widetilde{\nu}}(g_n(e_n)) = \ell^*(\widetilde{\nu})\right\}\right) \\
=&P\left[\lim_{n\to\infty} L(g_n)(E_n) = \ell^*(\widetilde{\nu})\right] \\
=&1.
\end{aligned}$$

■

No Capítulo 4, descrevemos uma regra de aprendizagem supervisionada no espaço métrico p-ádico d-dimensional $(\mathbb{Z}_p^d, \|\cdot\|_p)$ e no Capítulo 5 provamos a consistência universal de essa regra para certa classe de espaços métricos. Pelo Teorema 2.3.3, encontrando uma injeção boreliana $\phi_p : \mathbb{R}_+ \to \mathbb{Q}_p$, podemos compor esta função com a regra de aprendizagem p-ádica para obter uma regra universalmente consistente em $\mathbb{R}_+$.

Observação 2.3.2 ((∗))**.** Aqui vamos provar o fato usado nos passos marcados com $(*)$ na prova do Teorema 2.3.3. Primeiro vamos provar que $\widetilde{\nu}^n = \widetilde{\mu}^n \circ F_n^{-1}$, para todo $n \in \mathbb{N}$. Pelo *Teorema de Carathéodory* (ver [52, Teo. E.1.15]), é suficiente provar a igualdade para conjuntos cilíndricos (conjuntos que são produtos $\prod_{i=1}^n B_i$, com $B_i \in \mathscr{B}_{\Gamma\times\{0,1\}}$ para $i = 1, ..., n$), pois essa família é uma *álgebra* que gera a σ-álgebra de Borel no espaço produto $(\Gamma \times \{0,1\})^n$ e portanto em esse espaço existe uma única medida de probabilidade boreliana, a medida produto $\widetilde{\nu}^n$, tal que $\widetilde{\nu}^n\left(\prod_{i=1}^n B_i\right) = \prod_{i=1}^n \widetilde{\nu}(B_i)$ para conjuntos borelianos $B_i \in \mathscr{B}_{\Gamma\times\{0,1\}}$ quaisquer. Assim, considere os conjuntos cilíndricos $\prod_{i=1}^n B_i$ com $B_i \in \mathscr{B}_{\Gamma\times\{0,1\}}$, $i \in [n]$, arbitrários.
Então, para $d_n = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)) \in (\Omega \times \{0,1\})^n$, temos

$$\begin{aligned}
\widetilde{\mu}^n\left(F_n^{-1}\left(\prod_{i=1}^n B_i\right)\right) &= \widetilde{\mu}^n(\{d_n \in (\Omega \times \{0,1\})^n : F(x_i, y_i) \in B_i, i \in [n]\}) \\
&= \widetilde{\mu}^n\left(\prod_{i=1}^n F^{-1}(B_i)\right) \\
&= P\left[((X_1, Y_1), (X_2, Y_2), ...(X_n, Y_n)) \in \prod_{i=1}^n F^{-1}(B_i)\right] \\
&= \prod_{i=1}^n P[(X_i, Y_i) \in F^{-1}(B_i)],
\end{aligned}$$

pois $(X_i, Y_i) \overset{i.i.d}{\sim} \widetilde{\mu}$, logo

$$\begin{aligned}\widetilde{\mu}^n\left(F_n^{-1}\left(\prod_{i=1}^n B_i\right)\right) &= \prod_{i=1}^n P[(X_i,Y_i)\in F^{-1}(B_i)]\\ &= \prod_{i=1}^n \widetilde{\mu}(F^{-1}(B_i))\\ &= \prod_{i=1}^n \widetilde{\nu}(B_i)\\ &= \widetilde{\nu}^n\left(\prod_{i=1}^n B_i\right).\end{aligned}$$

Finalmente, o resultado também vale colocando $n=\infty$, com $F_\infty((x_1,y_1),(x_2,y_2),...) = (F(x_1,y_1),F(x_2,y_2),...) \in (\Gamma\times\{0,1\})^\infty$ e os conjuntos cilíndricos sendo da forma $\prod_{i=1}^\infty B_i$ com $B_i \in \mathscr{B}_{\Gamma\times\{0,1\}}$, $\forall i \in \mathbb{N}$, onde, salvo para um número finito de índices $i\in\mathbb{N}$, temos $B_i = \Gamma\times\{0,1\}$.

2.4 CLASSIFICAÇÃO MULTICLASSE

Nessa seção, vamos ver como aplicar os conceitos vistos para classificação binária no caso de *classificação multiclasse* e também que o Teorema 2.3.3 vale neste último caso.

Na *classificação multiclasse* teremos mais de dois rótulos para os elementos de Ω, ou seja, o conjunto de rótulos será da forma $\mathcal{Y} = \{\xi_i\}_{i=1}^m$, onde o inteiro $m>2$ é o número de classes.

No espaço boreliano produto $\Omega\times\mathcal{Y}$, se $\widetilde{\mu}$ é a lei dos dados rotulados e se procedemos como no caso binário, a lei dos dados não rotulados μ estará dada, para cada $B\in\mathscr{B}_\Omega$, por

$$\mu(B) \;=\; \widetilde{\mu}(\pi_\Omega^{-1}(B))$$

onde $\pi_\Omega : \Omega\times\mathcal{Y}\to\Omega$ é a projeção canônica, mas

$$\begin{aligned}\widetilde{\mu}(\pi_\Omega^{-1}(B)) &= \widetilde{\mu}(B\times\mathcal{Y})\\ &= \widetilde{\mu}\left(\dot{\bigcup_{i=1}^m} B\times\{\xi_i\}\right)\\ &= \sum_{i=1}^m \widetilde{\mu}\left(B\times\{\xi_i\}\right)\end{aligned}$$

e fazendo $\mu_i(B) = \widetilde{\mu}(B\times\{\xi_i\})$, para todo $i\in[m]$, e $B\subset\Omega$ subconjunto boreliano de Ω, temos

$$\mu(B) = \sum_{i=1}^m \mu_i(B).$$

As medidas μ_i, satisfazem $\mu_i(B) \leq \mu(B)$, para todo B subconjunto boreliano de Ω, logo pelo Teorema 2.1.1, existem as derivadas de *Radon-Nikodým* $\eta_i = \frac{d\mu_i}{d\mu} : \Omega \to [0,1]$. Um classificador multiclasse boreliano, é qualquer função boreliana $g : \Omega \to \mathcal{Y}$, que tem erro de classificação dado por

$$\begin{aligned} \text{erro}_{\widetilde{\mu}}(g) &= P[g(X) \neq Y] \\ &= \widetilde{\mu}(\{(x,y) \in \Omega \times \mathcal{Y} : g(x) \neq y\}). \end{aligned}$$

O erro de Bayes, é definido de forma análoga ao caso binário

$$\ell^*(\widetilde{\mu}) = \inf_{g \in \mathfrak{B}(\Omega,\mathcal{Y})} \text{erro}_{\widetilde{\mu}}(g),$$

onde $\mathfrak{B}(\Omega,\mathcal{Y})$ é o conjunto de todos os classificadores borelianos de Ω em $\mathcal{Y}$. No caso binário, chamando $\eta_1(x) = \eta(x) = P[Y = 1|X = x]$ e $\eta_0(x) = 1 - \eta(x) = P[Y = 0|X = x]$, para todo $x \in \Omega$, um classificador de Bayes pode se escrever da seguinte maneira

$$g^*_{\widetilde{\mu}}(x) = \underset{i \in \{0,1\}}{\text{argmax}}\, \eta_i(x),$$

ou seja, um classificador de Bayes devolve o rótulo que tem maior *probabilidade condicional* de ser o correto, dado que $X = x$.

Do mesmo modo, no caso multiclasse um classificador de Bayes é da forma

$$g^*_{\widetilde{\mu}}(x) = \underset{\xi_i \in \mathcal{Y}}{\text{argmax}}\, \eta_i(x)$$

com $\eta_i(x) = P[Y = \xi_i|X = x]$ para $x \in \Omega$.

Para verificar o resultado do Teorema 2.3.3 no caso multiclasse, considere os espaços borelianos padrão Ω e Γ, $\Omega \times \mathcal{Y}$ equipado com a medida de probabilidade $\widetilde{\mu}$, $f : \Omega \to \Gamma$ injeção boreliana, $F : \Omega \times \mathcal{Y} \to \Gamma \times \mathcal{Y}$ tal que $(x,y) \mapsto (f(x),y)$ e a medida em $\Gamma \times \mathcal{Y}$, $\widetilde{\nu} = \widetilde{\mu} \circ F^{-1}$.

De forma análoga a como foi feito com $\widetilde{\mu}$, para $\widetilde{\nu}$, obtemos a lei dos dados não rotulados $\nu = \widetilde{\nu} \circ \pi_\Gamma^{-1}$, com $\pi_\Gamma : \Gamma \times \mathcal{Y} \to \Gamma$ projeção canônica, as medidas $\nu_i, i \in [m]$ e as funções de regressão $\iota_i = \frac{d\nu_i}{d\nu}$. Como $\nu_i = \mu_i \circ f^{-1}$, pelo Teorema 2.3.2 e fazendo mudança de variáveis, temos $\eta_i = \iota_i \circ f$, μ-q.t.p para todo $i \in [m]$, assim, para cada $i \in [m]$, teremos um subconjunto boreliano A_{η_i} tal que $\mu(A_{\eta_i}) = 1$ e $\eta_i = \iota_i \circ f$ em A_{η_i}, logo o conjunto $A_\eta := \cap_{i=1}^m A_{\eta_i} \in \mathscr{B}_\Omega$ tem medida $\mu(A_\eta) = 1$ e satisfaz

$$\begin{aligned} g^*_{\widetilde{\mu}}(x) &= \underset{\xi_i \in \mathcal{Y}}{\text{argmax}}\, \eta_i(x) \\ &= \underset{\xi_i \in \mathcal{Y}}{\text{argmax}}\, \iota_i(f(x)) \\ &= g^*_{\widetilde{\nu}}(f(x)), \forall x \in A_\eta, \end{aligned}$$

ou seja, $g^*_{\widetilde{\mu}} = g^*_{\widetilde{\nu}} \circ f$, μ-q.t.p.

Este último implica que $\ell^*(\widetilde{\nu}) = \ell^*(\widetilde{\mu})$, logo trocando $\{0,1\}$ por $\mathcal{Y}$ nas definições 2.2.4, 2.2.5 de consistência e consistência universal, respectivamente, e na Proposição 2.3.3, a prova do resultado análogo ao Teorema 2.3.3 é a mesma, assim no caso multiclasse, se a regra de aprendizagem em Γ, $\mathcal{L}$, é universalmente consistente (respectivamente universalmente fortemente consistente), então a regra de aprendizagem em Ω, $\mathcal{L}^f$, também será universalmente consistente (respectivamente universalmente fortemente consistente).

2.5 DIMENSÃO DE NAGATA E O TEOREMA DA DIFERENCIAÇÃO DE LEBESGUE-BESICOVITCH

Finalizamos o capítulo de preliminares dedicando algumas linhas ao conceito de dimensão de *Nagata* para um espaço métrico [4, 48] e à relação dessa dimensão, via o resultado de David Preiss [54], com a validez do Teorema da Diferenciação de Lebesgue-Besicovitch [6] no espaço, pois esses conceitos serão *chaves* na hora de provar a consistência da regra de aprendizagem que desenvolveremos nos capítulos 4 e 5.

Começamos com um par de definições.

Definição 2.5.1 (Multiplicidade de uma família de conjuntos)**.** Dizemos que uma família, $\gamma \subset \mathcal{P}(\Omega)$, de subconjuntos de Ω, tem multiplicidade $\leq \delta \in \mathbb{N}_0$, se a interseção de mais de δ elementos de γ sempre é vazia, ou seja, se

$$\forall x \in \Omega, \ \sum_{V \in \gamma} \mathbb{I}_V(x) \leq \delta.$$

Definição 2.5.2 (Dimensão de Nagata [4, 18])**.** Seja $\delta \in \mathbb{N}_0$. Dizemos que o espaço métrico Ω tem *dimensão de Nagata* $\leq \delta$ sobre a escala[2] $s \in (0, +\infty]$, se toda família *finita* γ, de bolas fechadas de raio $< s$, admite uma subfamília $\gamma' \subseteq \gamma$ de multiplicidade $\leq \delta + 1$ que cobre todos os centros das bolas em γ, e escrevemos: $\dim^s_{\text{Nag}}(\Omega) \leq \delta$.

Observação 2.5.1. É claro que se um espaço métrico Ω tem dimensão de Nagata $\dim^s_{\text{Nag}}(\Omega) \leq \delta$, então terá $\dim^s_{\text{Nag}}(\Omega) \leq \beta$, para qualquer $\beta \in \mathbb{N}$, $\beta > \delta$. Assim, um espaço Ω terá dimensão de Nagata $\dim^s_{\text{Nag}}(\Omega) = \delta \in \mathbb{N}_0$, se $\dim^s_{\text{Nag}}(\Omega) \leq \delta$ e $\dim^s_{\text{Nag}}(\Omega) \not\leq \delta - 1$.

A definição acima as vezes é dada na sua forma equivalente.

Proposição 2.5.1 (Dimensão de Nagata finita [18])**.** *Um espaço métrico (Ω, ρ) tem $\dim^s_{Nag}(\Omega) \leq \delta \in \mathbb{N}_0$ com $s \in (0, +\infty]$, se e somente se, para todo $x \in \Omega$ e para pontos $x_1, x_2, \ldots, x_{\delta+2} \in \bar{B}(x, r)$ quaisquer com $r < s$, existem $i, j \in [\delta + 2]$, $i \neq j$, tais que $\rho(x_i, x_j) \leq \max\{\rho(x_i, x), \rho(x, x_j)\}$.*

Demonstração. ($\Rightarrow$) Suponha que $\dim^s_{\text{Nag}}(\Omega) \leq \delta$, com $s \in (0+\infty]$. Para $x \in \Omega$ e $0 < r < s$ quaisquer, considere uma sequência arbitrária de $\delta + 2$ pontos $x_1, x_2, \ldots, x_{\delta+2} \in \bar{B}(x, r)$

[2] O caso $s = +\infty$, significa que os raios das bolas na família γ podem ser quaisquer.

e defina $r_i := \rho(x_i, x) \leq r$, para $i \in [\delta + 2]$, logo por hipótese, a família $\gamma = \{\bar{B}(x_i, r_i),\ i \in [\delta + 2]\}$ possui uma subfamília $\gamma' \subset \gamma$ de multiplicidade $\leq \delta + 1$ que cobre o conjunto de centros $C := \{x_1, x_2, \ldots, x_{\delta+2}\}$. Como $x \in \bar{B}(x_i, r_i)$, para todo $i \in [\delta + 2]$, vemos que a γ não tem multiplicidade $\leq \delta + 1$, logo necessariamente $\gamma' \subsetneq \gamma$ e assim $\sharp(\gamma') \leq \delta + 1$, portanto, existe pelo menos uma bola de γ' que contém dois elementos do conjunto C, digamos $x_j \in \bar{B}(x_i, r_i) \in \gamma'$ para alguns $i, j \in [\delta + 2]$ com $i \neq j$, logo necessariamente, $\rho(x_i, x_j) \leq r_i = \rho(x_i, x) \leq \max\{\rho(x_i, x), \rho(x, x_j)\}$.
($\Leftarrow$) Seja γ uma família finita de bolas fechadas de raio $< s$. Se a multiplicidade de γ é $\leq \delta + 1$, então ela mesma é uma subfamília γ' satisfazendo a Definição 2.5.2. Suponha que γ não tem multiplicidade $\leq \delta + 1$. Então, existe $x \in \Omega$ tal que $\sum_{B \in \gamma} \mathbb{I}_B(x) \geq \delta + 2$, logo, escolhendo $\delta + 2$ bolas de γ contendo x e denotando por x_i, r_i, $i \in [\delta + 2]$ os respectivos centros e raios junto com $r = \max_{i \in [\delta+2]} \rho(x_i, x)$, temos $r < s$ e $x_1, x_2, \ldots, x_{\delta+2} \in \bar{B}(x, r)$. Assim, por hipótese, existem $i, j \in [\delta+2]$, $i \neq j$, tais que $\rho(x_i, x_j) \leq \max\{\rho(x_i, x), \rho(x, x_j)\}$. Sem perda de generalidade vamos supor $\rho(x_i, x_j) \leq \rho(x_i, x) \leq r_i$, pois supor $\rho(x_i, x_j) \leq \rho(x, x_j)$ leva ao mesmo resultado. Nesse caso $x_j \in \bar{B}(x_i, r_i)$ e assim a bola centrada em x_j pode ser removida da família γ, obtendo uma subfamília $\gamma' \subsetneq \gamma$ que cobre todos os centros das bolas em γ e tendo cardinal $\sharp(\gamma') = \sharp(\gamma) - 1$. Se γ' não tem multiplicidade $\leq \delta + 1$, repetimos o processo um número finito de vezes, até conseguir uma subfamília com multiplicidade $\leq \delta + 1$ que cobre todos os centros das bolas em γ e assim $\dim^{s}_{\mathrm{Nag}}(\Omega) \leq \delta$. ■

Na seguinte proposição, obtemos a dimensão de Nagata para os membros de uma classe de espaços métricos à qual pertence o corpo com valor absoluto dos números p-ádicos $(\mathbb{Q}_p, |\cdot|_p)$, que é a classe dos espaços *ultramétricos*.

Definição 2.5.3 (Espaço Ultramétrico)**.** O espaço métrico (Ω, ρ) será um espaço *ultramétrico*, se a métrica ρ é uma *ultramétrica*, isto é, se ρ satisfaz a *desigualdade triangular forte*:

$$\rho(x, y) \leq \max\{\rho(x, z), \rho(z, y)\},\ \forall x, y, z \in \Omega.$$

Proposição 2.5.2. *O espaço (Ω, ρ) é ultramétrico, se e somente se,* $\dim^{+\infty}_{Nag}(\Omega) = 0$.

Demonstração. ($\Rightarrow$) Suponha que Ω é um espaço ultramétrico, então, para $x \in \Omega$ e $r > 0$ quaisquer, se consideramos uma sequência arbitrária de $\delta + 2 = 2$ pontos $x_1, x_2 \in \bar{B}(x, r)$, necessariamente temos $\rho(x_1, x_2) \leq \max\{\rho(x, x_1), \rho(x, x_2)\}$, logo $\dim^{+\infty}_{\mathrm{Nag}}(\Omega) \leq 0$, e assim $\dim^{+\infty}_{\mathrm{Nag}}(\Omega) = 0$, pois $\dim^{+\infty}_{\mathrm{Nag}}(\Omega) \in \mathbb{N}_0$.
($\Leftarrow$) Suponha que o espaço métrico (Ω, ρ) tem $\dim^{+\infty}_{\mathrm{Nag}}(\Omega) = 0$. Considere $x, y, z \in \Omega$ todos distintos e $r \geq \max\{\rho(x, z), \rho(z, y)\}$ arbitrários, assim $x, y \in \bar{B}(z, r)$ e da Proposição 2.5.1 temos necessariamente $\rho(x, y) \leq \max\{\rho(x, z), \rho(z, y)\}$, isto é, o espaço métrico (Ω, ρ) é ultramétrico. ■

Exemplo 2.5.1. Da Proposição 2.5.1, vemos que $\dim_{\text{Nag}}^{+\infty}(\mathbb{R}) = 1$, pois, dados $x \in \mathbb{R}$, $r > 0$ e $x_1, x_2, x_3 \in [x-r, x+r]$ todos distintos e arbitrários, redefinindo os índices se for necessário, podemos supor sem perda de generalidade que $x_1 < x_2 < x_3$ e assim, chamando $|\cdot| := |\cdot|_\infty$ e se $x_2 \leq x$, então $|x_1 - x_2| \leq |x_1 - x| \leq \max\{|x_1 - x|, |x_2 - x|\}$ e se $x_2 > x$, temos $|x_2 - x_3| \leq |x_3 - x| \leq \max\{|x_2 - x|, |x_3 - x|\}$, de onde vemos que $\dim_{\text{Nag}}^{+\infty}(\mathbb{R}) \leq 1$. Por outro lado, para $x \in \mathbb{R}$ e $r > 0$ quaisquer, escolhendo $x_1, x_2 \in [x-r, x+r]$ tais que $x_1 < x < x_2$ e $|x_1 - x| \neq |x_2 - x|$, temos $|x_1 - x_2| > \max\{|x_1 - x|, |x_2 - x|\}$, ou seja, $(\mathbb{R}, |\cdot|_\infty)$ *não é* não-arquimediano e portanto $\dim_{\text{Nag}}^{+\infty}(\mathbb{R}) \neq 0$, logo $\dim_{\text{Nag}}^{+\infty}(\mathbb{R}) = 1$.

No artigo [18, Exemplo 4.5], mostram que a dimensão de Nagata para $(\mathbb{R}^d, \|\cdot\|_2)$, é finita e usando um argumento similar, que todo espaço normado finito-dimensional tem dimensão de Nagata finita.

No artigo [47], Nagata menciona que o problema de calcular a dimensão de Nagata do espaço euclidiano $(\mathbb{R}^d, \|\cdot\|_2)$ é um problema "*possivelmente aberto*", e também menciona, sem uma prova, que $\dim_{\text{Nag}}(\mathbb{R}) = 1$ e $\dim_{\text{Nag}}(\mathbb{R}^2) = 5$, onde não faz menção à escala s.

O resultado de Preiss, usa o conceito de espaço métrico de dimensão σ-finita no sentido de Nagata. No mesmo espírito do conceito de espaço de medida σ-finita, onde podemos representar o espaço como a união de uma família enumerável de conjuntos mensuráveis de *medida finita*, no conceito de espaço com dimensão σ-finita no sentido de Nagata, podemos representar o espaço como a união de uma família enumerável de *subespaços* de *dimensão de Nagata finita*.

Definição 2.5.4 (Dimensão de Nagata de um subespaço [18])**.** Dizemos que um subconjunto A do espaço métrico Ω tem *dimensão de Nagata* $\leq \delta \in \mathbb{N}_0$ sobre a escala $s \in (0, +\infty]$ *dentro* de Ω, se cada família finita γ de bolas fechadas em Ω com centros em A e raio $< s$, admite uma subfamília $\gamma' \subseteq \gamma$ de multiplicidade $\leq \delta + 1$ em Ω que cobre todos os centros das bolas em γ, e escrevemos: $\dim_{\text{Nag}}^{s}(A, \Omega) \leq \delta$.

No resultado de Preiss, o conceito de *métrica* finito-dimensional está dada em função de uma propriedade que deve satisfazer uma família finita de bolas fechadas, propriedade que aqui também usaremos e chamaremos de *família desconexa*.

Definição 2.5.5 (Família desconexa)**.** Sejam (Ω, ρ) um espaço métrico, I um conjunto de índices. Dizemos que a família de bolas fechadas $\{\bar{B}(x_i, r_i), i \in I\}$ é uma família *desconexa*, se para todo $i, j \in I$, com $i \neq j$, temos $x_i \notin \bar{B}(x_j, r_j)$ e $x_j \notin \bar{B}(x_i, r_i)$. Em outras palavras, o centro de toda bola da família não pertence às outras bolas ou, para todo $i, j \in I$, tais que $i \neq j$, temos $\rho(x_i, x_j) > \max\{r_i, r_j\}$.

Similar ao caso da definição de dimensão de Nagata finita para um espaço, aqui também podemos escrever a definição da dimensão para um subespaço na sua forma equivalente.

Proposição 2.5.3 (Subespaço finito-dimensional no sentido de Nagata [18])**.** *Seja A um subconjunto do espaço métrico Ω. Então,* $\dim^s_{Nag}(A,\Omega) \leq \delta$*, se e somente se, toda família finita desconexa de bolas fechadas em Ω de raio $< s$ com centros em A, tem multiplicidade $\leq \delta + 1$.*

Demonstração. ($\Rightarrow$) Seja γ uma família finita e desconexa de bolas fechadas em Ω com centros em A e raio $< s$. Como $\dim^s_{\text{Nag}}(A,\Omega) \leq \delta$, então a família γ admite uma subfamília γ' de multiplicidade $\leq \delta + 1$ que cobre todos os centros das bolas em γ, mas como γ é desconexa, necessariamente $\gamma' = \gamma$.
($\Leftarrow$) Queremos verificar a Definição 2.5.4. Seja $\gamma = \{B_i,\, i \in I\}$ uma família finita de bolas fechadas em Ω com centros em A e raio $< s$. Denotemos por C o conjunto dos centros das bolas em γ. Para uma subfamília $\gamma' = \{B_i,\, i \in I'\} \subseteq \gamma$, defina o conjunto $C_{\gamma'} = C \cap \cup_{i\in I'} B_i$, de todos os elementos de C cobertos pela subfamília γ'. Como temos um número finito de subfamílias de γ possíveis, entre todas as subfamílias desconexas[3] de γ, podemos escolher uma subfamília $\gamma^* = \{B_i,\, i \in I^*\} \subseteq \gamma$ (que por hipótese possui multiplicidade $\leq \delta + 1$), tal que

$$\sharp(C_{\gamma^*}) = \max_{\substack{\gamma' \subseteq \gamma \\ \gamma' \text{ desconexa}}} \sharp(C_{\gamma'}),$$

isto é, que maximiza o número de centros das bolas em γ cobertos por uma subfamília desconexa. Agora vamos mostrar que $C_{\gamma^*} = C$, e assim, γ^* será uma subfamília de γ de *multiplicidade* $\leq \delta + 1$ que cobre todos os centros das bolas em γ, verificando desse modo a Definição 2.5.4.

É claro que $C_{\gamma^*} \subseteq C$, assim, para chegar numa contradição, vamos supor que $C \not\subseteq C_{\gamma^*}$. Nessa situação, existe uma bola fechada $B_{i_c} \in \gamma$ cujo centro $c \in C \setminus C_{\gamma^*}$, ou seja, $c \notin \cup_{i\in I^*} B_i$. Defina $D = C \cap B_{i_c}$, o conjunto de centros dos elementos de γ que pertencem à bola B_{i_c}. Removemos todas as bolas de γ^* com centros em $D \setminus \{c\}$, e no lugar delas adicionamos B_{i_c} para obter a subfamília $\gamma^{**} \subseteq \gamma$. A nova subfamília γ^{**} é claramente desconexa e satisfaz:

$$\sharp(C_{\gamma^{**}}) = \sharp(C_{\gamma^*} \cup \{c\}) = \sharp(C_{\gamma^*}) + 1,$$

o que contradiz a maximalidade de $\sharp(C_{\gamma^*})$. Portanto $C_{\gamma^*} = C$. ■

Agora estamos em condições de definir formalmente o conceito de espaço métrico de dimensão σ-finita no sentido de Nagata.

Definição 2.5.6 (Espaço métrico de dimensão σ-finita)**.** Dizemos que o espaço métrico Ω possui dimensão *σ-finita* no sentido de Nagata, se existem sequências $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{P}(\Omega)$, $\{s_n\}_{n\in\mathbb{N}} \subset (0,+\infty]$ e $\{\delta_n\}_{n\in\mathbb{N}} \subseteq \mathbb{N}_0$, tais que $\dim^{s_n}_{\text{Nag}}(A_n,\Omega) \leq \delta_n$, para todo $n \in \mathbb{N}$ e $\Omega = \bigcup_{n=1}^{\infty} A_n$.

[3] Uma subfamília contendo apenas uma bola, é trivialmente uma subfamília desconexa, logo o conjunto de subfamílias desconexas de uma família finita de bolas fechadas, é finito e não vazio.

Seja $f : \mathbb{R} \to \mathbb{R}$ uma função integrável segundo a medida de Lebesgue na reta. No ano 1904 [42], Henri Lebesgue demonstrou que a derivada da integral de f em x com relação à medida de Lebesgue é igual a $f(x)$, para μ - q.t $x \in \mathbb{R}$, resultado que é conhecido como o Teorema da diferenciação de Lebesgue. Depois, no ano 1945, Abram Besicovitch [6], estende o resultado de Lebesgue para qualquer medida boreliana localmente finita[4] sobre o espaço euclidiano $\mathbb{R}^d$, resultado que é conhecido como o Teorema da diferenciação de Lebesgue-Besicovitch. Baseado nesses resultados, para um espaço métrico qualquer, definimos a seguinte condição.

Definição 2.5.7 (Condição de Lebesgue-Besicovitch Forte **(LBF)** [6])**.** Dizemos que o espaço métrico Ω satisfaz a condição de **L**ebesgue-**B**esicovitch **F**orte **(LBF)** para a medida boreliana localmente finita μ, se para toda função μ-integrável $f \in L^1(\mu)$, temos

$$\lim_{r\to 0^+} \frac{1}{\mu(\bar{B}(x,r))} \int_{\bar{B}(x,r)} |f(z) - f(x)|\, d\mu(z) = 0, \text{ para } \mu\text{ - q.t } x \in \Omega.$$

Observação 2.5.2. O sobrenome *forte* na condição de Lebesgue-Besicovitch, é para diferenciar o tipo de convergência (para zero) da variável aleatória

$$\frac{1}{\mu(\bar{B}(X,r))} \int_{\bar{B}(X,r)} |f(z) - f(X)|\, d\mu(z).$$

Se a convergência é *quase certa*, temos a condição **(LBF)** para Ω e μ: $\forall f \in L^1(\mu)$,

$$\frac{1}{\mu(\bar{B}(X,r))} \int_{\bar{B}(X,r)} |f(z) - f(X)|\, d\mu(z) \xrightarrow[r\to 0^+]{q.c} 0,$$

e se a convergência é em *probabilidade*, temos a condição de Lebesgue-Besicovitch *fraca* para Ω e μ, **(LBFr)**: $\forall f \in L^1(\mu)$,

$$\frac{1}{\mu(\bar{B}(X,r))} \int_{\bar{B}(X,r)} |f(z) - f(X)|\, d\mu(z) \xrightarrow[r\to 0^+]{P} 0,$$

ou de maneira equivalente,

$$\forall \epsilon > 0, \lim_{r\to 0^+} \mu\left(\left\{x \in \Omega : \frac{1}{\mu(\bar{B}(x,r))} \int_{\bar{B}(x,r)} |f(z) - f(x)|\, d\mu(z) > \epsilon\right\}\right) = 0.$$

Ao longo do texto, usaremos apenas a condição **(LBF)**.

Finalmente estamos em condições de enunciar, sem prova, o resultado de Preiss que relaciona os espaços de dimensão σ-finita no sentido de Nagata com a condição **(LBF)**.

Teorema 2.5.1 (Preiss [54])**.** *Se Ω é um espaço métrico completo e separável, então Ω satisfaz a condição **(LBF)** para cada medida boreliana localmente finita, se e somente se, Ω possui dimensão σ-finita no sentido de Nagata.*

[4] Uma medida boreliana sobre o espaço Ω, μ, é localmente finita, se para todo $x \in \Omega$ existe um aberto $x \in O_x \subseteq \Omega$, tal que $\mu(O_x) < +\infty$.

Antes de finalizar o capítulo, vamos fazer alguns comentários sobre a prova do resultado acima. No artigo de 1983, Preiss não deu a prova como tal do Teorema 2.5.1, em lugar disso, ele apenas esboçou as ideias básicas da prova em umas poucas linhas.

No ano 2006 num artigo de 61 páginas [4], Assouad and Gromard, elaboram uma prova completa da *suficiência* para espaços métricos que possuem dimensão de Nagata *finita*, de onde deduzimos facilmente o caso de dimensão σ-finita (**(LBF)** $\Leftarrow$ dimensão σ-finita); e com relação à recíproca, a *necessidade*, ela teve uma prova completa elaborada no ano 2018 na tese doutoral [40, Kumari].

3 O CORPO DOS NÚMEROS p-ÁDICOS

Neste capítulo veremos uma breve introdução aos números p-ádicos. Sobre o corpo $(\mathbb{Q}, +, \cdot)$, a topologia usual e a p-ádica são definidas a partir de um *valor absoluto*. Para começar, veremos o conceito de valor absoluto em um corpo qualquer, obtendo algumas propriedades que são válidas no contexto geral, como o *completamento* de corpos com valor absoluto. Depois, junto com uma detalhada prova do *Teorema de Ostrowski*, será discutido o caso de valores absolutos no corpo $(\mathbb{Q}, +, \cdot)$ dos números racionais e seus completamentos, em particular o completamento que dá lugar aos números p-ádicos, para finalmente determinar a *representação canônica* dos elementos de $\mathbb{Q}_p$. As principais referências neste capítulo serão [29, 36, 56].

3.1 VALOR ABSOLUTO EM UM CORPO

Primeiro lembramos que um corpo $(\mathbb{K}, +, \cdot)$ é uma *estrutura algébrica* formada pelo conjunto $\mathbb{K}$ e duas operações binárias sobre ele com neutro aditivo $0_{\mathbb{K}}$ e neutro multiplicativo $1_{\mathbb{K}}$, onde $(\mathbb{K}, +)$ e $(\mathbb{K} \setminus \{0_{\mathbb{K}}\}, \cdot)$ são *grupos abelianos*. A partir de agora e para simplificar a escrita, vamos denotar um corpo qualquer $(\mathbb{K}, +, \cdot)$ por $\mathbb{K}$ e seus elementos, $1_{\mathbb{K}}$ e $0_{\mathbb{K}}$, serão denotados simplesmente por 1 e 0, respectivamente. Também, ao longo do texto vamos denotar o conjunto dos números naturais mais o número $0 \in \mathbb{Z}$, por $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Nesta seção, vamos expor alguns resultados válidos em um corpo $\mathbb{K}$ qualquer.

Definição 3.1.1 (Valor absoluto em $\mathbb{K}$)**.** Um valor absoluto no corpo $\mathbb{K}$, é uma função $|\cdot| : \mathbb{K} \to \mathbb{R}_+$ satisfazendo:

(i) $|x| = 0$ se e somente se $x = 0$.

(ii) $|xy| = |x||y|$, para todo $x, y \in \mathbb{K}$.

(iii) $|x + y| \leq |x| + |y|$, para todo $x, y \in \mathbb{K}$ (Desigualdade triangular).

Um valor absoluto é chamado de *não-arquimediano* se satisfaz a *desigualdade triangular forte*:

(iv) $|x + y| \leq \max\{|x|, |y|\}$, para todo $x, y \in \mathbb{K}$.

Se um valor absoluto não satisfaz (iv), então é chamado de valor absoluto *arquimediano*.

Exemplo 3.1.1. Se consideramos $\mathbb{K} = \mathbb{Q}$, é imediato verificar que o valor absoluto usual é de fato um valor absoluto em $\mathbb{Q}$ e que é *arquimediano*, pois basta tomar $x, y \in \mathbb{Q}$, tais que $xy > 0$ para verificar que (iv) não é satisfeita.

Exemplo 3.1.2. Para $\mathbb{K}$ corpo qualquer, considere a função $|\cdot|_0 : \mathbb{K} \to \mathbb{R}_+$ definida por $|0|_0 = 0$ e $|x|_0 = 1$ para todo $x \neq 0$. É imediato verificar que $|\cdot|_0$ é um valor absoluto não-arquimediano, chamado valor *absoluto trivial*.

Exemplo 3.1.3. Em $\mathbb{Q}$, também consideraremos valores absolutos que dependem de um número primo $p > 1$, denotados por $|\cdot|_p$ e chamados de valores absolutos p-ádicos. São valores absolutos não-arquimedianos e que como veremos nas próximas seções (Teorema 3.2.2 de Ostrowski), qualquer valor absoluto em $\mathbb{Q}$ será equivalente (as métricas induzidas pelos valores absolutos geram a mesma topologia) com o valor absoluto usual ou com o valor absoluto trivial ou será equivalente com um valor absoluto p-ádico, para algum número primo $p > 1$. Nos últimos dois casos, o valor absoluto será do tipo não-arquimediano.

3.1.1 Propriedades

Agora vamos revisar as principais propriedades básicas dos corpos com valor absoluto. Antes de continuar, é bom lembrar que em um corpo $(\mathbb{K}, +, \cdot)$ qualquer, para $n \in \mathbb{N}$, temos

$$n \cdot 1 = \underbrace{1 + \cdots + 1}_{n \text{ vezes}} \in \mathbb{K}$$

e denotaremos esse elemento simplesmente por n.

Observação 3.1.1. Na Álgebra, existe o conceito de *característica de um corpo* $\mathbb{K}$, e que simplesmente é o menor número $p \in \mathbb{N}$ tal que

$$p \cdot 1 = 0.$$

Se esse inteiro existe, necessariamente é um *número primo* e caso ele não existir, dizemos que a característica do corpo é 0 (zero).

Além do corpo dos números p-ádicos, os corpos mais conhecidos: $\mathbb{C}, \mathbb{R}, \mathbb{Q}$, são de (ou têm) característica 0 e o corpo quociente, $\mathbb{Z}/p\mathbb{Z}$, onde $p > 1$ é um número primo, possui característica p.

Existem exemplos interessantes de corpos com valor absoluto de característica $p \neq 0$, e um tópico de pesquisa futuro, pode ser estudar a aprendizagem estatística nesses espaços e descobrir/procurar propriedades úteis derivadas dessas particulares estruturas. Mais detalhes sobre álgebra elementar em [34].

Ao longo do texto, vamos considerar apenas corpos com característica zero.

Proposição 3.1.1 (Proposição 1.6 de [36])**.** *Para todo $x, y \in \mathbb{K}$, com $\mathbb{K}$ corpo com valor absoluto $|\cdot|$ e $y \neq 0$, temos:*

(1) $|1| = |-1| = 1$,

(2) $|x| = |-x|$,

(3) *Se* $|x^n| = 1$ *para algum* $n \in \mathbb{N}$, *então* $|x| = 1$,

(4) $|x/y| = |x|/|y|$,

(5) $|n| \leq n$, *para todo* $n \in \mathbb{N}$.

Demonstração. Sejam $x, y \in \mathbb{K}$ com $y \neq 0$, então

(1) $|1| = |1 \cdot 1| = |1|^2$, de onde $|1| = 1$. O resultado para -1 é análogo.

(2) $|-x| = |(-1) \cdot x| = 1 \cdot |x|$.

(3) Se para algum $n \in \mathbb{N}$ temos $1 = |x^n| = |x|^n$, extraindo raiz n-ésima, é claro que $|x| = 1$.

(4) Primeiro notamos que $1 = |y \cdot y^{-1}| = |y||y^{-1}|$, de onde $|y^{-1}| = 1/|y|$. Logo $|x/y| = |x \cdot y^{-1}| = |x||y^{-1}| = |x|/|y|$.

(5) Usando o item (1) e a desigualdade triangular, temos $|n| = |\underbrace{1 + \cdots + 1}_{n \text{ vezes}}| \leq n \cdot |1| = n$.

■

Agora, veremos uma condição necessária e suficiente para que um valor absoluto em um corpo seja *não-arquimediano*.

Proposição 3.1.2 (Proposição 2.2.2 de [29])**.** *Seja* $\mathbb{K}$ *um corpo com valor absoluto* $|\cdot|$. *Então o valor absoluto* $|\cdot|$ *é não-arquimediano se e somente se* $|n| \leq 1$, *para todo* $n \in \mathbb{Z}$.

Demonstração. ($\Rightarrow$). Suponha que $|\cdot|$ é *não-arquimediano*. Provaremos por indução que $|n| \leq 1$ para todo $n \in \mathbb{Z}$.
Base da indução: $|1| = 1 \leq 1$.
Hipótese de indução: Suponha que $|m| \leq 1$, para todo $m \in \{1, 2, ..., n-1\}$.
Vamos provar que $|n| \leq 1$. Pela desigualdade $|1| \leq 1$ e a hipótese de indução, obtemos $|n| = |(n-1)+1| \leq \max\{|n-1|, |1|\} = 1$ e assim $|n| \leq 1$, para todo $n \in \mathbb{N}$. Dado que $|-n| = |n|$, obtemos $|n| \leq 1$, para todo $n \in \mathbb{Z}$.
($\Leftarrow$) Suponha agora que $|n| \leq 1$, para todo $n \in \mathbb{Z}$ e sejam $x, y \in \mathbb{K}$ arbitrários, vamos provar que $|x+y| \leq \max\{|x|, |y|\}$. Se $x = 0$ ou $y = 0$, a desigualdade é claramente satisfeita. Suponha então que $0 \neq x, y \in \mathbb{K}$. Dividindo a desigualdade por $|y|$, temos,

$$|x+y| \leq \max\{|x|, |y|\}, \forall\, 0 \neq x, y \in \mathbb{K} \Leftrightarrow \left|\frac{x}{y} + 1\right| \leq \max\left\{\left|\frac{x}{y}\right|, 1\right\}, \forall\, 0 \neq x, y \in \mathbb{K}$$

e também claramente temos,

$$\left|\frac{x}{y} + 1\right| \leq \max\left\{\left|\frac{x}{y}\right|, 1\right\}, \forall\, 0 \neq x, y \in \mathbb{K} \Leftrightarrow |z+1| \leq \max\{|z|, 1\}, \forall\, 0 \neq z \in \mathbb{K}.$$

Assim, para finalizar a demonstração é suficiente provar a última desigualdade.
Sejam $n \in \mathbb{N}$ e $0 \neq z \in \mathbb{K}$, quaisquer. Utilizando o Teorema do Binômio de Newton, temos

$$\begin{aligned} |z+1|^n = |(z+1)^n| &= \left| \sum_{k=0}^{n} \binom{n}{k} z^k \right| \\ &\leq \sum_{k=0}^{n} \left| \binom{n}{k} \right| |z|^k \\ &\leq \sum_{k=0}^{n} |z|^k, \quad \text{pois } \binom{n}{k} \in \mathbb{N} \\ &\leq (n+1) \max\{|z|^n, 1\}. \end{aligned}$$

Finalmente, extraindo raiz n-ésima, obtemos

$$|z+1| \leq \sqrt[n]{n+1} \max\{|z|, 1\}$$

e fazendo n tender a infinito, temos $\lim_{n\to\infty} \sqrt[n]{n+1} = 1$ e

$$|z+1| \leq \max\{|z|, 1\}.$$

■

Observação 3.1.2. A proposição acima, vai ser útil para entender a diferença entre valores absolutos arquimedianos e não-arquimedianos. O resultado anterior pode se escrever da seguinte maneira: O valor absoluto $|\cdot|$ é arquimediano se e só se satisfaz a propriedade arquimediana: Dados $x, y \in \mathbb{K}$, $x \neq 0$, existe um inteiro $n \in \mathbb{N}$ tal que $|nx| > |y|$.

Com efeito, suponha que o valor absoluto $|\cdot|$ é arquimediano. Então, existe um inteiro $n_0 \in \mathbb{N}$ tal que $|n_0| > 1$ e assim obtemos que $|n_0|^m \to \infty$ quando $m \to \infty$, de onde inferimos que existe um inteiro $m_0 \in \mathbb{N}$ tal que $|n_0^{m_0}| > |y|/|x|$ ou $|n_0^{m_0} x| > |y|$, isto é, o valor absoluto $|\cdot|$ satisfaz a propriedade arquimediana com o inteiro $n_0^{m_0}$. Reciprocamente, suponha que o valor absoluto $|\cdot|$ satisfaz a propriedade arquimediana e sejam $x, y \in \mathbb{K}$ tais que $x \neq 0$ e $|y| > |x|$. Então, existe um inteiro $n_0 \in \mathbb{N}$ tal que $|n_0| > |y|/|x| > 1$, isto é, o valor absoluto $|\cdot|$ é arquimediano.

3.1.2 Topologia

Assim como nos espaços vetoriais as normas induzem métricas, no caso dos corpos com valor absoluto temos a mesma situação. A prova da seguinte proposição é imediata.

Proposição 3.1.3. *Seja $\mathbb{K}$ um corpo com valor absoluto $|\cdot|$. A função $\rho : \mathbb{K} \times \mathbb{K} \to \mathbb{R}_+$ definida para $x, y \in \mathbb{K}$ mediante*

$$\rho(x, y) = |x - y|,$$

é uma métrica em $\mathbb{K}$.

Quando o valor absoluto $|\cdot|$ é não-arquimediano, é imediato verificar que a métrica induzida também satisfaz a desigualdade triangular forte da Definição 2.5.3, isto é, para $x, y, z \in \mathbb{K}$

$$\rho(x, z) \leq \max\{\rho(x, y), \rho(y, z)\},$$

e portanto, um corpo com valor absoluto não-arquimediano é um espaço *ultramétrico*.

A seguir, veremos que a desigualdade triangular forte, dota os espaços não-arquimedianos de algumas propriedades especiais.

Proposição 3.1.4 (Proposição 2.3.2 de [29])**.** *Seja $\mathbb{K}$ um corpo dotado de um valor absoluto não-arquimediano $|\cdot|$. Se $x, y \in \mathbb{K}$ são tais que $|x| \neq |y|$, então*

$$|x + y| = \max\{|x|, |y|\}.$$

Demonstração. Como $x, y \in \mathbb{K}$ são tais que $|x| \neq |y|$, primeiro vamos supor $|x| > |y|$, logo

$$|x + y| \leq \max\{|x|, |y|\} = |x|$$

e além disso, como $x = (x + y) - y$, temos:

$$|x| \leq \max\{|x + y|, |y|\}.$$

Como $|x| > |y|$, pela desigualdade acima, $|y|$ não pode ser o $\max\{|x + y|, |y|\}$, logo necessariamente $\max\{|x + y|, |y|\} = |x + y|$ e assim $|x| \leq |x + y|$, de onde obtemos $|x| = |x + y|$. O caso $|y| > |x|$ é análogo. ■

Um corolário imediato é o seguinte

Corolário 3.1.0.1. *Em um espaço (métrico) ultramétrico, todos os triângulos são isósceles e o comprimento da base não supera o comprimento dos lados.*

Agora, veremos alguns conceitos de espaços métricos no contexto da métrica induzida por um valor absoluto.

Definição 3.1.2 (Bolas abertas e bolas fechadas)**.** Sejam $\mathbb{K}$ um corpo com valor absoluto $|\cdot|$, $a \in \mathbb{K}$ e $r \in \mathbb{R}_+$. Definimos a bola aberta centrada em a de raio r por

$$B(a, r) = \{x \in \mathbb{K} : |x - a| < r\}$$

e a bola fechada centrada em a de raio r por

$$\bar{B}(a, r) = \{x \in \mathbb{K} : |x - a| \leq r\}.$$

Quando o valor absoluto é não-arquimediano, as bolas ganham propriedades que não vemos no caso arquimediano.

Proposição 3.1.5 (Proposição 2.3.4 de [29]). *Seja $\mathbb{K}$ um corpo dotado de um valor absoluto não-arquimediano $|\cdot|$. Então, para $r, s > 0$, $a, b \in \mathbb{K}$ quaisquer, temos*

(i) *Se $b \in B(a, r)$, então $B(b, r) = B(a, r)$.*

(ii) *$B(a, r) = B(b, r)$ ou $B(a, r) \cap B(b, r) = \emptyset$.*

(iii) *O conjunto $B(a, r)$ é aberto e fechado na topologia induzida por $|\cdot|$.*

(iv) *$B(a, r) \cap B(b, s) \neq \emptyset \Leftrightarrow B(a, r) \subset B(b, s)$ ou $B(b, s) \subset B(a, r)$.*

As propriedades (i)-(iv) também valem para as bolas fechadas.

Demonstração. Sejam $a, b \in \mathbb{K}$, então:

(i) Seja $r > 0$ e suponha $b \in B(a, r)$. Isto último é equivalente a dizer que $|b - a| < r$. Seja $x \in B(b, r)$, então $|x - b| < r$, logo

$$|x - a| \leq \max\{|x - b|, |b - a|\} < r$$

isto é, $x \in B(a, r)$ e assim $B(b, r) \subset B(a, r)$. De forma absolutamente análoga, provamos que $B(a, r) \subset B(b, r)$, de onde temos a igualdade. Como veremos mais para frente, o resultado em (i) em um fato chave na hora de formular um algoritmo de classificação no corpo não-arquimediano dos números p-ádicos $\mathbb{Q}_p$.

(ii) Se $r > 0$ e $B(a, r) \cap B(b, r) \neq \emptyset$, então existe $z \in B(a, r) \cap B(b, r)$, logo, por (i), temos

$$B(a, r) = B(z, r) = B(b, r).$$

(iii) Seja $r > 0$, então $B(a, r)$ é um conjunto aberto em qualquer espaço métrico e portanto apenas falta provar que a bola aberta é um conjunto fechado, ou seja $\overline{B(a, r)} = B(a, r)$. Seja $x \in \overline{B(a, r)}$, então toda bola centrada em x terá interseção não vazia com $B(a, r)$. Seja $0 < s \leq r$, como $B(x, s) \cap B(a, r) \neq \emptyset$, temos que existe $b \in B(x, s) \cap B(a, r)$, que satisfaz as condições

$$|b - a| < r \text{ e } |b - x| < s \leq r.$$

Com isto, mais a desigualdade triangular forte, temos

$$|x - a| \leq \max\{|x - b|, |b - a|\} < r,$$

ou seja, $x \in B(a, r)$ e assim $\overline{B(a, r)} = B(a, r)$.

(iv) ($\Rightarrow$) Sem perda de generalidade, podemos supor $r \leq s$. Se $B(a, r) \cap B(b, s) \neq \emptyset$, então existe $c \in B(a, r) \cap B(b, s)$ tal que, pelo item (i),

$$B(c, r) = B(a, r) \text{ e } B(c, s) = B(b, s).$$

Logo

$$B(a,r) = B(c,r) \subset B(c,s) = B(b,s)$$

e supondo $r > s$, obtemos a outra inclusão.

($\Leftarrow$) Trivial, porque a interseção será uma das duas bolas que são não vazias.

Finalmente, trocando $<$ por $\leq$ nos argumentos dos itens anteriores, obtemos os mesmos resultados para as bolas fechadas. ■

Como estamos interessados principalmente na topologia induzida por um valor absoluto, a seguir definimos o conceito de equivalência entre valores absolutos em um corpo, mas antes de ir com a definição, relembramos que duas métricas ρ_1 e ρ_2 sobre um conjunto Ω são equivalentes, o que é denotado por $\rho_1 \sim \rho_2$; se elas induzem a mesma topologia sobre Ω.

Definição 3.1.3 (Equivalência de valores absolutos sobre um corpo)**.** Dizemos que dois valores absolutos $|\cdot|_1$ e $|\cdot|_2$ sobre um corpo $\mathbb{K}$ são equivalentes, o que vamos denotar por $|\cdot|_1 \sim |\cdot|_2$, se eles induzem métricas equivalentes.

Algumas propriedades da equivalência de valores absolutos em um corpo que serão úteis na próxima proposição que caracteriza os valores absolutos equivalentes, são dadas no seguinte lema.

Lema 3.1.1. *Sejam $|\cdot|_1$ e $|\cdot|_2$, valores absolutos sobre o corpo $\mathbb{K}$. Se $|\cdot|_1 \sim |\cdot|_2$, então, para $x \in \mathbb{K}$, temos*

$$|x|_1 < 1 \Leftrightarrow |x|_2 < 1,$$

$$|x|_1 = 1 \Leftrightarrow |x|_2 = 1,$$

$$|x|_1 > 1 \Leftrightarrow |x|_2 > 1.$$

Proposição 3.1.6 (Proposição 1.10 de [36])**.** *Sejam $|\cdot|_1$ e $|\cdot|_2$, valores absolutos sobre o corpo $\mathbb{K}$. Então $|\cdot|_1 \sim |\cdot|_2$ se e somente se, existe um número real $\alpha > 0$, tal que*

$$|x|_2 = |x|_1^{\alpha}, \ \forall x \in \mathbb{K}. \tag{3}$$

Demonstração. ($\Rightarrow$) Suponha que $|\cdot|_1 \sim |\cdot|_2$. Se $|\cdot|_1$ é trivial, então não é difícil verificar que $|\cdot|_2$ também é trivial e assim (3) é satisfeita para qualquer $\alpha > 0$. Se $|\cdot|_1$ é não trivial, então existe um $z \in \mathbb{K}$ tal que $|z|_1 \neq 1$. Trocando z por z^{-1} se for necessário, podemos supor que $|z|_1 < 1$. Defina

$$\alpha = \frac{\ln(|z|_2)}{\ln(|z|_1)},$$

onde ln é a função logaritmo natural. Do Lema 3.1.1, como os valores absolutos são equivalentes, temos $|z|_2 < 1$ e assim ambos logaritmos do quociente que define α são negativos, portanto $\alpha > 0$. Agora vamos provar que α satisfaz (3). Vamos considerar o

caso de $x \in \mathbb{K}$ tal que $|x| < 1$, pois os casos $|x| = 1$ e $|x| > 1$, seguem do Lema 3.1.1. Também, considere o conjunto

$$Q = \{q = n/m \in \mathbb{Q} : n, m \in \mathbb{N}, |x|_1^q < |z|_1\}. \tag{4}$$

Para todo $q \in Q$,

$$|x|_1^m < |z|_1^n \Rightarrow \left|\frac{x^m}{z^n}\right|_1 < 1.$$

Assim, pelo Lema 3.1.1,

$$\left|\frac{x^m}{z^n}\right|_2 < 1,$$

de onde $|x|_2^m < |z|_2^n$ e $|x|_2^q < |z|_2$. O mesmo argumento é válido se intercambiamos os papeis de $|\cdot|_1$ e $|\cdot|_2$, obtendo assim que o conjunto Q, pode se escrever como

$$Q = \{q = n/m \in \mathbb{Q} : n, m \in \mathbb{N}, |x|_2^q < |z|_2\}. \tag{5}$$

Tomando logaritmos nas condições do conjunto Q nas expressões (4) e (5), obtemos que as condições

$$q > \frac{\ln(|z|_1)}{\ln(|x|_1)}, \quad q > \frac{\ln(|z|_2)}{\ln(|x|_2)} \tag{6}$$

devem ser ambas satisfeitas e elas estão bem definidas pois todos os logaritmos envolvidos são negativos. Assim, necessariamente devemos ter

$$\frac{\ln(|z|_1)}{\ln(|x|_1)} = \frac{\ln(|z|_2)}{\ln(|x|_2)},$$

pois caso contrário, teriamos que existe um número racional entre esses dois números e assim apenas uma das condições (6) seria satisfeita. Finalmente, definimos $\alpha > 0$ por

$$\alpha := \frac{\ln(|z|_1)}{\ln(|x|_1)} = \frac{\ln(|z|_2)}{\ln(|x|_2)}$$

e depois de um rápido cálculo, verificamos que $\alpha > 0$ assim definido, satisfaz (3).

($\Leftarrow$) Agora, suponha que existe $\alpha > 0$ tal que $|x|_2 = |x|_1^\alpha$, $\forall x \in \mathbb{K}$. Então, para todo $a \in \mathbb{K}$ e $r > 0$, temos

$$|x - a|_2 < r \Leftrightarrow |x - a|_1 < r^{\frac{1}{\alpha}}$$

ou seja, $B_2(a, r) = B_1(a, r^{\frac{1}{\alpha}})$ e assim, ambos valores absolutos geram a mesma topologia e portanto são equivalentes. ■

3.1.3 Construção do completamento de um corpo com valor absoluto

Nessa seção, vamos estudar o processo para completar corpos com valor absoluto, estudo que será feito com detalhes, porque a completude do corpo $\mathbb{Q}_p$ será importante para

que o algoritmo que desenvolveremos no próximo capítulo: esteja *bem definido*, *funcione*, e seja *universalmente consistente*. O nosso caso de interesse, é o completamento do corpo com valor absoluto $(\mathbb{Q}, |\cdot|_p)$, mas como o procedimento é padrão, estudaremos o completamento para um corpo com valor absoluto qualquer.

Primeiro, lembremos o conceito de sequência de Cauchy num espaço métrico.

Definição 3.1.4. Uma sequência $\{a_n\}_{n\in\mathbb{N}} \subset \Omega$ num espaço métrico (Ω, ρ), é uma sequência de Cauchy, se

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} : \rho(a_m, a_n) < \epsilon,\ \forall m, n \geq N$$

ou de forma equivalente

$$\lim_{m,n\to\infty} \rho(a_m, a_n) = 0.$$

As sequências de Cauchy, são sequências $\{a_n\}_{n\in\mathbb{N}}$ tais que a medida que $n, m \in \mathbb{N}$ crescem os termos a_n, a_m da sequência ficam cada vez mais próximos. As sequências convergentes também possuem esse comportamento, pois os termos ficam próximos do limite quando n cresce e portanto ficam próximos uns dos outros, assim as sequências convergentes também são sequências de Cauchy. Por outro lado, em alguns espaços métricos, como por exemplo $(\mathbb{Q}, |\cdot|_\infty)$ com $|\cdot|_\infty$ o valor absoluto usual (mais para frente vamos saber o significado do ∞ na notação), nem toda sequência de Cauchy é convergente, assim, se num espaço métrico, *toda sequência de Cauchy é convergente*, então dizemos que esse espaço métrico é *completo*.

Como já vimos, um corpo com valor absoluto é um caso particular de espaço métrico e portanto eles podem ou não ser completos. A partir de um corpo com valor absoluto $(\mathbb{K}, |\cdot|)$ não necessariamente completo com relação à métrica induzida pelo valor absoluto, vamos construir um outro corpo que denotamos por $\widehat{\mathbb{K}}$, contendo $\mathbb{K}$ e dotado de um valor absoluto induzido pelo valor absoluto $|\cdot|$, tal que $\widehat{\mathbb{K}}$ seja um corpo com valor absoluto *completo*. Dos cursos elementares de análise, sabemos que no caso do conjunto dos números racionais munido do valor absoluto usual, $|\cdot|_\infty$, o completamento que obtemos é o corpo com valor absoluto $(\mathbb{R}, |\cdot|_\infty)$. Nas próximas seções, aplicaremos o mesmo procedimento geral de completamento com o conjunto dos números racionais, mas dessa vez dotado de outro tipo de valores absolutos, os valores absolutos p-ádicos: $|\cdot|_p$.

No procedimento padrão para completar espaços métricos, o conceito central é o de *sequência de Cauchy*, pois os elementos de $\widehat{\mathbb{K}}$ são precisamente classes de equivalência de sequências de Cauchy em $\mathbb{K}$. No caso de corpos com valor absoluto, uma sequência de Cauchy é uma sequência que satisfaz:

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} : |a_m - a_n| < \epsilon,\ \forall m, n \geq N$$

ou de forma equivalente

$$\lim_{m,n\to\infty} |a_m - a_n| = 0.$$

No caso de um corpo com valor absoluto não-arquimediano, verificar se uma sequência é de Cauchy, é um pouco mais simples.

Proposição 3.1.7 (Lema 3.2.1 de [29]). *Seja* $\{a_n\}_{n\in\mathbb{N}} \subset \mathbb{K}$ *uma sequência no corpo com valor absoluto não-arquimediano* $(\mathbb{K}, |\cdot|)$. *Então,* $\{a_n\}_{n\in\mathbb{N}}$ *é uma sequência de Cauchy se e somente se*

$$\lim_{n\to\infty} |a_{n+1} - a_n| = 0.$$

Demonstração. Usando a desigualdade triangular forte recursivamente, para $m, n \in \mathbb{N}$, temos,

$$|a_m - a_n| \leq \max\{|a_{n+1} - a_n|, |a_{n+2} - a_{n+1}|, \cdots, |a_{n+r} - a_{n+r-1}|\}$$

de onde vemos que $\{a_n\}_{n\in\mathbb{N}}$ é de Cauchy, se e somente se, $\lim_{n\to\infty} |a_{n+1} - a_n|_p = 0$ ■

No seguinte exemplo, vemos que a proposição acima não vale no caso *arquimediano*.

Exemplo 3.1.4. Considere a sequência de *somas parciais* $S = \{s_n\}_{n\in\mathbb{N}}$, definida por:

$$s_n = \sum_{i=1}^{n} \frac{1}{i}.$$

É claro que S não é convergente, pois a série $\sum_{i=1}^{\infty} \frac{1}{i}$ diverge, assim, da completude de $\mathbb{R}$, inferimos que S não é uma sequência de Cauchy. Mas, por outro lado:

$$\lim_{n\to\infty} |s_{n+1} - s_n|_\infty = \lim_{n\to\infty} \frac{1}{n+1} = 0,$$

e portanto, temos que a Proposição 3.1.7 não vale no caso *arquimediano*.

É fácil conferir que a soma $\{a_n\}_{n\in\mathbb{N}} \pm \{b_n\}_{n\in\mathbb{N}} = \{a_n \pm b_n\}_{n\in\mathbb{N}}$ assim como o produto $\{a_n\}_{n\in\mathbb{N}} \cdot \{b_n\}_{n\in\mathbb{N}} = \{a_n \cdot b_n\}_{n\in\mathbb{N}}$ de sequências de Cauchy, são novamente sequências de Cauchy, portanto, o conjunto de sequências de Cauchy em $\mathbb{K}$, denotado por $\{\mathbb{K}\}$, é um anel comutativo cujo elemento neutro aditivo é a sequência de zeros:

$$\widehat{0} := \{0, 0, 0, ...\}$$

e o neutro multiplicativo é a sequência de uns:

$$\widehat{1} := \{1, 1, 1, ...\}.$$

Ainda, o anel $\{\mathbb{K}\}$ não é um corpo, pois possui divisores do zero. Para ver isso, por exemplo considere as sequências convergentes(e portanto de Cauchy)

$$\{0, 0, 1, 0, ...\} \cdot \{0, 1, 0, 0, ...\} = \widehat{0}.$$

Para cada $a \in \mathbb{K}$, como no caso $\widehat{0}$ e $\widehat{1}$, definimos:

$$\widehat{a} := \{a, a, a, ...\}$$

que ao ser uma sequência convergente, é de Cauchy, e assim $\{\mathbb{K}\}$ contém um subanel isomorfo com $\mathbb{K}$. Um conjunto importante no procedimento padrão de completamento de espaços métricos, é o conjunto $\mathcal{N}$ de sequências nulas.

Definição 3.1.5. Seja $\mathbb{K}$ um corpo com valor absoluto $|\cdot|$. Uma sequência nula em $\mathbb{K}$, é uma sequência $\{a_n\}_{n\in\mathbb{N}}$ satisfazendo:

$$\lim_{n\to\infty} |a_n| = 0.$$

O conjunto de sequências nulas em $\mathbb{K}$, será denotado por $\mathcal{N}$:

$$\mathcal{N} = \{\{a_n\}_{n\in\mathbb{N}} \subset \mathbb{K} : \lim_{n\to\infty} |a_n| = 0\}.$$

Primeiro, um par de propriedades básicas de $\mathcal{N}$.

Proposição 3.1.8. *Nas condições acima:*

(i) $\mathcal{N} \subset \{\mathbb{K}\}$.

(ii) $\mathcal{N}$ *é um ideal bilateral de* $\{\mathbb{K}\}$.

Demonstração. **(i)** Seja $\{a_n\}_{n\in\mathbb{N}} \in \mathcal{N}$. Como $\lim_{n\to\infty} |a_n| = 0$ e temos $|a_m - a_n| \leq |a_m| + |a_n|$, vemos que $\lim_{m,n\to\infty} |a_m - a_n| = 0$, de onde $\{a_n\}_{n\in\mathbb{N}} \in \{\mathbb{K}\}$.

(ii) Primeiro lembramos que $\mathcal{N}$ é um ideal(unilateral esquerdo) de $\{\mathbb{K}\}$, se é um subanel de $\{\mathbb{K}\}$ tal que para cada $x \in \mathcal{N}$ e $a \in \{\mathbb{K}\}$, temos $xa \in \mathcal{N}$. No nosso caso, se $\{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}} \in \mathcal{N}$, é claro que $\{a_n \pm b_n\}_{n\in\mathbb{N}}, \{a_n b_n\}_{n\in\mathbb{N}} \in \mathcal{N}$, de onde vemos que é um subanel e que se $\{a_n\}_{n\in\mathbb{N}} \in \mathcal{N}$ e $\{b_n\}_{n\in\mathbb{N}}$ é uma sequência limitada(em particular quando é de Cauchy), então $\{a_n b_n\}_{n\in\mathbb{N}} \in \mathcal{N}$, sendo assim $\mathcal{N}$ um ideal bilateral por causa da comutatividade em $\mathbb{K}$.

■

Chamemos $\widehat{\mathbb{K}} := \{\mathbb{K}\}/\mathcal{N}$ o anel quociente. Os elementos de $\widehat{\mathbb{K}}$, são *classes de equivalência* de elementos de $\{\mathbb{K}\}$ e duas sequências de Cauchy serão *equivalentes* se a diferença das sequências é uma sequência nula, isto é, $\{a_n\}_{n\in\mathbb{N}} \sim \{b_n\}_{n\in\mathbb{N}}$ se $\lim_{n\to\infty} |a_n - b_n| = 0$ e denotaremos a classe de equivalência de uma sequência de Cauchy $\{a_n\}_{n\in\mathbb{N}}$ por $[\{a_n\}]$. Finalmente, vamos considerar $\mathbb{K}$ como um subconjunto de $\widehat{\mathbb{K}}$ identificando $a \in \mathbb{K}$ com a classe de equivalência $[\widehat{a}] \in \widehat{\mathbb{K}}$.

Teorema 3.1.2 (Teorema 1.19 de [36])**.** $\widehat{\mathbb{K}}$ *é um corpo.*

Demonstração. Em $\widehat{\mathbb{K}}$, defina as operações sobre as classes:

$$[\{a_n\}] + [\{b_n\}] := [\{a_n + b_n\}]$$

e

$$[\{a_n\}] \cdot [\{b_n\}] = [\{a_n \cdot b_n\}].$$

É fácil conferir que a definição acima não depende do representante de cada classe e que as novas operações herdam as boas propriedades das operações sobre $\mathbb{K}$, assim com essas operações sobre as classes, $\widehat{\mathbb{K}}$ é um anel comutativo com neutro aditivo $[\widehat{0}]$ e neutro multiplicativo $[\widehat{1}]$. Portanto, para provar que $\widehat{\mathbb{K}}$ é um corpo, só falta mostrar que para toda classe não nula existe uma classe que é seu inverso multipicativo. Seja $\mathcal{C} \in \widehat{\mathbb{K}}$ uma classe de equivalência não nula, isto é, $\mathcal{C} \neq [\widehat{0}] = \mathcal{N}$ e seja $\{a_n\}_{n\in\mathbb{N}} \in \mathcal{C}$ uma sequência de Cauchy que vamos utilizar como respresentante da classe. Como $\lim_{n\to\infty} |a_n| \neq 0$, podemos encontrar um $\delta > 0$ e uma subsequência $\{a_{n_k}\}_{k\in\mathbb{N}}$ de $\{a_n\}_{n\in\mathbb{N}}$ tais que:

$$|a_{n_k}| \geq \delta, \ \forall k \in \mathbb{N}.$$

A sequência $\{a_{n_k}\}_{k\in\mathbb{N}}$, por ser uma subsequência de uma sequência de Cauchy, também é uma sequência de Cauchy e ainda: $\{a_{n_k}\}_{k\in\mathbb{N}} \sim \{a_k\}_{k\in\mathbb{N}}$. Com efeito, dado $\epsilon > 0$, como $\{a_n\}_{n\in\mathbb{N}}$ é de Cauchy,

$$\exists N \in \mathbb{N} : |a_k - a_{n_k}| < \epsilon, \ \forall k, n_k \geq N$$

e como é claro que para todo $k \in \mathbb{N}$, $n_k \geq k$, inferimos que para todo $k \geq N$, temos $|a_k - a_{n_k}| < \epsilon$, ou seja

$$\lim_{k\to\infty} |a_k - a_{n_k}| = 0.$$

Pelo fato anterior, podemos usar a subsequência como representante da classe, isto é, $\mathcal{C} = [\{a_k\}] = [\{a_{n_k}\}]$. Para cada $k \in \mathbb{N}$, defina $\widetilde{a}_k = a_{n_k}^{-1}$, então a sequência $\{\widetilde{a}_k\}_{k\in\mathbb{N}}$ é uma sequência de Cauchy. Com efeito, para $k, l \in \mathbb{N}$, temos

$$0 \leq |\widetilde{a}_k - \widetilde{a}_l| = \left|a_{n_k}^{-1} - a_{n_l}^{-1}\right| = \frac{|a_{n_k} - a_{n_l}|}{|a_{n_k}||a_{n_l}|} \leq \frac{|a_{n_k} - a_{n_l}|}{\delta^2}$$

e tomando limite

$$0 \leq \lim_{k,l\to\infty} |\widetilde{a}_k - \widetilde{a}_l| \leq \lim_{k,l\to\infty} \frac{|a_{n_k} - a_{n_l}|}{\delta^2} = 0,$$

pois $\{a_{n_k}\}_{k\in\mathbb{N}}$ é de Cauchy. Assim, $\{\widetilde{a}_k\}_{k\in\mathbb{N}} \in \{\mathbb{K}\}$ e denotaremos a classe dela por $\mathcal{D} = [\{\widetilde{a}_k\}] \in \widehat{K}$, logo

$$\mathcal{D} \cdot \mathcal{C} = \mathcal{C} \cdot \mathcal{D} = [\{a_{n_k} \cdot \widetilde{a}_k\}] = [\{1, 1, 1, ...\}] = [\widehat{1}],$$

ou seja $\mathcal{D} = \mathcal{C}^{-1}$ e a prova está completa. ■

Agora vamos estender o valor absoluto $|\cdot|$ de $\mathbb{K}$ para $\widehat{K}$.

Definição 3.1.6. Para $\mathcal{C} \in \widehat{K}$, defina

$$|\mathcal{C}| = \lim_{n\to\infty} |a_n|$$

onde $\{a_n\}_{n\in\mathbb{N}}$ é uma sequência de Cauchy em $\mathcal{C}$.

É um exercício simples mostrar que o limite acima sempre existe e que não depende do representante da classe. Com esse fato, podemos provar:

Proposição 3.1.9 (Proposição 1.21 de [36])**.** $|\cdot|$ *é um valor absoluto em* $\widehat{K}$*.*

Demonstração. Precisamos provar as propriedades da Definição 3.1.1.

(i) Se $\mathcal{C} = \widehat{[0]}$, então qualquer representante vai ser uma sequência nula, logo $|\mathcal{C}| = 0$. Se $\mathcal{C} \neq \widehat{[0]}$ e como vimos na prova do Teorema 3.1.2, podemos encontrar uma representante da classe $\{a_n\}_{n\in\mathbb{N}}$ e um número real $\delta > 0$, tais que:

$$|a_n| \geq \delta, \ \forall n \in \mathbb{N}.$$

Assim, $|\mathcal{C}| = \lim_{n\to\infty} |a_n| \geq \delta > 0$.

(ii) Sejam $\mathcal{C} = [\{a_n\}]$ e $\mathcal{D} = [\{b_n\}]$ elementos de $\widehat{\mathbb{K}}$. Pelas propriedades dos limites nos números reais, temos:

$$\begin{aligned} |\mathcal{C}\cdot\mathcal{D}| &= \lim_{n\to\infty} |a_n b_n| = \lim_{n\to\infty} |a_n||b_n| \\ &= \lim_{n\to\infty} |a_n| \lim_{n\to\infty} |b_n| = |\mathcal{C}||\mathcal{D}|. \end{aligned}$$

De forma análoga obtemos a propriedade da desigualdade triangular.

■

Finalmente, agora que já provamos que $(\widehat{\mathbb{K}}, |\cdot|)$ é um corpo com valor absoluto, vamos provar que $\widehat{\mathbb{K}}$ é de fato, completo.

Teorema 3.1.3 (Teorema 1.22 de [36])**.** $\widehat{\mathbb{K}}$ *é completo com relação ao valor absoluto* $|\cdot|$ *e* $\mathbb{K}$ *é um subconjunto denso em* $\widehat{\mathbb{K}}$*.*

Demonstração. Começamos provando a segunda parte do Teorema, pois ela vai ser utilizada na prova da primeira parte. Seja $\mathcal{C} \in \widehat{\mathbb{K}}$ e $\{a_k\}_{k\in\mathbb{N}}$ uma sequência de Cauchy em $\mathbb{K}$ representando $\mathcal{C}$. Para cada $n \in \mathbb{N}$ fixo, considere a sequência constante $\widehat{a_n} = \{a_n, a_n, ...\}$. Então, a sequência $\{a_k - a_n\}_{k\in\mathbb{N}}$ representa a classe $\mathcal{C} - [\widehat{a_n}]$ e lembrando que $\{a_k\}_{k\in\mathbb{N}}$ é de Cauchy, temos

$$\lim_{n\to\infty} |\mathcal{C} - [\widehat{a_n}]| = \lim_{k,n\to\infty} |a_k - a_n| = 0$$

provando assim, que $\mathbb{K}$ é um subconjunto denso de $\widehat{\mathbb{K}}$.

Agora considere a sequência de Cauchy em $\widehat{\mathbb{K}}$, $\{\mathcal{C}_n\}_{n\in\mathbb{N}} = \{\mathcal{C}_1, \mathcal{C}_2, ...\}$. Como $\mathbb{K}$ é denso em $\widehat{\mathbb{K}}$, para cada $\mathcal{C}_n$ existe um elemento $a_n \in \mathbb{K}$ tal que

$$|\mathcal{C}_n - [\widehat{a_n}]| < \frac{1}{n}, \tag{7}$$

ou seja, $\lim_{n\to\infty} |\mathcal{C}_n - [\widehat{a_n}]| = 0$, de onde vemos que $\{\mathcal{C}_n - [\widehat{a_n}]\}_{n\in\mathbb{N}}$ é uma sequência nula e portanto é de Cauchy em $\widehat{\mathbb{K}}$. Observe que

$$\{[\widehat{a_n}]\}_{n\in\mathbb{N}} = \{\mathcal{C}_n\}_{n\in\mathbb{N}} - \{\mathcal{C}_n - [\widehat{a_n}]\}_{n\in\mathbb{N}} \tag{8}$$

o que mostra que $\{[\widehat{a_n}]\}_{n\in\mathbb{N}}$ é uma sequência de Cauchy em $\widehat{\mathbb{K}}$, mas como todos os elementos pertencem a $\mathbb{K}$, a própria $\{a_n\}_{n\in\mathbb{N}}$ é uma sequência de Cauchy em $\mathbb{K}$. Denotemos a classe de equivalência de $\{a_n\}_{n\in\mathbb{N}}$ por $\mathcal{C}$. De (7) e (8), segue que $\{\mathcal{C} - [\widehat{a_n}]\}_{n\in\mathbb{N}}$ e $\{\mathcal{C}_n - [\widehat{a_n}]\}_{n\in\mathbb{N}}$, são sequências nulas em $\widehat{\mathbb{K}}$, e assim a diferença delas

$$\{\mathcal{C} - \mathcal{C}_n\}_{n\in\mathbb{N}} = \{\mathcal{C} - [\widehat{a_n}]\}_{n\in\mathbb{N}} - \{\mathcal{C}_n - [\widehat{a_n}]\}_{n\in\mathbb{N}}$$

também é uma sequência nula em $\widehat{\mathbb{K}}$, isto é

$$\lim_{n\to\infty} |\mathcal{C} - \mathcal{C}_n| = 0$$

o que é equivalente a dizer que $\mathcal{C} = \lim_{n\to\infty} \mathcal{C}_n$ e assim, a sequência $\{\mathcal{C}_n\}_{n\in\mathbb{N}}$ é convergente em $\widehat{\mathbb{K}}$. ■

3.1.4 Álgebra em corpos com valor absoluto não-arquimediano

Na hora de trabalhar com os números p-ádicos, vamos utilizar as propriedades topológicas mais do que as algébricas desses números, mas no caso geral de um corpo com valor absoluto *não-arquimediano*, a estrutura de corpo tem uma forte relação com o valor absoluto, gerando propriedades bastante diferentes das usuais (do caso arquimediano), propriedades que nós esperamos poder usar na aprendizagem estatística nesses corpos.

Observação 3.1.3. Primeiro: Observamos que se o valor absoluto $|\cdot|$ é não-arquimediano, a bola fechada unitária centrada na origem, $\bar{B}(0,1)$, é um subanel de $\mathbb{K}$. Com efeito, $0, 1 \in \bar{B}(0,1)$ e se $x, y \in \bar{B}(0,1)$, então $|x| \leq 1$ e $|y| \leq 1$, logo

$$|x + y| \leq \max\{|x|, |y|\} \leq 1$$

e

$$|xy| = |x||y| \leq 1$$

o que mostra que $x + y \in \bar{B}(0,1)$, $x \cdot y \in \bar{B}(0,1)$ e assim $\bar{B}(0,1)$ é um subanel de $\mathbb{K}$. **Segundo:** A bola aberta unitária centrada na origem, $B(0,1)$, é um ideal de $\bar{B}(0,1)$. Com efeito, seguindo o mesmo raciocínio acima, vemos que $B(0,1)$ é fechado para a soma, e para $x \in B(0,1)$ e $y \in \bar{B}(0,1)$, quaisquer, temos $|x| < 1$, $|y| \leq 1$ e

$$|xy| = |x||y| \leq |x| < 1.$$

Assim, $xy \in B(0,1)$ e $B(0,1)$ é um ideal de $\bar{B}(0,1)$.

Agora, damos a definição de anel e ideal de valorização.

Definição 3.1.7 (Anel e ideal de valorização)**.** Seja $\mathbb{K}$ um corpo munido de um valor absoluto não arquimediano $|\cdot|$. Definimos o *anel de valorização* de $|\cdot|$, como o subanel

$$\mathcal{O} = \bar{B}(0,1) = \{x \in \mathbb{K} : |x| \leq 1\}.$$

Também, definimos o *ideal de valorização* de $|\cdot|$, como o ideal

$$\mathcal{P} = B(0,1) = \{x \in \mathbb{K} : |x| < 1\}.$$

A seguinte proposição, que será enunciada sem prova, mostra algumas propriedades do par $\mathcal{O}$, $\mathcal{P}$. Detalhes são dados em [29].

Proposição 3.1.10. *Seja $\mathbb{K}$, dotado de um valor absoluto não-arquimediano $|\cdot|$. Então*

(i) *$\mathbb{K}$ é o corpo de frações de $\mathcal{O}$.*

(ii) *Se $x \in \mathcal{O}$ e $x \notin \mathcal{P}$, então x é inversível em $\mathcal{O}$, ou seja, $\mathcal{O}$ é um anel local com ideal maximal $\mathcal{P}$.*

Para finalizar, na situação da parte (ii) da proposição acima, temos que o anel quociente entre $\mathcal{O}$ e o ideal $\mathcal{P}$ é na verdade um corpo.

Definição 3.1.8. Sob as condições da Definição 3.1.7, o corpo residual de $\mathbb{K}$, é o corpo quociente

$$\kappa = \mathcal{O}/\mathcal{P}.$$

3.2 VALORES ABSOLUTOS EM $\mathbb{Q}$

Agora vamos nos focar em aplicar o visto de maneira geral, no caso dos números racionais $\mathbb{Q}$. Mas primeiro, devemos notar que se estamos interesados na topologia induzida por um valor absoluto em $\mathbb{Q}$, não teremos alternativas diferentes às obtidas com os valores absolutos dos exemplos dados na Seção 3.1:

(1) O valor absoluto trivial, $|\cdot|_0$

(2) O valor absoluto usual, $|\cdot|_\infty$

(3) E para cada inteiro primo $p > 1$, o valor absoluto p-ádico $|\cdot|_p$.

Chegou a hora de definir o valor absoluto de tipo p-ádico em $\mathbb{Q}$. Dados $p \in \mathbb{N}$, um número primo e $n \in \mathbb{Z}$, é claro que podemos escrever de forma única

$$n = p^v n',$$

onde $p \nmid n'$. Como v é determinado por p e n, escrevemos $v_p(n) = v$. Mais precisamente:

Definição 3.2.1 (Valorização p-ádica)**.** Dado $n \in \mathbb{Z}$, $n \neq 0$ e $p \in \mathbb{N}$ número primo, defina a valorização p-ádica de n, por $v_p(n)$, satisfazendo a relação:

$$n = p^{v_p(n)} n', \text{ onde } p \nmid n'.$$

Se $0 \neq x = a/b \in \mathbb{Q}$, defina a valorização p-ádica de x por

$$v_p(x) = v_p(a) - v_p(b),$$

junto com a convenção $v_p(0) = +\infty$.

Observação 3.2.1. Não é difícil verificar que a definição acima, para um número racional $x \in \mathbb{Q}$, *não depende da representação* de x como quociente de inteiros[1] e que a valorização p-ádica de um elemento $x \in \mathbb{Q}$, $x \neq 0$, fica determinada pela condição

$$x = p^{v_p(x)} \frac{a'}{b'}$$

onde $p \nmid a'b'$

Exemplo 3.2.1. Para visualizar o cálculo da valorização de um número racional, vamos ver alguns exemplos:

(1) $v_5(180)$: Como $180 = 5 \cdot 36$ e $5 \nmid 36$, temos $v_5(180) = 1$.

(2) $v_2(111/56)$: O inteiro 111 não é divisível por 2 e $56 = 2^3 \cdot 7$, logo $v_2(111/56) = -3$.

A seguir, um par de propriedades da valorização p-ádica.

Lema 3.2.1 (Lema 2.1.1 de [29])**.** *Para todo $x, y \in \mathbb{Q}$ e $p \in \mathbb{N}$ número primo, temos*

(i) $v_p(xy) = v_p(x) + v_p(y)$

(ii) $v_p(x + y) \geq \min\{v_p(x), v_p(y)\}$,

com as convenção $v_p(0) = +\infty$.

Demonstração. **(i)** Sejam $x, y \in \mathbb{Q}$ números racionais não nulos, então pela Observação 3.2.1, existem $a', b', c', d' \in \mathbb{Z}$, não divisíveis por p, tais que

$$x = p^{v_p(x)} \frac{a'}{b'} \text{ e } y = p^{v_p(y)} \frac{c'}{d'}, \text{ com } p \nmid a'b'c'd'$$

logo,

$$xy = p^{(v_p(x)+v_p(y))} \frac{a'c'}{b'd'}$$

com $p \nmid (a'c')(b'd')$, ou seja $v_p(xy) = v_p(x) + v_p(y)$.

[1] Se $x = \frac{a}{b} = \frac{c}{d} \neq 0$, então $v_p(a) - v_p(b) = v_p(c) - v_p(d)$.

(ii) Sejam $x, y \in \mathbb{Q}$ no formato do item (i). Primeiro suponha que $v_p(x) \leq v_p(y)$, o caso contrário é análogo. Então, podemos escrever

$$\begin{aligned} x + y &= p^{v_p(x)} \left(\frac{a'}{b'} + p^{(v_p(y)-v_p(x))} \frac{c'}{d'} \right) \\ &= p^{\min\{v_p(x),v_p(y)\}} \left(\frac{a'd' + p^{(v_p(y)-v_p(x))} b'c'}{b'd'} \right). \end{aligned}$$

Como antes $p \nmid b'd'$ e se $v_p(y) - v_p(x) \geq 1$, temos que $p \nmid (a'd' + p^{(v_p(y)-v_p(x))} b'c')$, de onde obtemos $v_p(x+y) = \min\{v_p(x), v_p(y)\}$. Se $v_p(y) - v_p(x) = 0$, então p pode dividir, ou não, ao inteiro $a'd' + b'c'$, nesse caso temos $v_p(x+y) \geq \min\{v_p(x), v_p(y)\}$, completando a prova.

■

Do Lema anterior, vemos que a valorização p-ádica, tem um comportamento "parecido" com o logaritmo de um valor absoluto, só que a desigualdade está na direção contrária. Com essa sugestão, temos o seguinte resultado.

Proposição 3.2.1 (Valor absoluto p-ádico)**.** *Considere $p \in \mathbb{N}$, um número primo. Defina a função $|\cdot|_p : \mathbb{Q} \to \mathbb{R}_+$, mediante*

$$|x|_p = p^{-v_p(x)},$$

com a convenção $p^{-\infty} = 0$, para ter $|0|_p = 0$. Então, $|\cdot|_p$, é um valor absoluto não-arquimediano em $\mathbb{Q}$, que chamamos de valor absoluto p-ádico.

Demonstração. Do Lema 3.2.1, as propriedades de valor absoluto não-arquimediano são satisfeitas trivialmente. ■

Antes de continuar, algumas observações.

Observação 3.2.2. Dado p número primo, da Proposição 3.1.6, vemos que para um número real $c > 1$ qualquer, o valor absoluto

$$|x|_c = c^{-v_p(x)}$$

é equivalente com o valor absoluto p-ádico $|\cdot|_p$, pois existe $\alpha > 0$ tal que $c = p^\alpha$, devido a que a função $f(x) = p^x$ satisfaz $f([0, +\infty[) = [1, +\infty[$.

Observação 3.2.3. Observe que, a diferença do que acontece no caso arquimediano

$$|p^n|_p = p^{-n} \overset{n\to\infty}{\longrightarrow} 0\,.$$

Observação 3.2.4. O valor absoluto p-ádico só pode tomar valores num conjunto enumerável, valores que pertencem ao conjunto

$$\mathrm{Im}(|\cdot|_p) = \{p^n : n \in \mathbb{Z}\} \cup \{0\}.$$

Observação 3.2.5. Para $x, y \in \mathbb{Z}$, temos a seguinte relação entre a distância de dois números e a congruência módulo p^n

$$x \equiv y \pmod{p^n} \Leftrightarrow |x-y|_p \leq p^{-n}.$$

Observação 3.2.6. Os valores absolutos p-ádicos $|\cdot|_{p_1}$ e $|\cdot|_{p_2}$, com $p_1, p_2 \in \mathbb{N}$ números primos distintos, não são equivalentes se $p_1 \neq p_2$. Com efeito, suponha que eles são equivalentes e considere a sequência $\{a_n\}_{n\in\mathbb{N}}$, com $a_n = (p_1/p_2)^n$. Observe que quando $n \to \infty$, $|a_n|_{p_1} \to 0$ e $|a_n|_{p_2} \to \infty$, mas, pela Proposição 3.1.6, existe $\alpha > 0$ tal que $|\cdot|_{p_2} = |\cdot|_{p_1}^{\alpha}$ e assim $|a_n|_{p_2} = |a_n|_{p_1}^{\alpha} \to 0$ quando $n \to \infty$, o que é uma contradição.

Agora estamos em condições para poder enunciar e provar o *Teorema de Ostrowski*.

Teorema 3.2.2 (Ostrowski (Teorema 3.1.2 de [29]))**.** *Todo valor absoluto não trivial em $\mathbb{Q}$, é equivalente a um dos valores absolutos $|\cdot|_p$, onde p é um número primo, ou "$p = \infty$".*

Demonstração. Seja $|\cdot|$ um valor absoluto não trivial sobre o corpo $\mathbb{Q}$. Temos os seguintes casos possíveis.

(a) Se $|\cdot|$ é *arquimediano*, então, pela Proposição 3.1.2, podemos escolher o menor inteiro positivo $n_0 \in \mathbb{N}$, tal que $|n_0| > 1$. Assim, é claro que existe um $\alpha > 0$ tal que

$$|n_0| = n_0^{\alpha}.$$

Para provar que podemos usar esse valor de α na Proposição 3.1.6, isto é, que para todo $x \in \mathbb{Q}$, $|x| = |x|_{\infty}^{\alpha}$, é suficiente provar que $|n| = n^{\alpha}$, para todo $n \in \mathbb{N}$. Seja $n \in \mathbb{N}$ qualquer e escrevamos n na base n_0, ou seja, no formato

$$n = a_0 + a_1 n_0 + a_2 n_0^2 + \cdots + a_k n_0^k,$$

com $0 \leq a_i \leq n_0 - 1$ e $a_k \neq 0$.

Observe que k é caracterizado pela desigualdade $n_0^k \leq n < n_0^{k+1}$, de onde obtemos $k\ln(n_0) \leq \ln(n) < (k+1)\ln(n_0)$, ou $k \leq \frac{\ln(n)}{\ln(n_0)} < k+1$, e assim

$$k = \left\lfloor \frac{\ln(n)}{\ln(n_0)} \right\rfloor$$

onde $\lfloor x \rfloor$ é a parte inteira de x. Tomando valor absoluto na expansão na base n_0 de n, temos

$$|n| = |a_0 + a_1 n_0 + a_2 n_0^2 + \cdots + a_k n_0^k| \leq |a_0| + |a_1| n_0^{\alpha} + |a_2| n_0^{2\alpha} + \cdots + |a_k| n_0^{k\alpha}.$$

Como n_0 é menor inteiro tal que $|n_0| > 1$, temos $|a_i| \leq 1$, para todo $0 \leq i \leq k$, de onde

$$\begin{aligned} |n| &\leq 1 + n_0^{\alpha} + n_0^{2\alpha} + \cdots + n_0^{k\alpha} \\ &= n_0^{k\alpha}(1 + n_0^{-\alpha} + n_0^{-2\alpha} + \cdots + n_0^{-k\alpha}) \\ &\leq n_0^{k\alpha} \sum_{i=0}^{\infty} n_0^{-i\alpha} = n_0^{k\alpha} \frac{n_0^{\alpha}}{n_0^{\alpha} - 1}. \end{aligned}$$

Logo, chamando $C = n_0^\alpha/(n_0^\alpha - 1)$, temos

$$|n| \leq C n_0^{k\alpha} \leq C n^\alpha.$$

Aplicando esse resultado no inteiro n^N, com $N \in \mathbb{N}$, obtemos

$$|n^N| \leq C n^{N\alpha}.$$

Logo, para $N \in \mathbb{N}$, temos

$$|n| \leq \sqrt[N]{C} n^\alpha.$$

Fazendo $N \to \infty$, obtemos $|n| \leq n^\alpha$, conseguindo assim a primeira desigualdade. Para mostrar a desigualdade oposta, voltamos olhar na igualdade

$$n = a_0 + a_1 n_0 + a_2 n_0^2 + \cdots + a_k n_0^k.$$

Como $n_0^k \leq n < n_0^{k+1}$, podemos escrever

$$n_0^{(k+1)\alpha} = |n_0^{(k+1)}| = |n + n_0^{(k+1)} - n| \leq |n| + |n_0^{(k+1)} - n|,$$

de onde

$$|n| \geq n_0^{(k+1)\alpha} - |n_0^{(k+1)} - n| \geq n_0^{(k+1)\alpha} - (n_0^{(k+1)} - n)^\alpha.$$

Como $n \geq n_0^k$, inferimos que

$$\begin{aligned} |n| &\geq n_0^{(k+1)\alpha} - (n_0^{(k+1)} - n_0^k)^\alpha \\ &= n_0^{(k+1)\alpha}\left(1 - \left(1 - \frac{1}{n_0}\right)^\alpha\right) \\ &= C' n_0^{(k+1)\alpha} > C' n^\alpha, \end{aligned}$$

com $C' = 1 - (1 - 1/n_0)^\alpha$ não depende de n. Procedendo igual que no caso anterior, inferimos que $|n| \geq n^\alpha$, de onde obtemos $|n| = n^\alpha$, provando assim que $|\cdot|$ é equivalente com o valor absoluto $|\cdot|_\infty$.

(b) Agora, suponha que o valor absoluto é não-arquimediano, então, pela Proposição 3.1.2, temos $|n| \leq 1$, para todo $n \in \mathbb{Z}$. Como estamos supondo que $|\cdot|$ é não trivial[2], podemos encontrar o menor inteiro positivo n_0 tal que $0 < |n_0| < 1$. Observe que n_0 necessariamente deve ser um número primo, pois caso contrário teríamos: $n_0 = a \cdot b$, com $|a| = |b| = 1$, pela minimalidade de n_0, contradizendo o fato $|n_0| < 1$. Seja $p = n_0$. Considere $n \in \mathbb{Z}$, um inteiro que não seja divisível por p. Vamos mostrar que $|n| = 1$. Com efeito, fazendo a divisão de n por p, obtemos

$$n = qp + r$$

[2] O valor absoluto trivial é definido por: $|0|_0 = 0$ e $|x|_0 = 1, \forall\, x \neq 0$, assim, se $|\cdot|$ é não trivial e não-arquimediano, pela Proposição 3.1.2, necessariamente existe um $n \in \mathbb{N}$ tal que $0 < |n| < 1$.

onde $0 < r < p$. Da minimalidade de p, temos necessariamente que $|r| = 1$ e $|qp| < 1$ pois $|q| \leq 1$ e $|p| < 1$. Como $|\cdot|$ é não-arquimediano, $|n| = \max\{|qp|, |r|\} = 1$. Para concluir, para $n \in \mathbb{Z}$ qualquer, escrevemos $n = p^v n'$, com $p \nmid n'$ e obtemos

$$|n| = |p|^v |n'| = |p|^v = c^{-v}$$

com $c = |p|^{-1} > 1$, e assim pela Observação 3.2.2, $|\cdot|$ é equivalente ao valor absoluto p-ádico $|\cdot|_p$.

■

Finalmente, veremos um resultado que utiliza todos os valores de p, incluindo $p = \infty$.

Proposição 3.2.2 (Fórmula do produto (Proposição 3.1.3 de [29]))**.** *Para $x \in \mathbb{Q}$, com $x \neq 0$, temos*

$$\prod_{p \leq \infty} |x|_p = 1,$$

onde $p \leq \infty$, significa que fazemos o produto sobre todos os primos de $\mathbb{Q}$, incluindo o "primo infinito".

Demonstração. Como todo número racional é quociente de inteiros e $|x| = |-x|$ para todo $x \in \mathbb{Q}$, é suficiente provar a fórmula para $x \in \mathbb{N}$. Nessa situação, podemos escrever $x = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_k^{\alpha_k}$. Logo

$$\begin{cases} |x|_q = 1, & \text{se } q \neq p_i \\ |x|_{p_i} = p_i^{-\alpha_i}, & \text{para } i = 1, 2, ..., k \\ |x|_\infty = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_k^{\alpha_k} \end{cases}$$

e fazendo o produto de todas essas quantidades, obtemos o resultado. ■

Observação 3.2.7. O simbolo "∞" no valor absoluto usual em $\mathbb{Q}$, é por conveniência na notação. Por exemplo, é útil para escrever de forma simples a *fórmula do produto* indexando facilmente todos os valores absolutos possíveis. Como o conjunto de números primos é infinito, *faz sentido pensar* no valor absoluto usual em $\mathbb{Q}$, pelo menos em termos de notação, como o valor absoluto associado a um "primo infinito", e assim todo valor absoluto não trivial em $\mathbb{Q}$ estaría associado a um número primo, seja finito ou infinito.

3.2.1 Completamentos de $\mathbb{Q}$

Como vimos no Teorema de Ostrowski 3.2.2, qualquer valor absoluto não trivial em $\mathbb{Q}$ será equivalente ou com o valor absoluto usual $|\cdot|_\infty$ ou com o valor absoluto p-ádico $|\cdot|_p$ para algum número primo p. No caso do valor absoluto usual, $(\mathbb{Q}, |\cdot|_\infty)$ não é um espaço métrico completo, logo usando o processo de completamento da Seção 3.1.3 obtemos o

corpo com valor absoluto dos números reais $(\mathbb{R}, |\cdot|_\infty)$. No caso dos valores absolutos p-ádicos temos a mesma situação, ou seja, $(\mathbb{Q}, |\cdot|_p)$ não é completo para qualquer p número primo. Resumimos esses fatos na seguinte proposição.

Proposição 3.2.3. *Seja $|\cdot|$ um valor absoluto não trivial sobre $\mathbb{Q}$. Então, o corpo com valor absoluto $(\mathbb{Q}, |\cdot|)$, não é um espaço métrico completo.*

Demonstração. Seja $|\cdot|$ um valor absoluto não trivial sobre $\mathbb{Q}$, e para chegar a uma contradição, vamos supor que o espaço métrico $(\mathbb{Q}, |\cdot|)$ é *completo*. Como em todo espaço métrico os *singletons* são conjuntos fechados, com eles podemos escrever $\mathbb{Q}$ como a união de uma família enumerável de conjuntos fechados:

$$\mathbb{Q} = \bigcup_{q\in\mathbb{Q}} \{q\}.$$

Claramente $int(\mathbb{Q}) = \mathring{\mathbb{Q}} = \mathbb{Q} \neq \emptyset$, logo, pelo *Teorema da Categoria de Baire*, necessariamente deve existir um $q_0 \in \mathbb{Q}$ tal que $int(\{q_0\}) \neq \emptyset$, ou seja, deve existir $r > 0$ tal que $B(q_0, r) = \{q_0\}$. Como $|\cdot|$ é não trivial, existe[3] $0 \neq z \in \mathbb{Q}$ tal que $|z| < r$, logo $q_0 + z \neq q_0$ e $|q_0 + z - q_0| = |z| < r$, de onde vemos que $q_0 + z \in B(q_0, r) = \{q_0\}$, o que é claramente uma contradição. Portanto, o espaço métrico $(\mathbb{Q}, |\cdot|)$ não pode ser completo. ■

Observação 3.2.8. Nas condições da proposição anterior, o que acontece no caso do valor absoluto trivial?. Aí temos completitude. Com efeito, sejam $\Omega \neq \emptyset$ e ρ_0 a métrica "*zero-um*" ou métrica *discreta* (a topologia gerada por essa métrica é a topologia discreta) sobre Ω: $\rho_0(x, y) = 1$, se $x \neq y$ e $\rho_0(x, x) = 0$, para todo $x, y \in \Omega$. O espaço métrico (Ω, ρ_0) é completo, pois dada uma sequência de Cauchy em Ω, $\{a_n\}_{n\in\mathbb{N}} \subset \Omega$, e $0 < \epsilon < 1$, temos que existe $N \in \mathbb{N}$ tal que $\rho_0(a_n, a_m) < \epsilon$, para todo $m, n \geq N$, ou seja, $a_n = a_m$, para todo $m, n \geq N$ e assim $a_n \to a_N \in \Omega$, quando $n \to \infty$. Mesmo sendo completos, os espaços métricos de topologia discreta são pouco interessantes, por essa razão, trabalheremos com métricas não triviais.

Da Proposição 3.2.3, vemos que tal como acontece no caso $|\cdot|_\infty$, para qualquer número primo $p > 1$, o espaço $(\mathbb{Q}, |\cdot|_p)$ não é completo. Portanto, do mesmo modo de como é feito no caso do valor absoluto $|\cdot|_\infty$, usaremos o processo de completamento da Seção 3.1.3 para completar $(\mathbb{Q}, |\cdot|_p)$, obtendo assim, o corpo dos números p-ádicos.

Definição 3.2.2 (O corpo dos números p-ádicos)**.** Definimos o corpo com valor absoluto dos números p-ádicos, que denotamos por $(\mathbb{Q}_p, |\cdot|_p)$, como o corpo com valor absoluto obtido ao completar com relação a $|\cdot|_p$, o corpo $(\mathbb{Q}, |\cdot|_p)$.

[3] Pela Proposição 3.1.6 e o Teorema 3.2.2 (Ostrowski), se $|\cdot| = |\cdot|_\infty^\alpha$, com $\alpha > 0$, escolhemos $z \in \mathbb{Q}$ tal que $0 < z < r^{1/\alpha}$, e se $|\cdot| = |\cdot|_p^\alpha$, escolhemos $z = p^{-h}$ tal que $h \in \mathbb{Z}$ e $p^h < r^{1/\alpha}$.

3.2.2 Representação dos números p-ádicos

Agora veremos algumas propriedades e a representação dos números p-ádicos. Primeiro, lembremos que os elementos de $\mathbb{Q}_p = \widehat{\mathbb{Q}}$, por ser um completamento, são classes de equivalência de sequências de Cauchy em $\mathbb{Q}$ com relação ao valor absoluto $|\cdot|_p$. Também é bom lembrar que $\mathbb{Q}$ pode ser identificado com o subcorpo de $\mathbb{Q}_p$ formado pelas classes de equivalência de sequências de Cauchy constantes. Para $a \in \mathbb{Q}_p$, seja $\{a_n\}_{n\in\mathbb{N}}$ uma sequência de Cauchy de números racionais representando a. Assim, o valor absoluto sobre $\mathbb{Q}_p$ é por definição

$$|a|_p = \lim_{n\to\infty} |a_n|_p.$$

Proposição 3.2.4 (Lema 3.2.5 de [29]). *Seja $0 \neq a \in \mathbb{Q}_p$ e $\{a_n\}_{n\in\mathbb{N}}$ uma sequência de Cauchy em a. Então, existe $N \in \mathbb{N}$ tal que*

$$|a|_p = \lim_{n\to\infty} |a_n|_p = |a_N|_p.$$

Demonstração. Como $a \neq 0$, temos que a sequência de Cauchy $\{a_n\}_{n\in\mathbb{N}}$ não é nula, isto é, $\lim_{n\to\infty} |a_n|_p \neq 0$. Como visto no Teorema 3.1.2, existem $\delta > 0$ e uma subsequência $\{a_{n_k}\}_{k\in\mathbb{N}}$ de $\{a_n\}_{n\in\mathbb{N}}$, *que também é uma representante de a*, tais que

$$|a_{n_k}|_p \geq \delta, \ \forall k \in \mathbb{N}.$$

Por outro lado, como $\{a_{n_k}\}_{k\in\mathbb{N}}$ também é de Cauchy, existe $M \in \mathbb{N}$ tal que

$$|a_{n_k} - a_{n_l}|_p < \delta, \ \forall k, l \geq M,$$

de onde vemos que para todo $k, l \geq M$

$$|a_{n_k} - a_{n_l}|_p < \delta \leq \max\{|a_{n_k}|_p, |a_{n_l}|_p\}.$$

Logo, pela Proposição 3.1.4, inferimos que necessariamente $|a_{n_k}|_p = |a_{n_l}|_p$ para todo $k, l \geq M$ e assim

$$|a|_p = \lim_{k\to\infty} |a_{n_k}|_p = |a_{n_M}|_p = |a_N|_p,$$

onde $N = n_M$. ■

Da proposição anterior, inferimos que o conjunto de valores possíveis para o valor absoluto $|\cdot|_p$ em $\mathbb{Q}_p$ é o mesmo que para o valor absoluto $|\cdot|_p$ em $\mathbb{Q}$, o conjunto $\mathrm{Im}(|\cdot|_p) = \{p^n : n \in \mathbb{Z}\} \cup \{0\}$. No caso do valor absoluto usual a situação é diferente, pois o conjunto de valores possíveis para $|\cdot|_\infty$ em $\mathbb{Q}$ é $\mathbb{Q}_+$ e em $\mathbb{R}$ é o intervalo real $[0, +\infty)$.

No que segue, vamos considerar as séries da forma

$$\sum_{i=-N}^{\infty} \alpha_i p^i = \alpha_{-N}p^{-N} + \alpha_{-N+1}p^{-N+1} + \cdots + \alpha_0 + \alpha_1 p + \alpha_2 p^2 + \cdots \tag{9}$$

onde $N \in \mathbb{Z}$, $0 < \alpha_{-N} < p$ e $0 \leq \alpha_i < p$ para todo $i > -N$. As somas parciais formam uma sequência de Cauchy, já que para todo $\epsilon > 0$, escolha $n_0 \in \mathbb{N}$ tal que $p^{-n_0} < \epsilon$ e para $m > n \geq n_0$, temos

$$\left| \sum_{i=-N}^{m} \alpha_i p^i - \sum_{i=-N}^{n} \alpha_i p^i \right|_p = \left| \sum_{i=n+1}^{m} \alpha_i p^i \right|_p \leq \max_{n+1 \leq i \leq m} \{|\alpha_i p^i|_p\} \leq p^{-(n+1)} < p^{-n_0} < \epsilon.$$

Assim, cada série da forma (9) será convergente em $\mathbb{Q}_p$. Reciprocamente, vamos provar que cada classe de equivalência de sequências de Cauchy em $\mathbb{Q}$ contém uma única representante canônica, que é uma sequência de somas parciais de uma série da forma (9). Para poder estabelecer esse fato, primeiro vamos definir o anel de inteiros p-ádicos

Definição 3.2.3 ($\mathbb{Z}_p$)**.** Definimos o anel dos inteiros p-ádicos, que denotamos por $\mathbb{Z}_p$, como

$$\mathbb{Z}_p := \bar{B}(0,1) = \{x \in \mathbb{Q}_p : |x|_p \leq 1\}.$$

Antes de continuar, uma pequena observação sobre a notação.

Observação 3.2.9. Para $p \in \mathbb{N}$, o símbolo $\mathbb{Z}_p$ pode representar dois conceitos diferentes dependendo da área onde ele é utilizado.
Na *aritmética modular*, o símbolo $\mathbb{Z}_p$ representa o conjunto de classes de equivalência da relação de *congruência módulo* p em $\mathbb{Z}$, definida por:

$$x \equiv y \,(\mathrm{mod}\, p) \Leftrightarrow x - y = m \cdot p,\ m \in \mathbb{Z}.$$

Nesse contexto, para p primo, $\mathbb{Z}_p$ munido da soma e multiplicação de classes, é um *corpo* com p elementos.
A outra área onde o símbolo é utilizado é na *análise p-ádica*, onde p também é um número primo. Aqui, como já foi definido, o símbolo $\mathbb{Z}_p$ representa o anel de inteiros p-ádicos, e para evitar confusões, o corpo de p elementos é denotado por $\mathbb{F}_p$.

Na linguagem da Definição 3.1.7, $\mathbb{Z}_p$ é o anel de valorização de $\mathbb{Q}_p$ e por ser uma bola num espaço métrico não-arquimediano, também é um conjunto aberto e fechado.

Teorema 3.2.3 (Proposição 3.3.2 de [29])**.** *O anel $\mathbb{Z}_p$ é um anel local cujo ideal maximal é*

$$p\mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p < 1\}.$$

Além disso

(i) *Dado $x \in \mathbb{Z}_p$ e $n \geq 1$, existe $\beta_n \in \mathbb{Z}$, $0 \leq \beta_n < p^n$, tal que $|x - \beta_n|_p \leq p^{-n}$. O inteiro β_n nessas condições, é único.*

(ii) *Dado $x \in \mathbb{Z}_p$, existe uma sequência de Cauchy $\{\beta_n\}_{n\in\mathbb{N}} \subset \mathbb{Z}$, tal que $\beta_n \to x$, satisfazendo*

(1) $\beta_n \in \mathbb{Z}$ *com* $0 \leq \beta_n < p^n$

(2) $\beta_{n+1} \equiv \beta_n \,(\mathrm{mod}\, p^n)$.

A sequência $\{\beta_n\}_{n\in\mathbb{N}}$ *nessas condições é única.*

Demonstração. Da Proposição 3.1.10, a primeira afirmação é imediata. Para ver a forma do ideal de valorização, observamos que

$$|x|_p < 1 \Leftrightarrow |x|_p \leq p^{-1} \Leftrightarrow \left|\frac{x}{p}\right|_p \leq 1 \Leftrightarrow x \in p\mathbb{Z}_p.$$

(i) Escolha $x \in \mathbb{Z}_p$ e $n \in \mathbb{N}$ quaisquer. Como $\mathbb{Q}$ é denso em $\mathbb{Q}_p$, então existe $a/b \in \mathbb{Q}$ *fração simplificada*[4], tal que

$$\left|x - \frac{a}{b}\right|_p \leq p^{-n} < 1.$$

Logo,

$$\left|\frac{a}{b}\right|_p \leq \max\{|x|_p, |x - a/b|_p\} \leq 1$$

de onde temos necessariamente que $p \nmid b$, pois se $b = p^v c$, com $v \in \mathbb{N}$ e $p \nmid c$, como por hipótese $p \nmid a$ (fração simplificada), teriamos necessariamente $|a/b|_p = p^v|a/c| = p^v > 1$, o que é uma contradição. Assim, como $p \nmid b$, existe $b' \in \mathbb{Z}$ tal que $bb' \equiv 1 \,(\mathrm{mod}\, p^n)$ de onde obtemos (Observação 3.2.5) $|bb' - 1|_p \leq p^{-n}$, e assim

$$\left|\frac{a}{b} - ab'\right|_p = \left|\frac{a}{b}(bb' - 1)\right|_p = \left|\frac{a}{b}\right|_p |bb' - 1|_p \leq p^{-n}.$$

Finalmente, como $ab' \in \mathbb{Z}$, *existe um único* $\beta_n \in \mathbb{Z}$ satisfazendo $0 \leq \beta_n < p^n$ e $ab' \equiv \beta_n \,(\mathrm{mod}\, p^n)$, logo

$$|ab' - \beta_n|_p \leq p^{-n}$$

e usando a desigualdade triangular forte, obtemos

$$|x - \beta_n|_p \leq p^{-n}.$$

(ii) Como (i) é válida para cada $n \in \mathbb{N}$, temos que existe uma sequência $\{\beta_n\}_{n\in\mathbb{N}}$ onde β_n é o único inteiro tal que $0 \leq \beta_n < p^n$ e $|x - \beta_n|_p \leq p^{-n}$, logo

$$|\beta_{n+1} - \beta_n|_p \leq \max\{|x - \beta_n|_p, |x - \beta_{n+1}|_p\} = p^{-n},$$

de onde vemos que $\beta_{n+1} \equiv \beta_n \,(\mathrm{mod}\, p^n)$, e pela Proposição 3.1.7, que $\{\beta_n\}_{n\in\mathbb{N}}$ é uma sequência de Cauchy que converge para x, quando $n \to \infty$. A unicidade de $\{\beta_n\}_{n\in\mathbb{N}}$ nessas condições, é clara.

■

[4] Uma fração de inteiros $\frac{a}{b}$ é chamada de fração simplificada, se $\mathrm{mdc}(a, b) = 1$.

Observação 3.2.10. Na prova do item **(i)** do Teorema 3.2.3, se escolhemos $\beta_n = ab'$, teremos um exemplo de uma sequência de números inteiros satisfazendo a condição $\beta_{n+1} \equiv \beta_n \,(\mathrm{mod}\, p^n)$ (ou seja, $\{\beta_n\}_{n\in\mathbb{N}}$ é de Cauchy), que converge para x mas que não necessariamente satisfaz $0 \leq \beta_n < p^n$, assim, além das sequências garantidas pelo Teorema 3.2.3 temos outras sequências de Cauchy realizando o mesmo elemento.

Também do item **(i)**, vemos que dado $x \in \mathbb{Z}_p$, sempre podemos encontrar um inteiro não negativo que seja arbitrariamente próximo de x, isso quer dizer que $\mathbb{N}_0$ é denso no espaço métrico $(\mathbb{Z}_p, |\cdot|_p)$, ou seja, $\overline{\mathbb{N}_0} = \mathbb{Z}_p$ e assim, $(\mathbb{Z}_p, |\cdot|_p)$ é um espaço métrico separável, que ao ser fechado em $(\mathbb{Q}_p, |\cdot|_p)$, é também um espaço métrico completo.

Agora estamos em condições de obter a representação canônica de um inteiro p-ádico como uma série de potências em p, ou seja, como a expressão (9) com $N = 0$. Considere $x \in \mathbb{Z}_p$. Do Teorema 3.2.3, existe uma sequência $\{\beta_n\}_{n\in\mathbb{N}} \subset \mathbb{Z}$, tal que

- $0 \leq \beta_n < p^n$
- $\beta_n \equiv x \,(\mathrm{mod}\, p^n)$
- $\beta_{n+1} \equiv \beta_n \,(\mathrm{mod}\, p^n)$.

O próximo passo é escrever os inteiros β_n na base p. Como $\beta_{n+1} \equiv \beta_n \,(\mathrm{mod}\, p^n)$, temos

$$\begin{cases} \beta_1 = \alpha_0, & 0 \leq \alpha_0 \leq p-1 \\ \beta_2 = \alpha_0 + \alpha_1 p, & 0 \leq \alpha_1 \leq p-1 \\ \beta_3 = \alpha_0 + \alpha_1 p + \alpha_2 p^2, & 0 \leq \alpha_2 \leq p-1 \end{cases}$$

e continuando com o processo, vemos que os inteiros β_n, formam uma sequência de somas parciais:

$$\beta_n = \sum_{i=0}^{n-1} \alpha_i p^i$$

que converge para $x \in \mathbb{Z}_p$, logo, podemos representar x como o limite $x = \lim_{n\to\infty} \beta_n$, isto é,

$$x = \alpha_0 + \alpha_1 p + \alpha_2 p^2 + \cdots + \alpha_n p^n + \cdots .$$

Primeiro verificamos que as expressões desse tipo são convergentes em $\mathbb{Z}_p$.

Lema 3.2.4 (Lema 3.3.8 de [29])**.** *Sejam $b_i \in \mathbb{Z}$, para $i \in \mathbb{N}$, quaisquer. Então, a série*

$$x = b_0 + b_1 p + b_2 p^2 + \cdots + b_n p^n + \cdots$$

é convergente em $\mathbb{Z}_p$.

Demonstração. É suficiente provar que a sequência de somas parciais é uma sequência de Cauchy. As somas parciais são $s_n = b_0 + b_1 p + b_2 p^2 + \cdots + b_n p^n$ e se $m = n + r > n$, e pela Proposição 3.1.7, temos

$$|s_{n+1} - s_n|_p = |b_{n+1} p^{n+1}|_p \leq p^{-(n+1)} \overset{n\to\infty}{\longrightarrow} 0$$

e assim s_n é convergente em $\mathbb{Z}_p$. ■

Como corolário imediato do Teorema 3.2.3 e o Lema acima, temos

Corolário 3.2.4.1. *Cada elemento $x \in \mathbb{Z}_p$ se escreve de maneira única como uma série de potências de p*

$$x = \alpha_0 + \alpha_1 p + \alpha_2 p^2 + \cdots + \alpha_n p^n + \cdots$$

onde $0 \leq \alpha_n < p$, para todo $n \in \mathbb{N}$.

Observando a expressão (9), vemos que para obter um $x \in \mathbb{Q}_p$ qualquer, é suficiente dividir um elemento de $\mathbb{Z}_p$ por uma potência de p.

Proposição 3.2.5. *Seja $x \in \mathbb{Q}_p$, então, existe $N \in \mathbb{N}_0$ tal que*

$$p^N x \in \mathbb{Z}_p.$$

Além disso, para $y = \sum_{n=0}^{\infty} \omega_n p^n \in \mathbb{Z}_p$, temos que $|y|_p = 1$ se e somente se $\omega_0 \neq 0$.

Demonstração. Para a primeira parte da proposição, se $x \in \mathbb{Q}_p$ é tal que $|x|_p \leq 1$, então, $N = 0$. Suponha agora que $|x|_p > 1$ e considere um inteiro $N \in \mathbb{N}$ tal que $1 < |x|_p \leq p^N$, então necessariamente $|p^N x|_p \leq 1$, ou seja $p^N x \in \mathbb{Z}_p$. Agora, para a segunda parte da proposição.
($\Rightarrow$) Suponha que $|y|_p = 1$, e também que $\omega_0 = 0$, então

$$y = \sum_{n=0}^{\infty} \omega_n p^n = p(\omega_1 + \omega_2 p + \dots),$$

de onde obtemos $|y|_p \leq p^{-1} < 1$, o que é uma contradição, portanto $\omega_0 \neq 0$.
($\Leftarrow$) Suponha $y = \sum_{n=0}^{\infty} \omega_n p^n$ com $\omega_0 \neq 0$ e $|y|_p < 1$, logo, $|y|_p = p^{-m}$ para algum $m \in \mathbb{N}$ e $|p^{-m} y|_p = 1$, assim de ($\Rightarrow$), $p^{-m} y = \sum_{n=0}^{\infty} \gamma_n p^n$ com $\gamma_0 \neq 0$ e $y = \sum_{n=0}^{\infty} \gamma_n p^{(n+m)}$. Portanto, da unicidade da representação em $\mathbb{Q}_p$, temos $\omega_n = 0$ para $n = 0, \dots, m-1$, e em particular $\omega_0 = 0$, o que é uma contradição. ■

Corolário 3.2.4.2. *Cada $x \in \mathbb{Q}_p$ não nulo, se escreve de maneira única como uma série de potências de p*

$$\begin{aligned} x &= \alpha_{-N} p^{-N} + \cdots + \alpha_0 + \alpha_1 p + \alpha_2 p^2 + \cdots + \alpha_n p^n + \cdots \\ &= \sum_{n=-N}^{\infty} \alpha_n p^n \end{aligned}$$

onde, $N \in \mathbb{Z}$, $a_{-N} \neq 0$, $0 \leq \alpha_i < p$ e $|x|_p = p^N$.

Demonstração. Seja $x \in \mathbb{Q}_p$ *não nulo*, da proposição anterior vemos que o conjunto $\{n \in \mathbb{Z} : p^n x \in \mathbb{Z}_p\}$ é não vazio. Considere $N \in \mathbb{Z}$ tal que[5]

$$N = \min\{n \in \mathbb{Z} : p^n x \in \mathbb{Z}_p\},$$

então $|p^N x|_p = 1$. Com efeito, como $|p^N x|_p \leq 1$, se $|p^N x|_p < 1$, então $p^N x \in p\mathbb{Z}_p$, ou seja existe $z \in \mathbb{Z}_p$ tal que $p^N x = pz$ e $p^{N-1}x = z$ o que contradiz a minimalidade de N, logo $p^N x = y$ com $|y|_p = 1$. Finalmente e também pela proposição anterior, para $y \in \mathbb{Z}_p$ tal que $|y|_p = 1$, temos $y = \sum_{n=0}^{\infty} \omega_n p^n$ com $\omega_0 \neq 0$, logo

$$x = p^{-N} y = \sum_{n=0}^{\infty} \omega_n p^{n-N} = \sum_{m=-N}^{\infty} \alpha_m p^m$$

onde $\alpha_m = \omega_{m+N}$ e $\alpha_{-N} = \omega_0 \neq 0$ e portanto

$$|x|_p = p^N = p^{-\min\{n \in \mathbb{Z} : \alpha_n \neq 0\}}.$$

■

No caso dos números reais, dado $b \in \mathbb{N}$, com $b > 1$, para $x > 0$ podemos escrever

$$x = \sum_{n=-N}^{\infty} \alpha_n b^{-n}$$

onde $0 \leq a_n < b - 1$ e $\alpha_{-N} \neq 0$, são os *dígitos de x na base b* e essa representação *não é única*. Se $N \geq 0$, então x tem parte inteira diferente de zero e representamos x em função desses digítos por

$$x = \alpha_{-N} \dots \alpha_{-2}\alpha_{-1}\alpha_0 \, . \, \alpha_1\alpha_2 \dots \alpha_n \dots .$$

Se $N < 0$, então $x \in (0, 1)$ e escrevemos

$$x = 0 \, . \, 00 \dots 0\alpha_{-N}\alpha_{-N+1}\alpha_{-N+2} \dots .$$

No caso p-ádico podemos fazer o mesmo. Se mantemos a convenção de que os dígitos correspondentes com as potências positivas de p fiquem do lado esquerdo da representação, para o número p-ádico

$$x = \sum_{n=-N}^{\infty} \alpha_n p^n$$

temos a *representação única*, para $N \geq 0$

$$x = \dots \alpha_n \dots \alpha_2\alpha_1\alpha_0 \, . \, \alpha_{-1}\alpha_{-2} \dots \alpha_{-N}$$

[5] Para $x \neq 0$ o mínimo sempre existe, pois se $p^n x \in \mathbb{Z}_p$, então $p^m x \in \mathbb{Z}_p$, $\forall\, m \geq n$, logo se o mínimo não existisse, necessariamente devemos ter $p^n x \in \mathbb{Z}_p$, $\forall n \in \mathbb{Z}_-$, ou seja, $|x|_p \leq p^n, \forall n \in \mathbb{Z}_-$, de onde vemos que necessariamente $x = 0$, obtendo uma contradição.

e para $N < 0$

$$x = \dots \alpha_{-N+2}\alpha_{-N+1}\alpha_{-N}0 \dots 00 \ . \ 0.$$

A representação como sequência de dígitos do número p-ádico x acima, é chamada de *expansão p-ádica canônica* de x. Note as diferenças nas representações como sequências de dígitos no caso real e p-ádico:

$$\dots \alpha_n \dots \alpha_2\alpha_1\alpha_0 \ . \ \alpha_{-1}\alpha_{-2} \dots \alpha_{-N} \quad \text{em } \mathbb{Q}_p \tag{10}$$

$$\alpha_{-N} \dots \alpha_{-2}\alpha_{-1}\alpha_0 \ . \ \alpha_1\alpha_2 \dots \alpha_n \dots \quad \text{em } \mathbb{R}. \tag{11}$$

Por último, vamos provar que o espaço métrico $(\mathbb{Z}_p, |\cdot|_p)$ é compacto, resultado que implica que $\mathbb{Q}_p$ é localmente compacto pois $\mathbb{Z}_p$ é uma vizinhança do zero. A prova usa o clássico argumento da diagonalização de Cantor.

Teorema 3.2.5 (Teorema 1.34 de [36])**.** *O espaço métrico* $(\mathbb{Z}_p, |\cdot|_p)$*, é compacto.*

Demonstração. Seja $\{x_n\}_{n\in\mathbb{N}}$ uma sequência em $\mathbb{Z}_p$. Escrevamos a expansão canônica de cada termo da sequência

$$x_n = \sum_{i=0}^{\infty} x_i^n p^i.$$

Como os dígitos $x_i^n \in \{0, 1, \cdots, p-1\}$, podemos encontrar $\alpha_0 \in \{0, 1, \cdots, p-1\}$ e uma subsequência infinita de $\{x_n\}_{n\in\mathbb{N}}$, denotada por $\{x_{0n}\}_{n\in\mathbb{N}}$, tal que o primeiro dígito de cada x_{0n}, para $n \in \mathbb{N}$, é α_0. Procedendo da mesma maneira, encontramos $\alpha_1 \in \{0, 1, \cdots, p-1\}$ e uma subsequência infinita $\{x_{1n}\}_{n\in\mathbb{N}}$ de $\{x_{0n}\}_{n\in\mathbb{N}}$ tal que os primeiros dois dígitos de cada x_{1n}, para $n \in \mathbb{N}$ são α_0 e α_1. Continuando com esse procedimento, obtemos os dígitos $\alpha_0, \alpha_1, \cdots$ junto com a sequência de sequências

$$x_{00}, x_{01}, x_{02}, \cdots, x_{0m}, \cdots,$$
$$x_{10}, x_{11}, x_{12}, \cdots, x_{1m}, \cdots,$$
$$x_{20}, x_{21}, x_{22}, \cdots, x_{2m}, \cdots,$$

tal que cada sequência é uma subsequência da sequência anterior e tal que cada elemento da j-ésima linha começa com os dígitos $\alpha_0, \alpha_1, ..., \alpha_j$. Para cada $j = 0, 1, ...$, temos

$$x_{jj} \in \{x_{j-1j}, x_{j-1j+1}, ...\}.$$

Portanto, a sequência diagonal $\{x_{00}, x_{11}, ...\}$ continua sendo uma subsequência da sequência original e que obviamente converge para $\alpha_0 + \alpha_1 p + \alpha_2 p^2 + \cdots \in \mathbb{Z}_p$, ou seja, toda sequência em $\mathbb{Z}_p$ possui uma subsequência convergente em $\mathbb{Z}_p$ e assim o anel de valorização dos inteiros p-ádicos é compacto. ■

3.2.3 O $\mathbb{Q}_p$-espaço vetorial $(\mathbb{Q}_p^d, +, \cdot)$

Na prática, os algoritmos de aprendizagem supervisionada são aplicados no espaço euclidiano, ou seja, as regras de aprendizagem trabalham com amostras rotuladas da forma $d_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathbb{R}^d \times \{0, 1\})^n$, onde os elementos x_i são membros do $\mathbb{R}$-espaço vetorial d-dimensional $(\mathbb{R}^d, +, \cdot)$, equipado com a métrica induzida por alguma norma.

Para trabalhar com *vetores* de números p-ádicos, procederemos de forma similar, assim, precisamos equipar o $\mathbb{Q}_p$-espaço vetorial $(\mathbb{Q}_p^d, +, \cdot)$, onde $+$ e $\cdot$ são as operações vetoriais componente a componente usuais, com uma estrutura métrica induzida por uma norma.

Procedendo de maneira análoga ao caso de $\mathbb{R}$-espaços vetoriais e trocando o valor absoluto $|\cdot|_\infty$ por $|\cdot|_p$, é possível mostrar que no $\mathbb{Q}_p$-espaço vetorial $\mathbb{Q}_p^d$, *todas as normas são equivalentes*[6] e que com qualquer uma delas, $\mathbb{Q}_p^d$ é um *espaço de Banach* graças à completitude de $\mathbb{Q}_p$. Ainda, como claramente o conjunto $\mathbb{Q}^d$ é denso no espaço $(\mathbb{Q}_p^d, \|\cdot\|)$, com $\|\cdot\|$ uma norma qualquer sobre $\mathbb{Q}_p^d$, temos que $(\mathbb{Q}_p^d, \|\cdot\|)$ é também um espaço métrico separável.

Mesmo que todas as normas gerem a mesma topologia em $\mathbb{Q}_p^d$, o *formato* de uma bola varía segundo a norma, por essa razão, no seguinte capítulo vamos trabalhar com a norma em $\mathbb{Q}_p^d$ que veremos a seguir, cuja métrica induzida será conveniente para nossos propósitos.

Definição 3.2.4 (Norma do tipo ℓ^∞ em $\mathbb{Q}_p^d$)**.** Sejam $d, p \in \mathbb{N}$ com p um número primo. No $\mathbb{Q}_p$-espaço vetorial $(\mathbb{Q}_p^d, +, \cdot)$, vamos considerar a função $\|\cdot\|_p^\infty : \mathbb{Q}_p^d \to \mathbb{R}_+$, definida para cada $a = (a_1, a_2, \ldots, a_d) \in \mathbb{Q}_p^d$ por

$$\|a\|_p^\infty := \max_{i \in [d]} |a_i|_p.$$

É imediato verificar que $\|\cdot\|_p^\infty$ é uma norma em $\mathbb{Q}_p^d$, e denotaremos por $\bar{B}_d(a, p^h)$ a bola fechada em $(\mathbb{Q}_p^d, \|\cdot\|_p^\infty)$ centrada em $a \in \mathbb{Q}_p^d$ de raio p^h, com $h \in \mathbb{Z}$. Ainda, observamos que para $d = 1$, temos $\|\cdot\|_p^\infty = |\cdot|_p$.
No caso de *não existir* confusão com alguma outra norma em $\mathbb{Q}_p^d$, para simplificar a notação, denotaremos a norma $\|\cdot\|_p^\infty$ simplesmente por: $\|\cdot\|_p$.

A norma $\|\cdot\|_p$ satisfaz a desigualdade triangular forte, pois para $a, b \in \mathbb{Q}_p^d$ arbitrários, existe $m \in [d]$ tal que

$$\|a + b\|_p = \max_{i \in [d]} |a_i + b_i|_p = |a_m + b_m|_p \leq \max\{|a_m|_p, |b_m|_p\} \leq \max\{\|a\|_p, \|b\|_p\},$$

[6] Duas normas são equivalentes se as métricas induzidas geram a mesma topologia. Na prova clássica em $\mathbb{R}$-espaços vetoriais, todas as normas são equivalentes com a norma $\|x\|_1 = \sum_{i=1}^d |x_i|_\infty$, e no caso p-ádico podemos usar $\|x\|_{1p} = \sum_{i=1}^d |x_i|_p$.

isto é, $\|\cdot\|_p$ é uma norma não-arquimediana em $\mathbb{Q}_p^d$, e portanto, $(\mathbb{Q}_p^d, \|\cdot\|_p)$ é um *espaço ultramétrico, completo e separável*.

4 UMA REGRA DE APRENDIZAGEM SUPERVISIONADA USANDO NÚMEROS p-ÁDICOS

Neste capítulo, vamos desenvolver uma regra de aprendizagem (Definição 2.2.1) sobre o cubo $[0,1]^d \subset \mathbb{R}_+^d$, com $d \geq 1$, que utiliza a *estrutura hierárquica* do espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$. Para conseguir nosso objetivo, primeiro veremos algumas propriedades geométricas das bolas no espaço *ultramétrico* $(\mathbb{Q}_p, |\cdot|_p)$, que vão nos permitir definir uma regra de aprendizagem nos inteiros p-ádicos. Depois, representaremos o conjunto $\mathbb{Z}_p$ como uma *árvore de busca* p-ária cheia[1] [16], $\mathcal{A}(\mathbb{Z}_p)$, para logo definir uma regra de aprendizagem sobre $(\mathbb{Z}_p, |\cdot|_p)$ aproveitando essa particular representação. Depois de estudar a natureza das bolas na topologia do espaço normado d-dimensional $(\mathbb{Q}_p^d, \|\cdot\|_p)$, com $d > 1$, que também é ultramétrico, estendemos a regra de aprendizagem sobre $\mathbb{Z}_p$ para o espaço $(\mathbb{Z}_p^d, \|\cdot\|_p)$ e finalmente com a ajuda da técnica de *redução de dimensionalidade boreliana*, (ver [53] e o Teorema 2.3.3), definimos uma regra de aprendizagem sobre $[0,1]^d \subset \mathbb{R}_+^d$, com $d \geq 1$, como a composição (Definição 2.3.1) de uma aplicação *injetora e boreliana*, $\Phi_p : [0,1]^d \to \mathbb{Z}_p^d$, e a regra de aprendizagem sobre $\mathbb{Z}_p^d$.

Ao longo do texto, vamos adotar a terminologia padrão sobre árvores de decisão utilizada na *Ciência de Dados/Ciências da Computação*. Detalhes sobre esses conceitos básicos podem ser encontrados nas seguintes referências, [32, 45, 57].

Antes de começar com o desenvolvimento da regra de aprendizagem nos números p-ádicos, vamos fazer um comentário sobre a face desta mesma regra num contexto mais geral.

Como visto no Capítulo 1 (Introdução), se seguimos a ordem cronológica dos diversos estágios da pesquisa, depois de estudar em detalhe as principais propriedades dos números p-ádicos, o passo seguinte foi desenvolver uma regra de aprendizagem que usufrui dessas particulares propriedades, para logo no seguinte estágio, estudar a consistência da nova regra.

Na hora de estudar a consistência, a regra de aprendizagem desenvolvida no espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$ *revelou a sua face mais geral*, e portanto resultou ser que a nova regra de aprendizagem, na verdade pode ser aplicada em qualquer domínio boreliano padrão (Ω, ρ), com ρ métrica sobre Ω, pois como veremos no Capítulo 5, para cada $k \in \mathbb{N}$, a regra no espaço $(\mathbb{Z}_p^d, \|\cdot\|_p)$ pode se escrever como a regra de aprendizagem do tipo plug-in, que chamaremos ${}^+k$-NN e que denotaremos por $\mathcal{L}_{+k\text{-NN}} = (g_{nk})_{n=1}^{\infty}$, definida para cada $n \in \mathbb{N}$, $1 \leq k \leq n$, $d_n = ((x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)) \in (\Omega \times \{0,1\})^n$ com

[1] Uma árvore m-ária, é uma árvore onde cada vértice possui no máximo m vértices filhos, e ela será *cheia*, se cada vértice pai tem exatamente m vértices filhos. Uma árvore m-ária de altura finita, só pode ser cheia se $m = 0$, portanto para $m \in \mathbb{N}$, qualquer árvore m-ária cheia, necessariamente é de altura infinita.

$\varsigma_n = (x_1, x_2, \ldots, x_n) \in \Omega^n$ e $x \in \Omega$, mediante

$$g_{nk}(d_n)(x) = \begin{cases} 1, & \eta_{nk}(d_n, x) \geq \frac{1}{2} \\ 0, & \text{caso contrário,} \end{cases}$$

onde as funções

$$\eta_{nk} : (\Omega \times \{0,1\})^n \times \Omega \to [0,1]$$

são definidas por

$$\eta_{nk}(d_n, x) := \frac{1}{N_k^{\varsigma_n}(x)} \sum_{x_i \in \bar{B}(x, r_{k\text{-NN}}^{\varsigma_n}(x))} y_i,$$

com

$$r_{k\text{-NN}}^{\varsigma_n}(x) := \min\left\{r > 0 : \sharp\left(\{i \in [n] : x_i \in \bar{B}(x,r)\}\right) \geq k\right\}$$

e

$$N_k^{\varsigma_n}(x) := \sharp\left(\{i \in [n] : x_i \in \bar{B}(x, r_{k\text{-NN}}^{\varsigma_n}(x))\}\right) \geq k.$$

Ou seja, se $(\Omega, \rho) = (\mathbb{Z}_p^d, \|\cdot\|_p)$, a regra de aprendizagem $\mathcal{L}_{+k\text{-NN}}$ é reduzida à regra definida pelos *simples* algoritmos que desenvolveremos neste capítulo.

4.1 PROPRIEDADES GEOMÉTRICAS DAS BOLAS EM $(\mathbb{Q}_p, |\cdot|_p)$

Para começar, vamos ver as propriedades das bolas no espaço ultramétrico $(\mathbb{Q}_p, |\cdot|_p)$ que vão nos permitir definir uma regra de aprendizagem nos inteiros p-ádicos $(\mathbb{Z}_p, |\cdot|_p)$. Na Proposição 3.1.5, vimos que num corpo com valor absoluto não-arquimediano $(\mathbb{K}, |\cdot|)$, as bolas possuem as seguintes propriedades, onde $r, s > 0$ e $a, b \in \mathbb{K}$:

(i) Se $b \in B(a,r)$, então $B(b,r) = B(a,r)$.

(ii) $B(b,r) = B(a,r)$ ou $B(b,r) \cap B(a,r) = \emptyset$.

(iii) O conjunto $B(a,r)$ é aberto e fechado na topologia induzida por $|\cdot|$.

(iv) $B(a,r) \cap B(b,s) \neq \emptyset \Leftrightarrow B(a,r) \subset B(b,s)$ ou $B(b,s) \subset B(a,r)$.

(v) As propriedades **(i)**-**(iv)** valem também para as bolas fechadas.

A propriedade **(i)**, diz que em uma bola aberta ou fechada, *todo elemento da bola é o centro da bola*. A propriedade **(ii)** diz que duas bolas com o mesmo raio ou são iguais ou são disjuntas, a propriedade **(iii)** diz que toda bola é ao mesmo tempo um conjunto aberto e fechado, e a propriedade **(iv)** diz que duas bolas quaisquer, ou são disjuntas ou uma está contida na outra. Isto acontece em qualquer espaço ultramétrico. No caso p-ádico, ainda temos.

Proposição 4.1.1. *Seja $a \in \mathbb{Q}_p$ arbitrário. Então*

(i) *Para todo $h \in \mathbb{Z}$, $\bar{B}(a, p^h) = B(a, p^{h+1})$.*

(ii) *Para $\epsilon > 0$ qualquer, existe $h \in \mathbb{Z}$ tal que*

$$\bar{B}(a, \epsilon) = \bar{B}(a, p^h).$$

(iii) *Para $x = \sum_{j=-N}^{\infty} \alpha_j p^j$ e $y = \sum_{j=-M}^{\infty} \gamma_j p^j$, números p-ádicos quaisquer, temos*

$$|x - y|_p = p^{-\min\{j \in \mathbb{Z} : \alpha_j \neq \gamma_j\}}$$

e para $h \in \mathbb{Z}$

$$|x - y|_p \leq p^h \Leftrightarrow \alpha_j = \gamma_j, \forall j \leq -h - 1.$$

Demonstração. **(i)** A inclusão $\bar{B}(a, p^h) \subset B(a, p^{h+1})$ é clara. Como o valor absoluto p-ádico toma valores no conjunto

$$\mathrm{Im}(|\cdot|_p) = \{0\} \cup \{p^h : h \in \mathbb{Z}\},$$

vemos que não pode existir um número p-ádico x tal que $p^h < |x|_p < p^{h+1}$, logo se $x \in B(a, p^{h+1})$ então $|x - a|_p < p^{h+1}$ e assim $|x - a|_p \leq p^h$, obtendo a igualdade.

(ii) Para $\epsilon > 0$, existe um único $h \in \mathbb{Z}$ tal que $p^h \leq \epsilon < p^{h+1}$. Se $p^h = \epsilon$ a prova acaba. Suponha que $p^h < \epsilon < p^{h+1}$. É claro que $\bar{B}(a, p^h) \subset \bar{B}(a, \epsilon)$ e para a outra desigualdade, vemos que se $x \in \bar{B}(a, \epsilon)$ então $|x - a|_p \leq \epsilon$, e pelo mesmo argumento utilizado no item (i), temos que necessariamente $|x - a|_p \leq p^h$, logo $x \in \bar{B}(a, p^h)$.

(iii) Sejam $x = \sum_{j=-N}^{\infty} \alpha_j p^j$ e $y = \sum_{j=-M}^{\infty} \gamma_j p^j$, números p-ádicos quaisquer com $M, N \in \mathbb{Z}$ e $\alpha_{-N}, \gamma_{-M} \neq 0$. Se $N \neq M$, então $|x|_p \neq |y|_p$, logo, pela Proposição 3.1.4,

$$|x - y|_p = \max\{|x|_p, |y|_p\} = p^{-\min\{-N, -M\}} = p^{-\min\{j \in \mathbb{Z} : \alpha_j \neq \gamma_j\}},$$

obtendo o resultado. Suponha $N = M$ e seja $j_0 = \min\{j \in \mathbb{Z} : \alpha_j \neq \gamma_j\} \geq -N$, então $\alpha_j = \gamma_j, \forall\, j \leq j_0 - 1$ e

$$\begin{aligned} x - y &= p^{j_0} \left(\sum_{j=j_0}^{\infty} \alpha_j p^{j-j_0} - \sum_{j=j_0}^{\infty} \gamma_j p^{j-j_0} \right) \\ &= p^{j_0} \left(\sum_{m=0}^{\infty} \omega_m p^m - \sum_{m=0}^{\infty} \xi_m p^m \right) \end{aligned} \tag{12}$$

onde $\omega_m = \alpha_{m+j_0}$ e $\xi_m = \gamma_{m+j_0}$ com $\omega_0 \neq \xi_0$.
Ainda,

$$\left| \sum_{m=0}^{\infty} \omega_m p^m - \sum_{m=0}^{\infty} \xi_m p^m \right|_p = \left| (\omega_0 - \xi_0) + p \left(\sum_{m=1}^{\infty} \omega_m p^{m-1} - \sum_{m=1}^{\infty} \xi_m p^{m-1} \right) \right|_p = 1$$

pois, por um lado, temos que o inteiro $\omega_0 - \xi_0 \neq 0$ e $|\omega_0 - \xi_0|_\infty \leq p-1$, de onde inferimos que $p \nmid (\omega_0 - \xi_0)$ obtendo assim $|\omega_0 - \xi_0|_p = 1$, e por outro,

$$\left| p \left(\sum_{m=1}^{\infty} \omega_m p^{m-1} - \sum_{m=1}^{\infty} \xi_m p^{m-1} \right) \right|_p \leq p^{-1} < 1,$$

assim, pela Proposição 3.1.4, $\left|\sum_{m=0}^{\infty} \omega_m p^m - \sum_{m=0}^{\infty} \xi_m p^m\right|_p = |\omega_0 - \xi_0|_p = 1$. Tomando valor absoluto em ambos lados da igualdade (12), obtemos $|x-y|_p = p^{-j_0}$. Por último, para $h \in \mathbb{Z}$, qualquer,

$$|x-y|_p \leq p^h \Leftrightarrow \min\{j \in \mathbb{Z} : \alpha_j \neq \gamma_j\} \geq -h \Leftrightarrow \alpha_j = \gamma_j, \forall j \leq -h-1.$$

■

Para simplificar a notação, definimos o conjunto D_b dos dígitos na base b.

Definição 4.1.1. Para $b \in \mathbb{N}$, $b > 1$, defina o conjunto de dígitos da base b, $D_b \subset \mathbb{N}_0$, por

$$D_b := \{0, 1, 2, \ldots, b-1\}.$$

Outra propriedade importante das bolas p-ádicas, é que elas podem se escrever como a união de p bolas fechadas de igual raio que, pela Proposição 3.1.5, são necessariamente conjuntos dois a dois disjuntos. Vamos mostrar esse fato com $\mathbb{Z}_p$.

Proposição 4.1.2. *Seja $h \in \mathbb{N}$ qualquer. Então*

$$\mathbb{Z}_p = \bigcup_{\alpha_{(h)} \in D_p^h}^{\cdot} \bar{B}([|\alpha_{(h)}|], p^{-h}),$$

onde $\alpha_{(h)} = (\alpha_0, \alpha_1, ..., \alpha_{h-1}) \in D_p^h = \prod_{i=1}^{h} D_p$ e $[|\alpha_{(h)}|] := \sum_{n=0}^{h-1} \alpha_n p^n \in \mathbb{Z}_p$.

Demonstração. Primeiro note que, pela Proposição 4.1.1 **(i)**,

$$\begin{aligned} \mathbb{Z}_p &= B(0,1) \cup S(0,1) \\ &= \bar{B}(0, p^{-1}) \cup S(0,1) \end{aligned}$$

onde $S(0,1)$ é a *esfera unitária centrada na origem*, mas, pela Proposição 3.2.5, temos

$$\begin{aligned} S(0,1) &= \{x \in \mathbb{Z}_p : |x|_p = 1\} \\ &= \left\{ x \in \mathbb{Z}_p : x = \sum_{n=0}^{\infty} \alpha_n p^n,\ \alpha_0 \neq 0 \right\} \\ &= \bigcup_{\alpha_0=1}^{p-1}{}^{\cdot} \left\{ x \in \mathbb{Z}_p : x - \alpha_0 = \sum_{n=1}^{\infty} \alpha_n p^n \right\} \\ &= \bigcup_{\alpha_0=1}^{p-1}{}^{\cdot} \left\{ x \in \mathbb{Z}_p : |x - \alpha_0|_p \leq p^{-1} \right\} \\ &= \bigcup_{\alpha_0=1}^{p-1}{}^{\cdot} \bar{B}\left(\alpha_0, p^{-1}\right), \end{aligned}$$

e portanto

$$\mathbb{Z}_p = \bigcup_{\alpha_0=0}^{p-1} \bar{B}\left(\alpha_0, p^{-1}\right),$$

onde as bolas da união acima, são conjuntos dois a dois disjuntos. Para $\alpha_0 \in D_p$, os elementos de $\mathbb{Z}_p$ que pertencem à bola $\bar{B}(\alpha_0, p^{-1})$, são os números cujo *último* dígito (a representação p-ádica é de direita para esquerda), o correspondente à potência p^0, seja α_0. Por sua vez, e usando o argumento anterior, para $\alpha_0 \in D_p$, podemos particionar a bola $\bar{B}(\alpha_0, p^{-1})$ por

$$\bar{B}\left(\alpha_0, p^{-1}\right) = \bigcup_{\alpha_1=0}^{p-1} \bar{B}\left(\alpha_0 + \alpha_1 p, p^{-2}\right),$$

obtendo assim uma partição de $\mathbb{Z}_p$ em bolas de raio p^{-2}

$$\begin{aligned}\mathbb{Z}_p &= \bigcup_{\alpha_0=0}^{p-1} \bigcup_{\alpha_1=0}^{p-1} \bar{B}\left(\alpha_0 + \alpha_1 p, p^{-2}\right) \\ &= \bigcup_{(\alpha_0,\alpha_1)\in D_p^2} \bar{B}\left(\alpha_0 + \alpha_1 p, p^{-2}\right).\end{aligned}$$

Continuando com este processo, para $h \in \mathbb{N}$, temos

$$\mathbb{Z}_p = \bigcup_{\alpha_{(h)}\in D_p^h} \bar{B}([|\alpha_{(h)}|], p^{-h}),$$

onde os conjuntos da união acima são dois a dois disjuntos, $\alpha_{(h)} := (\alpha_0, \alpha_1, ..., \alpha_{h-1}) \in D_p^h$ e $[|\alpha_{(h)}|] := \sum_{n=0}^{h-1} \alpha_n p^n$. ■

Observação 4.1.1. Para $h \in \mathbb{N}$ qualquer, o produto cartesiano $D_p^h = \prod_{i=1}^{h} D_p$ possui $\sharp(D_p^h) = p^h$ elementos. Logo, para cada $h \in \mathbb{N}$, da Proposição 4.1.2, temos uma partição de $\mathbb{Z}_p$ em p^h bolas de raio p^{-h} que podemos indexar utilizando o conjunto D_p^h.

Assim uma bola $\bar{B}([|\alpha_{(h)}|], p^{-h})$ será indexada usando a representação numérica de $[|\alpha_{(h)}|] \in \mathbb{Z}_p$ no formato (10), isto é, será indexada com a sequência de h dígitos $\alpha_{h-1}\cdots\alpha_1\alpha_0$, sequência que chamaremos de *identificador* da bola $\bar{B}([|\alpha_{(h)}|], p^{-h})$.

4.2 REPRESENTAÇÃO DE $\mathbb{Z}_p$ COMO UMA ÁRVORE p-ÁRIA CHEIA

A métrica induzida em $\mathbb{Q}$ pelo valor absoluto p-ádico, está estreitamente relacionada com a aritmética de $(\mathbb{Q}, +, \cdot)$ e o valor de p, por isso, como veremos na Observação 4.2.1 e de forma diferente de como acontece no caso real[2], o corpo $(\mathbb{Q}_p, +, \cdot)$ não é um *corpo ordenado*, isto é, não é possível definir em $\mathbb{Q}_p$ uma ordem total que *respeite as operações de corpo*. Portanto, tentar uma representação gráfica de $\mathbb{Z}_p$ no estilo que acostumamos fazer

[2] A topologia de $\mathbb{R}$ é por defeito a *topologia da ordem*, como também se chama à topologia induzida por $|\cdot|_\infty$.

no espaço euclidiano (desenhar o gráfico de uma bola no espaço $(\mathbb{R}^3, \|\cdot\|_2)$, por exemplo), carece de sentido, pois não podemos falar de eixos coordenados no sentido (euclidiano) usual.

Uma alternativa para visualizar "uma cópia" de $\mathbb{Z}_p$ no espaço euclidiano é dada em [36, Secção 2.3]. Ali, os autores constroem *modelos euclidianos* de $\mathbb{Z}_p$, onde esses modelos são subconjuntos de $\mathbb{R}^d$ *homeomorfos*[3] a $\mathbb{Z}_p$.

Agora, se o que procuramos é representar de forma gráfica a *estrutura hierarquica* do espaço $(\mathbb{Z}_p, |\cdot|_p)$, da Proposição 4.1.2 vemos que podemos representar essa estrutura mediante uma árvore p-ária cheia [16]. Deve ficar claro que essa não é uma representação geométrica no sentido mencionado acima, porém, ela é uma representação que permite registrar e visualizar a *hierarquia* dos conjuntos $\bar{B}(a, p^{-h}) \subset \mathbb{Z}_p$ dentro de $\mathbb{Z}_p$ e fornece um método para determinar o elemento da partição de $\mathbb{Z}_p$, em bolas fechadas de raio p^{-h}, onde pertence um elemento $x \in \mathbb{Z}_p$ qualquer, ou seja, permite construir uma *árvore de busca* que, junto com um critério de decisão, podemos utilizar para definir uma *regra de aprendizagem* em $(\mathbb{Z}_p, |\cdot|_p)$.

Antes de continuar, vamos fazer uma observação pendente.

Observação 4.2.1 ($(\mathbb{Q}_p, +, \cdot)$ não é um corpo ordenado)**.** Primeiro lembremos que $(\mathbb{K}, +, \cdot, \preccurlyeq)$ é um corpo ordenado, se $(\mathbb{K}, +, \cdot)$ é um corpo e $\preccurlyeq$ é uma ordem total sobre $\mathbb{K}$ (todo par de elementos de $\mathbb{K}$ podem ser comparados usando $\preccurlyeq$) satisfazendo, para todo $a, b, c \in \mathbb{K}$,

(i) Se $a \preccurlyeq b$, então $a + c \preccurlyeq b + c$,

(ii) Se $0_{\mathbb{K}} \preccurlyeq a$ e $0_{\mathbb{K}} \preccurlyeq b$, então $0_{\mathbb{K}} \preccurlyeq a \cdot b$.

Assim, se supomos que em $(\mathbb{Q}_p, +, \cdot)$ existe uma ordem total $\preccurlyeq$ e definindo a *relação de ordem parcial estrita*, $\prec$, por $a \prec b \Leftrightarrow a \preccurlyeq b$ e $a \neq b$, para 0 e 1, temos duas alternativas:

$1 \prec 0$: Nesse caso, pela propriedade **(i)** acima, somando -1 em ambos lados temos $0 \prec -1$, assim, de **(ii)** multiplicando em ambos lados de $1 \prec 0$ por -1, temos $-1 = 1(-1) \prec 0(-1) = 0$, o que é uma contradição.

$0 \prec 1$: Nesse caso, somando -1 em ambos lados, temos $-1 \prec 0$. Por outro lado, se somamos 1 de maneira recursiva em $0 \prec 1$, obtemos

$$0 \prec 1 \prec 2 \prec \cdots \prec p-1 \prec p,$$

logo, de **(ii)**

$$0 \prec 1 \prec (p-1)p^j, \ \forall j \in \mathbb{N}_0.$$

[3] Dois espaços topologicos são homeomorfos, se existe um homeomorfismo entre eles. Um homeomorfismo entre os espaços topológicos E e F, é qualquer função bijetora, $f : E \to F$, tal que f e f^{-1} são funções contínuas.

Assim,

$$0 \prec \sum_{j=0}^{\infty}(p-1)p^j = -1$$

o que novamente é uma contradição.

Portanto, concluímos que $(\mathbb{Q}_p, +, \cdot)$, não pode ser um corpo ordenado.

Para poder definir a estrutura de árvore com a qual representaremos os inteiros p-ádicos, precisamos da noção de *altura* de uma bola em $\mathbb{Z}_p$.

Definição 4.2.1 (Altura de uma bola em $\mathbb{Z}_p$)**.** Seja $B = \bar{B}(a, p^{-h}) \subset \mathbb{Z}_p$, uma bola qualquer. O inteiro não negativo $h = h(B)$, será chamado de *altura de* B, e faz menção à partição de $\mathbb{Z}_p$ em bolas fechadas de raio p^{-h}, partição que será chamada de *partição de* $\mathbb{Z}_p$ *à altura* h.

Agora, para determinar a bola da partição à altura h na qual pertence um elemento $x \in \mathbb{Z}_p$ qualquer, temos a seguinte observação.

Observação 4.2.2. Para $h \in \mathbb{N}$ qualquer e $x = \sum_{j=0}^{\infty}\alpha_j p^j \in \mathbb{Z}_p$ na sua expansão canônica, da Proposição 4.1.2, vemos que para determinar o elemento da partição de $\mathbb{Z}_p$ à altura h onde pertence x, é suficiente conhecer os últimos h dígitos de x (lembrar como é a leitura de um número p-ádico), ou seja, é suficiente conhecer os dígitos $\alpha_{h-1}\ldots\alpha_1\alpha_0$, *na mesma ordem dos índices*. Assim, o identificador da bola fechada de raio p^{-h} à qual pertence x será denotado por $x_{(h)} := \alpha_{h-1}\ldots\alpha_1\alpha_0$.

Agora, vamos usar os conceitos anteriores para representar os elementos das partições de $\mathbb{Z}_p$ à altura $h \in \mathbb{N}$, como *vértices* ou *nós de uma árvore de busca* p-ária cheia, onde a estrutura hierarquica será representada pelas arestas entre vértices. Assim, usando uma amostra rotulada $d_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{Z}_p \times \{0, 1\})^n$ com os números $x_i \in \mathbb{Z}_p$ na sua representação canônica, podemos facilmente implementar computacionalmente o Algoritmo 1, para construir uma sub-árvore de altura finita, e o Algoritmo 2, que é um algoritmo de busca na árvore já construida, para obter com a combinação de ambos uma regra de aprendizagem sobre o espaço ultramétrico $(\mathbb{Z}_p, |\cdot|_p)$.

Definição 4.2.2 (Representação de $\mathbb{Z}_p$ como árvore p-ária cheia [16])**.** Definimos a *árvore de* $\mathbb{Z}_p$, ou simplesmente *a árvore p-ádica*, que denotaremos por $\mathcal{A}(\mathbb{Z}_p)$, como a árvore p-ária de altura infinita enraizada no vértice $V(\mathbb{Z}_p)$ (representando a $\mathbb{Z}_p$), cujos vértices à altura $h \in \mathbb{N}$ representam as bolas de raio p^{-h} da partição à altura h de $\mathbb{Z}_p$, e onde desenhamos uma *aresta* $U \to V$ entre os vértices U e V, sempre que $h(B_V) = h(B_U) + 1$ e $B_V \subset B_U$, onde B_U e B_V são as bolas representadas pelos vértices U e V, respectivamente.

Observação 4.2.3. Pela Proposição 4.1.2, vemos que na árvore $\mathcal{A}(\mathbb{Z}_p)$, um *vértice pai* na altura $h \in \mathbb{N}$, tem p *vértices filhos* na altura $h+1$ e assim, $\mathcal{A}(\mathbb{Z}_p)$ é uma árvore p-ária

com vértice raiz representando a $\mathbb{Z}_p$ e onde cada vértice tem exatamente p vértices filhos, ou seja, é uma árvore p-ária *cheia*. Na Figura 1, temos uma representação gráfica de $\mathcal{A}(\mathbb{Z}_2)$, onde os vértices na altura h são rotulados com os correspondentes identificadores de comprimento h da Observação 4.1.1.

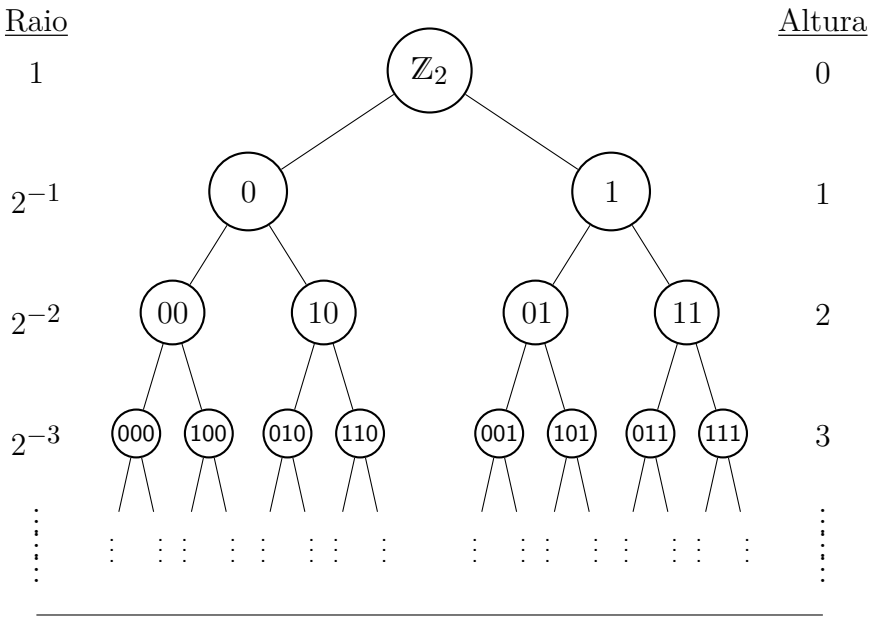

Figura 1 – Representação de $(\mathbb{Z}_2, |\cdot|_2)$ como árvore 2-ária de altura infinita.

Graças à unicidade da representação canônica nos números p-ádicos, temos que cada $x = \sum_{j=0}^{\infty} \alpha_j p^j \in \mathbb{Z}_p$ corresponde *biunivocamente* (ver [16]) com um *caminho infinito* na árvore p-ádica $\mathcal{A}(\mathbb{Z}_p)$ começando no vértice raiz, $V(\mathbb{Z}_p)$. Denotemos por $V(\gamma_{(h)})$ o vértice à altura h representando à bola com identificador $\gamma_{(h)} = \gamma_{h-1} \dots \gamma_1 \gamma_0$. Então, $x \in \mathbb{Z}_p$ como acima, pode ser *codificado* pelo caminho infinito

$$V(\mathbb{Z}_p) \to V(x_{(0)}) \to V(x_{(1)}) \to \cdots \to V(x_{(h)}) \to \cdots$$

e assim, para determinar o vértice na altura $h \in \mathbb{N}$ do caminho que codifica x, precisamos buscar o h-ésimo dígito de x entre os p filhos do vértice $V(x_{(h-1)})$. Também, temos que duas bolas B_U e B_V, que são representadas pelos vértices U e V, satisfazem $B_V \subset B_U$ se e somente se, existe um caminho na árvore $\mathcal{A}(\mathbb{Z}_p)$ desde o vértice U até o vértice V. A distância em $\mathbb{Z}_p$ pode ser visualizada na árvore $\mathcal{A}(\mathbb{Z}_p)$ levando em conta que, pelo item **(iii)** da Proposição 4.1.1, se $x, y \in \mathbb{Z}_p$ são tais que $|x - y|_p = p^{-h}$, temos que os primeiros h dígitos de x e y coincidem e no dígito $h+1$ (o dígito correspondente à potência p^h da expansão canônica), diferem, logo h representa a altura da árvore p-ádica até onde os caminhos de x e y coincidem.

Exemplo 4.2.1. Sejam $x = \dots 00101000$ e $y = \dots 1010110$, inteiros 2-ádicos. Na Figura 2, traçamos os caminhos de x e y na árvore $\mathcal{A}(\mathbb{Z}_2)$ e como $|x - y|_2 = 2^{-1}$, vemos que os caminhos de x e y se separam a partir da altura $h = 1$.

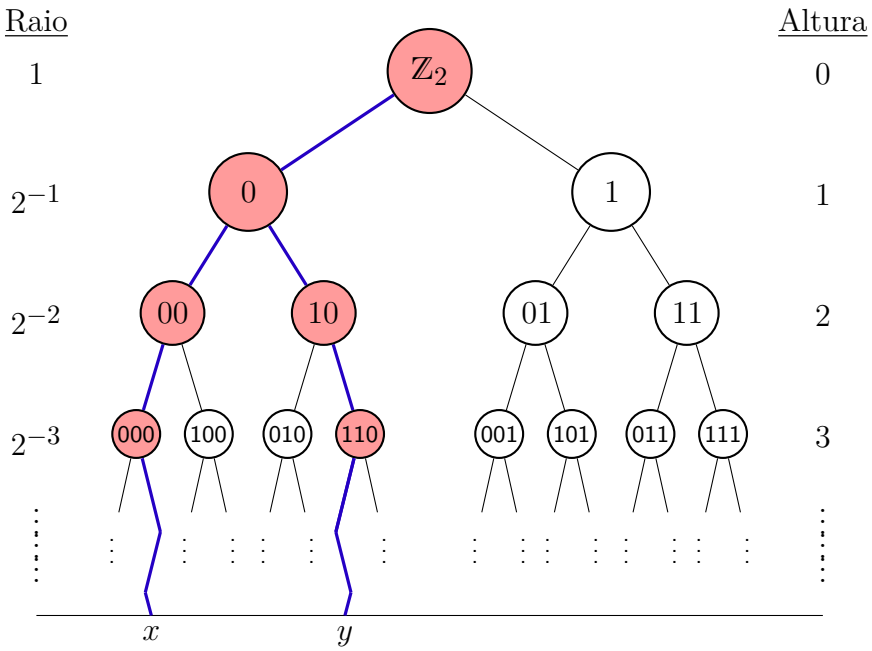

Figura 2 – Caminhos infinitos na árvore $\mathcal{A}(\mathbb{Z}_2)$, associados com $x, y \in \mathbb{Z}_2$ do Exemplo 4.2.1.

É claro que computacionalmente podemos trabalhar apenas com árvores de altura finita. Para utilizar essa estrutura na aprendizagem de máquina supervisionada, estaremos interessados nas sub-árvores de $\mathcal{A}(\mathbb{Z}_p)$ enraizadas em $V(\mathbb{Z}_p)$ que sejam a união dos caminhos traçados, até uma altura finita, pelos elementos de uma amostra rotulada. Vejamos um exemplo.

Exemplo 4.2.2. Seja $H = 3$, a altura máxima de nossa árvore e considere a amostra rotulada, $d_6 \in (\mathbb{Z}_2 \times \{0,1\})^6$:

$$d_6 = ((\ldots 0100, \mathbf{0}), (\ldots 1100, \mathbf{0}), (\ldots 0010, \mathbf{1}), (\ldots 1010, \mathbf{1}), (\ldots 0001, \mathbf{0}), (\ldots 1001, \mathbf{1})).$$

Para cada elemento da amostra, traçamos o caminho na árvore $\mathcal{A}(\mathbb{Z}_2)$ para cada altura $0 \leq h \leq 3$. Um par de informações úteis que podemos obter na hora de *visitar* cada vértice por parte dos elementos da amostra e que computacionalmente podemos estocar em um par de *variáveis inteiras*, digamos $n_v, s_v \in \mathbb{N}_0$; são o número de elementos da amostra que pertencem à bola representada pelo vértice e a soma dos rótulos[4] desses elementos, respectivamente. A situação é representada na Figura 3, onde o par en cada vértice representa às variáveis (n_v, s_v).

Observação 4.2.4. A árvore da Figura 3, possui vértices "vazios", ou seja, que não possuim as informações (n_v, s_v), isto é porque esses vértices representam bolas de $\mathbb{Z}_2$ que

[4] No caso de classificação binária $\{0,1\}$, no par (n_v, s_v), s_v representa o número de rótulos 1 nos n_v elementos da amostra que pertencem à respectiva bola representada por esse vértice. Assim, no caso de classificação binária, podemos usar (n_v, s_v) para registrar a votação em cada vértice.

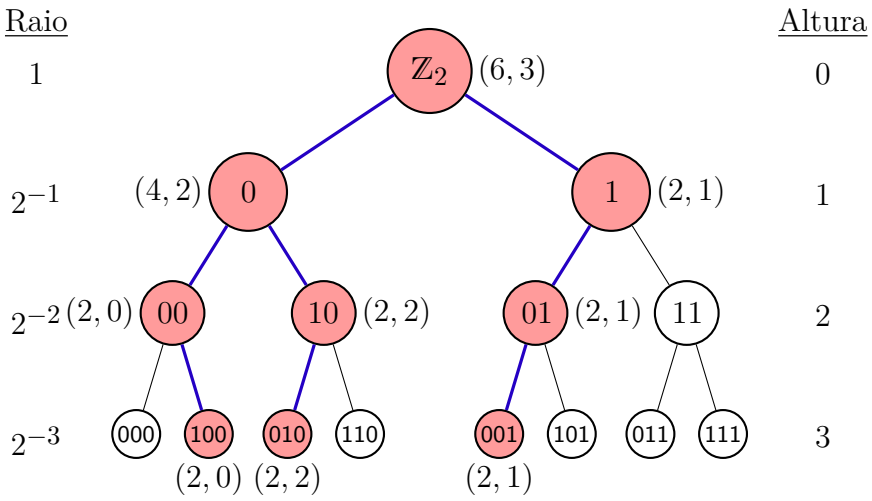

Figura 3 – Caminhos até a altura $H = 3$ na árvore $\mathcal{A}(\mathbb{Z}_2)$, associados com os elementos da amostra do Exemplo 4.2.2.

não contém elementos da amostra. Como computacionalmente não faz sentido salvar na memória esses vértices, vamos *podar* ou remover eles da árvore. Na Figura, 4 temos a árvore contendo apenas a união dos caminhos traçados pelos elementos de d_6 até $h = 3$.

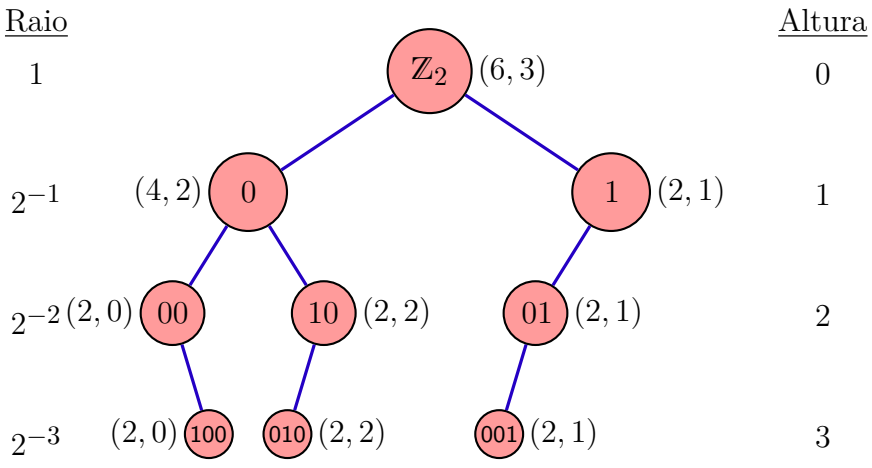

Figura 4 – Árvore do Exemplo 4.2.2 sem vértices vazios.

Como a intenção é tentar implementar computacionalmente esse tipo de estruturas, na seguinte definição vamos resumir as ideias anteriores *assumindo um entorno de trabalho computacional*.

Definição 4.2.3 ($\mathcal{A}_p^H(d_n)$: Árvore de decisão p-ádica de altura H gerada por d_n)**.** Sejam $p > 1$ um número primo, $d_n \in (\mathbb{Z}_p \times \{0,1\})^n$ uma amostra rotulada e $H \in \mathbb{N}$. Definimos a *árvore de decisão p-ádica de altura H gerada pela amostra rotulada* d_n, que denotaremos por $\mathcal{A}_p^H(d_n)$, como a árvore formada pela união dos caminhos até a altura $h = H$, começando pelo vértice raiz, $V(\mathbb{Z}_p)$, traçados na árvore $\mathcal{A}(\mathbb{Z}_p)$ pelos elementos da amostra

d_n, estocando em cada vértice as variáveis inteiras (n_v, s_v), sendo o número de elementos da amostra que pertencem à bola representada pelo vértice e a soma dos rótulos desses elementos, respectivamente.

Finalmente, e a modo de exemplo, a Figura 4 é uma representação gráfica da árvore de decisão 2-ádica $\mathcal{A}_2^3(d_6)$.

4.3 UMA REGRA DE APRENDIZAGEM SOBRE $(\mathbb{Z}_p^d, \|\cdot\|_p)$, COM $d \geq 1$

Agora vamos traduzir/utilizar as ideias precedentes para definir uma regra de aprendizagem sobre o espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$, com $\|(a_1, a_2, \ldots, a_d)\|_p = \max_{i\in[d]} |a_i|_p$ como na Definição 3.2.4. Primerio, traduziremos as ideias anteriores em dois algoritmos, Algoritmo 1 e 2, que definem uma regra de aprendizagem em $(\mathbb{Z}_p, |\cdot|_p)$. Uma vez feito isso e levando em conta o formato das bolas no espaço normado $(\mathbb{Q}_p^d, \|\cdot\|_p)$, estendemos de forma natural a regra de aprendizagem para o espaço $(\mathbb{Z}_p^d, \|\cdot\|_p)$, com $d > 1$.

4.3.1 Uma regra de aprendizagem sobre $(\mathbb{Z}_p, |\cdot|_p)$

Agora vamos formalizar as ideias predecentes para enunciar uma regra de aprendizagem sobre $(\mathbb{Z}_p, |\cdot|_p)$. Considere $H \in \mathbb{N}$ e $d_n = ((x_1, y_1), \ldots (x_n, y_n)) \in (\mathbb{Z}_p \times \{0, 1\})^n$, uma amostra rotulada de elementos de $\mathbb{Z}_p$.

Primeiro, para *certo* $H \in \mathbb{N}$, que vamos especificar mais para frente, vamos construir a árvore $\mathcal{A}_p^H(d_n)$ como feito no Exemplo 4.2.2 e depois vamos definir um *critério de decisão* sobre ela, obtendo assim, uma regra de aprendizagem no espaço ultramétrico $(\mathbb{Z}_p, |\cdot|_p)$.

Para construir a árvore $\mathcal{A}_p^H(d_n)$, podemos traçar a trajetória de cada elemento x_i do par $(x_i, y_i) \in d_n$, para $i \in [n]$, mediante, por exemplo, o ingresso de registros num dicionário[5], onde para cada altura $1 \leq h \leq H$, a *chave* de cada registro é apenas um dígito (o h-ésimo dígito de algúm elemento da amostra) e os *valores* do registro são as quantidades (n_v, s_v) junto com um conjuto de vértices filhos, que serão chamados de *campos de cada vértice*. Assim, cada vértice terá um campo n_v, um campo s_v e um campo $filhos$.

Formalmente, para cada vértice na altura $1 \leq h \leq H$ da árvore $\mathcal{A}_p^H(d_n)$ com identificador $\alpha_{(h)} \in D_p^h$, $V(\alpha_{(h)})$, o campo $n_v = n_v(\alpha_{(h)})$ representa o número de elementos da amostra que pertencem à bola de identificador $\alpha_{(h)}$ e $s_v = s_v(\alpha_{(h)})$ é a soma dos rótulos

[5] Além dos dicionários, em linguagens de programação como Python, podemos utilizar classes ou outro tipo de estruturas para implementar os Algoritmos 1 e 2. No Capítulo 6, fazemos algumas provas numéricas, onde a linguagem de programação usada é Python 3 e a implementação do classificador p-ádico foi feita utilizando classes.

desses elementos, ou seja,

$$n_v(\alpha_{(h)}) = \sum_{i=1}^{n} \mathbb{I}_{\bar{B}([|\alpha_{(h)}|],p^{-h})}(x_i),$$
$$s_v(\alpha_{(h)}) = \sum_{i=1}^{n} y_i \mathbb{I}_{\bar{B}([|\alpha_{(h)}|],p^{-h})}(x_i).$$

Com as informações definidas acima, o *pseudocódigo* para construir a árvore $\mathcal{A}_p^H(d_n)$ (ou simplesmente para treinar o classificador em $\mathbb{Z}_p$), é o Algoritmo 1, onde a sintaxe para denotar os diversos *campos* de cada vértice, é uma *mistura* feita a partir dos conceitos de dicionários e classes da linguagem Python.

Algoritmo 1 Construção árvore $\mathcal{A}_p^H(d_n)$

Input: $d_n = ((x_1, y_1), \dots (x_n, y_n)) \in (\mathbb{Z}_p \times \{0,1\})^n$, $H \in \mathbb{N}$, com $x_i = \sum_{j=0}^{\infty} \alpha_j^i p^j$, na forma canônica para cada $i \in [n]$.
Output: $\mathcal{A}$

$\mathcal{A}.raiz \leftarrow [(n_v, s_v), filhos]$ (Cria vértice raiz da árvore $\mathcal{A}$)
$\mathcal{A}.raiz.n_v \leftarrow n$ (Inicializa campos do vértice raiz)
$\mathcal{A}.raiz.s_v \leftarrow \sum_{i=1}^{n} y_i$
$\mathcal{A}.raiz.filhos \leftarrow \{\}$
for $i = 1$ **to** n **do**
  $vertice \leftarrow \mathcal{A}.raiz$
  for $j = 0$ **to** $H - 1$ **do**
    if $\alpha_j^i \in vertice.filhos$ **then**
      $vertice \leftarrow vertice.filhos[\alpha_j^i]$
      $vertice.n_v \leftarrow vertice.n_v + 1$ (Atualiza os campos (n_v, s_v))
      $vertice.s_v \leftarrow vertice.s_v + y_i$
    else
      $vertice.filhos[\alpha_j^i] \leftarrow [(n_v, s_v), filhos]$ (Cria vértice com chave α_j^i)
      $vertice \leftarrow vertice.filhos[\alpha_j^i]$
      $vertice.n_v \leftarrow 1$ (Inicializa os campos do novo vértice)
      $vertice.s_v \leftarrow y_i$
      $vertice.filhos \leftarrow \{\}$
    end if
  end for
end for

Observação 4.3.1. Usando as ideias precedentes, dada uma árvore $\mathcal{A}_p^H(d_n)$, se $x \in \mathbb{Z}_p$ pertence a uma bola da *partição de* $\mathbb{Z}_p$ *à altura* $1 \leq h \leq H$ que contém elementos da amostra d_n, então essa bola estará representada na árvore $\mathcal{A}_p^H(d_n)$ por um vértice com campos $(n_v(x_{(h)}), s_v(x_{(h)}))$, assim, podemos considerar esses $n_v(x_{(h)})$ elementos da amostra como *vizinhos próximos de* x *à altura* h, pois pela Proposição 3.1.5, no mundo não-arquimediano todo elemento de uma bola é o centro da bola. Além disso, a medida

que subimos na árvore $\mathcal{A}_p^H(d_n)$ pelo caminho traçado por x, o raio das bolas representadas por esses vértices diminui, por isso, quanto maior é a altura de um vértice no caminho de x, os vizinhos dele na bola correspondente serão vizinhos *cada vez mais próximos*.

Agora, para rotular um elemento $x \in \mathbb{Z}_p$ usando a árvore $\mathcal{A}_p^H(d_n)$, precisamos de um *critério de decisão*, isto é, um critério para decidir em qual momento da execução do algoritmo de busca na árvore $\mathcal{A}_p^H(d_n)$ devemos fazer a votação para obter o rótulo predito de x. Da Observação 4.3.1, podemos usar como critério de decisão, que a votação para rotular um elemento $x \in \mathbb{Z}_p$ seja realizada *à maior altura possível* no caminho percorrido por x, sempre que se respeite o número mínimo (quórum) de elementos da amostra para fazer a votação, assim, adicionando um hiperparâmetro $k \in \mathbb{N}$, com $1 \leq k \leq n$, podemos utilizar esse número como *critério de decisão* da seguinte maneira:

(1) Procuramos o menor $H \in \mathbb{N}$ tal que *todo* vértice, $V(\alpha_{(H)})$, à altura $h = H$ da árvore $\mathcal{A}_p^H(d_n)$, satisfaça $n_v(\alpha_{(H)}) < k$, se $k > 1$ ou a condição óbvia, $n_v(\alpha_{(H)}) = 1$ no caso $k = 1$. Construir para o mesmo k uma árvore de altura maior, $H' > H$, só provocará maior gasto de tempo e memória sem melhora do desempenho, pois as predições serão as mesmas que para H. Uma árvore $\mathcal{A}_p^H(d_n)$ com $H \in \mathbb{N}$ com essas características, será chamada de árvore de *altura minimal*.

(2) Para rotular $x \in \mathbb{Z}_p$, traçamos o caminho de x em $\mathcal{A}_p^H(d_n)$ subindo na árvore sempre que em cada vértice no caminho de x, o número de vizinhos da amostra seja pelo menos k, isto é, sempre que seja satisfeita a condição $n_v(x_{(h)}) \geq k$. Denotemos por $1 \leq h_f(x) \leq H$, a *menor altura* onde o vértice visitado por x satisfaz $n_v(x_{(h_f(x))}) < k$. Note que nas condições do item **(1)**, a altura $1 \leq h_f(x) \leq H$ sempre existe no caso $k > 1$. Nessa situação, o vértice no caminho traçado por x à altura $h_k(x) = h_f(x) - 1$, por hipótese satisfaz $n_v(x_{(h_k(x))}) \geq k$, logo, ao fazer a votação com as informações desse vértice, estaremos fazendo a votação com os elementos da amostra que pertencem à bola centrada em x de menor raio[6] que contém *pelo menos k elementos de* d_n.

(3) Se ao traçar o caminho de $x \in \mathbb{Z}_p$ na árvore $\mathcal{A}_p^H(d_n)$ nos encontramos com a situação de que o vértice visitado por x à altura $1 \leq h(x) < H$ satisfaz $n_v(x_{(h(x))}) \geq k$ e a bola $\bar{B}(x, p^{-(h(x)+1)})$ não tem representante na árvore $\mathcal{A}_p^H(d_n)$, então isso quer dizer que a quantidade de elementos da amostra d_n que pertencem à bola $\bar{B}(x, p^{-(h(x)+1)})$ será *zero*, que é menor do que $k \geq 1$, portanto nesse caso calculamos o voto majoritário com as informações da bola $\bar{B}(x, p^{-h(x)})$ e assim estaremos fazendo a votação com os elementos da amostra que pertencem à bola centrada em x de menor raio que contém *pelo menos k elementos de* d_n.

[6] Especificações serão dadas no Capítulo 5.

Utilizando o Algoritmo 1 para obter uma árvore $\mathcal{A}_p^H(d_n)$ de altura minimal como no item **(1)**, o pseudocódigo que resume os itens **(2)** e **(3)** para rotular um elemento $x \in \mathbb{Z}_p$, é o Algoritmo 2.

Algoritmo 2 Classificação em $\mathbb{Z}_p$

Input: Recebe $x = \sum_{j=0}^{\infty} \alpha_j p^j \in \mathbb{Z}_p$, $\mathcal{A} = \mathcal{A}_p^H(d_n)$ e $k \in \mathbb{N}$.
Output: $\mathbb{I}_{[\frac{1}{2},1]}(vertice_{atual}.s_v / vertice_{atual}.n_v)$.

$vertice_{atual} \leftarrow \mathcal{A}.raiz$
for $j = 0$ **to** $H - 1$ **do**
  if $\alpha_j \in vertice_{atual}.filhos$ **then**
    $vertice \leftarrow vertice_{atual}.filhos[\alpha_j]$
  else
    $vertice \leftarrow [(0,0), \{\}]$
  end if
  if $vertice.n_v < k$ **then**
    break
  else
    $vertice_{atual} \leftarrow vertice$
  end if
end for

Resumindo: para *treinar* o classificador em $(\mathbb{Z}_p, |\cdot|_p)$, que chamaremos simplesmente de *classificador p-ádico*, devemos obter a árvore $\mathcal{A}_p^H(d_n)$ de altura minimal mediante o Algoritmo 1, e para *predizer* o rótulo de um elemento $x \in \mathbb{Z}_p$, utilizamos esses resultados como entrada do Algoritmo 2 para logo obter a predição do rótulo de x.

Observação 4.3.2. Embora os algoritmos 1 e 2 sejam a implementação da regra ^+k-NN, eles não valem para o caso euclidiano, pois eles surgem das propriedades das bolas na topologia dos números p-ádicos, propriedades que não são válidas no caso euclidiano.

4.3.2 Extensão para $(\mathbb{Z}_p^d, \|\cdot\|_p)$, com $d > 1$ da regra de aprendizagem sobre $(\mathbb{Z}_p, |\cdot|_p)$

Agora vamos estender o algoritmo de classificação p-ádico para o caso d-dimensional, onde $d > 1$. Para fazer isso, da mesma maneira de como é feito no caso euclidiano, vamos dotar o $\mathbb{Q}_p$-espaço vetorial $(\mathbb{Q}_p^d, +, \cdot)$ com a métrica induzida por uma norma.

Na Seção 3.2.3, vimos que no espaço $\mathbb{Q}_p^d$ *todas as normas são equivalentes* e, graças à completitude de $\mathbb{Q}_p$, com qualquer uma delas $\mathbb{Q}_p^d$ é um *espaço de Banach* que também é um espaço métrico separável.

Mesmo que todas as normas gerem a mesma topologia em $\mathbb{Q}_p^d$, o formato de uma bola varía segundo a norma. Para obter uma versão d-dimensional da Proposição 4.1.2, que é o fato chave que permite definir o algoritmo em $(\mathbb{Z}_p, |\cdot|_p)$, vamos usar a norma $\|\cdot\|_p$ da Definição 3.2.4, pois como veremos na seguinte proposição, com ela uma bola no espaço produto $(\mathbb{Q}_p^d, \|\cdot\|_p)$ é simplesmente o produto de bolas no espaço $(\mathbb{Q}_p, |\cdot|_p)$.

Proposição 4.3.1 (Formato das bolas em $(\mathbb{Q}_p^d, \|\cdot\|_p)$)**.** *Sejam $a = (a_1, a_2, \ldots, a_d) \in \mathbb{Q}_p^d$, e $h \in \mathbb{Z}$, quaisquer. Então*

$$\bar{B}_d(a, p^h) = \prod_{i=1}^{d} \bar{B}(a_i, p^h).$$

Demonstração.

$$\begin{aligned}
\bar{B}_d(a, p^h) &= \{x \in \mathbb{Q}_p^d : \|x - a\|_p \leq p^h\} \\
&= \{x \in \mathbb{Q}_p^d : \max_{i \in [d]} |x_i - a_i|_p \leq p^h\} \\
&= \{x \in \mathbb{Q}_p^d : |x_i - a_i|_p \leq p^h, \forall 1 \leq i \leq d\} \\
&= \prod_{i=1}^{d} \{x_i \in \mathbb{Q}_p : |x_i - a_i|_p \leq p^h\} \\
&= \prod_{i=1}^{d} \bar{B}(a_i, p^h).
\end{aligned}$$

■

Da proposição anterior, é imediato conferir que os items **(i)** e **(ii)** da Proposição 4.1.1 e o fato de que todo elemento da bola é o centro da bola, continuam valendo para as bolas em $(\mathbb{Q}_p^d, \|\cdot\|_p)$. O item **(iii)** da Proposição 4.1.1 também vale para uma noção de *dígito d-dimensional* que veremos na próxima seção, e que chamaremos de *dígito de um vetor p-ádico.*

Agora estamos em condições de enunciar uma versão para o caso d-dimensional da Proposição 4.1.2.

Proposição 4.3.2. *Sejam $d, h \in \mathbb{N}$ quaisquer. No espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$, temos*

$$\mathbb{Z}_p^d = \bigcup_{\alpha_{(h)}^{(d)} \in (D_p^h)^d} \bar{B}_d\left([|\alpha_{(h)}^{(d)}|], p^{-h}\right),$$

onde $\alpha_{(h)}^{(d)} := \left(\alpha_{(h)}^1, \alpha_{(h)}^2, ..., \alpha_{(h)}^d\right) \in (D_p^h)^d = \prod_{i=1}^d D_p^h$ *e*

$$[|\alpha_{(h)}^{(d)}|] := \left([|\alpha_{(h)}^1|], [|\alpha_{(h)}^2|], \ldots, [|\alpha_{(h)}^d|]\right) = \left(\sum_{n=0}^{h-1} \alpha_n^1 p^n, \sum_{n=0}^{h-1} \alpha_n^2 p^n, \ldots, \sum_{n=0}^{h-1} \alpha_n^d p^n\right) \in \mathbb{Z}_p^d.$$

Quando for necessário, usaremos a notação simplificada, $\alpha_{(h)}^i = \alpha_{h-1}^i \ldots \alpha_1^i \alpha_0^i$ e nessa versão, $\alpha_{(h)}^{(d)}$, denotará o identificador da bola $\bar{B}_d\left([|\alpha_{(h)}^{(d)}|], p^{-h}\right)$.

Demonstração. Da Proposição 4.1.2, para $h \in \mathbb{N}$ qualquer, temos

$$\begin{aligned}\mathbb{Z}_p^d &= \prod_{i=1}^{d}\left[\bigcup_{\alpha_{(h)}^i \in D_p^h} \bar{B}([|\alpha_{(h)}^i|], p^{-h})\right] \\ &= \bigcup_{\alpha_{(h)}^1 \in D_p^h}\ \bigcup_{\alpha_{(h)}^2 \in D_p^h} \cdots \bigcup_{\alpha_{(h)}^d \in D_p^h} \prod_{i=1}^{d} \bar{B}([|\alpha_{(h)}^i|], p^{-h}) \\ &= \bigcup_{(\alpha_{(h)}^1, \alpha_{(h)}^2, \ldots, \alpha_{(h)}^d) \in (D_p^h)^d} \prod_{i=1}^{d} \bar{B}([|\alpha_{(h)}^i|], p^{-h})\end{aligned}$$

que combinado com a Proposição 4.3.1, fornece

$$\mathbb{Z}_p^d = \bigcup_{\alpha_{(h)}^{(d)} \in (D_p^h)^d} \bar{B}_d\left([|\alpha_{(h)}^{(d)}|], p^{-h}\right)$$

onde $\alpha_{(h)}^{(d)} := \left(\alpha_{(h)}^1, \alpha_{(h)}^2, ..., \alpha_{(h)}^d\right) \in (D_p^h)^d$ e $[|\alpha_{(h)}^{(d)}|] := \left([|\alpha_{(h)}^1|], [|\alpha_{(h)}^2|], \ldots, [|\alpha_{(h)}^d|]\right) \in \mathbb{Z}_p^d$. ■

A seguir, faremos algumas observações sobre as consequências da Proposição 4.3.2.

Observação 4.3.3 (Número de elementos da partição de $\mathbb{Z}_p^d$ à altura h)**.** Para $h \in \mathbb{N}$ qualquer, o produto cartesiano $(D_p^h)^d = \prod_{i=1}^{d} D_p^h$ possui $\sharp\left((D_p^h)^d\right) = p^{hd}$ elementos. Logo, para cada $h \in \mathbb{N}$, da Proposição 4.3.2, temos uma partição de $\mathbb{Z}_p^d$ em p^{hd} bolas de raio p^{-h} que podemos indexar utilizando os identificadores $\alpha_{(h)}^{(d)}$ na versão simplificada.

Observação 4.3.4. Para $h \in \mathbb{N}$ e pela Proposição 4.3.2, o vetor p-ádico

$$x = \left(\sum_{n=0}^{\infty} \alpha_n^1 p^n, \sum_{n=0}^{\infty} \alpha_n^2 p^n, \ldots, \sum_{n=0}^{\infty} \alpha_n^d p^n\right) \in \mathbb{Z}_p^d,$$

pertence à bola da partição de $\mathbb{Z}_p^d$ em bolas fechadas de raio p^{-h}, que tem identificador

$$x_{(h)}^{(d)} = (x_{(h)}^1, x_{(h)}^2, \ldots, x_{(h)}^d) = (\alpha_{h-1}^1 \ldots \alpha_1^1 \alpha_0^1, \alpha_{h-1}^2 \ldots \alpha_1^2 \alpha_0^2, \ldots, \alpha_{h-1}^d \ldots \alpha_1^d \alpha_0^d).$$

Observação 4.3.5 (Árvore p^d-ária $\mathcal{A}(\mathbb{Z}_p^d)$)**.** No caso d-dimensional, o conceito de altura de uma bola em $\mathbb{Z}_p^d$ é análogo ao da Definição 4.2.1, mas neste caso a altura de uma bola faz menção à partição de $\mathbb{Z}_p^d$ em p^{hd} bolas fechadas de raio p^{-h}. Também, de forma análoga a como foi feito para $\mathcal{A}(\mathbb{Z}_p)$, podemos definir a árvore $\mathcal{A}(\mathbb{Z}_p^d)$, como a árvore p^d-ária cheia, enraizada no vértice $V(\mathbb{Z}_p^d)$, que representa $\mathbb{Z}_p^d$, cujos vértices representam as bolas no espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$ e as regras para desenhar uma aresta entre vértices são as mesmas que na Definição 4.2.2.

Observação 4.3.6 (Codificando um caminho infinito na arvore $\mathcal{A}(\mathbb{Z}_p^d)$)**.** Como feito anteriormente, cada

$$x = \left(\sum_{n=0}^{\infty} \alpha_n^1 p^n, \sum_{n=0}^{\infty} \alpha_n^2 p^n, \ldots, \sum_{n=0}^{\infty} \alpha_n^d p^n \right) \in \mathbb{Z}_p^d,$$

corresponde *biunivocamente* com um *caminho infinito* na árvore p-ádica d-dimensional, $\mathcal{A}(\mathbb{Z}_p^d)$, começando no vértice raiz, $V(\mathbb{Z}_p^d)$. Denotemos por $V(\gamma_{(h)}^{(d)})$, o vértice à altura h representando à bola com identificador

$$\gamma_{(h)}^{(d)} = (\gamma_{h-1}^1 \ldots \gamma_1^1 \gamma_0^1, \gamma_{h-1}^2 \ldots \gamma_1^2 \gamma_0^2, \ldots, \gamma_{h-1}^d \ldots \gamma_1^d \gamma_0^d).$$

Então, $x \in \mathbb{Z}_p^d$ como acima, pode ser *codificado* pelo caminho infinito

$$V(\mathbb{Z}_p^d) \to V(x_{(0)}^{(d)}) \to V(x_{(1)}^{(d)}) \to \cdots \to V(x_{(h)}^{(d)}) \to \cdots$$

e assim, para determinar o vértice na altura $h \in \mathbb{N}$ do caminho que codifica x, precisamos buscar entre os p^d filhos do vértice $V(x_{(h-1)}^{(d)})$, a d-tupla cuja i-ésima componente, com $i \in [d]$, é o h-ésimo dígito de $x_i = \sum_{n=0}^{\infty} \alpha_n^i p^n$, isto é, buscamos a tupla $(\alpha_{h-1}^1, \alpha_{h-1}^2, \ldots, \alpha_{h-1}^d)$. Também temos que duas bolas B_U e B_V, representadas pelos vértices U e V, satisfazem $B_V \subset B_U$ se e somente se, existe um caminho na árvore $\mathcal{A}(\mathbb{Z}_p^d)$ desde o vértice U até o vértice V.

Agora vamos dar um nome à tupla de dígitos que determina cada vértice do caminho de um vetor na árvore $\mathcal{A}(\mathbb{Z}_p^d)$.

Definição 4.3.1 (Dígito de um vetor de $(\mathbb{Q}_p^d, \|\cdot\|_p)$)**.** Seja $x = (x_1, x_2, ..., x_d) \in \mathbb{Q}_p^d$, onde os números $x_i = \sum_{j=-N_i}^{\infty} \alpha_j^i p^j$ estão na sua expansão canônica. Para $j \in \mathbb{Z}$, definimos o j-ésimo dígito do vetor p-ádico x, que denotaremos por $\mathcal{D}_p^j(x)$, como a d-tupla cuja i-ésima componente é o dígito correspondente com a j-ésima potência na expansão de x_i:

$$\mathcal{D}_p^j(x) = (\alpha_j^1, \alpha_j^2, \ldots, \alpha_j^d) \in D_p^d,$$

onde se para certo $x_i = \sum_{j=-N_i}^{\infty} \alpha_j^i p^j$ temos $j < -N_i$, então $\alpha_j^i = 0$.

Exemplo 4.3.1. Considere o vetor 5-ádico,

$$x = (\ldots 41223\,.\,1224, \ldots 11103\,.\,002, \ldots 22334\,.\,33) \in \mathbb{Q}_5^3.$$

A norma e os dígitos de x para alguns valores de $j \in \mathbb{Z}$: $\|x\|_5 = |\ldots 41223\,.\,1224|_5 = 5^4$, $\mathcal{D}_5^{-4}(x) = (4, 0, 0)$, $\mathcal{D}_5^{-1}(x) = (1, 0, 3)$, $\mathcal{D}_5^0(x) = (3, 3, 4)$ e $\mathcal{D}_5^3(x) = (1, 1, 2)$.

Com essa noção de dígito d-dimensional, podemos provar a versão vetorial da Proposição 4.1.1.

Proposição 4.3.3. *No espaço ultramétrico* $(\mathbb{Q}_p^d, \|\cdot\|_p)$*, com* $d \geq 1$*, para* $a \in \mathbb{Q}_p^d$ *arbitrário, temos*

(i) *Para todo* $h \in \mathbb{Z}$*,* $\bar{B}_d(a, p^h) = B_d(a, p^{h+1})$*.*

(ii) *Para* $\epsilon > 0$ *qualquer, existe* $h \in \mathbb{Z}$ *tal que*

$$\bar{B}_d(a, \epsilon) = \bar{B}_d(a, p^h).$$

(iii) *Para* $x, y \in \mathbb{Q}_p^d$*, e* $h \in \mathbb{Z}$*, quaisquer*

$$\|x - y\|_p = p^{-\min\{j \in \mathbb{Z}: \mathcal{D}_p^j(x) \neq \mathcal{D}_p^j(y)\}}$$

e

$$\|x - y\|_p \leq p^h \Leftrightarrow \mathcal{D}_p^j(x) = \mathcal{D}_p^j(y), \forall j \leq -h - 1.$$

Demonstração. **(i),(ii)** A prova dos itens **(i)** e **(ii)** é imediata, pois uma bola no espaço ultramétrico $(\mathbb{Q}_p^d, \|\cdot\|_p)$ é produto de bolas do mesmo raio no espaço $(\mathbb{Q}_p, |\cdot|_p)$.

(iii) Sejam $x, y \in \mathbb{Q}_p^d$, cujas componentes satisfazem, para todo $i \in [d]$ $x_i = \sum_{j=-N_i}^{\infty} \alpha_j^i p^j$ e $y_i = \sum_{j=-M_i}^{\infty} \gamma_j^i p^j$. Primeiro observamos que existe $m \in [d]$, tal que

$$\|x - y\|_p = \max_{i \in [d]} |x_i - y_i| = |x_m - y_m|_p \geq |x_i - y_i|_p, \forall i \in [d].$$

Logo, pelo item **(iii)** da Proposição 4.1.1, $|x_i - y_i|_p = p^{-\min\{j \in \mathbb{Z}: \alpha_j^i \neq \gamma_j^i\}}$, para todo $i \in [d]$ e

$$j_m := \min\{j \in \mathbb{Z} : \alpha_j^m \neq \gamma_j^m\} \leq \min\{j \in \mathbb{Z} : \alpha_j^i \neq \gamma_j^i\}, \forall i \in [d],$$

assim

$$\alpha_j^i = \gamma_j^i, \forall j \leq j_m - 1, \forall i \in [d],$$

isto é,

$$\mathcal{D}_p^j(x) = \mathcal{D}_p^j(y), \forall j \leq j_m - 1$$

e $\mathcal{D}_p^{j_m}(x) \neq \mathcal{D}_p^{j_m}(y)$, pois $\alpha_{j_m}^m \neq \gamma_{j_m}^m$. Portanto

$$\|x - y\|_p = p^{-\min\{j \in \mathbb{Z}: \mathcal{D}_p^j(x) \neq \mathcal{D}_p^j(y)\}}.$$

Agora, seja $h \in \mathbb{Z}$. Então

$$\|x - y\|_p \leq p^h \Leftrightarrow \max_{i \in [d]} |x_i - y_i|_p \leq p^h \Leftrightarrow |x_i - y_i|_p \leq p^h, \forall i \in [d],$$

e novamente pelo item **(iii)** da Proposição 4.1.1

$$|x_i - y_i|_p \leq p^h, \forall i \in [d] \Leftrightarrow \alpha_j^i = \gamma_j^i, \forall j \leq -h - 1, \forall i \in [d],$$

ou seja, $\|x - y\|_p \leq p^h \Leftrightarrow \mathcal{D}_p^j(x) = \mathcal{D}_p^j(y), \forall j \leq -h - 1$.

■

Observação 4.3.7. Pela Proposição 4.3.3, e do mesmo modo que no espaço $(\mathbb{Z}_p, |\cdot|_p)$, a distância em $\mathbb{Z}_p^d$ pode ser visualizada na árvore $\mathcal{A}(\mathbb{Z}_p^d)$, pois se $x, y \in \mathbb{Z}_p^d$ são tais que $\|x-y\|_p = p^{-h}$, temos que os primeiros h dígitos coincidem, $\mathcal{D}_p^j(x) = \mathcal{D}_p^j(y)$, $0 \leq j \leq h-1$, e no próximo dígito diferem, $\mathcal{D}_p^h(x) \neq \mathcal{D}_p^h(y)$, logo h representa a altura da árvore p-ádica d-dimensional até onde os caminhos de x e y coincidem.

Finalmente, com essa noção de dígito d-dimensional, podemos definir, como feito com a árvore $\mathcal{A}_p^H(d_n)$, a árvore de decisão p-ádica d-dimensional de altura minimal H gerada por d_n, que denotamos por $\mathcal{A}_{p^d}^H(d_n)$[7], e construir essa árvore aplicando o Algoritmo 1, onde as chaves dos vértices na altura h, dessa vez serão os dígitos de dimensão d, $\mathcal{D}_p^j(x_i) \in D_p^d$, dos elementos $x_i \in \mathbb{Z}_p^d$ da amostra rotulada $d_n \in (\mathbb{Z}_p^d \times \{0,1\})^n$. Assim, para construir a árvore $\mathcal{A}_{p^d}^H(d_n)$, só precisamos trocar α_j^i por $\mathcal{D}_p^j(x_i)$ no Algoritmo 1, e para rotular elementos de $\mathbb{Z}_p^d$, no Algoritmo 2 precisamos trocar α_j por $\mathcal{D}_p^j(x)$. Para futuras referências, escrevemos essas modificações para o caso $d \geq 1$ nos algoritmos 3 e 4.

Algoritmo 3 Construção árvore $\mathcal{A}_{p^d}^H(d_n)$

Input: $d_n = ((x_1, y_1), \ldots (x_n, y_n)) \in (\mathbb{Z}_p^d \times \{0,1\})^n$, e $H \in \mathbb{N}$, com $x_i = (x_1^i, x_2^i, \ldots, x_d^i)$ e $x_l^i = \sum_{j=0}^{\infty} \alpha_j^{l,i} p^j$, na forma canônica para cada $l \in [d]$ e $i \in [n]$.
Output: $\mathcal{A}$

$\mathcal{A}.raiz \leftarrow [(n_v, s_v), filhos]$ (Cria vértice raiz da árvore $\mathcal{A}$)
$\mathcal{A}.raiz.n_v \leftarrow n$ (Inicializa campos do vértice raiz)
$\mathcal{A}.raiz.s_v \leftarrow \sum_{i=1}^{n} y_i$
$\mathcal{A}.raiz.filhos \leftarrow \{\}$
for $i = 1$ **to** n **do**
  $vertice \leftarrow \mathcal{A}.raiz$
  for $j = 0$ **to** $H-1$ **do**
    if $\mathcal{D}_p^j(x_i) \in vertice.filhos$ **then**
      $vertice \leftarrow vertice.filhos[\mathcal{D}_p^j(x_i)]$
      $vertice.n_v \leftarrow vertice.n_v + 1$ (Atualiza os campos (n_v, s_v))
      $vertice.s_v \leftarrow vertice.s_v + y_i$
    else
      $vertice.filhos[\mathcal{D}_p^j(x_i)] \leftarrow [(n_v, s_v), filhos]$ (Cria vértice com chave $\mathcal{D}_p^j(x_i)$)
      $vertice \leftarrow vertice.filhos[\mathcal{D}_p^j(x_i)]$
      $vertice.n_v \leftarrow 1$ (Inicializa os campos do novo vértice)
      $vertice.s_v \leftarrow y_i$
      $vertice.filhos \leftarrow \{\}$
    end if
  end for
end for

[7] A potência p^d na notação, é pelo fato que a árvore gerada por uma amostra é uma subárvore de $\mathcal{A}(\mathbb{Z}_p^d)$, que por sua vez é uma árvore p^d-ária cheia.

Algoritmo 4 Classificação em $\mathbb{Z}_p^d$

Input: Recebe $x = (x_1, x_2, \ldots, x_d)$ com $x_i = \sum_{j=0}^{\infty} \alpha_j^i p^j \in \mathbb{Z}_p$, $i \in [d]$ na forma canônica, $\mathcal{A} = \mathcal{A}_{p^d}^H(d_n)$, H e $k \in \mathbb{N}$.
Output: $\mathbb{I}_{[\frac{1}{2},1]}(vertice_{atual}.s_v / vertice_{atual}.n_v)$.

$vertice_{atual} \leftarrow \mathcal{A}.raiz$
for $j = 0$ **to** $H - 1$ **do**
  if $\mathcal{D}_p^j(x) \in vertice_{atual}.filhos$ **then**
    $vertice \leftarrow vertice_{atual}.filhos[\mathcal{D}_p^j(x)]$
  else
    $vertice \leftarrow [(0,0), \{\}]$
  end if
  if $vertice.n_v < k$ **then**
    break
  else
    $vertice_{atual} \leftarrow vertice$
  end if
end for

Em conclusão: a regra de aprendizagem no espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$, com $d \in \mathbb{N}$, definida pelos algoritmos 3 e 4, será chamada simplesmente de *Regra de Aprendizagem p-ádica*, e será denotada por $\mathcal{L}_{(k,p)} = (g_n^{(k,p)})_{n=1}^{\infty}$.

Observação 4.3.8. Ressaltamos que embora cada nó da árvore $\mathcal{A}(\mathbb{Z}_p^d)$ possui p^d filhos, para $d \in \mathbb{N}$ *grande*, não temos uma *explosão* de nós na hora de almacenar computacionalmente a árvore $\mathcal{A}_{p^d}^H(d_n)$, pois para sua construção consideramos apenas os nós dados pelos caminhos traçados pelos elementos da amostra d_n. Graças a isso, e como veremos nos experimentos numéricos do Capítulo 6, os algoritmos 3 e 4 não *sofrem* dramaticamente os efeitos de dimensões elevadas.

4.4 UMA REGRA DE APRENDIZAGEM NO ESPAÇO $[0,1]^d \subset \mathbb{R}_+^d$ BASEADA NA REGRA DE APRENDIZAGEM p-ÁDICA

Agora veremos a última seção do capítulo. Aqui, utilizando a técnica de *redução de dimensionalidade boreliana* (ver [52, 53]), vamos definir uma regra de aprendizagem sobre $[0,1]^d \subset \mathbb{R}^d$ baseada na regra p-ádica definida na seção anterior.

Para conseguir nosso objetivo, observamos que a métrica p-ádica está fortemente ligada a aritmética de $\mathbb{Q}$, por essa razão, primeiro veremos alguns resultados relativos à aritmética de $\mathbb{R}_+$ que vão nos permitir construir uma aplicação *injetora* e *boreliana*, $\phi_p : (\mathbb{R}_+, |\cdot|_\infty) \to (\mathbb{Q}_p, |\cdot|_p)$. Logo, vamos estender ϕ_p de forma natural para o espaço $\mathbb{R}_+^d$, obtendo a aplicação injetora e boreliana $\Phi_p : \mathbb{R}_+^d \to \mathbb{Q}_p^d$, aplicação cuja restrição ao cubo $[0,1]^d$, iremos compor com a regra de aprendizagem p-ádica na forma que indica a Definição 2.3.1, para assim obter uma regra de aprendizagem sobre o cubo $[0,1]^d$, com

$d \geq 1$, que usufrui da estrutura hierárquica do espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$.

Vamos começar com um resultado simples que mostra que a *expansão na base* $b > 1$ de qualquer número real define uma série convergente ou de soma finita, e assim, graças a esse resultado, podemos manipular algebricamente essas expressões sem medo de errar. Os resultados gerais serão feitos para uma base $b > 1$ qualquer.

Proposição 4.4.1. *Seja* $b \in \mathbb{N}$, $b > 1$. *Para cada* $\{\alpha_n\}_{n=-N}^{\infty} \subset D_b$, *com* $N \in \mathbb{Z}$, *temos* $\sum_{n=-N}^{\infty} \alpha_n b^{-n} < \infty$, *no espaço métrico* $(\mathbb{R}, |\ |_\infty)$.

Demonstração. Considere $\{\alpha_n\}_{n=-N}^{\infty} \subset D_b$ arbitrária. Para $M \in \mathbb{Z}$, $M \geq -N+1$, temos

$$S_M = \sum_{n=-N}^{M} \alpha_n b^{-n} \leq (b-1) \sum_{n=-N}^{\infty} b^{-n}.$$

Se $N \leq 0$, então

$$\sum_{n=-N}^{\infty} b^{-n} \leq \sum_{n=0}^{\infty} b^{-n} = \frac{b}{b-1} =: C$$

e se $N > 0$,

$$\sum_{n=-N}^{\infty} b^{-n} = \sum_{n=-N}^{-1} b^{-n} + \sum_{n=0}^{\infty} b^{-n} = \sum_{n=-N}^{-1} b^{-n} + \frac{b}{b-1} =: C.$$

Em qualquer caso, temos

$$S_M \leq (b-1) \sum_{n=-N}^{\infty} b^{-n} \leq (b-1)C,$$

logo

$$\sum_{n=-N}^{\infty} \alpha_n b^{-n} = \lim_{M \to \infty} S_M \leq (b-1)C < \infty.$$

■

Outro resultado útil e simples é o seguinte.

Lema 4.4.1. *Para* $b \in \mathbb{N}$, $b > 1$ *e para qualquer* $N \in \mathbb{Z}$, *temos*

$$b^{-N} = \sum_{n=N+1}^{\infty} (b-1) b^{-n}.$$

Demonstração. Considere $M \geq N+1$ e $S_M := \sum_{n=N+1}^{M} (b-1)b^{-n}$, logo, expandindo a soma obtemos $S_M = b^{-N} - b^{-M}$, de onde

$$|S_M - b^{-N}|_\infty = b^{-M} \overset{M \to \infty}{\longrightarrow} 0.$$

■

Observação 4.4.1. Da Proposição 4.4.1, vemos que toda série $\sum_{n=-N}^{\infty} \alpha_n b^{-n}$ define um número real não negativo, e por outro lado, também temos que todo número real não negativo possui uma expansão na base $b > 1$ nesse formato. Assim, como estamos procurando uma aplicação de $\mathbb{R}_+ \to \mathbb{Q}_p$, se levamos em conta as expressões (10) e (11) podemos facilmente relacionar um número real com um número p-ádico mediante a relação $\bar{\phi}_p \subset \mathbb{R}_+ \times \mathbb{Q}_p$, definida por

$$\bar{\phi}_p \left(\sum_{n=-N}^{\infty} \alpha_n p^{-n} \right) = \sum_{n=-N}^{\infty} \alpha_n p^n.$$

A relação $\bar{\phi}_p$ *não é uma função*, pois do Lema 4.4.1, vemos que $1 = \sum_{n=1}^{\infty}(p-1)p^{-n}$ e portanto, $\bar{\phi}_p(1) = 1 \neq \sum_{n=1}^{\infty}(p-1)p^n = \bar{\phi}_p(\sum_{n=1}^{\infty}(p-1)p^{-n})$, já que a representação nos números p-ádicos, é única. Contudo, e pela mesma razão, a relação inversa $\bar{\phi}_p^{-1}$, sim é uma função, $\bar{\phi}_p^{-1} : \mathbb{Q}_p \to \mathbb{R}_+$. Como todo número real positivo possui uma expansão na base $p > 1$, a função $\bar{\phi}_p^{-1}$ é sobrejetora, mas pelo Lema 4.4.1, vemos que não é injetora. O problema da ambiguidade de $\bar{\phi}_p$, surge na situação do Lema 4.4.1, ou seja, com as sequências de dígitos $\{\alpha_n\}_{n=-N}^{\infty} \subset D_p$ que a partir de certo termo são constantes com valor igual a $p-1$, isto é, quando existe $M \in \mathbb{N}$ tal que $\alpha_n = p-1$ para todo $n \geq M$.

Na próxima proposição, veremos que as sequências de dígitos que provocam a ambiguidade na relação $\bar{\phi}_p$, são as sequências satisfazendo $\alpha_n \to p-1$, quando $n \to \infty$.

Proposição 4.4.2. *Seja (Ω, τ) um espaço topológico de Hausdorff, e $A \subset \Omega$ um conjunto finito. Se $\{\alpha_n\}_{n\in\mathbb{N}}$ é uma sequência de elementos de A e $r \in A$, então,*

$$\alpha_n \to r \Leftrightarrow \exists M \in \mathbb{N} : \alpha_n = r, \forall n \geq M.$$

Demonstração. Sejam $\{\alpha_n\}_{n\in\mathbb{N}} \subset A$ e $r \in A$ tais que $\alpha_n \to r \in A$, como Ω é Hausdorff, para cada $x \in A \setminus \{r\}$, existem abertos $O_{rx}, O_x \in \tau$, com $r \in O_{rx}, x \in O_x$ tais que $O_{rx} \cap O_x = \emptyset$. Defina $O_r = \cap_{x\in A\setminus\{r\}} O_{rx} \in \tau$, então $r \in O_r$ e $O_r \cap (A \setminus \{r\}) \subset O_r \cap \cup_{x\in A\setminus\{r\}} O_x = \emptyset$, de onde, $\{r\} = A \cap O_r$. Logo, como $r \in O_r \in \tau$, existe $M \in \mathbb{N}$ tal que $\alpha_n \in O_r$, $\forall n \geq M$, más como $\{\alpha_n\}_{n\in\mathbb{N}} \subset A$ necessariamente devemos ter $\alpha_n = r$, $\forall n \geq M$. A reciproca é imediata. ■

Agora estamos em condições de enunciar e provar um resultado que permite escrever de maneira única a expansão de $x > 0$ na base $b > 1$ e que permite evitar a ambiguidade em $\bar{\phi}_p$.

Proposição 4.4.3. *Fixe $b \in \mathbb{N}$, $b > 1$. Para cada $x \in \mathbb{R}$, $x > 0$, existe um único $N \in \mathbb{Z}$ e uma única sequência $\{\alpha_n\}_{n=-N}^{\infty} \subset D_b$, satisfazendo $\alpha_{-N} \neq 0$ e $\alpha_n \not\to b-1$, tais que $x = \sum_{n=-N}^{\infty} \alpha_n b^{-n}$.*

Demonstração. **Existência:** Para $b \in \mathbb{N}$, $b > 1$, a família $\{[b^n, b^{n+1})\}_{n\in\mathbb{Z}}$, é uma partição enumerável do conjunto $(0, +\infty)$, logo, para um número real, $x > 0$, existe um único $N \in \mathbb{Z}$ tal que $x \in [b^N, b^{N+1})$, ou seja, $b^N \leq x < b^{N+1}$. Assim, ao calcular a expansão de x na base b, obtemos uma sequência de dígitos $\{\alpha_n\}_{n=-N}^{\infty} \subset D_b$, tal que $x = \sum_{n=-N}^{\infty} \alpha_n b^{-n}$ e $\alpha_{-N} \neq 0$. Agora, se x é um número irracional ou um número racional com *período diferente* de $b-1$, pela Proposição 4.4.2, é claro que a sequência $\{\alpha_n\}_{n=-N}^{\infty} \subset D_b$ que define x, necessariamente satisfaz $\alpha_n \not\to b-1$. Por outro lado, se x tem período $b-1$, isto é, $\alpha_n \to b-1$, existe $M \in \mathbb{Z}$, $M \geq -N$ tal que $\alpha_n = b-1$, para todo $n \geq M$.

Logo, pelo Lema 4.4.1, temos

$$\begin{aligned} x &= \sum_{n=-N}^{M-1} \alpha_n b^{-n} + \sum_{n=M}^{\infty} \alpha_n b^{-n} \\ &= \sum_{n=-N}^{M-1} \alpha_n b^{-n} + \sum_{n=M}^{\infty} (b-1) b^{-n} \\ &= \sum_{n=-N}^{M-1} \alpha_n b^{-n} + b^{-(M-1)} \\ &= \sum_{n=-N}^{M-2} \alpha_n b^{-n} + (\alpha_{M-1} + 1) b^{-(M-1)} \end{aligned}$$

com $1 \leq \alpha_{M-1} + 1 \leq b$. Depois de reduzir termos semelhantes, a sequência de dígitos que define a última expansão de x claramente converge para zero, e assim para todo $x \in \mathbb{R}_+$, existe uma sequência em D_b que não converge para $b-1$ e que define uma expansão de x. **Unicidade:**, suponha que $\{\alpha_n\}_{n=-N}^{\infty}, \{\gamma_n\}_{n=-N}^{\infty} \subset D_b$ são sequências diferentes tais que $\alpha_n, \gamma_n \not\to b-1$ e seja $k = \min\{n \in \mathbb{Z} : \alpha_n \neq \gamma_n\}$, logo, essas sequências definem os números $x, y \in \mathbb{R}_+$

$$x = \sum_{n=-N}^{k-1} \alpha_n b^{-n} + \alpha_k b^{-k} + \sum_{n=k+1}^{\infty} \alpha_n b^{-n}$$

e

$$y = \sum_{n=-N}^{k-1} \alpha_n b^{-n} + \gamma_k b^{-k} + \sum_{n=k+1}^{\infty} \gamma_n b^{-n}.$$

Suponha $\alpha_k > \gamma_k$, pois $\alpha_k \neq \gamma_k$, logo

$$\begin{aligned} y &= \sum_{n=-N}^{k-1} \alpha_n b^{-n} + \gamma_k b^{-k} + \sum_{n=k+1}^{\infty} \gamma_n b^{-n} \\ &< \sum_{n=-N}^{k-1} \alpha_n b^{-n} + \gamma_k b^{-k} + \sum_{n=k+1}^{\infty} (b-1) b^{-n} \\ &= \sum_{n=-N}^{k-1} \alpha_n b^{-n} + \gamma_k b^{-k} + b^{-k} \\ &= \sum_{n=-N}^{k-1} \alpha_n b^{-n} + (\gamma_k + 1) b^{-k} \\ &\leq \sum_{n=-N}^{k-1} \alpha_n b^{-n} + \alpha_k b^{-k} \end{aligned}$$

e portanto

$$y < \sum_{n=-N}^{k-1} \alpha_n b^{-n} + \alpha_k b^{-k} \leq x,$$

ou seja, $x \neq y$. O caso $\alpha_k < \gamma_k$ é análogo, portanto a representação nessas condições, é única. ■

4.4.1 Injeção Borel mensurável $\Phi_p : \mathbb{R}_+^d \to \mathbb{Q}_p^d$

Nessa seção, usando as ideias da Observação 4.4.1 e a Proposição 4.4.3, vamos construir uma função injetora e boreliana, $\phi_p : \mathbb{R}_+ \to \mathbb{Q}_p$, para logo estender ela de forma natural para o caso d-dimensional, com $d \geq 1$, obtendo a função $\Phi_p : \mathbb{R}_+^d \to \mathbb{Q}_p^d$ que também será uma função injetora e Borel mensurável.

Definição 4.4.1. Seja $p \in \mathbb{N}$ um número primo. Definimos a função $\varphi_p : \mathbb{Q}_p \to \mathbb{R}_+$ mediante

$$\varphi_p\left(\sum_{n=-N}^{\infty} \alpha_n p^n\right) = \sum_{n=-N}^{\infty} \alpha_n p^{-n},$$

e o subconjunto de $\mathbb{Q}_p$

$$\mathscr{D}_p := \left\{ \sum_{n=-N}^{\infty} \alpha_n p^n \in \mathbb{Q}_p : \{\alpha_n\}_{n=-N}^{\infty} \subset D_p,\ N \in \mathbb{Z},\ \alpha_n \not\to p-1 \right\}.$$

Pela Observação 4.4.1, vemos que a função $\varphi_p = \bar{\phi}_p^{-1}$ é sobrejetora mas *não é injetora*, porém, pela Proposição 4.4.3, vemos que a restrição de φ_p ao conjunto $\mathscr{D}_p$, a função $\varphi_p|_{\mathscr{D}_p} : \mathscr{D}_p \to \mathbb{R}_+$, é uma bijeção. Assim, a função inversa $(\varphi_p|_{\mathscr{D}_p})^{-1} : \mathbb{R}_+ \to \mathbb{Q}_p$ é uma função injetora.

Definição 4.4.2. Seja $p \in \mathbb{N}$ um número primo, definimos a função $\phi_p : \mathbb{R}_+ \to \mathbb{Q}_p$ por

$$\phi_p = (\varphi_p|_{\mathscr{D}_p})^{-1}.$$

Na seguinte proposição, vamos ver algumas propriedades simples da função ϕ_p.

Proposição 4.4.4. *Para ϕ_p e para o conjunto $\mathscr{D}_p$ da Definição 4.4.1, temos.*

(i) $\mathrm{Im}(\phi_p) = \mathscr{D}_p$ *é denso em* $\mathbb{Q}_p$.

(ii) *Para todo* $x \in \mathbb{R}_+$, $|\phi_p(x)|_p \leq x$

(iii) *Para todo* $x, y \in \mathbb{R}_+$, $|\phi_p(x) - \phi_p(y)|_p \leq \max\{x, y\}$.

Demonstração. **(i)** Como todo número p-ádico é o limite de sequências de somas parciais da forma $S_M = \sum_{n=-N}^{M} \alpha_n p^n$, com $\alpha_n \in D_p$, $M, N \in \mathbb{Z}$ e $M \geq -N$, pela unicidade da representação nos números p-ádicos e pela Proposição 4.4.2, temos que a sequência de dígitos que define cada S_M tende para zero, isto implica

$$\mathbb{Q}_p = \overline{\left\{ \sum_{n=-N}^{\infty} \alpha_n p^n \in \mathbb{Q}_p : \{\alpha_n\}_{n=-N}^{\infty} \subset D_p, \alpha_n \to 0 \right\}} \subseteq \overline{\mathscr{D}_p} \subseteq \mathbb{Q}_p.$$

(ii) Seja $x > 0$. Pela Proposição 4.4.3, existe um único $N \in \mathbb{Z}$ e uma única sequência $\{\alpha_n\}_{n=-N}^{\infty} \subset D_p$, tais que $x = \sum_{n=-N}^{\infty} \alpha_n p^{-n}$, $\alpha_{-N} \neq 0$ e $\alpha_n \not\to p-1$. Logo $x \geq p^N$ e portanto

$$|\phi_p(x)|_p = p^N \leq x.$$

(iii) É imediata a partir da desigualdade triangular forte e do item **(ii)**.

■

No próximo resultado vamos provar que a função ϕ_p é Borel mensurável, e portanto, graças ao Teorema 2.3.2, teremos que ϕ_p é um isomorfismo boreliano de $\mathbb{R}_+$ sobre $\mathscr{D}_p$.

Proposição 4.4.5. *A função $\phi_p : \mathbb{R}_+ \to \mathbb{Q}_p$, é injetora e Borel mensurável.*

Demonstração. A função ϕ_p é injetora por construção. Para provar que ϕ_p é Borel mensurável, observamos que as bolas formam uma base enumerável para a topologia de $(\mathbb{Q}_p, |\cdot|_p)$, logo, pela Proposição A.1.1, precisamos verificar que a imagem recíproca de uma bola aberta em $\mathbb{Q}_p$ mediante ϕ_p seja um subconjunto boreliano de $\mathbb{R}_+$. Seja $a \in \mathbb{Q}_p$, $a = \sum_{n=-N}^{\infty} \alpha_n p^n$, com $N \in \mathbb{Z}$, $\alpha_{-N} \neq 0$ e $h \in \mathbb{Z}$. Se $h > N$, graças ao item **(ii)** da Proposição 3.1.5, é fácil verificar que $B(a, p^h) = B(0, p^h)$, pois $a \in B(0, p^h)$ já que $|a|_p = p^N < p^h$; e também temos

$$\begin{aligned}
\phi_p^{-1}(B(0, p^h)) &= \{x \in \mathbb{R}_+ : |\phi_p(x)|_p < p^h\} \\
&= \{x \sum_{n=-M}^{\infty} \alpha_n p^{-n} \in \mathbb{R}_+ : M \in \mathbb{Z}, M \leq h-1, \alpha_n \not\to p-1\} \\
&= \left[0, p^h\right),
\end{aligned}$$

que é um subconjunto boreliano de $\mathbb{R}_+$. Suponha então $h \leq N$ e defina o *truncamento* de a até o índice $-h \geq -N$ por $a^{(-h)} := \sum_{n=-N}^{-h} \alpha_n p^n \in \mathscr{D}_p$, e a imagem de $a^{(-h)}$ via ϕ_p^{-1} por $\bar{a}^{(-h)} := \phi_p^{-1}(a^{(-h)}) = \sum_{n=-N}^{-h} \alpha_n p^{-n} \in \mathbb{R}_+$. Primeiro observamos que se $b \in B(a, p^h)$, com $h \leq N$, temos

$$|b-a|_p < p^h \leq p^N = |a|_p \Longrightarrow |b-a|_p \neq |a|_p,$$

logo, da Proposição 3.1.4 inferimos que $|b|_p = \max\{|b-a|_p, |a|_p\} = |a|_p$, e assim, obtemos que $b \in B(a, p^h)$ *necessariamente* deve ser da forma $b = \sum_{n=-N}^{\infty} \beta_n p^n$ com $\beta_{-N} \neq 0$. Então, para $h \leq N$, pelo fato observado acima e usando o item **(iii)** da Proposição 4.1.1, temos

$$\begin{aligned}
\phi_p^{-1}\left(B(a,p^h)\right) &= \left\{x \in \mathbb{R}_+ : |\phi_p(x) - a|_p < p^h\right\} \\
&= \left\{x \in \mathbb{R}_+ : \phi_p(x) = \sum_{n=-N}^{\infty} \gamma_n p^n, \gamma_{-N} \neq 0, |\phi_p(x) - a|_p \leq p^{h-1}\right\} \\
&= \left\{x = \sum_{n=-N}^{\infty} \gamma_n p^{-n} \in \mathbb{R}_+ : \gamma_n = \alpha_n, n \leq -h, \gamma_n \not\to p-1\right\},
\end{aligned}$$

de onde vemos que se $x \in \phi_p^{-1}\left(B(a,p^h)\right)$, então $0 \leq x - \bar{a}^{(-h)} = \sum_{n=-h+1}^{\infty} \gamma_n p^{-n}$, com $\gamma_n \not\to p-1$ e assim $0 \leq x - \bar{a}^{(-h)} < p^h$, ou seja $x \in \left[\bar{a}^{(-h)}, \bar{a}^{(-h)} + p^h\right)$ e portanto

$$\phi_p^{-1}\left(B(a,p^h)\right) \subset \left[\bar{a}^{(-h)}, \bar{a}^{(-h)} + p^h\right).$$

Por outro lado, se $x \in \left[\bar{a}^{(-h)}, \bar{a}^{(-h)} + p^h\right)$, temos $0 \leq x - \bar{a}^{(-h)} < p^h$, então pela Proposição 4.4.3, existe $\{\gamma_n\}_{n=-h+1}^{\infty} \subset D_p$ tal que $0 \leq x - \bar{a}^{(-h)} = \sum_{n=-h+1}^{\infty} \gamma_n p^{-n}$ com $\gamma_n \not\to p-1$ (e γ_{-h+1} não necessariamente diferente de zero), logo

$$x = \sum_{n=-N}^{\infty} \beta_n p^{-n}, \beta_n = \alpha_n, n \leq -h, \beta_n = \gamma_n, n \geq -h+1, \beta_n \not\to p-1.$$

Assim,

$$\begin{aligned}
|\phi_p(x) - a|_p &= \left| \sum_{n=-h+1}^{\infty} \gamma_n p^n - \sum_{n=-h+1}^{\infty} \alpha_n p^n \right|_p \\
&\leq \max\left\{ \left| \sum_{n=-(h-1)}^{\infty} \alpha_n p^n \right|_p, \left| \sum_{n=-(h-1)}^{\infty} \gamma_n p^n \right|_p \right\} < p^h,
\end{aligned}$$

de onde vemos que $\phi_p(x) \in B(a,p^h)$ e portanto $\left[\bar{a}^{(-h)}, \bar{a}^{(-h)} + p^h\right) \subset \phi_p^{-1}\left(B(a,p^h)\right)$, obtendo assim

$$\phi_p^{-1}\left(B(a,p^h)\right) = \left[\bar{a}^{(-h)}, \bar{a}^{(-h)} + p^h\right), \tag{13}$$

que é um subconjunto boreliano de $\mathbb{R}_+$. ■

Dos cursos básicos de Análise, sabemos que toda função contínua entre espaços mensuráveis borelianos é também uma função Borel mensurável, porém, também sabemos que nem toda função boreliana é contínua. Com relação à continuidade da função ϕ_p, temos a seguinte proposição.

Proposição 4.4.6. *A função ϕ_p é contínua no conjunto*

$$\mathscr{C}_p = \{0\} \cup \left\{ \sum_{n=-N}^{\infty} \alpha_n p^{-n} \in \mathbb{R}_+ : \{\alpha_n\}_{n=-N}^{\infty} \subset D_p, N \in \mathbb{Z}, \alpha_n \not\to 0, \alpha_n \not\to p-1 \right\}$$

e não é contínua nos pontos do conjunto

$$\mathscr{N}_p = \left\{ \sum_{n=-N}^{\infty} \alpha_n p^{-n} \in \mathbb{R}_+ : \{\alpha_n\}_{n=-N}^{\infty} \subset D_p, N \in \mathbb{Z}, \alpha_{-N} \neq 0, \alpha_n \to 0 \right\}.$$

Demonstração. Primeiro vamos provar a continuidade de ϕ_p em $a = 0 \in \mathbb{R}_+$. Como $\phi_p(0) = 0$, da prova da Proposição 4.4.5, temos que para cada $h \in \mathbb{Z}$

$$\phi_p^{-1}(B(\phi_p(0), p^h)) = \left[0, p^h\right) = B_{\mathbb{R}_+}(0, p^h),$$

logo, para cada $0 < \epsilon = p^h$, existe $0 < \delta = p^h$, tal que $\phi_p(B_{\mathbb{R}_+}(0,\delta)) \subseteq B(\phi_p(0), p^h)$, provando a continuidade em $a = 0$, pois $\mathrm{Im}(|\cdot|_p) = \{0\} \cup \{p^h : h \in \mathbb{Z}\}$.

Agora, considere $a = \sum_{n=-N}^{\infty} \alpha_n p^{-n} \in \mathscr{C}_p$, $\alpha_{-N} \neq 0$ e $h \in \mathbb{Z}$. Primeiro vamos supor que $h > N$. Nesse caso, temos $B(\phi_p(a), p^h) = B(0, p^h)$, pois $|\phi_p(a)|_p = p^N < p^h$, logo

$$\phi_p^{-1}(B(\phi_p(a), p^h)) = \left[0, p^h\right),$$

onde $0 < p^N \leq a < p^{N+1} \leq p^h$, assim, considerando $0 < \delta < (p^{N+1} - a)$, temos $\phi_p(B_{\mathbb{R}_+}(a,\delta)) \subseteq B(\phi_p(a), p^h)$.
Agora vamos supor que $h \leq N$, nesse caso

$$\phi_p^{-1}(B(\phi_p(a), p^h)) = \left[a^{(-h)}, a^{(-h)} + p^h\right),$$

onde $a^{(-h)} = \sum_{n=-N}^{-h} \alpha_n p^{-n} = \phi_p^{-1}(\phi_p(a)^{(-h)})$. Como $\alpha_n \not\to 0, p-1$, temos que para todo inteiro $h \leq N$, $0 < a^{(-h)} < a < a^{(-h)} + p^h$, logo para passar no teste de continuidade basta escolher $0 < \delta < \min\{a^{(-h)} + p^h - a, a - a^{(-h)}\}$, para obter $\phi_p(B_{\mathbb{R}_+}(a,\delta)) \subseteq B(\phi_p(a), p^h)$.

Finalmente provamos que ϕ_p não é contínua nos números reais positivos cuja expansão na base p satisfaz $\alpha_n \to 0$. Seja $a \in \mathscr{N}_p$, das Proposições 4.4.2 e 4.4.3, temos que existem $M, N \in \mathbb{Z}$, com $M \leq N$, tais que $a = \sum_{n=-N}^{-M} \alpha_n p^{-n}$. Considere $\epsilon = p^M$, então $a^{(-M)} = a$, e

$$\phi_p^{-1}(B(\phi_p(a), p^M)) = [a, a + p^M),$$

logo $\forall \delta > 0$, qualquer $x \geq 0$ tal que $a - \delta < x < a$, satisfaz $\phi_p(x) \notin B(\phi_p(a), p^M)$, de onde $\phi_p(B_{\mathbb{R}_+}(a,\delta)) \not\subseteq B(\phi_p(a), p^M)$; completando a prova. ■

Observação 4.4.2. A função ϕ_p definida acima não é a única maneira *natural* de associar um real positivo com um número p-ádico. A função ϕ_p, envia

$$x = \alpha_{-N}\alpha_{-N+1}\dots\alpha_{-1}\alpha_0 \,.\, \alpha_1\alpha_2\alpha_3\dots$$

na base p para o número p-ádico

$$\phi_p(x) = \dots\alpha_2\alpha_1\alpha_0 \,.\, \alpha_{-1}\dots\alpha_{-N+1}\alpha_{-N}.$$

Assim, por exemplo $\phi_p(1) = 1$ e $\phi_p(p) = p^{-1}$.
Para $x = \sum_{n=-N}^{\infty}\alpha_n p^{-(n+1)} = p^{-1}\sum_{n=-N}^{\infty}\alpha_n p^{-n}$ ou $px = \sum_{n=-N}^{\infty}\alpha_n p^{-n}$, defina

$$\psi_p(x) := \phi_p(px) = \sum_{n=-N}^{\infty}\alpha_n p^n.$$

Como a multiplicação por um elemento de $\mathbb{R}_+$ ($x \mapsto px$) é uma função contínua (e em particular boreliana), então ψ_p também é uma função Borel mensurável por ser a composição de funções borelianas. A função ψ_p, envia

$$x = \alpha_{-N}\alpha_{-N+1}\dots\alpha_{-1} \,.\, \alpha_0\alpha_1\alpha_2\dots$$

na base p para

$$\psi_p(x) = \dots\alpha_2\alpha_1\alpha_0 \,.\, \alpha_{-1}\dots\alpha_{-N+1}\alpha_{-N},$$

isto é, envia a parte inteira de $x \in \mathbb{R}_+$ para a parte *fracionária* de $\psi_p(x)$ e a parte fracionária de x para a parte *inteira* de $\psi_p(x)$. Aqui, temos $\psi_p(1) = \phi_p(p) = p^{-1}$.

Finalmente, estendemos a função ϕ_p para uma função injetora e boreliana no espaço $\mathbb{R}_+^d$.

Proposição 4.4.7. *A função $\Phi_p : \mathbb{R}_+^d \to \mathbb{Q}_p^d$, definida por*

$$\Phi_p(x_1, x_2, \dots, x_d) = (\phi_p(x_1), \phi_p(x_2), \dots, \phi_p(x_d)),$$

é injetora e Borel mensurável.

Demonstração. Como a função ϕ_p é injetora, a injetividade de Φ_p segue imediatamente. Considere as projeções canônicas $\pi_i^p : \mathbb{Q}_p^d \to \mathbb{Q}_p$ e $\pi_i : \mathbb{R}_+^d \to \mathbb{R}_+$, para $i \in [d]$. Pelo Corolário A.1.0.1, como $\mathbb{R}_+$ e $\mathbb{Q}_p$ são espaços métricos separáveis, temos $\mathscr{B}_{\mathbb{R}_+^d} = \bigotimes_{i=1}^d \mathscr{B}_{\mathbb{R}_+}$ e $\mathscr{B}_{\mathbb{Q}_p^d} = \bigotimes_{i=1}^d \mathscr{B}_{\mathbb{Q}_p}$, logo, fazendo $X = \mathbb{R}_+^d$, $\mathscr{A}_X = \mathscr{B}_{\mathbb{R}_+^d}$, $Y_i = \mathbb{Q}_p$, $Y = \mathbb{Q}_p^d$ e $\mathscr{A}_Y = \mathscr{B}_{\mathbb{Q}_p^d}$ na Proposição A.1.3, temos que Φ_p é Borel mensurável, se e somente se, $\Phi_i^p := \pi_i^p \circ \Phi_p$ é boreliana, $\forall i \in [d]$. Mas, para $i \in [d]$

$$\begin{aligned}
\Phi_i^p(x_1, x_2, \dots, x_d) &= (\pi_i^p \circ \Phi_p)(x_1, x_2, \dots, x_d) \\
&= \pi_i^p(\phi_p(x_1), \phi_p(x_2), \dots, \phi_p(x_d)) \\
&= \phi_p(x_i) \\
&= (\phi_p \circ \pi_i)(x_1, x_2, \dots, x_d)
\end{aligned}$$

para todo $(x_1, x_2, \ldots, x_d) \in \mathbb{R}_+^d$, logo $\Phi_i^p = \phi_p \circ \pi_i$ é a composição de uma função Borel mensurável, ϕ_p, e uma função contínua, π_i, portanto ela é Borel mensurável, $\forall i \in [d]$, e assim Φ_p é uma função boreliana. ■

Finalizamos o capítulo com algumas observações.

Observação 4.4.3 (Como usar a regra de aprendizagem em $[0,1]^d$)**.** Se dispomos de uma amostra rotulada $d_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in ([0,1]^d \times \{0,1\})^n$, para usar a regra de aprendizagem dada pela Definição 2.3.1, como a composição da regra p-ádica com a aplicação $\Phi_p|_{[0,1]^d}$ (que por simplicidade denotaremos por Φ_p), primeiro obtemos a amostra rotulada p-ádica $e_n = ((\Phi_p(x_1), y_1), (\Phi_p(x_2), y_2), \ldots, (\Phi_p(x_n), y_n)) \in (\mathbb{Z}_p^d \times \{0,1\})^n$ para logo construir a árvore $\mathcal{A}_{p^d}^H(e_n)$ de altura minimal H usando o Algoritmo 3. Uma vez treinado o classificador, o rótulo predito para o vetor $x \in [0,1]^d$, será o rótulo do vetor $\Phi_p(x)$ predito pelo Algoritmo 4 com entradas $\mathcal{A}_{p^d}^H(e_n)$ e H. Detalhes sobre esse procedimento usando conjuntos de dados disponíveis na *web*, serão expostos no Capítulo 6.

Observação 4.4.4 (Família de regras em $[0,1]^d$ baseadas na regra p-ádica)**.** Uma fórmula simples para gerar *novas regras de aprendizagem* no cubo $[0,1]^d$, é usar a estrutura linear do espaço normado $(\mathbb{Q}_p^d, \|\cdot\|_p)$ combinada com o *isomorfismo boreliano* Φ_p.

Com efeito, é claro que toda matriz quadrada de ordem $d \in \mathbb{N}$ com entradas em $\mathbb{Q}_p$, $M \in \mathrm{M}_d(\mathbb{Q}_p)$, define uma *aplicação linear*, $T_M : \mathbb{Q}_p^d \to \mathbb{Q}_p^d$, mediante o produto matriz-vetor usual: $T_M(x) := Mx \in \mathbb{Q}_p^d$; aplicação que é contínua (e portanto boreliana) pois é uma aplicação linear entre dois $\mathbb{Q}_p$-espaços vetoriais normados (com normas compatíveis com $|\cdot|_p$) cujo domínio é um espaço de dimensão finita (e portanto podemos definir $\|\cdot\|_p$) onde todas as normas são equivalentes. Logo, como T_M conecta dois espaços vetoriais da mesma dimensão, a aplicação T_M será injetora, se e somente se, T_M é sobrejetora, e portanto, qualquer matriz $M \in \mathcal{Z} := \{M \in \mathrm{M}_d(\mathbb{Z}_p) : \det(M) \neq 0\}$[8], define uma aplicação $T_M : \mathbb{Q}_p^d \to \mathbb{Q}_p^d$ *injetora e Borel mensurável*, cuja restrição $T_M := T_M|_{\mathbb{Z}_p^d} : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$ também será uma injeção boreliana, gerando assim, as regras $\mathcal{L}_{(k,p)}^{T_M}$ em $\mathbb{Z}_p^d$ e as regras $\mathcal{L}_{(k,p)}^{T_M \circ \Phi_p}$ no cubo $[0,1]^d$, para $M \in \mathcal{Z}$.

Nessa última observação, vamos construir uma medida boreliana de probabilidade sobre o conjunto $\mathcal{Z}$, de matrizes inversíveis de inteiros p-ádicos, medida que será utilizada na hora de provar a consistência das *florestas p-ádicas aleatórias* da Seção 5.3.

Observação 4.4.5 (Uma medida de probabilidade boreliana sobre $\mathcal{Z} \subseteq \mathrm{M}_d(\mathbb{Z}_p)$)**.** Agora, vamos construir uma medida de probabilidade boreliana, ν, no espaço métrico separável $(\mathcal{Z}, \|\cdot\|_p)$, onde $\mathcal{Z} := \{M \in \mathrm{M}_d(\mathbb{Z}_p) : \det(M) \neq 0\}$ e a norma matricial é definida para cada $M = (M_{ij})_{i,j=1}^d \in \mathrm{M}_d(\mathbb{Q}_p)$, por $\|M\|_p = \max_{i,j \in [d]} |M_{ij}|_p$; obtendo assim, o espaço boreliano de probabilidade *separável* $(\mathcal{Z}, \mathscr{B}_{\mathcal{Z}}, \nu)$.

[8] Para uma matriz $M \in \mathrm{M}_d(\mathbb{K})$, onde $\mathbb{K}$ é um corpo qualquer, o determinante, $\det(M)$ e o traço, $\mathrm{tr}(M)$, são calculados da forma usual.

Como **primeiro passo**, observamos que não é difícil verificar que a aplicação linear $\Psi : \mathrm{M}_d(\mathbb{Q}_p) \to \mathbb{Q}_p^{d^2}$, definida para cada $M = (M_{ij})_{i,j=1}^d \in \mathrm{M}_d(\mathbb{Q}_p)$, por:

$$\Psi((M_{ij})_{i,j=1}^d) := (M_{11}, M_{12}, \ldots, M_{1d}, M_{21}, \ldots, M_{2d}, \ldots, M_{d1}, \ldots, M_{dd}) \in \mathbb{Q}_p^{d^2},$$

é um *isomorfismo de espaços vetoriais normados*, isto é, Ψ é uma aplicação linear e bijetora que preserva a norma: $\|\Psi(M)\|_p = \|M\|_p, \forall M \in \mathrm{M}_d(\mathbb{Q}_p)$, assim, os espaços $(\mathrm{M}_d(\mathbb{Q}_p), \|\cdot\|_p)$ e $(\mathbb{Q}_p^{d^2}, \|\cdot\|_p)$ são *indistinguíveis* como espaços vetoriais normados. Também observamos que $\mathrm{M}_d(\mathbb{Z}_p) = \Psi^{-1}(\mathbb{Z}_p^{d^2})$ é um subconjunto compacto de $\mathrm{M}_d(\mathbb{Q}_p)$, pois é a imagem contínua de um compacto.

O **segundo passo** será definir uma medida de probabilidade boreliana no espaço $(\mathrm{M}_d(\mathbb{Z}_p), \|\cdot\|_p)$, usando uma medida em $(\mathbb{Z}_p^{d^2}, \|\cdot\|_p)$ combinada com o isomorfismo Ψ.

Com efeito, é um exercício rotineiro verificar que para $m \in \mathbb{N}$ qualquer, o grupo aditivo (abeliano) formado por vetores de inteiros p-ádicos, $(\mathbb{Z}_p^m, +)$, é um *grupo topológico compacto* com a topologia induzida pela norma, $\tau_{\|\cdot\|_p} \subseteq \mathcal{P}(\mathbb{Z}_p^m)$, isto é, $(\mathbb{Z}_p^m, +, \tau_{\|\cdot\|_p})$ é um *grupo* munido de uma topologia tal que as operações de grupo: $+ : \mathbb{Z}_p^m \times \mathbb{Z}_p^m \to \mathbb{Z}_p^m$ definida por $(x, y) \mapsto x + y$ e $- : \mathbb{Z}_p^m \to \mathbb{Z}_p^m$, definida por $x \mapsto -x$; são funções contínuas, logo, de um resultado clássico da análise (ver, [30, Alfred Haar], [28, Seção 11.1], [31, Capítulo XI], [52, Seção I.2]), temos que existe a[9] medida de *Haar* em $(\mathbb{Z}_p^m, +, \tau_{\|\cdot\|_p})$, que denotaremos por μ_{H_m}. A medida boreliana de *probabilidade*[10] $\mu_{H_m} : \mathscr{B}_{\mathbb{Z}_p^m} \to [0, 1]$, possui as seguintes propriedades:

Invariância ante translações:

$$\mu_{H_m}(a + B) = \mu_{H_m}(B + a) = \mu_{H_m}(B), \forall a \in \mathbb{Z}_p^m, \forall B \in \mathscr{B}_{\mathbb{Z}_p^m},$$

onde $a + B = \{a + b \in \mathbb{Z}_p^m : b \in B\}$.

Positividade nos conjuntos abertos:

$$\mu_{H_m}(O) > 0, \forall O \in \tau_{\|\cdot\|_p}, O \neq \emptyset.$$

Assim, $(\mathbb{Z}_p^m, \mathscr{B}_{\mathbb{Z}_p^m}, \mu_{H_m})$ é um espaço boreliano de probabilidade completo e separável.

[9] Medida que é única, a menos de multiplicações por um número positivo.

[10] Uma medida de Haar sobre um grupo topológico compacto é finita e positiva, assim ao normalizar, obtemos uma medida boreliana de probabilidade que continua sendo uma medida de Haar.

Para μ_{H_m}, pela Proposição 4.3.2, temos:

$$\begin{aligned}
1 &= \mu_{H_m}(\mathbb{Z}_p^m) \\
&= \sum_{\alpha_{(h)}^{(m)} \in (D_p^h)^m} \mu_{H_m}\left(\bar{B}_m([|\alpha_{(h)}^{(m)}|], p^{-h})\right) \\
&= \sum_{\alpha_{(h)}^{(m)} \in (D_p^h)^m} \mu_{H_m}\left([|\alpha_{(h)}^{(m)}|] + \bar{B}_m(0, p^{-h})\right) \\
&= \sum_{\alpha_{(h)}^{(m)} \in (D_p^h)^m} \mu_{H_m}\left(\bar{B}_m(0, p^{-h})\right) \\
&= p^{hm} \cdot \mu_{H_m}\left(\bar{B}_m(0, p^{-h})\right),
\end{aligned}$$

de onde:

$$\mu_{H_m}\left(\bar{B}_m(a, p^{-h})\right) = \mu_{H_m}\left(a + \bar{B}_m(0, p^{-h})\right) = \mu_{H_m}\left(\bar{B}_m(0, p^{-h})\right) = p^{-hm},$$

para $m, h \in \mathbb{N}$ e $a \in \mathbb{Z}_p^m$, arbitrários. Quando $m = 1$, chamamos $\mu_H := \mu_{H_1}$ à medida de Haar normalizada sobre $\mathbb{Z}_p$, assim, da Proposição 4.3.1 e como as σ-álgebras envolvidas são borelianas, inferimos que $\mu_{H_m} = \mu_H^m = \otimes_{i=1}^m \mu_H$. Em particular, para $d \in \mathbb{N}$, temos $\mu_{H_{d^2}} = \mu_H^{d^2} = \otimes_{i=1}^{d^2} \mu_H$ e portanto a aplicação Ψ^{-1} restrita a $\mathbb{Z}_p^{d^2}$, $\Psi^{-1} : \mathbb{Z}_p^{d^2} \to \mathrm{M}_d(\mathbb{Z}_p)$, define uma medida boreliana de probabilidade em $\mathrm{M}_d(\mathbb{Z}_p)$, que chamaremos μ_{M_d}, mediante: $\mu_{\mathrm{M}_d} = \mu_{H_{d^2}} \circ \Psi$, obtendo assim o espaço boreliano de probabilidade completo e separável, $(\mathrm{M}_d(\mathbb{Z}_p), \mathscr{B}_{\mathrm{M}_d(\mathbb{Z}_p)}, \mu_{\mathrm{M}_d})$.

Finalmente, o **terceiro passo** é definir uma medida em $\mathcal{Z}$ a partir de μ_{M_d}. Primeiro observamos que $\mathcal{Z} = \mathrm{M}_d(\mathbb{Z}_p) \cap \mathrm{GL}_d(\mathbb{Q}_p)$, onde $\mathrm{GL}_d(\mathbb{Q}_p) = \{M \in \mathrm{M}_d(\mathbb{Q}_p) : \det(M) \neq 0\}$ é conhecido como o *grupo linear geral* de grau d sobre o corpo $\mathbb{Q}_p$. De maneira análoga a como é feito com $\mathrm{GL}_d(\mathbb{R})$, é possível mostrar que o conjunto $\mathrm{GL}_d(\mathbb{Q}_p)$ é aberto em $\mathrm{M}_d(\mathbb{Q}_p)$, logo, $\mathcal{Z}$ é *aberto* em $\mathrm{M}_d(\mathbb{Z}_p)$ e $\Psi(\mathcal{Z}) \neq \emptyset$, é aberto em $\mathbb{Z}_p^{d^2}$, assim, pela positividade nos abertos da medida $\mu_H^{d^2}$, temos:

$$0 < \mu_H^{d^2}(\Psi(\mathcal{Z})) = \mu_{\mathrm{M}_d}(\mathcal{Z}) \leq 1.$$

Normalizando, obtemos a medida boreliana de probabilidade, $\nu := \frac{1}{\mu_{\mathrm{M}_d}(\mathcal{Z})}\mu_{\mathrm{M}_d}$, e portanto, $(\mathcal{Z}, \mathscr{B}_{\mathcal{Z}}, \nu)$ é um espaço boreliano de probabilidade separável.

Concluímos a observação dedicando algumas linhas ao processo de amostragem de uma *matriz aleatória*, $Z \in \mathrm{M}_d(\mathbb{Z}_p)$, tal que $Z \sim \mu_{\mathrm{M}_d}$. Pelo fato de Ψ ser um isomorfismo de espaços vetoriais normados, temos que $Z \sim \mu_{\mathrm{M}_d}$, se e somente se, $X = \Psi(Z) \sim \mu_H^{d^2}$, por isso vamos nos focar na amostragem de pontos $X \in \mathbb{Z}_p^{d^2}$, tais que $X \sim \mu_H^{d^2}$. Considere a partição à altura $h \in \mathbb{N}$ de $\mathbb{Z}_p^{d^2}$:

$$\mathbb{Z}_p^{d^2} = \bigcup_{\alpha_{(h)}^{(d^2)} \in (D_p^h)^{d^2}} \bar{B}_{d^2}\left([|\alpha_{(h)}^{(d^2)}|], p^{-h}\right).$$

Logo, um vetor $X = (X_1, X_2, \ldots, X_{d^2}) \in \mathbb{Z}_p^{d^2}$, de variáveis aleatórias $X_i \overset{i.i.d}{\sim} \mu_H$, vai pertencer a qualquer uma das p^{hd^2} bolas da partição, com probabilidade p^{-hd^2}. Por outro lado, se $X \in \bar{B}_{d^2}\left([|\alpha_{(h)}^{(d^2)}|], p^{-h}\right)$, então os primeiros h dígitos de X: $\mathcal{D}_p^j(X)$, $0 \leq j \leq h-1$, coincidem com os primeiros h dígitos do vetor $[|\alpha_{(h)}^{(d^2)}|] \in \mathbb{Z}_p^{d^2}$. Assim, como desde o ponto de vista computacional podemos trabalhar apenas com números p-ádicos cuja *expansão canônica* seja *finita*, e se não nos importam os dígitos $\mathcal{D}_p^j(X)$, para $j \geq h$; então, podemos fazer a amostragem de X considerando as variáveis aleatórias $X_i \overset{i.i.d}{\sim} \mu_H$, sendo da forma $X_i = \sum_{j=0}^{h-1} A_j^i p^j$ para $i \in [d^2]$. Nessas condições, fazer a amostragem de uma variável $X_i \sim \mu_H$ na altura h, é equivalente com escolher de forma *aleatória* e com distribuição uniforme, qualquer um dos p^h números $x_i \in \mathbb{N}_0$ tal que $0 \leq x_i \leq p^h - 1$ e assim, fazer a amostragem de um vetor $X \sim \mu_H^{d^2}$ na altura h, é equivalente a escolher de forma *aleatória, uniforme e independente*, números inteiros $0 \leq x_i \leq p^h - 1$, para $i \in [d^2]$. Finalmente, fazer a amostragem de uma matriz $Z = \Psi^{-1}(X) \sim \mu_{\mathrm{M}_d}$ na altura $h \in \mathbb{N}$, é equivalente a escolher de forma aleatória, com distribuição uniforme e de modo independente, números $z_{ij} \in \mathbb{N}_0$ tais que $0 \leq z_{ij} \leq p^h - 1$, para $i, j \in [d]$, obtendo assim, uma *instância* $z = (z_{ij})_{i,j=1}^d$ da variável aleatória $Z \sim \mu_{\mathrm{M}_d}$. Se somado ao anterior, temos que $\det(z) \neq 0$, então z será uma instância da variável aleatória $Z \sim \nu$.

5 CONSISTÊNCIA DA REGRA DE APRENDIZAGEM p-ÁDICA E DA APRENDIZAGEM ENSEMBLE

Esse capítulo, é dedicado à consistência universal da regra de aprendizagem p-ádica e à de certa classe de regras de aprendizagem do tipo ensemble.

O primeiro passo, é obter uma expressão matemática da regra de aprendizagem p-ádica que faça possível a análise de consistência. Depois, enunciaremos e provaremos o resultado principal do texto, resultado que *implica* que a regra de aprendizagem por trás dos algoritmos 3 e 4, é uma regra *universalmente consistente* em certa classe de espaços métricos, classe que tem o espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$ como um dos seus membros: nos espaços métricos de dimensão σ-finita no sentido de *Nagata* [4, 48, 49].

Depois de provar um resultado de consistência da *família* de regras de aprendizagem dada pelo voto majoritário entre um número finito de membros de uma família consistente de regras; definimos uma família aleatória de regras de aprendizagem, todas universalmente consistentes no espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$, com as quais podemos obter famílias de regras de aprendizagem do tipo ensemble que terão uma componente aleatória e serão universalmente consistentes em $(\mathbb{Z}_p^d, \|\cdot\|_p)$ (Teorema 2.3.3), "ao estilo" *Random Forest* (ver [14]), ou seja, teremos uma espécie de *Floresta p-ádica Aleatória*, ou *p-adic Random Forest*, mas que será *universalmente consistente*.

5.1 EXPRESSÃO MATEMÁTICA DA REGRA DE APRENDIZAGEM p-ÁDICA

No Capítulo 4, definimos a regra de aprendizagem p-ádica mediante os algoritmos 3 e 4. Para poder analisar a consistência dessa regra, precisamos de uma expressão matemática que a represente e seja possível de manipular usando a linguagem da teoria da medida.

Primeiro lembremos como faz a classificação a regra $\mathcal{L}_{(k,p)} = (g_n^{(k,p)})_{n=1}^{\infty}$. Suponha que temos uma amostra rotulada $d_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathbb{Z}_p^d \times \{0,1\})^n$, com $\varsigma_n = (x_1, x_2, \ldots, x_n) \in \Omega^n$ e a árvore $\mathcal{A}_{p^d}^H(d_n)$ de altura mínimal construída usando o Algoritmo 3. Para classificar o vetor $x \in \mathbb{Z}_p^d$, primeiro traçamos o caminho dele na árvore $\mathcal{A}_{p^d}^H(d_n)$ usando o Algoritmo 4, *até chegar* ao primeiro nível de altura $0 \leq h_f(x) \leq H$, tal que o número de elementos da amostra na bola $\bar{B}_d(x, p^{-h_f(x)})$, seja menor do que k, ou seja, até que a variável n_v satisfaça,

$$n_v(x_{(h_f(x))}^{(d)}) = \sharp\left(\{i \in [n] : x_i \in \bar{B}_d(x, p^{-h_f(x)})\}\right) < k, \text{ com}$$
$$n_v(x_{(h_f(x)-1)}^{(d)}) = \sharp\left(\{i \in [n] : x_i \in \bar{B}_d(x, p^{-(h_f(x)-1)})\}\right) \geq k,$$

para depois calcular o rótulo de x fazendo a votação com os rótulos dos elementos da amostra que pertencem à bola $\bar{B}_d(x, p^{-h_k(x)})$, onde $h_k(x) = h_f(x) - 1$, isto é, o rótulo predito pela regra p-ádica para x será $y = 1$, se a quantidade de elementos x_i da amostra

que pertencem à bola $\bar{B}_d(x, p^{-h_k(x)})$ e que possuem rótulo $y_i = 1$, é maior ou igual[1] que a metade da quantidade total de elementos da amostra que pertencem à bola, e $y = 0$ no caso contrário. Então, podemos escrever

$$g_n^{(k,p)}(d_n)(x) = \begin{cases} 1, & \frac{s_v(x^{(d)}_{(h_k(x))})}{n_v(x^{(d)}_{(h_k(x))})} \geq \frac{1}{2} \\ 0, & \text{caso contrário} \end{cases}$$

ou

$$g_n^{(k,p)}(d_n)(x) = \begin{cases} 1, & \frac{1}{n_v(x^{(d)}_{(h_k(x))})} \sum_{x_i \in \bar{B}_d(x, p^{-h_k(x)})} y_i \geq \frac{1}{2} \\ 0, & \text{caso contrário.} \end{cases}$$

Por outro lado, lembrando que não existe $a \in \mathbb{Z}_p^d$, tal que

$$p^{-h_f(x)} < \|a\|_p < p^{-h_f(x)+1} = p^{-h_k(x)}$$

temos necessariamente

$$\begin{aligned} p^{-h_k(x)} &= \min\left\{r > 0 : \sharp\left(\{i \in [n] : x_i \in \bar{B}_d(x, r)\}\right) \geq k\right\} \\ &=: r^{\varsigma_n}_{k\text{-NN}}(x), \end{aligned}$$

ou seja, a regra de aprendizagem p-ádica pode se escrever como a regra do tipo *plug-in* onde o rótulo predito para o vetor $x \in \mathbb{Z}_p^d$ é o voto majoritário computado com os rótulos dos elementos da amostra ς_n, que pertecem à bola $\bar{B}_d(x, r^{\varsigma_n}_{k\text{-NN}}(x))$.
Ainda, como

$$n_v(x^{(d)}_{(h_k(x))}) = \sharp\left(\{i \in [n] : x_i \in \bar{B}_d(x, r^{\varsigma_n}_{k\text{-NN}}(x))\}\right),$$

vemos que a regra $\mathcal{L}_p$ é determinada pelo raio $r^{\varsigma_n}_{k\text{-NN}}(x)$, valor que pode ser definido em qualquer espaço métrico[2], por essa razão, vamos definir a *nova regra de aprendizagem supervisionada* para esse nível de generalidade.

Definição 5.1.1 (Regra de aprendizagem ^+k-NN)**.** Seja Ω um espaço métrico. Para $k \in \mathbb{N}$, definimos em Ω a regra de aprendizagem ^+k-NN, que denotamos por $\mathcal{L}_{+k\text{-NN}}$, como a regra do tipo plug-in definida pela família $\mathcal{F}_{+k\text{-NN}} = (\eta_{nk})_{n \in \mathbb{N}}$, onde para cada $n \in \mathbb{N}$ com $n \geq k$, as funções

$$\eta_{nk} : (\Omega \times \{0,1\})^n \times \Omega \to [0,1]$$

são definidas para cada $d_n = ((x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)) \in (\Omega \times \{0,1\})^n$ e $x \in \Omega$, com $\varsigma_n = (x_1, x_2, \ldots, x_n) \in \Omega^n$, por

$$\eta_{nk}(d_n, x) := \frac{1}{N_k^{\varsigma_n}(x)} \sum_{x_i \in \bar{B}(x, r^{\varsigma_n}_{k\text{-NN}}(x))} y_i,$$

[1] No caso de empate, podemos retornar qualquer dos valores $\{0,1\}$. Por simplicidade, no caso de empate, a regra p-ádica retornará o rótulo 1.

[2] Independentemente das boas ou más propriedades que a regra possa ter em alguns espaços métricos.

onde

$$r_{k\text{-NN}}^{\varsigma_n}(x) := \min\left\{r > 0 : \sharp\left(\{i \in [n] : x_i \in \bar{B}(x,r)\}\right) \geq k\right\}$$

e

$$N_k^{\varsigma_n}(x) := \sharp\left(\{i \in [n] : x_i \in \bar{B}(x, r_{k\text{-NN}}^{\varsigma_n}(x))\}\right) \geq k,$$

isto é, a regra $\mathcal{L}_{+k\text{-NN}} = (g_{nk})_{n=1}^{\infty}$ é dada para cada $n \geq k$, $d_n \in (\Omega \times \{0,1\})^n$ e $x \in \Omega$, por,

$$g_{nk}(d_n)(x) = \begin{cases} 1, & \frac{1}{N_k^{\varsigma_n}(x)} \sum_{x_i \in \bar{B}(x, r_{k\text{-NN}}^{\varsigma_n}(x))} y_i \geq \frac{1}{2} \\ 0, & \text{outro caso.} \end{cases}$$

Antes de continuar, dedicamos algumas linhas à relação do raio $r_{k\text{-NN}}^{\varsigma_n}(x)$ com um importante resultado na área de aprendizagem de máquina supervisionada, conhecido como *Lema de Cover-Hart*.

5.1.1 O raio $r_{k\text{-NN}}^{\varsigma_n}(x)$ e o Lema de Cover-Hart

Num espaço métrico (Ω, ρ), o valor $r_{k\text{-NN}}^{\varsigma_n}(x)$ é por definição o menor raio de uma bola fechada centrada em $x \in \Omega$ que contém pelo menos k elementos da amostra ς_n. Assim, podemos ordenar[3] os elementos da amostra em função da sua distância até o ponto x, mediante:

$$\rho(x, x_{(1)}(x)) \leq \rho(x, x_{(2)}(x)) \leq \cdots \leq \rho(x, x_{(k)}(x)) \leq \cdots \leq \rho(x, x_{(n)}(x)),$$

onde para $k \in [n]$, $x_{(k)}(x)$ é o k-ésimo vizinho mais próximo de x na amostra ς_n:

$$\rho(x, x_{(k)}(x)) = r_{k\text{-NN}}^{\varsigma_n}(x).$$

No ano 1967, Cover e Hart (ver páginas 23 e 26 de [22]), provaram que para $k \in \mathbb{N}$ *fixo*, se μ é uma medida de probabilidade boreliana sobre o espaço métrico *separável* (Ω, ρ), e se modelamos ς_n mediante a amostra aleatória, $\sigma_n = (X_1, X_2, \ldots, X_n) \in \Omega^n$, de variáveis $X_i \overset{i.i.d}{\sim} \mu$, então, para $X \sim \mu$, variável aleatória independente de σ_n, temos

$$\rho(X, X_{(k)}(X)) \xrightarrow[n\to\infty]{q.c} 0.$$

O resultado acima, diz que para $(\mu^\infty \otimes \mu)$-q.t $(\varsigma_\infty, x) \in \Omega^\infty \times \Omega$, o k-ésimo vizinho mais próximo de x no segmento inicial ς_n, do caminho amostral ς_∞, tende para x quando $n \to \infty$, ou seja, para $(\mu^\infty \otimes \mu)$-q.t $(\varsigma_\infty, x) \in \Omega^\infty \times \Omega$, temos

$$r_{k\text{-NN}}^{\varsigma_n}(x) \xrightarrow{n\to\infty} 0.$$

[3] No caso de empate nas distâncias até $x \in \Omega$, o critério para ordenar os vizinhos pode ser qualquer, por exemplo, utilizando o índice: se $\rho(x, x_i) = \rho(x, x_j)$ com $i < j$, na sequência de desdigualdades consideramos $\rho(x, x_i) \leq \rho(x, x_j)$.

O resultado acima ainda vale se consideramos uma sequência de valores de k, $\{k_n\}_{n\in\mathbb{N}} \subset \mathbb{N}$, tal que $k_n \leq n$ para todo $n \in \mathbb{N}$ e

$$\lim_{n\to\infty} \frac{k_n}{n} = 0. \tag{14}$$

Uma prova desse fato no caso euclidiano, pode ser encontrada no Lema 5.1 de [24], prova que continua valendo para espaços métricos separáveis quaisquer.

Antes de continuar, precisamos da seguinte definição.

Definição 5.1.2 (Suporte de uma medida boreliana)**.** Seja μ uma medida boreliana sobre um espaço métrico Ω. Definimos o *suporte* de μ, que denotaremos por $\text{supp}(\mu)$, como o conjunto formado pelos elementos $x \in \Omega$ satisfazendo:

$$\forall \epsilon > 0,\ \mu(B(x,\epsilon)) > 0.$$

O seguinte resultado, cuja demonstração é mais ou menos padrão, é uma parte do Lema de Cover-Hart em [22] e será enunciado sem prova.

Proposição 5.1.1 (Cover e Hart [22])**.** *Seja μ uma medida boreliana de probabilidade sobre um espaço métrico separável. Então*

$$\mu(\text{supp}(\mu)) = 1.$$

A versão do Lema de Cover-Hart que será apresentada aqui, é a versão para espaços métricos separáveis considerando uma sequência de valores de k satisfazendo (14). Uma prova do seguinte resultado pode ser consultada no Lema 3.2.2 de [40].

Teorema 5.1.1 (Lema de Cover-Hart [22, 40])**.** *Sejam μ uma medida de probabilidade boreliana sobre o espaço métrico separável (Ω,ρ), $\sigma_n = (X_1, X_2, \ldots, X_n) \in \Omega^n$, uma amostra aleatória de variáveis $X_i \overset{i.i.d}{\sim} \mu$ e $X \sim \mu$ variável aleatória independente de σ_n. Denotemos por $X_{(k)}(X)$ o k-ésimo vizinho mais próximo de X na amostra aleatória σ_n, obtendo assim $\rho(X, X_{(k)}(X)) = r^{\sigma_n}_{k\text{-}NN}(X)$. Se $\{k_n\}_{n\in\mathbb{N}} \subset \mathbb{N}$ é uma sequência de números $k_n \leq n$ satisfazendo $\lim_{n\to\infty} k_n/n = 0$, então $r^{\sigma_n}_{k_n\text{-}NN}(x) \xrightarrow[n\to\infty]{q.c} 0$, $\forall x \in \text{supp}(\mu)$ e*

$$r^{\sigma_n}_{k_n\text{-}NN}(X) \xrightarrow[n\to\infty]{q.c} 0.$$

Observação 5.1.1 (k-NN v/s ^+k-NN)**.** Nessa observação vamos dar um par de argumentos que tentam justificar a semelhança no nome da nova regra de aprendizagem, ^+k-NN, com o da regra k-NN.

Primeiro observamos que ambas regras estão *emparentadas* pelo raio $r^{\varsigma_n}_{k\text{-NN}}(x)$, pois utilizando o contexto e a notação do Lema de Cover-Hart, vemos que ambas regras determinam o rótulo predito de $x \in \Omega$, mediante o voto majoritário dos rótulos *de* elementos da amostra $\varsigma_n \in \Omega^n$ que pertencem à bola $\bar{B}(x, r^{\varsigma_n}_{k\text{-NN}}(x))$, porém, elas têm uma *sutil diferença* não trivial. Por um lado, a regra k-NN computa o voto majoritário com os rótulos dos k elementos da amostra ς_n mais próximos de x, isto é, com os rótulos dos elementos do conjunto

$$\mathcal{N}_k(\varsigma_n, x) = \left\{x_{(1)}, x_{(2)}, \ldots, x_{(k)}\right\} \subset \varsigma_n \cap \bar{B}(x, r_{k\text{-NN}}^{\varsigma_n}(x)),$$

mediante

$$g_{n(k\text{-NN})}(d_n)(x) = \begin{cases} 1, & \frac{1}{k}\sum_{x_i \in \mathcal{N}_k(\varsigma_n, x)} y_i \geq \frac{1}{2} \\ 0, & \text{caso contrário,} \end{cases}$$

e por outro, a regra ${}^{+}k$-NN, calcula o voto majoritário com os rótulos dos $N_k^{\varsigma_n}(x) \geq k$ elementos da amostra ς_n mais próximos de x, ou seja, com os rótulos dos elementos do conjunto

$$\mathcal{N}_{+k\text{-NN}}(\varsigma_n, x) = \left\{x_{(1)}, x_{(2)}, \ldots, x_{(k)}, \ldots, x_{(N_k^{\varsigma_n}(x))}\right\} = \varsigma_n \cap \bar{B}(x, r_{k\text{-NN}}^{\varsigma_n}(x)),$$

mediante

$$g_{nk}(d_n)(x) = \begin{cases} 1, & \frac{1}{N_k^{\varsigma_n}(x)}\sum_{x_i \in \mathcal{N}_{+k\text{-NN}}(\varsigma_n, x)} y_i \geq \frac{1}{2} \\ 0, & \text{caso contrário,} \end{cases}$$

onde claramente temos $\mathcal{N}_k(\varsigma_n, x) \subset \mathcal{N}_{+k\text{-NN}}(\varsigma_n, x)$.

5.2 CONSISTÊNCIA DA REGRA DE APRENDIZAGEM ${}^{+}k$-NN

Nessa seção vamos apresentar o resultado principal do texto, resultado que via o Corolário 2.2.1.2 e o Teorema da Convergência Dominada, implica que a regra ${}^{+}k$-NN é *consistente* em qualquer espaço métrico separável que satisfaça a condição de Lebesgue-Besicovitch forte **(LBF)** para a distribuição dos dados não rotulados (Definição 2.5.7), definição que vamos enunciar novamente.

Definição 5.2.1 (Condição de Lebesgue-Besicovitch Forte **(LBF)** [6])**.** Dizemos que o espaço métrico Ω satisfaz a condição de **L**ebesgue-**B**esicovitch **F**orte **(LBF)** para a medida boreliana localmente finita μ, se para toda função μ-integrável $f \in L^1(\mu)$, temos

$$\lim_{r \to 0^+} \frac{1}{\mu(\bar{B}(x,r))} \int_{\bar{B}(x,r)} |f(z) - f(x)|\, d\mu(z) = 0, \text{ para } \mu\text{-q.t } x \in \Omega.$$

Antes de ir ao "*prato*" principal, vamos fazer algumas observações sobre a prova do resultado.

Na prova de consistência da regra ${}^{+}k$-NN, por um lado, são combinadas de maneira não trivial as ideias do Teorema 2.2 de [17, Cérou-Guyader] e do Teorema 2.1 de [23, Devroye], ambos resultados referentes ao classificador k-NN, e por outro, também são incorporadas as ideias desenvolvidas na prova do Teorema 6.1 de [24, Devroye-Györfi-Lugosi], resultado que trata da consistência de *regras de partição*, que é a classe de regras de aprendizagem à qual pertencem a regra do histograma e as árvores de decisão no espaço euclidiano.

Teorema 5.2.1 (Consistência da regra de aprendizagem ^{+}k-NN)**.** *Seja (Ω, ρ) um espaço métrico separável, com $\widetilde{\mu}$ medida boreliana de probabilidade sobre $\Omega \times \{0,1\}$ caracterizada pelo par (η, μ) e suponha que Ω satisfaz a condição* ***(LBF)*** *para μ.*

Denotando uma amostra rotulada aleatória de variáveis $(X_i, Y_i) \overset{i.i.d}{\sim} \widetilde{\mu}$ por $D_n = ((X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)) \in (\Omega \times \{0,1\})^n$, com $\sigma_n = (X_1, X_2, \ldots, X_n)$, e se $\{k_n\}_{n\in\mathbb{N}} \subset \mathbb{N}$ é uma sequência de inteiros satisfazendo as condições: $\forall n \in \mathbb{N}$, $k_n \leq n$ e quando $n \to \infty$, $k_n \to \infty$ e $k_n/n \to 0$; então, a família de funções $\mathcal{F}_{+_{k\text{-}NN}} = (\eta_{nk_n})_{n=1}^{\infty}$, definidas para cada $(d_n, x) \in (\Omega \times \{0,1\})^n \times \Omega$ por:

$$\eta_{nk_n}(d_n, x) := \frac{1}{N_{k_n}^{\varsigma_n}(x)} \sum_{x_i \in \bar{B}(x, r_{k_n\text{-}NN}^{\varsigma_n}(x))} y_i,$$

satisfaz

$$\mathbb{E}\left[|\eta_{nk_n}(D_n, x) - \eta(x)|\right] \xrightarrow{n\to\infty} 0, \ \textit{para} \ \ \mu\text{-}q.t \ \ x \in \Omega.$$

Demonstração. Procurando dar ênfase na dependência da variável aleatória σ_n, vamos chamar um par de quantidades já conhecidas por seus novos nomes: $r_x^k(\sigma_n) := r_{k\text{-NN}}^{\sigma_n}(x)$ e $N_x^k(\sigma_n) := N_k^{\sigma_n}(x)$.

Para $n \in \mathbb{N}$, defina a função auxiliar $\tilde{\eta}_n : \Omega^n \times \Omega \to [0,1]$, no par $(\varsigma_n, x) \in \Omega^n \times \Omega$, por:

$$\tilde{\eta}_n(\varsigma_n, x) = \mathbb{E}\left[\eta(X) \,|\, X \in \bar{B}(x, r_x^{k_n}(\varsigma_n))\right]$$

assim, pela desigualdade triangular, para $x \in \text{supp}(\mu)$ qualquer, obtemos:

$$\mathbb{E}[|\eta_{nk_n}(D_n, x) - \eta(x)|] \leq \underbrace{\mathbb{E}[|\eta_{nk_n}(D_n, x) - \tilde{\eta}_n(\sigma_n, x)|]}_{\text{(I)}} + \underbrace{\mathbb{E}[|\tilde{\eta}_n(\sigma_n, x) - \eta(x)|]}_{\text{(II)}}.$$

A tarefa agora será provar que os termos **(I)** e **(II)** tendem para zero quando $n \to \infty$, sobre um subconjunto do supp(μ) com medida 1.

(I) $\mathbb{E}[|\eta_{nk_n}(D_n, x) - \tilde{\eta}_n(\sigma_n, x)|]$:

Denotemos por $(c.s)$ o uso da desigualdade de *Cauchy-Schwarz*, então

$$\begin{aligned}
\mathbb{E}\left[|\eta_{nk_n}(D_n, x) - \tilde{\eta}_n(\sigma_n, x)|\right] &= \mathbb{E}\left[\left|\frac{1}{N_x^{k_n}(\sigma_n)} \sum_{X_i \in \bar{B}(x, r_x^{k_n}(\sigma_n))} Y_i - \tilde{\eta}_n(\sigma_n, x)\right|\right] \\
&= \mathbb{E}\left[\frac{|z_x(D_n) - N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n, x)|}{N_x^{k_n}(\sigma_n)}\right] \\
&\overset{(c.s)}{\leq} \sqrt{\mathbb{E}\left[\frac{|z_x(D_n) - N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n, x)|^2}{N_x^{k_n}(\sigma_n)^2}\right]}, \qquad (15)
\end{aligned}$$

onde

$$z_x(D_n) := \sum_{X_i \in \bar{B}(x, r_x^{k_n}(\sigma_n))} Y_i.$$

Assim, condicionando sobre as variáveis aleatórias $N_x^{k_n}(\sigma_n)$ e $r_x^{k_n}(\sigma_n)$, obtemos

$$\mathbb{E}[|\eta_{nk_n}(D_n,x)-\tilde{\eta}_n(\sigma_n,x)|] \leq \sqrt{\mathbb{E}\left[\mathbb{E}\left[\frac{|z_x(D_n)-N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n,x)|^2}{N_x^{k_n}(\sigma_n)^2} \mid N_x^{k_n}(\sigma_n), r_x^{k_n}(\sigma_n)\right]\right]}.$$

Condicionalmente sobre $r_x^{k_n}(\sigma_n)$ e $N_x^{k_n}(\sigma_n)$, a variável aleatória $z_x(D_n)$ é o número de rótulos $Y_i \in \{0,1\}$ valendo $Y_i = 1$ dentro dos $N_x^{k_n}(\sigma_n)$ rótulos aleatórios dos elementos $X_i \in \bar{B}(x, r_x^{k_n}(\sigma_n))$, com probabilidade de *sucesso* p_i:

$$\begin{aligned}
p_i &= P\left[Y_i = 1 | X_i \in \bar{B}(x, r_x^{k_n}(\sigma_n))\right] \\
&= \frac{P\left[Y_i = 1, X_i \in \bar{B}(x, r_x^{k_n}(\sigma_n))\right]}{P\left[X_i \in \bar{B}(x, r_x^{k_n}(\sigma_n))\right]} \\
&= \frac{P\left[(X_i, Y_i) \in \bar{B}(x, r_x^{k_n}(\sigma_n)) \times \{1\}\right]}{\mu\left(\bar{B}(x, r_x^{k_n}(\sigma_n))\right)} \\
&= \frac{\tilde{\mu}\left(\bar{B}(x, r_x^{k_n}(\sigma_n)) \times \{1\}\right)}{\mu\left(\bar{B}(x, r_x^{k_n}(\sigma_n))\right)} \\
&= \frac{\mu_1\left(\bar{B}(x, r_x^{k_n}(\sigma_n))\right)}{\mu\left(\bar{B}(x, r_x^{k_n}(\sigma_n))\right)} \\
&= \frac{1}{\mu\left(\bar{B}(x, r_x^{k_n}(\sigma_n))\right)} \int_{\bar{B}(x, r_x^{k_n}(\sigma_n))} \eta(z)\, d\mu(z) \\
&= \mathbb{E}\left[\eta(X) \mid X \in \bar{B}(x, r_x^{k_n}(\sigma_n)\right] \\
&= \tilde{\eta}_n(\sigma_n, x),
\end{aligned}$$

isto é, condicionando sobre $N_x^{k_n}(\sigma_n)$ e $r_x^{k_n}(\sigma_n)$, a variável $z_x(D_n)$ é uma variável aleatória *Binomial* com parâmetros $N_x^{k_n}(\sigma_n)$ e $\tilde{\eta}_n(\sigma_n, x)$, logo, condicionalmente, $z_x(D_n)$ tem valor esperado

$$\mathbb{E}\left[z_x(D_n) \mid N_x^{k_n}(\sigma_n), r_x^{k_n}(\sigma_n)\right] = N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n, x)$$

e variância

$$\mathbb{E}\left[|z_x(D_n) - N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n,x)|^2 \mid N_x^{k_n}(\sigma_n), r_x^{k_n}(\sigma_n)\right] = N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n,x)(1-\tilde{\eta}_n(\sigma_n,x)).$$

Em consequência,

$$\begin{aligned}
&\mathbb{E}\left[\frac{|z_x(D_n) - N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n,x)|^2}{N_x^{k_n}(\sigma_n)^2} \mid N_x^{k_n}(\sigma_n), r_x^{k_n}(\sigma_n)\right] \\
&= \frac{1}{N_x^{k_n}(\sigma_n)^2}\mathbb{E}\left[|z_x(D_n) - N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n,x)|^2 \mid N_x^{k_n}(\sigma_n), r_x^{k_n}(\sigma_n)\right] \\
&= \frac{1}{N_x^{k_n}(\sigma_n)^2} N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n,x)(1-\tilde{\eta}_n(\sigma_n,x)) \\
&= \frac{\tilde{\eta}_n(\sigma_n,x)(1-\tilde{\eta}_n(\sigma_n,x))}{N_x^{k_n}(\sigma_n)} \\
&\overset{(*)}{\leq} \frac{1}{4N_x^{k_n}(\sigma_n)} \\
&\leq \frac{1}{4k_n},
\end{aligned}$$

onde o passo marcado com $(*)$, é porque a função real $t \mapsto t(1-t)$ possui um máximo global em $t = 1/2$.

Finalmente, aplicando valor esperado, obtemos

$$\mathbb{E}\left[\frac{|z_x(D_n) - N_x^{k_n}(\sigma_n)\tilde{\eta}_n(\sigma_n,x)|^2}{N_x^{k_n}(\sigma_n)^2}\right] \leq \frac{1}{4k_n},$$

desigualdade que combinada com (15), fornece:

$$\mathbb{E}\left[|\eta_{nk_n}(D_n,x) - \tilde{\eta}_n(\sigma_n,x)|\right] \leq \frac{1}{2\sqrt{k_n}} \xrightarrow{n\to\infty} 0,\ \forall x \in \operatorname{supp}(\mu).$$

(II) $\mathbb{E}[|\tilde{\eta}_n(\sigma_n,x) - \eta(x)|]$:

Como Ω satisfaz a condição **(LBF)** para μ e $\eta \in L^1(\mu)$, temos

$$\lim_{r\to 0^+} \frac{1}{\mu(\bar{B}(x,r))}\int_{\bar{B}(x,r)} |\eta(z) - \eta(x)|\, d\mu(z) = 0, \text{ para } \mu\text{-q.t } x \in \Omega. \tag{16}$$

Seja $A_\eta \subset \Omega$ o conjunto de pontos onde (16) vale, então $\mu(A_\eta) = 1$, e como da Proposição 5.1.1 temos $\mu(\operatorname{supp}(\mu)) = 1$, finalmente inferimos $\mu(A_\eta \cap \operatorname{supp}(\mu)) = 1$.

Para $x \in S_\eta := A_\eta \cap \operatorname{supp}(\mu)$, temos

$$\begin{aligned}
\mathbb{E}[|\tilde{\eta}_n(\sigma_n,x) - \eta(x)|] &\leq \mathbb{E}\left[\frac{1}{\mu\left(\bar{B}(x, r_x^{k_n}(\sigma_n))\right)}\int_{\bar{B}(x,r_x^{k_n}(\sigma_n))} |\eta(z) - \eta(x)|\, d\mu(z)\right] \\
&\leq \mathbb{E}\left[\sup_{0<r\leq r_x^{k_n}(\sigma_n)} \frac{1}{\mu(\bar{B}(x,r))}\int_{\bar{B}(x,r)} |\eta(z) - \eta(x)|\, d\mu(z)\right].
\end{aligned} \tag{17}$$

Para simplificar a notação, defina $G(x,r) := \frac{1}{\mu(\bar{B}(x,r))}\int_{\bar{B}(x,r)} |\eta(z) - \eta(x)|\, d\mu(z)$. Escolha $\epsilon > 0$ arbitrário, de (16), a condição **(LBF)** para η, existe $\delta > 0$ tal que para todo $0 < r \leq \delta$, temos $G(x,r) < \epsilon$.

Logo, para a expressão (17), temos

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0<r\leq r_x^{k_n}(\sigma_n)} \frac{1}{\mu(\bar{B}(x,r))}\int_{\bar{B}(x,r)} |\eta(z)-\eta(x)|\,d\mu(z)\right]\\
&= \mathbb{E}\left[\mathbb{I}_{\{r_x^{k_n}(\sigma_n)\leq\delta\}} \sup_{0<r\leq r_x^{k_n}(\sigma_n)} G(x,r)\right] + \mathbb{E}\left[\mathbb{I}_{\{r_x^{k_n}(\sigma_n)>\delta\}} \sup_{0<r\leq r_x^{k_n}(\sigma_n)} G(x,r)\right]\\
&\leq \epsilon + \mathbb{E}\left[\mathbb{I}_{\{r_x^{k_n}(\sigma_n)>\delta\}} \sup_{0<r\leq r_x^{k_n}(\sigma_n)} G(x,r)\right]\\
&\leq \epsilon + \mathbb{E}\left[\mathbb{I}_{\{r_x^{k_n}(\sigma_n)>\delta\}} \sup_{r>0} G(x,r)\right]\\
&\leq \epsilon + P\left[r_x^{k_n}(\sigma_n)>\delta\right],
\end{aligned}$$

pois, como $x \in S_\eta \subset \text{supp}(\mu)$, necessariamente $\sup_{r>0} G(x,r) \leq 1$ e $\mathbb{E}\left[\mathbb{I}_{\{r_x^{k_n}(\sigma_n)>\delta\}}\right] = P\left[r_x^{k_n}(\sigma_n)>\delta\right]$.

Resumindo

$$\mathbb{E}[|\tilde{\eta}_n(\sigma_n,x)-\eta(x)|] \leq \epsilon + P\left[r_x^{k_n}(\sigma_n)>\delta\right]$$

e pelo Teorema 5.1.1 (Lema de Cover-Hart), $r_x^{k_n}(\sigma_n) \xrightarrow[n\to\infty]{q.c} 0$, para qualquer $x \in \text{supp}(\mu)$, logo, necessariamente $r_x^{k_n}(\sigma_n) \xrightarrow[n\to\infty]{P} 0$ e em particular, $P\left[r_x^{k_n}(\sigma_n)>\delta\right] \xrightarrow{n\to\infty} 0$, assim como $\epsilon > 0$ é arbitrário, concluímos:

$$\mathbb{E}[|\tilde{\eta}_n(\sigma_n,x)-\eta(x)|] \xrightarrow{n\to\infty} 0, \text{ para } x \in S_\eta.$$

Portanto, como a conclusão sobre o termo **(I)** vale para qualquer elemento $x \in \text{supp}(\mu)$, $S_\eta \subset \text{supp}(\mu)$ e $\mu(S_\eta) = 1$, finalmente obtemos

$$\mathbb{E}\left[|\eta_{nk_n}(D_n,x)-\eta(x)|\right] \xrightarrow{n\to\infty} 0, \text{ para } \mu\text{-q.t } x \in \Omega.$$

■

Observação 5.2.1. A técnica da demonstração do Teorema 5.2.1 não serve para provar a consistência do classificador k-NN, pois podem existir mais de k elementos de σ_n a uma distância menor ou igual que $r_{k\text{-NN}}^{\sigma_n}(x)$ do ponto $x \in \Omega$ a ser classificado, e portanto, para classificar x utilizando o k-NN precisamos fazer a *escolha* de k desses vizinhos, o que torna a prova muito mais difícil que no caso do classificador ${}^+k$-NN onde não precisamos escolher.

Finalizamos essa seção com três simples corolários, onde por padrão assumiremos *todas as hipóteses* do Teorema 5.2.1.

Corolário 5.2.1.1. *A regra de aprendizagem $\mathcal{L}_{+_{k\text{-}NN}}$ é consistente em qualquer espaço métrico separável satisfazendo a condição **(LBF)** para a lei de distribuição dos dados não rotulados.*

Demonstração. Como mencionado no preâmbulo, pelo Corolário 2.2.1.2, a regra de aprendizagem $\mathcal{L}_{+k\text{-NN}}$ será consistente com a medida $\widetilde{\mu}$ para $k = k_n$, se

$$\mathbb{E}\left[|\eta_{nk_n}(D_n, X) - \eta(X)|\mathbb{I}_{\Omega\setminus\eta_{1/2}}(X)\right] \leq \mathbb{E}\left[|\eta_{nk_n}(D_n, X) - \eta(X)|\right] \xrightarrow{n\to\infty} 0,$$

mas pelo Teorema da Convergência Dominada de Lebesgue e pelo Teorema 5.2.1, temos

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[|\eta_{nk_n}(D_n, X) - \eta(X)|\right] &= \lim_{n\to\infty} \int_{S_\eta} \mathbb{E}\left[|\eta_{nk_n}(D_n, x) - \eta(x)|\right] \, d\mu(x) \\ &= \int_{S_\eta} \lim_{n\to\infty} \mathbb{E}\left[|\eta_{nk_n}(D_n, x) - \eta(x)|\right] \, d\mu(x) \\ &= 0,\end{aligned}$$

onde $S_\eta \subset \text{supp}(\mu)$ é o conjunto com medida $\mu(S_\eta) = 1$ na prova do Teorema 5.2.1. ■

Corolário 5.2.1.2. *A regra de aprendizagem $\mathcal{L}_{+k\text{-NN}}$ é universalmente consistente em qualquer espaço métrico completo e separável que possua dimensão σ-finita no sentido de Nagata.*

Demonstração. Do Teorema 2.5.1 [54, Preiss], vemos que num espaço completo e separável de dimensão σ-finita no sentido de Nagata, a condição **(LBF)** vale para qualquer medida boreliana localmente finita, e em particular, vale para qualquer medida boreliana de probabilidade, assim do Corolário 5.2.1.1, inferimos que a regra ${}^{+}k$-NN será *universalmente consistente* em qualquer espaço métrico completo e separável com dimensão σ-finita no sentido de Nagata. ■

Corolário 5.2.1.3. *As regras de aprendizagem p-ádicas, $\mathcal{L}_{(k,p)}$, $\mathcal{L}^{T_M}_{(k,p)}$ e as regras de aprendizagem no cubo $[0,1]^d$, $\mathcal{L}^{\Phi_p}_{(k,p)}$, $\mathcal{L}^{T_M\circ\Phi_p}_{(k,p)}$ onde a transformação linear $T_M : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$ é definida por uma matriz*[4] *$M \in \mathcal{Z} = \{M \in \mathrm{M}_d(\mathbb{Z}_p) : \det(M) \neq 0\}$ como na Observação 4.4.4, são universalmente consistentes no espaço ultramétrico $\mathbb{Z}_p^d$ e no cubo $[0,1]^d$, respectivamente, para qualquer $d \in \mathbb{N}$.*

Demonstração. Como o espaço métrico completo e separável, $(\mathbb{Z}_p^d, \|\cdot\|_p)$, é não-arquimediano, da Proposição 2.5.2 concluímos que ele tem $\dim^{+\infty}_{\text{Nag}}(\mathbb{Z}_p^d) = 0$, e em particular, que ele tem dimensão σ-finita no sentido de Nagata, logo, pelo Corolário 5.2.1.2, a regra ${}^{+}k$-NN aplicada no espaço $(\mathbb{Z}_p^d, \|\cdot\|_p)$ (a regra de aprendizagem p-ádica, $\mathcal{L}_{(k,p)}$), é universalmente consistente. Por último, graças ao Teorema 2.3.3, as regras de aprendizagem $\mathcal{L}^{T_M}_{(k,p)}$ também são universalmente consistentes em $(\mathbb{Z}_p^d, \|\cdot\|_p)$ e as regras de aprendizagem $\mathcal{L}^{\Phi_p}_{(k,p)}$ e $\mathcal{L}^{T_M\circ\Phi_p}_{(k,p)}$, são universalmente consistentes no cubo $[0,1]^d$, para $T_M : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$, transformação linear (e portanto contínua) definida pela matriz *inversível* $M \in \mathcal{Z} = \{M \in \mathrm{M}_d(\mathbb{Z}_p) : \det(M) \neq 0\}$. ■

[4] A matriz $M \in \mathrm{M}_{d_2\times d_1}(\mathbb{K})$ de entradas no corpo $\mathbb{K}$, define uma transformação linear entre $\mathbb{K}$-espaços vetoriais $T_M : \mathbb{K}^{d_1} \to \mathbb{K}^{d_2}$, mediante o produto matriz-vetor usual, $T_M(x) = Mx$, para todo $x \in \mathbb{K}^{d_1}$.

5.3 CONSISTÊNCIA DA APRENDIZAGEM ENSEMBLE

Finalizamos o capítulo com a prova da consistência da família de regras de classificação dada pelo voto majoritário entre membros de uma família consistente de regras de classificação binária, resultado que permitirá definir classificadores dados pelo voto majoritário entre *classificadores p-ádicos aleatórios* que definem uma família universalmente consistente de regras de classificação no espaço ultramétrico $(\mathbb{Z}_p^d, \|\cdot\|_p)$, gerando assim (inspirados no termo cunhado por Leo Breiman [14]) uma espécie de *florestas de árvores p-ádicas aleatórias*[5] ou *p-adic random forests* que serão universalmente consistentes em $(\mathbb{Z}_p^d, \|\cdot\|_p)$, famílias de regras que combinadas com a função Φ_p, geram novas famílias universalmente consistentes de regras de aprendizagem no cubo $[0,1]^d$.

No que segue, Ω é um espaço boreliano padrão e $\widetilde{\mu}$ denota a lei dos dados rotulados no espaço $\Omega \times \{0,1\}$. Começamos definindo o conceito de *família consistente de regras de aprendizagem*.

Definição 5.3.1 (Família consistente de regras de aprendizagem)**.** Seja $\mathcal{F} = \{\mathcal{L}(z)\}_{z\in\mathcal{Z}}$, uma família de regras de aprendizagem sobre Ω, *indexada* pelo espaço boreliano de probabilidade *separável*[6] $(\mathcal{Z}, \nu)$, tal que para cada $z \in \mathcal{Z}$, temos $\mathcal{L}(z) = (g_n(z))_{n=1}^{\infty}$, onde as aplicações

$$\mathcal{Z} \times (\Omega \times \{0,1\})^n \times \Omega \ni (z, d_n, x) \mapsto g_n(z)(d_n)(x) \in \{0,1\},$$

são borelianas, e suponha que as variáveis aleatórias $D_n \sim \widetilde{\mu}^n$, $(X,Y) \sim \widetilde{\mu}$ e $Z \sim \nu$, são independentes.

Dizemos que a família $\mathcal{F}$ é consistente com a medida $\widetilde{\mu}$, se

$$P[g_n(Z)(D_n)(X) \neq Y] \xrightarrow{n\to\infty} \ell^*(\widetilde{\mu}),$$

onde a probabilidade é relativa à medida produto $\nu \otimes \widetilde{\mu}^n \otimes \widetilde{\mu}$; e que $\mathcal{F}$ é *universalmente* consistente sobre Ω, se ela é consistente com qualquer medida boreliana de probabilidade sobre $\Omega \times \{0,1\}$.

Observação 5.3.1. Da definição acima, observamos que uma família $\mathcal{F} = \{\mathcal{L}(z)\}_{z\in\mathcal{Z}} = \{(g_n(z))_{n=1}^{\infty}\}_{z\in\mathcal{Z}}$ é consistente com a medida $\widetilde{\mu}$, se a variável aleatória no espaço $\mathcal{Z}$, $P[g_n(Z)(D_n)(X) \neq Y \,|\, Z]$, converge para o erro de Bayes em *probabilidade*, ou seja,

$$P[g_n(Z)(D_n)(X) \neq Y \,|\, Z] \xrightarrow[n\to\infty]{P} \ell^*(\widetilde{\mu}).$$

Logo, se $\mathcal{F}$ é uma família tal que as regras $\mathcal{L}(z)$ são consistentes com a medida $\widetilde{\mu}$ *para todo* $z \in \mathcal{Z}$ ou para ν-q.t $z \in \mathcal{Z}$, então $\mathcal{F}$ será necessariamente consistente, pois nesses

[5] No Capítulo 4, definimos a regra de aprendizagem p-ádica mediante uma árvore de decisão $\mathcal{A}_{p^d}^H(d_n)$, portanto, no contexto dos métodos ensemble, chamamos as regras de aprendizagem p-ádicas simplesmente de árvores p-ádicas.

[6] Para nossos propósitos, é suficiente que $\mathcal{Z}$ seja um espaço boreliano de probabilidade cuja topologia possua uma base enumerável de abertos, como por exemplo, um espaço métrico *separável*, mas não exigiremos que $\mathcal{Z}$ seja completo.

casos, teremos que a variável aleatória no espaço $\mathcal{Z}$, $P[g_n(Z)(D_n)(X) \neq Y \,|\, Z]$, converge *pontualmente* ou converge *quase certamente* para o erro $\ell^*(\widetilde{\mu})$, respectivamente, implicando assim, a convergência em probabilidade para $\ell^*(\widetilde{\mu})$.

Antes de ir ao resultado principal dessa seção, vamos escrever de uma forma equivalente a definição de consistência para uma família de regras de aprendizagem.

Lema 5.3.1. *Sejam (μ, η), o par que caracteriza a medida $\widetilde{\mu}$ e $(\mathcal{Z}, \nu)$ um espaço boreliano de probabilidade separável, com $Z \sim \nu$, $D_n \sim \widetilde{\mu}^n$, $X \sim \mu$, variáveis aleatórias independentes. Então, a família de regras de aprendizagem sobre Ω, $\mathcal{F} = \{\mathcal{L}(z)\}_{z\in\mathcal{Z}}$, com $\mathcal{L}(z) = (g_n(z))_{n=1}^{\infty}$; é consistente com a medida $\widetilde{\mu}$, se e somente se,*

$$\mathbb{E}[(1-2\eta(X))\, \mathbb{I}_{\{g_n=1\}}(Z, D_n, X)\mathbb{I}_{\{\eta<1/2\}}(X)] \xrightarrow{n\to\infty} 0 \tag{18}$$

e

$$\mathbb{E}[(2\eta(X)-1)\, \mathbb{I}_{\{g_n=0\}}(Z, D_n, X)\mathbb{I}_{\{\eta>1/2\}}(X)] \xrightarrow{n\to\infty} 0. \tag{19}$$

Demonstração. Seja $\ell^* = \ell^*(\widetilde{\mu})$ o erro de Bayes para $\widetilde{\mu}$. Seguindo as ideias da prova do Teorema 2.1.2 e condicionando sobre $Z = z$, $D_n = d_n$ e $X = x$, obtemos

$$\begin{aligned} P[g_n(Z)(D_n)(X) \neq & Y \,|\, Z = z, D_n = d_n, X = x] - P[g^*(X) \neq Y \,|\, X = x] \\ &= |2\eta(x) - 1||g_n(z)(d_n)(x) - g^*(x)|\mathbb{I}_{\{\eta\neq 1/2\}}(x) \end{aligned}$$

logo, tomando valor esperado no espaço $\mathcal{Z} \times (\Omega \times \{0,1\})^n \times \Omega$, temos

$$P[g_n(Z)(D_n)(X) \neq Y] - \ell^* = \mathbb{E}[|2\eta(X) - 1||g_n(Z)(D_n)(X) - g^*(X)|\mathbb{I}_{\{\eta\neq 1/2\}}(X)],$$

onde o valor esperado acima pode se escrever como a soma de

$$\mathbb{E}[(1-2\eta(X))\mathbb{I}_{\{g_n=1\}}(Z, D_n, X)\mathbb{I}_{\{\eta<1/2\}}(X)] \geq 0$$

e

$$\mathbb{E}[(2\eta(X)-1)\mathbb{I}_{\{g_n=0\}}(Z, D_n, X)\mathbb{I}_{\{\eta>1/2\}}(X)] \geq 0,$$

assim, a família $\mathcal{F}$ é consistente com a medida $\widetilde{\mu}$, se e somente se,

$$\mathbb{E}[(1-2\eta(X))\, \mathbb{I}_{\{g_n=1\}}(Z, D_n, X)\mathbb{I}_{\{\eta<1/2\}}(X)] \xrightarrow{n\to\infty} 0$$

e

$$\mathbb{E}[(2\eta(X)-1)\, \mathbb{I}_{\{g_n=0\}}(Z, D_n, X)\mathbb{I}_{\{\eta>1/2\}}(X)] \xrightarrow{n\to\infty} 0.$$

■

Agora vamos definir formalmente o classificador dado pelo voto majoritário de classificadores *binários* sobre Ω.

Definição 5.3.2 (Voto majoritário)**.** Seja $m \in \mathbb{N}$. Sobre Ω, definimos o classificador dado pelo *voto majoritário* dos classificadores $g^{(i)}$, para $i \in [m]$ e que denotaremos por g^m; como o classificador do tipo *plug-in* definido para cada $x \in \Omega$, mediante

$$g^m(x) = \begin{cases} 1, & \eta^m(x) \geq \frac{1}{2} \\ 0, & \text{outro caso,} \end{cases}$$

onde

$$\eta^m(x) := \frac{1}{m} \sum_{i=1}^{m} g^{(i)}(x),$$

é uma aproximação da função de regressão desconhecida, η.

A seguir, o resultado principal da seção.

Proposição 5.3.1. *Sejam $m, f : \mathbb{N} \to \mathbb{N}$ funções arbitrárias, $\mathcal{F} = \{(g_n(z))_{n=1}^{\infty}\}_{z \in \mathcal{Z}}$, com $(\mathcal{Z}, \nu)$ espaço boreliano de probabilidade separável; uma família de regras de classificação binária sobre Ω que é consistente com a medida $\widetilde{\mu}$ e suponha que $Z^{\infty} := (Z_1, Z_2, \dots,) \in \mathcal{Z}^{\infty}$ é tal que as variáveis aleatórias $Z^{\infty} \sim \nu^{\infty}$, $D_n \sim \widetilde{\mu}^n$ e $X \sim \mu$, são independentes. Considere a família de regras de classificação $\mathcal{F}^{m,f} = \{(g_n^{m,f}(z^{\infty}))_{n=1}^{\infty}\}_{z^{\infty} \in \mathcal{Z}^{\infty}}$, onde para cada $z^{\infty} = (z_1, z_2, \dots) \in \mathcal{Z}^{\infty}$ e para cada amostra $d_n \in (\Omega \times \{0,1\})^n$, o classificador $g_n^{m,f}(z^{\infty})(d_n)$ é dado pelo voto majoritário dos classificadores $g_n(z_{f(i)})(d_n)$, $i \in [m(n)]$. Então, a família $\mathcal{F}^{m,f}$ é consistente com a medida $\widetilde{\mu}$, isto é*

$$P[g_n^{m,f}(Z^{\infty})(D_n)(X) \neq Y] \xrightarrow{n \to \infty} \ell^*(\widetilde{\mu}).$$

Demonstração. Pelo Lema 5.3.1, é suficiente provar que

$$(1 - 2\eta(X))\, \mathbb{I}_{\{g_n^{m,f}=1\}}(Z^{\infty}, D_n, X) \mathbb{I}_{\{\eta<1/2\}}(X) \xrightarrow[n \to \infty]{P} 0$$

e

$$(2\eta(X) - 1)\, \mathbb{I}_{\{g_n^{m,f}=0\}}(Z^{\infty}, D_n, X) \mathbb{I}_{\{\eta>1/2\}}(X) \xrightarrow[n \to \infty]{P} 0.$$

As provas das condições acima são análogas, portanto aqui vamos escrever apenas a prova da primeira condição. Seja $\epsilon > 0$. Pela desigualdade de *Markov (M)*, temos

$$\begin{aligned} P[\epsilon < (1 - 2\eta(X))\, \mathbb{I}_{\{g_n^{m,f}=1\}}(Z^{\infty}, D_n, X) \mathbb{I}_{\{\eta<1/2\}}(X)] & \\ = P\left[\frac{1}{m(n)} \sum_{i=1}^{m(n)} \mathbb{I}_{B_i}(Z^{\infty}, D_n, X) \geq 1/2\right] & \\ \overset{(M)}{\leq} \frac{2}{m(n)} \sum_{i=1}^{m(n)} \mathbb{E}[\mathbb{I}_{B_i}(Z^{\infty}, D_n, X)] & \\ = \frac{2}{m(n)} \sum_{i=1}^{m(n)} P[B_i], & \end{aligned}$$

onde os conjuntos $B_i \subset \mathcal{Z}^\infty \times (\Omega \times \{0,1\})^n \times \Omega$, são definidos por

$$B_i = \{(z^\infty, d_n, x) : \epsilon < (1-2\eta(x))\ \mathbb{I}_{\{g_n=1\}}(\pi_{f(i)}(z^\infty), d_n, x)\mathbb{I}_{\{\eta<1/2\}}(x)\},$$

com $\pi_j : \mathcal{Z}^\infty \to \mathcal{Z}$, $(z_1, z_2, \dots) \mapsto z_j$, $j \in \mathbb{N}$, sendo as projeções canônicas. Como por hipótese $Z_i \overset{i.i.d}{\sim} \nu$, temos que para todo[7] $i \in \mathbb{N}$

$$P[B_i] = P[\epsilon < (1-2\eta(X))\ \mathbb{I}_{\{g_n=1\}}(Z, D_n, X)\mathbb{I}_{\{\eta<1/2\}}(X)]$$

assim, como a família $\mathcal{F}$ é consistente com a medida $\widetilde{\mu}$, pelo Lema 5.3.1, $\mathcal{F}$ satisfaz as condições (18) e (19), de onde inferimos

$$\begin{aligned} &P[\epsilon < (1-2\eta(X))\ \mathbb{I}_{\{g_n^{m,f}=1\}}(Z^\infty, D_n, X)\mathbb{I}_{\{\eta<1/2\}}(X)] \\ &\quad \leq 2P[\epsilon < (1-2\eta(X))\ \mathbb{I}_{\{g_n=1\}}(Z, D_n, X)\mathbb{I}_{\{\eta<1/2\}}(X)] \xrightarrow{n\to\infty} 0. \end{aligned}$$

Como $\epsilon > 0$ é arbitrário, temos que a regra $\mathcal{F}^{m,f}$ satisfaz a condição (18) e, procedendo de forma análoga, também satisfaz a condição (19), portanto, a família $\mathcal{F}^{m,f}$ é consistente com a medida $\tilde{\mu}$. ■

Finalmente, abordamos o caso da família de regras de aprendizagem dada pelo voto majoritário de classificadores definidos pelas regras da Observação 4.4.4: $\mathcal{F}_{(k,p)} = \left\{\mathcal{L}_{(k,p)}^{T_M}\right\}_{M\in\mathcal{Z}}$ ou $\mathcal{F}_{(k,p)}^{\Phi_p} = \left\{\mathcal{L}_{(k,p)}^{T_M\circ\Phi_p}\right\}_{M\in\mathcal{Z}}$; famílias indexadas pelo espaço boreliano de probabilidade separável da Observação 4.4.5, denotado por $(\mathcal{Z}, \nu)$, onde $\mathcal{Z} = \{M \in \mathrm{M}_d(\mathbb{Z}_p) : \det(M) \neq 0\}$ e $\nu = \frac{1}{\mu_{\mathrm{M}_d}(\mathcal{Z})}\mu_{\mathrm{M}_d}$, com μ_{M_d} medida boreliana de probabilidade em $\mathrm{M}_d(\mathbb{Z}_p)$. Do Corolário 5.2.1.3 e da Observação 5.3.1, inferimos que *as famílias* $\mathcal{F}_{(k,p)}$ e $\mathcal{F}_{(k,p)}^{\Phi_p}$, são *universalmente consistentes* no espaço $\mathbb{Z}_p^d$ e no cubo $[0,1]^d$, respectivamente, assim, pela Proposição 5.3.1, a família dada pelo voto majoritário de um número finito de classificadores definidos pelas regras $\mathcal{L}_{(k,p)}^{T_M}$ ou $\mathcal{L}_{(k,p)}^{T_M\circ\Phi_p}$, será universalmente consistente em $\mathbb{Z}_p^d$ ou no cubo $[0,1]^d$, respectivamente. Resumimos esses fatos no último resultado do capítulo.

Corolário 5.3.1.1 (Consistência da Floresta p-ádica Aleatória)**.** *Sejam $m, f : \mathbb{N} \to \mathbb{N}$, funções arbitrárias, $d, p \in \mathbb{N}$, $p > 1$ número primo e $(\mathcal{Z}, \nu)$, com $\mathcal{Z} = \{M \in \mathrm{M}_d(\mathbb{Z}_p) : \det(M) \neq 0\}$ e $\nu = \frac{1}{\mu_{\mathrm{M}_d}(\mathcal{Z})}\mu_{\mathrm{M}_d}$. Considere a família de regras de aprendizagem $\mathcal{F}_{(k,p)}^{m,f} = \{\mathcal{L}_{(k,p)}^{m,f}(M^\infty)\}_{M^\infty\in\mathcal{Z}^\infty} = \{(g_n^{m,f}(M^\infty))_{n=1}^\infty\}_{M^\infty\in\mathcal{Z}^\infty}$, indexada pelo espaço $(\mathcal{Z}^\infty, \nu^\infty)$, onde para cada $M^\infty \in \mathcal{Z}^\infty$ e $d_n \in (\Omega \times \{0,1\})^n$, o classificador $g_n^{m,f}(M^\infty)(d_n)$ é dado pelo voto majoritário dos classificadores definidos pelas regras $\mathcal{L}_{(k,p)}(M_i)$, $i \in [m(n)]$, na amostra d_n. Então, a família $\mathcal{F}_{(k,p)}^{m,f}$ é universalmente consistente no espaço Ω, onde $\mathcal{L}_{(k,p)}(M_i) = \mathcal{L}_{(k,p)}^{T_{M_i}}$, se $\Omega = \mathbb{Z}_p^d$ ou $\mathcal{L}_{(k,p)}(M_i) = \mathcal{L}_{(k,p)}^{T_{M_i}\circ\Phi_p}$, se $\Omega = [0,1]^d$.*

[7] Para $i \in \mathbb{N}$, a função $G_i : \mathcal{Z}^\infty \times (\Omega \times \{0,1\})^n \times \Omega \to \mathcal{Z} \times (\Omega \times \{0,1\})^n \times \Omega$, definida mediante: $(z^\infty, d_n, x) \mapsto (\pi_{f(i)}(z^\infty), d_n, x)$, é boreliana (contínua), satisfazendo $(\nu^\infty \otimes \widetilde{\mu}^n \otimes \mu) \circ G_i^{-1} = \nu \otimes \widetilde{\mu}^n \otimes \mu$ e $B_i = G_i^{-1}(\{(z, d_n, x) : \epsilon < (1-2\eta(x))\ \mathbb{I}_{\{g_n=1\}}(z, d_n, x)\mathbb{I}_{\{\eta<1/2\}}(x)\})$.

6 EXPERIMENTOS NUMÉRICOS

Neste capítulo, vamos realizar testes numéricos com a regra de classificação desenvolvida no Capítulo 4: a regra $\mathcal{L}_k^{\Phi_p} := \mathcal{L}_{(k,p)}^{\Phi_p}$ para $k, p \in \mathbb{N}$, com p um número primo.

Quando utilizamos um algoritmo de aprendizagem de máquina supervisionada para que *aprenda* como rotular um determinado conjunto de dados, medimos a habilidade do algoritmo para aprender esses dados mediante uma ou algumas *métricas de avaliação* cuja escolha depende do problema que buscamos resolver, portanto, medir o desempenho de um classificador de forma *absoluta* com relação a alguma de essas métricas, não faz sentido. Assim, para ver em ação a nova regra de classificação, consideraremos uma família de 8 conjuntos de dados rotulados disponíveis de forma gratuita na web, onde para cada um desses conjuntos, vamos medir a habilidade de $\mathcal{L}_k^{\Phi_p}$ para rotular os dados utilizando alguma métrica de avaliação adequada que será *sugerida* pela natureza do problema que deu origem ao conjunto de dados; e como ponto de comparação, usaremos o desempenho (sobre o mesmo conjunto de dados e sob as mesmas métricas de avaliação) de cinco regras de classificação bem conhecidas e amplamente utilizadas, concluíndo com um *teste* para medir a *significância estatística* da diferença, com relação à principal métrica de avaliação escolhida, entre os modelos de classificação concorrentes.

Esse capítulo está formado por uma seção de assuntos *preliminares* e uma seção de *resultados*. Na seção de *preliminares*, além de definir a nossa estratégia de comparação de modelos de aprendizagem de máquina supervisionada, veremos todo o necessário para poder fazer dita comparação; e na seção de *resultados*, e para cada um dos conjuntos de dados escolhidos, faremos uma pequena análise do problema que permita determinar as métricas de avaliação adequadas para a comparação, seguida dos resultados e conclusões do caso.

6.1 PRELIMINARES

Agora vamos ver todo o necessário para poder realizar a comparação dos modelos de aprendizagem de máquina sobre um conjunto de dados rotulados.

Primeiro veremos a *estratégia de comparação* entre algoritmos de aprendizagem de máquina supervisionada que adotaremos neste trabalho, seguida de um breve resumo das métricas de avaliação e dos algoritmos de classificação utilizados na comparação; além da estratégia utilizada para escolher a métrica de avaliação adequada em função das exigências do problema que dá origem aos dados considerados.

6.1.1 Estratégia de comparação

Vamos definir os passos que seguiremos para comparar o desempenho dos algoritmos de classificação em um conjunto de dados.

1. Pré-processamento de dados

Para cada um dos conjuntos de dados, primeiro é dada uma breve descrição da origem do conjunto e do problema de classificação subjacente, para logo fazer um exame preliminar com a intenção de verificar se o conjunto possui *dados não válidos*[1]; seguido do pré-processamento mais adequado para tratar essas eventuais dificuldades, obtendo assim um conjunto de dados útil para ser utilizado pelos algoritmos de classificação.

Também é feita uma breve análise sobre as classes presentes no conjunto de dados, e no caso de existir mais de duas classes, o problema é transformado num problema de *classificação binária* escolhendo uma das classes como alvo, a qual será rotulada com 1, e rotulando o restante das classes com 0. Além do anterior, e usando a informação básica do conjunto de dados, determinamos a métrica de avaliação mais adequada para medir o desempenho dos classificadores, métricas que serão enunciadas na Seção 6.1.2.

Por último, dividimos o conjunto de dados resultante do pré-processamento em conjuntos de *treinamento* e *teste*, reservando o conjunto de teste para ser utilizado no passo 3.

2. Otimização de hiperparâmetros

Suponha que cada modelo que será utilizado na comparação possui um conjunto de *hiperparâmetros*. Em cada modelo, diferentes combinações dos hiperparâmetros geram diferentes classificadores.

Para cada um dos modelos considerados, *escolhemos* um conjunto de valores possíveis para cada um dos seus hiperparâmetros, formando o que chamaremos de *grade de hiperparâmetros*; obtendo assim um conjunto de *representantes do modelo*, um por cada combinação hiperparamétrica.

A continuação, fazemos $c_v \in \mathbb{N}$ divisões do conjunto de *treinamento* obtido no final do passo 1, em subconjuntos de treinamento e validação, e treinamos/validamos cada representante do modelo em cada uma de essas divisões, registrando a sua *pontuação média* com relação à métrica de avaliação definida no passo 1. Essa técnica é chamada de *validação cruzada*, ou *cross-validation* [15].

Por último, para cada modelo encontramos seu *melhor representante* na grade, como sendo aquele representante que registra a maior pontuação média ao longo das c_v validações cruzadas.

3. Avaliação dos melhores representantes

Uma vez que dispomos dos melhores representantes para cada modelo do passo 2, treinamos/testamos cada um desses classificadores nos conjuntos de treinamento e teste

[1] Chamaremos *dado não válido*, a qualquer dado do tipo `null`, `NaN` ou qualquer formato que não possua uma conversão clara para dados do tipo `int` ou `float`.

obtidos no final do passo 1, e registramos as pontuações para várias métricas de avaliação.

4. Análise bayesiana de resultados experimentais

Normalmente, a análise comparativa de modelos termina no passo 3. Os resultados de saída da função `GridSearchCV` da biblioteca `sklearn` de *Python*, não fornecem informações sobre a certeza das diferenças entre os modelos. Isto é, não diz se os resultados anteriores são estatisticamente significativos. Para avaliar isso, é necessário realizar um teste estatístico. Especificamente, para contrastar o desempenho de dois modelos, devemos comparar estatisticamente suas pontuações com relação a uma métrica de avaliação específica.

Para cada modelo, realizamos $c_v \times r$ validações cruzadas, isto é, $c_v \in \mathbb{N}$ validações cruzadas repetindo o processo $r \in \mathbb{N}$ vezes, onde cada vez as c_v validações são feitas com um novo reordenamento aleatório do conjunto original. Assim, obtemos $c_v \cdot r$ amostras (pontuações da métrica de avaliação) para cada modelo. No entanto, as pontuações dos modelos não são independentes: todos os modelos são avaliados nas mesmas $c_v \cdot r$ divisões em conjuntos de treinamento/validação, aumentando a correlação entre o desempenho dos modelos. Uma vez que algumas partições dos dados podem tornar a distinção das classes particularmente fácil ou difícil de encontrar para todos os modelos, as pontuações deles irão *co-variar* e o seu desempenho dependerá muito da divisão do conjunto. Como consequência, se é assumida a independência entre amostras estaremos subestimando a variância calculada em nossos testes estatísticos, aumentando o número de erros falsos positivos, ou seja, detectando uma diferença significativa entre modelos quando ela não existe [46]. Vários testes estatísticos com correção de variância foram desenvolvidos para esses casos. Aqui consideramos o teste t corrigido de *Nadeau e Bengio* [5, 46] sob a estrutura estatística *bayesiana*.

Dados dois modelos a comparar, suponha que estamos interessados em saber se o primeiro modelo é significativamente melhor que o segundo quando comparamos a média de suas pontuações nas validações cruzadas. Uma forma de ter uma noção da significância da diferença na pontuação média de dois modelos, é usar uma *estimativa bayesiana* para calcular a probabilidade de que o primeiro modelo seja melhor que o segundo. A estimativa bayesiana produzirá uma distribuição *seguida* pela média μ das diferenças no desempenho de dois modelos.

A estimativa bayesiana pode ser realizada de várias formas para responder à nossa questão, mas aqui será usada a abordagem sugerida em [5].

Considere a variável aleatória t de Student:

$$t \sim St\left(\mu; n-1, \bar{x}, (\frac{1}{n} + \frac{n_{test}}{n_{train}})\hat{\sigma}^2\right)$$

onde $n = n_{train} + n_{test}$, n_{train} e n_{test} são o número de instâncias usadas como treinamento e teste nas validações cruzadas, $\bar{x}$ é a média das diferenças de pontuações observadas dos

modelos, e $\hat{\sigma}^2$ é a variância dessas diferenças nas pontuações.

Podemos calcular a probabilidade de que o primeiro modelo seja melhor que o segundo calculando a área abaixo da curva da função de distribuição de probabilidade da variável t desde zero até ∞. E também o inverso: podemos calcular a probabilidade de que o segundo modelo seja melhor que o primeiro calculando a área abaixo da curva desde $-\infty$ até zero.

Região de equivalência prática: ROPE

Às vezes podemos estar interessados em determinar as probabilidades de que nossos modelos tenham um desempenho *equivalente*, onde "equivalente" é definido num sentido *prático*. Em [5], salvo que o problema indique outra coisa, sugerem usar como padrão, que dois classificadores são práticamente equivalentes se a média de suas pontuações diferem no máximo em 1%. Assim, nessa situação a região de equivalência prática (ROPE: Region of Practical Equivalence) é o intervalo ROPE $= [-0.01, 0.01]$. Aqui, salvo que o problema indique o contrário, será utilizada essa região de equivalência prática. Assim, para calcular a probabilidade de dois classificadores serem praticamente equivalentes, calculamos a área abaixo da curva da função de densidade de probabilidade da variável t no intervalo ROPE, e por exemplo, se essa probabilidade é maior do que 0.95, podemos considerar que os modelos tem um desempenho praticamente equivalente no Dataset. Por último, se ROPE $= [-a, a]$, com $a > 0$, podemos calcular a probabilidade de que o primeiro modelo tenha *melhor* desempenho que o segundo, calculando a área abaixo da curva da função de densidade de probabilidade da variável t, desde a até ∞, e calcular a probabilidade de que o desempenho do primeiro modelo seja *pior* que o do segundo, calculando a área abaixo da mesma curva desde $-\infty$ até $-a$. Portanto, dado um par de modelos de aprendizagem e dada a região de equivalência prática, ROPE, podemos obter três probabilidades para comparar o par de modelos: a probabilidade de que o modelo 1 seja *pior* do que o modelo 2, a probabilidade de que o modelo 1 seja *melhor* do que o modelo 2, e a probabilidade de que os modelos 1 e 2 sejam práticamente equivalentes. Finalmente, o resultado do teste de significância estatística, serão essas três probabilidades para cada par de modelos usados na comparação.

6.1.2 Classes de dados preditos e métricas de avaliação

No contexto do problema da *classificação binária* da Seção 2.1, o conjunto de rótulos, $\mathcal{Y}$, possui apenas dois elementos ou classes que, independentemente do nome com qual são representadas, sempre podem ser codificadas de forma genérica pelos rótulos $\mathcal{Y} = \{0, 1\}$. Dependendo do problema concreto, uma das classes pode ser mais relevante que a outra para o problema de classificação, nesse caso, a classe mais relevante será

codificada com o rótulo 1 e será chamada de *classe positiva* e a classe com rótulo 0 será chamada de *classe negativa*.

Suponha que $Y_{\text{pred}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n) \in \{0,1\}^n$ é o vetor de rótulos *preditos* por um classificador para n observações $\varsigma_n = (x_1, x_2, \ldots, x_n) \in \Omega^n$, e $Y_{\text{atual}} = (y_1, y_2, \ldots, y_n) \in \{0,1\}^n$ é o vetor de rótulos *verdadeiros* ou *atuais* para os elementos de ς_n. Para avaliar as predições feitas por um classificador, simplesmente comparamos os rótulos de Y_{pred} com os rótulos correspondentes de Y_{atual}, assim, em problemas de classificação binária, um rótulo ou *valor* predito por um classificador pode ser catalogado como:

(i) Verdadeiro positivo (VP): quando o classificador prediz a classe como *positiva* e o valor atual é da classe *positiva*.

(ii) Verdadeiro negativo (VN): quando o classificador prediz a classe como *negativa* e o valor atual é da classe *negativa*.

(iii) Falso positivo (FP): quando o classificador prediz a classe como *positiva* e o valor atual é da classe *negativa*.

(iv) Falso negativo (FN): quando o classificador prediz a classe como *negativa* e o valor atual é da classe *positiva*.

Assim, dados os vetores de $\{0,1\}^n$, Y_{pred} e Y_{atual}, se contamos os membros de cada categoria acima obtemos os seguintes números, que chamaremos com o mesmo nome da categoria que representam:

$$\begin{aligned}
\text{VP} &:= \sharp(\{i \in [n] : \bar{y}_i = 1 \text{ e } y_i = 1\}) \\
\text{VN} &:= \sharp(\{i \in [n] : \bar{y}_i = 0 \text{ e } y_i = 0\}) \\
\text{FP} &:= \sharp(\{i \in [n] : \bar{y}_i = 1 \text{ e } y_i = 0\}) \\
\text{FN} &:= \sharp(\{i \in [n] : \bar{y}_i = 0 \text{ e } y_i = 1\}).
\end{aligned}$$

Para facilitar a análise do desempenho de um classificador e visualizar de melhor maneira esses números, podemos dispor eles numa matriz de (2×2), chamada *matriz de confusão*:

		Valores Preditos: $\bar{y}_i$	
		Negativo	Positivo
Valores Atuais: y_i	Negativo	**VN**	**FP**
	Positivo	**FN**	**VP**

Na prática, não basta apenas contar a quantidade de acertos que o classificador teve para decidir se ele é bom ou não. Dependendo do problema estudado, métricas

diferentes devem ser utilizadas para essa avaliação. Existem muitas métricas de avaliação para classificadores, mas aqui utilizaremos algumas que usam apenas os valores preditos descritos na matriz de confusão.

Agora veremos de *forma breve* as métricas que serão utilizadas para avaliar a qualidade dos classificadores que usaremos na comparação, e no caso do leitor precisar de mais detalhes, uma análise sistemática para *vinte e quatro* métricas de avaliação em problemas de classificação pode ser encontrada em [58].

Acurácia :

É considerada uma das métricas mais simples e importantes. Essa métrica simplesmente avalia o percentual de acertos, ou seja, é obtida pela razão entre a quantidade de acertos e o total de instâncias:

$$\text{acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}}.$$

Precisão :

Essa métrica é a proporção de observações com rótulo *predito* igual a 1 cujo rótulo *atual* seja 1, ou seja, é a proporção de *verdadeiros positivos* sobre o número de valores preditos como positivos pelo classificador:

$$\text{precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}.$$

Valores altos (próximos de 1) nessa métrica de avaliação indicam um baixo número de falsos positivos com relação ao número de verdadeiros positivos.

Sensibilidade (Recall) :

Essa métrica, também conhecida como *recall*, é a proporção de observações com rótulo *atual* igual a 1 cujo rótulo *predito* seja 1, ou seja, é a proporção de *verdadeiros positivos* sobre o número de observações com rótulo atual positivo:

$$\text{recall} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$

Valores altos (próximos de 1) nessa métrica de avaliação indicam um baixo número de falsos negativos.

F_β **- score** :

F_β - score ou simplesmente F_β, é a média harmônica *ponderada* da precisão e do recall usando o fator $\beta > 0$, fator que representa a importância do recall por sobre a precisão na hora da avaliação do classificador, isto é, β é escolhido de tal maneira que a importância considerada para o *recall* seja β vezes a importância da precisão:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precisão} \cdot \text{recall}}{\beta^2 \cdot \text{precisão} + \text{recall}}.$$

Para β, um par de valores comuns são $\beta = 0.5$, quando é mais importante a precisão do que o recall, e $\beta = 2$, quando é mais importante o recall do que a precisão. No caso em que ambas métricas de avaliação sejam igual de importantes para avaliar o desempenho do classificador, usamos $\beta = 1$.

Vejamos um pequeno exemplo.

Exemplo 6.1.1. Considere a matriz de confusão seguinte:

		Valores Preditos: $\bar{y}_i$	
		Negativo	Positivo
Valores Atuais: y_i	Negativo	966266	8974
	Positivo	5386	82980

Com base na matriz de confusão acima, temos:

$$\text{acurácia} = \frac{82980 + 966266}{82980 + 8974 + 5386 + 966266} \approx 0.9864,$$

$$\text{precisão} = \frac{82980}{82980 + 8974} \approx 0.9024,$$

$$\text{recall} = \frac{82980}{82980 + 5386} \approx 0.9390,$$

$$F_1 \approx 2 \cdot \frac{0.9024 \cdot 0.9390}{0.9024 + 0.9390} \approx 0.9203.$$

6.1.3 Regras de aprendizagem usadas na comparação

Agora veremos um breve resumo de cada uma das regras de aprendizagem que utilizaremos nos testes numéricos.

6.1.3.1 O classificador de k vizinhos mais próximos: k-NN

O classificador de k-vizinhos mais próximos, é um algoritmo de classificação baseado em distâncias que funciona da seguinte forma: dada uma amostra rotulada d_n, que chamaremos de *conjunto de treinamento* e uma instância de teste x, o algoritmo primeiro encontra os k vizinhos mais próximos de x no conjunto de treinamento, para logo determinar a classe de x pelo voto majoritário entre esses k vizinhos. Esse algoritmo possui dois hiperparâmetros: o número de vizinhos mais próximos $k \in \mathbb{N}$, e a distância usada para encontrar esse k vizinhos, onde no caso de $\mathbb{R}^m$, a distância mais amplamente usada é a euclidiana. Uma diferença do k-NN com os outros algoritmos de classificação que usaremos aqui, tais como a árvore de decisão, é que no k-NN a etapa de treinamento é caracterizada apenas pelo armazenamento das instâncias, ou seja, no final das contas não

existe treinamento. Assim, a função que será usada para gerar o rótulo é definida na hora de avaliar a instância de teste, por esse motivo, a regra de classificação k-NN é um regra de aprendizagem do tipo local.

Formalmente, seja (Ω, ρ) um espaço métrico e como antes, $d_n \in (\Omega \times \{0,1\})^n$ uma amostra rotulada que é uma instância da variável aleatória $D_n \sim \widetilde{\mu}^n$ com $\widetilde{\mu}$ medida boreliana de probabilidade em $\Omega \times \{0,1\}$. Para $x \in \Omega$, o classificador k-NN ordena de forma crescente os elementos de d_n em função da distância até x, obtendo

$$\{x_{(1)}, x_{(2)}, \ldots, x_{(k)}, \ldots, x_{(n)}\}$$

onde $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \cdots \leq \rho(x, x_{(n)})$, logo considerando apenas a subamostra

$$\mathcal{N}_k(x) = \{x_{(1)}, x_{(2)}, \ldots, x_{(k)}\},$$

a regra k-NN devolve o voto majoritário entre os correspondentes k rótulos

$$\{y_{(1)}, y_{(2)}, \ldots, y_{(k)}\}.$$

O valor de k é tomado sendo um inteiro *ímpar* para evitar os empates. De forma mais geral, a regra k-NN pertence a um tipo de regras *plug-in* que são geradas por famílias de funções $\mathcal{F} = (\eta_n)_{n \in \mathbb{N}}$, definidas por $\eta_n : (\Omega \times \{0,1\})^n \times \Omega \to [0,1]$ e

$$\eta_n(d_n)(x) = \sum_{i=1}^{n} y_i W_{ni}(d_n, x)$$

onde os números $0 \leq W_{ni}(d_n, x) \leq 1$ satisfazendo $\sum_{i=1}^{n} W_{ni}(d_n, x) = 1, \forall d_n, x$, são chamados pesos.

Logo, a regra k-NN é definida por

$$g_n(d_n)(x) = \begin{cases} 1, & \sum_{i=1}^{n} y_i W_{ni}(d_n, x) \geq \frac{1}{2} \\ 0, & \text{caso contrário,} \end{cases}$$

onde os pesos satisfazem

$$W_{ni}(d_n, x) = 1/k \Leftrightarrow x_i \in \{x_{(1)}, x_{(2)}, \ldots, x_{(k)}\},$$

obtemos assim a forma usual da regra:

$$g_n(d_n)(x) = \begin{cases} 1, & \frac{1}{k} \sum_{i=1}^{k} y_{(i)} \geq \frac{1}{2} \\ 0, & \text{caso contrário.} \end{cases}$$

6.1.3.2 Árvore de decisão

O classificador de Árvore de decisão, é uma regra de aprendizagem supervisionada no *espaço euclidiano* que constrói uma árvore de busca binária baseada no conjunto de

treinamento. Aqui vamos mostrar uma versão da árvore de decisão que é a mais usada na prática e é desenvolvida em [13].

Dada uma amostra rotulada, a ideia principal da árvore de decisão é dividir o domínio em dois subconjuntos usando uma função (otimal) de decisão binária baseada no conjunto de treinamento. Para dividir o domínio, é escolhida uma característica (coordenada) de acordo com uma condição de homogeneidade de classe baseada na amostra. Esse processo é repetido para cada um dos subconjutos de maneira recursiva e o algoritmo para de rodar quando a subamostra contida em cada subconjunto contém apenas instâncias da mesma classe ou quando mais divisões não melhorem a condição de homegeneidade de classes. Para cada subconjunto final, é associado um rótulo de classe baseado nos rótulos das instâncias da subamostra contida no subconjunto. Para classificar uma nova observação, ela irá "subindo" na árvore até chegar a um subconjunto final, obtendo como predição o rótulo associado a esse subconjunto.

Desde o ponto de vista geométrico, as divisões feitas pela árvore particionam o espaço euclidiano em hiper-retângulos, (possivelmente finitos ou com lados de comprimento infinito) paralelos aos eixos, baseados no conjunto de treinamento onde cada hiper-retângulo tem associado um rótulo baseado na subamostra contida nele. Assim, a árvore de decisão prediz o rótulo para uma nova instância considerando o hiper-retângulo onde a nova instância pertence.

Agora vamos definir formalmente a regra de classificação dada pela árvore de decisão. Primeiro definimos a noção de homogeneidade de classe em termos da *entropia*, e a otimalidade da decisão binária para a divisão do domínio na coordenada escolhida é definida em termos do *ganho de informação maximal*.

Considere o espaço euclidiano $(\mathbb{R}^d, \|\cdot\|_2)$ e $d_n \in (\mathbb{R}^d \times \{0,1\})^n$ uma amostra rotulada que é uma instância da variável aleatória $D_n \sim \widetilde{\mu}^n$ com $\widetilde{\mu}$ medida boreliana de probabilidade em $\mathbb{R}^d \times \{0,1\}$. Também, como os elementos da amostra d_n possuem coordenadas, $x_i \in \mathbb{R}^d$, vamos denotar esses pontos por:

$$x_i = (x_1^i, x_2^i, \ldots, x_d^i).$$

Definição 6.1.1. Dada uma amostra rotulada $d_n \in (\mathbb{R}^d \times \{0,1\})^n$, definimos a *entropia* H de d_n, por

$$H(d_n) := -(p_0 \log(p_0) + p_1 \log(p_1))$$

onde

$$p_0 = \frac{\sharp(\{(x_i, y_i) \in d_n : y_i = 0\})}{n}$$

e

$$p_1 = \frac{\sharp(\{(x_i, y_i) \in d_n : y_i = 1\})}{n}$$

A árvore de decisão, procura recursivamente dividir o conjunto de treinamento em duas subamostras de tal maneira de que a entropia das duas subamostras seja minimizada

com relação à entropia da amostra completa. O ganho de informação mede a mudança na entropia.

Definição 6.1.2. Seja $d_n \in (\mathbb{R}^d \times \{0,1\})^n$ uma amostra rotulada. Para $j \in [d]$ e $a \in \mathbb{R}$, defina as subamostras

$$d_n^{j,a-} := \{(x_i, y_i) \in d_n : x_j^i \leq a\}$$

e

$$d_n^{j,a+} := \{(x_i, y_i) \in d_n : x_j^i > a\}.$$

Definimos o ganho de informação $G_{j,a}$ para d_n na coordenada j com delimitador a, por

$$G_{j,a}(d_n) := H(d_n) - \left(\frac{\sharp(d_n^{j,a-})}{n} H(d_n^{j,a-}) + \frac{\sharp(d_n^{j,a+})}{n} H(d_n^{j,a+}) \right).$$

Uma vez definidos os conceitos principais, podemos definir os passos dados pela árvore de decisão para os conjuntos de treinamento e teste.

Para o conjunto de treinamento

(1) Fixe um valor limiar $\alpha \geq 0$ sendo o ganho de informação mínimo aceitado.

(2) Calcule a entropia H para o conjunto de treinamento d_n.

(3) De forma exaustiva, determinar a *melhor divisão*:

$$(j^*, a^*) = \arg\max_{(j,a)} G_{j,a}(d_n)$$

(4) Divida $d_n = d_n^{j^*,a^*-} \cup d_n^{j^*,a^*+}$

(5) Repetir os passos (2) até (4) para cada uma das subamostras $d_n^{j^*,a^*-}$ e $d_n^{j^*,a^*+}$ recursivamente até alguma das seguintes condições de parada ser satisfeita:

(i) Ambas subamostras $d_n^{j^*,a^*-}$ e $d_n^{j^*,a^*+}$ contêm apenas instâncias de uma única classe.

(ii) Os ganhos de informação satisfazem $G_{j,a}(d_n^{j^*,a^*-}) \leq \alpha$ e $G_{j,a}(d_n^{j^*,a^*+}) \leq \alpha$, para todo $j \in [d]$ e $a \in \mathbb{R}$.

(6) Após finalizar, o conjunto de treinamento deve ser particionado em $m \in \mathbb{N}$ subconjuntos

$$d_n = d_n^1 \cup d_n^2 \cup \cdots \cup d_n^m$$

com as *melhores divisões*

$$\{(j_1, a_1), (j_2, a_2), \ldots, (j_{m-1}, a_{m-1})\}$$

e associamos a cada subconjunto d_n^i o rótulo dado pelo voto majoritário dos rótulos de seus elementos.

Para o conjunto de teste

Dada uma nova instância, determine com qual subamostra d_n^i a instância está associada seguindo as decisões dadas pelas *melhores divisões*

$$\{(j_1, a_1), (j_2, a_2), \ldots, (j_{m-1}, a_{m-1})\}$$

calculadas no treinamento, retornando o rótulo associado à subamostra d_n^i.

Pela natureza recursiva da árvore de decisão, o procedimento pode ser fácilmente visualizado e modelado por uma árvore de busca binária, por isso o nome. O conjunto de treinamento completo começa no nó raíz que é dividido em dois nós filhos, correspondentes às subamostras do conjunto de treinamento obtidas pelas divisões binárias ótimas. Cada nó filho, por sua vez, é dividido em dois nós e esse processo é feito de forma recursiva até alguma das condições de parada ser satisfeita para todos os nós inferiores. Após a parada, cada nó no nível final torna-se uma *folha* correspondente ao rótulo da classe dominante. Para predizer o rótulo de uma nova instância a árvore de decisão passa a instância através da árvore e a classe predita é a classe associada à folha onde essa instância pertence.

6.1.3.3 Random Forest

A regra de aprendizagem da *Floresta Aleatória*, *Random Forest*, foi desenvolvida por *Leo Breiman* [14], e generaliza a regra da árvore de decisão. A *primeira versão* da regra dada em [14], primeiro usa o método de *bootstrap aggregation* ou *bagging*, para tomar m subamostras de tamanho n, com substituição[2], do conjunto de treinamento $d_n \in (\mathbb{R}^d \times \{0,1\})^n$ e constroi uma árvore de decisão para cada subamostra:

$$\begin{aligned}
d_n^1 &= \left\{(x_1^1, y_1^1), (x_2^1, y_2^1), \ldots, (x_n^1, y_n^1)\right\} \\
d_n^2 &= \left\{(x_1^2, y_1^2), (x_2^2, y_2^2), \ldots, (x_n^2, y_n^2)\right\} \\
&\vdots \\
d_n^m &= \left\{(x_1^m, y_1^m), (x_2^m, y_2^m), \ldots, (x_n^m, y_n^m)\right\},
\end{aligned}$$

árvores que denotamos por $A_1, A_2, \ldots, A_m$. Depois, para a construção de cada árvore de decisão, em cada nó ou iteração, utiliza *apenas* $1 \leq w \leq d$ coordenadas selecionadas aleatoriamente, com w hiperparâmetro de entrada fixo, para encontrar a melhor divisão do conjunto de treinamento do passo anterior. Em outras palavras, com relação ao passo (3) do processo de construção de uma árvore de decisão, um subconjunto $C \subset [d]$ com w elementos escolhidos de forma aleatória e a melhor divisão satisfaz $j \in C$ e $a \in \mathbb{R}$. Para uma nova instância $x \in \mathbb{R}^d$, cada árvore de decisão faz a sua predição, $A_i(x)$, e a floresta

[2] Versões posteriores utilizam subamostras de tamanho diferente de n.

aleatória retorna o voto majoritário dos rótulos $A_i(x)$, $i \in [m]$. Na etapa de treinamento, a floresta aleatória depende de dois hiper-parâmetros fornecidos pelo usuário: m que é o número de estimadores ou de árvores na floresta e w, que é o número de coordenadas que serão escolhidas aleatoriamente para construir cada árvore. Na prática, a floresta aleatória tem um bom poder preditivo e em geral seu desempenho é bom em muitas áreas e aplicações. Por outro lado, desde o ponto de vista teórico, o poder preditivo da floresta aleatória não está justificado. De fato, em [7] é mostrado que a floresta aleatória não é universalmente consistente. Esse fato aparentemente contraditório, ilustra uma distinção comum entre teoria e prática para regras de aprendizagem supervisionada. Com frequência, uma regra de aprendizagem com propriedades teóricas comprovadas, como ser universalmente consistente, pode não produzir previsões precisas no cotidiano ou sua complexidade computacional pode ser muito alta na prática. Por outro lado, uma regra de aprendizagem simples e intuitiva, sem fatos teóricos que garantam sua confiabilidade, pode ter excelente poder preditivo e pode ser computacionalmente eficiente.

6.1.3.4 Support Vector Machine: SVM

As *máquinas de vetores de suporte* (Support Vector Machine ou *SVM*) são modelos de aprendizagem de máquina supervisionada para problemas de classificação e regressão, desenvolvidos nos laboratórios *AT&T Bell* por uma equipe liderada por *Vladimir Vapnik* [21]. Para problemas de classificação binária, um modelo SVM gera um classificador *não probabilístico*, utilizando a informação contida em uma amostra rotulada, construindo um *hiperplano* no espaço das instâncias que pode ser usado para separar as classes. Intuitivamente, se as classes podem ser separadas linearmente, uma boa separação entre classes é alcançada pelo hiperplano que possui a *maior distância* até os pontos (de qualquer classe) *mais próximos* do conjunto de treinamento; pois em geral, quanto maior seja essa distância, menor será o erro de generalização do classificador associado. A distância de separação entre classes define a *margem funcional*. A Figura 5, mostra um exemplo em $\mathbb{R}^2$ de um problema linearmente separável e o hiperplano que separa ambas classes, onde os três pontos nos limites da *margem funcional*, são chamados de *vetores de suporte*. Nesse caso, o classificador resultante, classifica uma nova amostra segundo o lado do hiperplano separador ao qual ela pertence.

No caso em que as classes não sejam linearmente separáveis, os modelos SVM podem obter bons resultados usando *Kernels não lineares*. Detalhes em [10].

6.1.3.4.1 Linear SVM e LSVC

Seja $d_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^d \times \{-1, 1\})^n$, uma amostra rotulada[3] que usaremos como conjunto de treinamento. O objetivo aqui é encontrar o *hiperplano de*

[3] Aqui por conveniência, as classes são codificadas por -1 e 1.

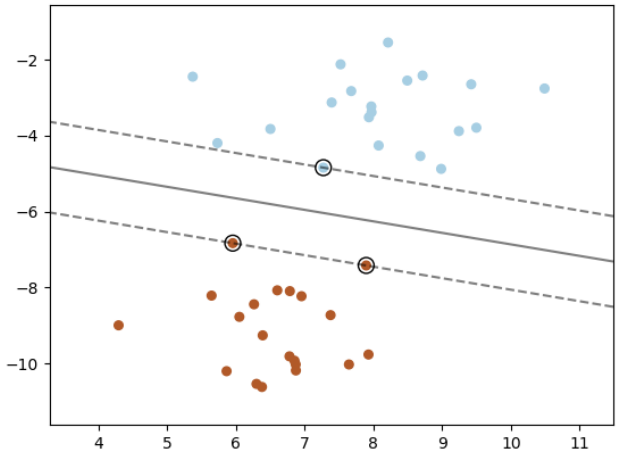

Figura 5 – Exemplo dados linearmente separáveis e margem funcional

margem máxima que separa o conjunto de pontos rotulados com $y = 1$ dos pontos rotulados com $y = -1$, que é definido de tal maneira que a distância entre o hiperplano e o ponto x_i mais próximo de cada classe, seja maximizada.

Um *hiperplano* é um conjunto de pontos $x \in \mathbb{R}^d$ satisfazendo uma equação do tipo:

$$w \cdot x + b = 0,$$

onde $w \in \mathbb{R}^d$ é o *vetor normal* ao hiperplano, $b \in \mathbb{R}$ e $\cdot$ é o *produto escalar* entre vetores de $\mathbb{R}^d$. Distinguimos dois casos.

Dados linearmente separáveis: LSVM

Nesse caso, podemos encontrar dois hiperplanos paralelos que separam as duas classes de dados, tal que a distância entre eles seja a maior possível. A região limitada por esses dois hiperplanos é chamada de *margem* e o hiperplano de *margem máxima* é o hiperplano paralelo que fica a meio caminho entre eles. Esses hiperplanos podem ser descritos pelo seguinte par de equações:

$$w \cdot x + b = 1 \text{ e } w \cdot x + b = -1.$$

Desde o ponto de vista geométrico, a distância entre esses hiperplanos paralelos é obtida usando a distância de um ponto a um hiperplano, e é igual a $\frac{2}{\|w\|_2}$, logo, para maximizar a distância entre os hiperplanos, devemos *minimizar* $\|w\|_2$. Também, devemos evitar que os dados *caiam* dentro da margem, para isso adicionamos as seguintes restrições:

$$w \cdot x_i + b \geq 1, \text{ se } y_i = 1$$

ou

$$w \cdot x_i + b \leq -1, \text{ se } y_i = -1,$$

as quais garantem que cada instância do conjunto de treinamento deve estar no lado correto da margem e que podem ser reescritas como:

$$y_i(w \cdot x_i + b) \geq 1, \forall i \in [n]. \tag{20}$$

Assim, podemos reunir as condições acima no seguinte problema de *otimização*:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.a.} \quad & y_i(w \cdot x_i + b) \geq 1, \forall i \in [n]. \end{aligned} \tag{21}$$

As soluções $w \in \mathbb{R}^d$ e $b \in \mathbb{R}$ do problema acima, determinam um classificador mediante a aplicação $x \mapsto \text{sgn}(w \cdot x + b)$, onde sgn, é a função *sinal*, que devolve 1 se o argumento é um número real positivo, 0 se o argumento é 0 e -1 caso o argumento é negativo; e esse classificador será chamado **LSVM**.

Uma consequência da descrição geométrica é que o hiperplano de *margem máxima* é completamente determinado por aquelas instâncias x_i que estão mais próximos dele. Essas instâncias x_i, são chamados de *vetores de suporte*.

Dados não linearmente separáveis: LSVC

Para estender o modelo *LSVM* para os casos onde os dados *não são* linearmente separáveis, considere a seguinte função:

$$\max\{0, 1 - y_i(w \cdot x_i + b)\}. \tag{22}$$

A função acima é zero se as restrições (20) são satisfeitas, isto é, se cada x_i do conjunto de treinamento está do lado correto da margem; e para uma instância que está do lado errado, o valor da função é *estritamente positivo*. O objetivo aqui é resolver:

$$\min_{w,b} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i(w \cdot x_i + b)\},$$

onde $C > 0$, é o hiperparâmetro de *regularização* que permite abordar problemas de classificação que não são linearmente separáveis. Podemos escrever o problema de otimização acima de forma similar ao problema (21), "*abrindo*" a função (22), obtendo assim o problema de otimização:

$$\begin{aligned} \min_{w,b,(\xi_1,\ldots,\xi_n)} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.a.} \quad & y_i(w \cdot x_i + b)) \geq 1 - \xi_i,\ \xi_i \geq 0, \forall i \in [n]. \end{aligned} \tag{23}$$

Do mesmo modo que no caso anterior, as soluções $w \in \mathbb{R}^d$ e $b \in \mathbb{R}$ do problema acima determinam um classificador *binário* mediante a aplicação $x \mapsto \text{sgn}(w \cdot x + b)$; e esse classificador será chamado **LSVC**.

Assim, para valores grandes do hiperparâmetro de regularização $C > 0$, o modelo LSVC tem um comportamento similar ao modelo LSVM, no caso de ser possível separar linearmente as classes, e ainda classificará mesmo que a separação linear não seja possível. Mais informações em [10, 21].

6.1.3.5 Regressão Logística

A *Regressão Logística* é uma extensão da Regressão Linear $Y = w \cdot X + b$, no caso em que a variável $Y \in \{0, 1\}$ possui uma distribuição Binomial/Bernoulli. Seja $d_n = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathbb{R}^d \times \{0, 1\})^n$, uma amostra rotulada que usaremos como conjunto de treinamento. O objetivo aqui é predizer a probabilidade da classe positiva $P[Y_i = 1 | X_i = x_i]$ mediante a estimativa

$$p(x_i) = \frac{1}{1 + e^{-(w \cdot x_i + b)}},$$

onde $w \in \mathbb{R}^d$ e $b \in \mathbb{R}$ são soluções do problema de otimização

$$\min_{w,b} C \sum_{i=1}^{n} (-y_i \ln(p(x_i)) - (1 - y_i) \ln(1 - p(x_i))) + r(w),$$

com $r(w)$ sendo o termo de regularização, e a constante $C > 0$ um hiperparâmetro de regularização similar ao definido para o classificador LSVC. Na implementação da Regressão Logística da biblioteca `sklearn` de Python, os termos de regularização disponíveis são $r(w) = \|w\|_1$, $r(w) = \frac{1}{2}\|w\|_2^2$ e $r(w) = \rho\|w\|_1 + \frac{(1-\rho)}{2}\|w\|_2^2$, com $\rho \in [0, 1]$ um hiperparâmetro escolhido pelo usuário. Uma vez determinados os parâmetros w, b, o rótulo predito pelo modelo de Regressão Logística é dado por:

$$g(x) = \begin{cases} 1, & p(x) \geq \frac{1}{2} \\ 0, & \text{outro caso.} \end{cases}$$

6.1.4 Escolhendo a métrica adequada

Ao usar aprendizagem de máquina supervisionada na tomada de decisões em problemas da vida real, desejamos construir classificadores que acertem o máximo possível. Uma maneira simples de medir a qualidade do desempenho de um algoritmo de classificação, é usando a *acurácia*. A acurácia pode ser considerada uma métrica que nos dá uma visão geral do desempenho, uma vez que ela mede o total de acertos considerando o total de observações. Entretanto, outras métricas podem ser importantes dependendo de como o problema foi modelado.

A métrica que utilizaremos para avaliar o desempenho, depende do comportamento que desejamos que possua o classificador que vamos construir; isto é, depende do comportamento do classificador que consideramos *mais benéfico* para nossos propósitos ou que gere *menos prejuízos*. Ilustramos essas ideias com um par de exemplos.

Exemplo 6.1.2. Suponha que um algoritmo faz a detecção automática de *spam* numa conta de correio eletrônico. Nesse caso, um *falso positivo* pode ser considerado um problema *mais crítico* (que uma mensagem importante seja considerada spam, pode causar prejuízos ao usuário da conta) do que um *falso negativo*. Logo, a melhor métrica para comparação entre diferentes métodos de detecção de spam, além da acurácia, seria a *precisão*.

Exemplo 6.1.3. Considere um algoritmo que *detecta falhas* em um avião comercial. Se uma peça apresenta problemas mas o algoritmo indica que ela funciona corretamente, estaremos em presença de um *falso negativo*. Isso colocaria as vidas dos passageiros e da tripulação em perigo e nessa situação um *falso negativo* seria um problema crítico. Portanto, na hora de construir um algoritmo para detectar falhas no avião devemos tentar minimizar os *falsos negativos*. Uma métrica de avaliação que poderia ser utilizada para comparar algoritmos diferentes seria o *recall*. Valores altos dessa métrica indicam altos valores de verdadeiros positivos mesmo quando se leva em conta o total de falsos negativos. Nesse caso, se na hora de treinar o algoritmo o conjunto de dados possui um número reduzido de instâncias da classe *positiva*, então a acurácia não é uma boa métrica de desempenho, pois por exemplo, o classificador que a cada instância associa a classe *negativa* terá uma alta acurácia sem nenhum poder preditivo, assim, aqui é mais adequado o uso do *recall* como métrica de avaliação do desempenho.

Na próxima seção de resultados, vamos comparar um conjunto de algoritmos de classificação ao longo de vários conjuntos de dados. Para cada conjunto, é feita uma análise básica e a métrica utilizada para obter os melhores representantes de cada algoritmo será indicada pela *natureza* do próprio problema que pretende ser abordado utilizando o conjunto de dados. Pelo visto nos exemplos acima, para poder escolher a métrica de avaliação adequada, devemos identificar que tipo de *erro* cometido pelo classificador é mais *grave* e também devemos levar em conta a *proporção de classes* na amostra rotulada.

Com relação à proporção de classes, em um problema de classificação binária, dada uma amostra rotulada de tamanho $n \in \mathbb{N}$, d_n, defina $N_m \in \mathbb{N}$, como o número de instâncias da classe minoritária na amostra e $N_M \in \mathbb{N}$, sendo o número de exemplos da classe majoritária, logo, $N_m + N_M = n$. Se $N_m/N_M \approx 1$, então dizemos que a amostra rotulada é *balanceada* ou *equilibrada*, e será uma amostra *desbalanceada*, no caso contrário. Quando os classificadores se enfrentam a uma amostra desbalanceada, tendem a classificar melhor a classe majoritária, mas na prática (ver [15, Capítulo 2]) esse efeito é mais pronunciado para amostras onde $N_m/N_M < 0.1$, ou dito de outro modo, onde a classe minoritária representa uma porcentagem inferior ao 10% da classe majoritária.

Para melhorar o desempenho dos classificadores na hora de identificar a classe minoritária em amostras fortemente desbalanceadas ($N_m/N_M < 0.1$), existem algumas técnicas de *balanceamento de classes* (ver [15, Capítulo 4]), onde são repetidos alguns exemplos da classe minoritária para equilibrar a proporção de classes, ou são gerados

novos exemplos mediante técnicas de interpolação; assim, o uso de essas técnicas implica manipular de maneira artificial o conjunto de treinamento para obter uma maior representação da classe minoritária e elas não resolvem o problema de fundo que é a escassa disponibilidade de exemplos de uma das classes.

Portanto, na Seção 6.2 de resultados, para poupar esforço computacional e devido a que o foco é o *desempenho relativo* entre os classificadores, só será aplicada alguma técnica de balanceamento de classes em amostras tais que $N_m/N_M < 0.1$, quando o desempenho da *maioria* dos classificadores utilizados na comparação tenha uma "*melhora significativa*" (superior ao 1%) na métrica de avaliação considerada; onde essa situação será analisada caso a caso.

Finalmente, pelo exposto na Seção 6.1.2 e seguindo as ideias de [15, 27, 33], para evitar casos *patológicos* na hora de determinar os hiperparâmetros ótimos, como por exemplo, escolher classificadores com recall = 1 e precisão ≈ 0; temos a seguinte *guia* para definir a métrica de avaliação adequada na hora de encontrar os hiperparâmetros ótimos para cada modelo:

Guia para definir a métrica adequada				
Classes de igual importância		**Classe positiva mais importante**		
$N_m/N_M \geq 0.1$	$N_m/N_M < 0.1$	FP mais prejuízo	FP $\equiv$ FN	FN mais prejuízo
acurácia	**G-mean**	$\boldsymbol{F_{0.5}}$	$\boldsymbol{F_1}$	$\boldsymbol{F_2}$

Tabela 1 – Guia para determinar a melhor métrica de avaliação [15].

onde, por exemplo, "FP mais prejuízo" significa que o prejuízo causado pelos falsos positivos é maior que o prejuízo gerado pelos falsos negativos, e a métrica **G-mean**, é a média geométrica do recall e a *especificidade*: **G-mean** $= \sqrt{\text{recall} \cdot \text{especificidade}}$, com, especificidade $= \frac{\text{VN}}{\text{VN+FP}}$. Nos conjuntos de dados considerados na seguinte seção, a classe alvo ou classe positiva será sempre mais importante que a negativa, portanto a métrica de avaliação **G-mean** não será utilizada aqui.

6.2 RESULTADOS

Nessa seção, vamos aplicar a estratégia de comparação à família formada pelos modelos clássicos mais a regra $\mathcal{L}_k^{\Phi_p}$, ao longo de 8 conjuntos de dados disponíveis de forma gratuita na web.

A implementação do classificador $\mathcal{L}_k^{\Phi_p}$ é feita na linguagem de programação *Python 3* e as implementações dos modelos clássicos utilizados na comparação é a fornecida pela biblioteca `sklearn`[4] do Python. Cada modelo possui um conjunto de hiperparâmetros, alguns dos quais serão considerados para a otimização e os restantes serão utilizados com seus valores padrão. Uma lista com as bibliotecas adicionais necessárias, com as funções

[4] Para maiores informações, visite o site https://scikit-learn.org.

específicas da biblioteca `sklearn` que implementam os algoritmos, junto com os conjuntos de hiperparâmetros considerados por cada modelo em cada conjunto de dados, pode ser encontrada no Apêndice A.2.

Lamentavelmente, no transcurso da pesquisa não tivemos acesso a um computador de boas características técnicas, o modesto computador disponível (veja as modestas características técnicas da máquina utilizada na Seção 6.2.9) foi utilizado principalmente para desenvolver o protótipo do algoritmo $\mathcal{L}_k^{\Phi_p}$, portanto, todos os cálculos mostrados aqui (salvo a Seção 6.2.9), foram realizados na nuvem utilizando uma conta gratuita na plataforma *Kaggle*[5], utilizando ao máximo os recursos computacionais mediante a ativação das opções de cálculo paralelo nas implementações dos algoritmos da biblioteca `sklearn`. Por outro lado, a implementação do algoritmo $\mathcal{L}_k^{\Phi_p}$ é apenas um protótipo sem praticamente nenhum tipo de otimização no código, por essa razão, na análise feita para cada conjunto de dados, não são considerados os tempos de cálculo como ponto de comparação. Uma breve análise comparativa com relação aos tempos de cálculo num conjunto de dados fixo e em condições *justas* para todos os modelos (tempos de cálculo que foram obtidos utilizando o nosso fiel e modesto colaborador), é feita na Seção 6.2.9. No resto do capítulo, denotaremos abreviadamente o nome dos modelos utilizados na comparação como na seguinte tabela.

Nomes abreviados	
DT	Árvore de Decisão
LR	Regressão Logística
LSVC	Máquinas de Vectores de Suporte
RF	Floresta Aleatória

Os códigos da implementação do classificador $\mathcal{L}_k^{\Phi_p}$ e da estrutura que permite executar a estratégia de comparação, podem ser encontradas no repositório:

`https://github.com/rmathdrum/p-adic-prototype`

6.2.1 Dataset 1: Dry Bean

Um sistema de visão computacional foi desenvolvido para distinguir sete diferentes variedades registradas de feijões secos com características semelhantes, a fim de obter uma classificação uniforme das sementes. Para o modelo de classificação, imagens de 13611 grãos de 7 diferentes tipos de feijões secos registrados foram tiradas com uma câmera de alta resolução. Imagens de feijões obtidas por sistema de visão computacional foram submetidas às etapas de segmentação e extração de características, totalizando 16 características; 12 dimensões e 4 variáveis de forma, foram obtidas a partir dos grãos.

[5] Para mais informações, visite o site: `https://www.kaggle.com`.

O conjunto de dados completo e a descrição acima, foram obtidos do site *UCI Machine Learning Repository*[6].

1. Pré-processamento de dados

A Tabela 2, resume as informações básicas do conjunto de dados, sem o vetor de rótulos, que vamos utilizar no que segue. Como o dataset não possui dados *não válidos*, podemos utilizar ele sem a necessidade de fazer pré-processamentos de limpeza.

Informações básicas	
Nº de instâncias	13611
Dimensão dos dados	16
Possui dados não válidos?	Não
Tipo de dados	`float64`

Tabela 2 – Informações básicas do Dataset 1

Agora com relação ao vetor de rótulos, as classes e suas porcentagens dentro da amostra, são mostrados na Tabela 3.

Classe	Porcentagem
`DERMASON`	26.05
`SIRA`	19.37
`SEKER`	14.89
`HOROZ`	14.17
`CALI`	11.98
`BARBUNYA`	9.71
`BOMBAY`	3.84

Tabela 3 – Multiclasses Dataset 1

Para reduzir o problema para um de classificação binária, e para facilitar o trabalho dos classificadores, vamos *escolher* uma das classes mais numerosas para identificar, assim escolhendo `SEKER` como classe alvo, codificamos o vetor de rótulos associando o rótulo 1 à classe `SEKER` e associando o rótulo 0 para as classes restantes. Dessa forma, o nosso *novo* vetor de rótulos, possui a distribuição de classes da Figura 6.

O número de observações da classe minoritária é aproximadamente o 17.5% do número de observações da classe majoritária ($N_m/N_M \approx 0.1749$), portanto, segundo a convenção da Seção 6.1.4, o conjunto de dados assim conformado é desbalanceado mas não é o suficiente para considerar o uso de alguma técnica de balance de classes.

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

[6] O conjunto de dados pode ser baixado desde o seguinte link: https://doi.org/10.24432/C50S4B.

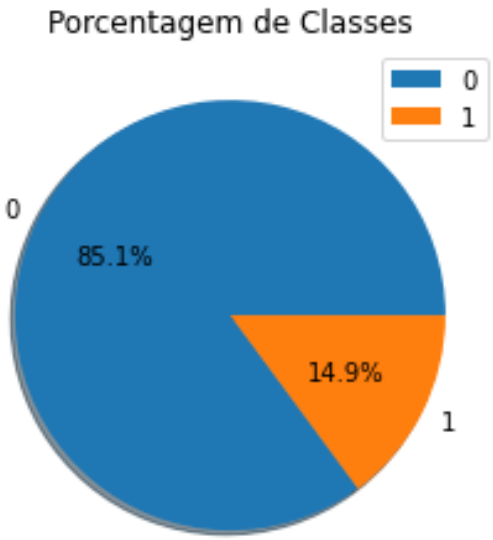

Figura 6 – Classes Dataset 1

Número de instâncias	
Conjunto de treinamento	9074
Conjunto de teste	4537

2. Otimização de hiperparâmetros

Agora que temos nosso conjunto de dados pronto, precisamos determinar a métrica de avaliação adequada para determinar os hiperparâmetros ótimos. Pelo discutido na Seção 6.1.4, a métrica de avaliação adequada aqui é a métrica $\boldsymbol{F_1}$, pois a classe positiva é mais importante que a classe negativa; e como não dispomos de informação adicional sobre qual tipo de erro provoca maior *prejuízo*, vamos supor que os falsos positivos e falsos negativos geram prejuízos equivalentes.

Assim, fazendo 5×20 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de avaliação F_1) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 4.

#	Modelo	$\boldsymbol{hp_1}$	$\boldsymbol{hp_2}$	$\boldsymbol{hp_3}$	$\boldsymbol{F_1}$	**acurácia**	**precisão**	**recall**
1	LR	$C = 10^3$	l_2	liblinear	**0.9485**	**0.9849**	0.9599	**0.9379**
2	RF	entropy	$H = 12$	$n_e = 100$	0.9482	0.9848	0.9600	0.9371
3	k-NN	$k = 8$	-	-	0.9441	0.9837	**0.9645**	0.9250
4	LSVC	$C = 10$	l_2	$i_s = 1$	0.9400	0.9824	0.9560	0.9245
5	DT	entropy	$H = 8$	best	0.9352	0.9809	0.9449	0.9263
6	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 1$	$H = 4$	0.9173	0.9756	0.9244	0.911

Tabela 4 – Melhores representantes de cada modelo no Dataset 1

As pontuações médias, com relação à métrica F_1, dos modelos mais promissores no conjunto de treinamento; *sugerem* que os modelos LR, RF, k-NN, LSVC e DT, podem ter desempenho equivalente e superior ao desempenho do modelo $\mathcal{L}_k^{\Phi_p}$.

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos e testamos os melhores representantes de cada modelo nos conjuntos de treinamento e teste obtidos no final do passo 1, obtendo assim os resultados da Tabela 5.

#	Modelo	hp_1	hp_2	hp_3	F_1	acurácia	precisão	recall
1	LR	$C = 10^3$	l_2	liblinear	**0.9568**	**0.9872**	**0.9626**	**0.9512**
2	RF	entropy	$H = 12$	$n_e = 100$	0.9512	0.9855	0.9512	**0.9512**
3	k-NN	$k = 8$	-	-	0.9451	0.9839	0.9617	0.9290
4	LSVC	$C = 10$	l_2	$i_s = 1$	0.9443	0.9837	0.9617	0.9275
5	DT	entropy	$H = 8$	best	0.9383	0.9817	0.9432	0.9334
6	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 1$	$H = 4$	0.9381	0.9815	0.9340	0.9423

Tabela 5 – Avaliação final dos representantes de cada modelo no Dataset 1

Os resultados da Tabela 5, sugerem quase o mesmo comportamento, só que agora o desempenho do modelo $\mathcal{L}_k^{\Phi_p}$ dessa vez fica mais próximo do desempenho dos outros modelos.

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 7 e a Tabela 6 com os resultados do *teste de significância* para cada par de modelos.

5. Conclusões finais para o Dataset 1

Da Figura 7 e da Tabela 6, vemos que o modelo $\mathcal{L}_k^{\Phi_p}$ tem uma alta probabilidade, em qualquer dos casos é maior do que 0.93, de ter um desempenho pior do que os modelos LR, RF, LSVC e k-NN, e uma probabilidade de 0.7076 de ter desempenho pior do que o DT. Por outro lado, os modelos, RF e LR tem uma probabilidade superior a 0.98 de ter desempenho equivalente, e os modelos LSVC e k-NN, possuem uma probabilidade de equivalência prática maior que 0.97.

6.2.2 Dataset 2: HTRU2

O conjunto de dados *HTRU2*, descreve uma amostra de candidatos a *pulsar* coletados durante o *High Time Resolution Universe Survey* [44]. Os pulsares são um tipo

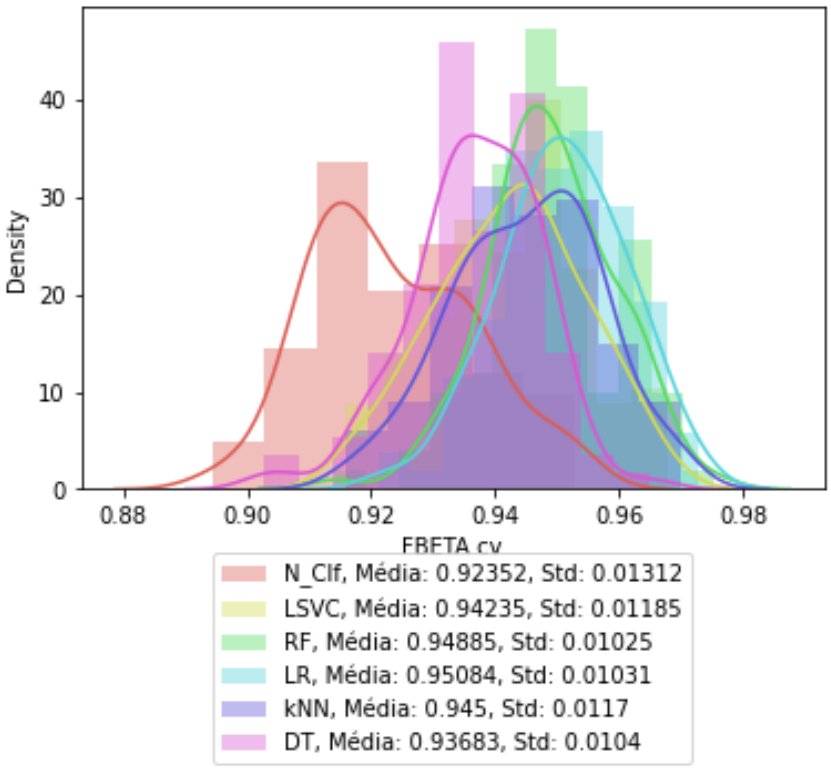

Figura 7 – Distribuição CV Dataset 1

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	LSVC	0.9382	0.0	0.0618
$\mathcal{L}_k^{\Phi_p}$	RF	0.9946	0.0	0.0054
$\mathcal{L}_k^{\Phi_p}$	LR	0.9990	0.0	0.0010
$\mathcal{L}_k^{\Phi_p}$	k-NN	0.9872	0.0	0.0128
$\mathcal{L}_k^{\Phi_p}$	DT	0.7076	0.0001	0.2923
LSVC	RF	0.2112	0.0001	0.7887
LSVC	LR	0.3154	0.0	0.6846
LSVC	k-NN	0.0271	0.0006	0.97242
LSVC	DT	0.0005	0.1651	0.8344
RF	LR	0.0124	0.0005	0.9872
RF	k-NN	0.0005	0.0684	0.9311
RF	DT	0.0	0.6765	0.3235
LR	k-NN	0.0	0.1230	0.8770
LR	DT	0.0	0.8234	0.1766
k-NN	DT	0.0001	0.3524	0.6475

Tabela 6 – Probabilidades Dataset 1

raro de estrela de nêutrons que a medida que rotam produzem padrões de emissões de rádio detectáveis aqui na Terra. Cada pulsar produz um padrão de emissão ligeiramente diferente, que varia ligeiramente com cada rotação. Assim, uma detecção de sinal potencial conhecida como "*candidato*" é calculada em média ao longo de muitas rotações do pulsar,

conforme determinado pelo comprimento de uma observação. Na ausência de informações adicionais, cada candidato poderia descrever um pulsar real. No entanto, na prática, quase todas as detecções são causadas por interferência de radiofrequência (RFI) e ruído, dificultando a localização de sinais legítimos. Aqui, os exemplos legítimos de pulsar são uma classe positiva minoritária, e os exemplos espúrios, a classe negativa majoritária, onde todos esses exemplos foram verificados por anotadores humanos.

O conjunto de dados completo e a descrição acima, foram obtidos do site *UCI Machine Learning Repository*[7], e mais detalhes sobre o conjunto de dados podem ser encontrados em [43, 44].

1. Pré-processamento de dados

A Tabela 7, resume as informações básicas do conjunto de dados, sem o vetor de rótulos, que vamos utilizar no que segue. Como o dataset não possui dados *não válidos*, podemos utilizar ele sem a necessidade de fazer pré-processamentos de limpeza.

Informações básicas	
Nº de instâncias	17898
Dimensão dos dados	8
Possui dados não válidos?	Não
Tipo de dados	`float64`

Tabela 7 – Informações básicas do Dataset 2

Agora com relação ao vetor de rótulos, a classe positiva com rótulo 1 representa um pulsar e a classe negativa com rótulo 0 representa ruído (RFI) onde a distribuição de classes é mostrada na Figura 8.

O número de observações da classe minoritária é aproximadamente o 10.08% do número de observações da classe majoritária ($N_m/N_M \approx 0.1008$), portanto, segundo a convenção da Seção 6.1.4, no conjunto de dados assim conformado não serão aplicadas técnicas de balance de classes.

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

Número de instâncias	
Conjunto de treinamento	11932
Conjunto de teste	5966

[7] O conjunto de dados pode ser baixado desde o seguinte link: https://doi.org/10.24432/C5DK6R.

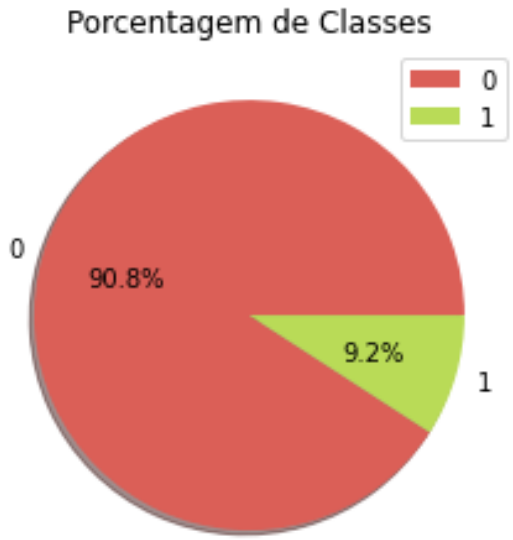

Figura 8 – Classes Dataset 2

2. Otimização de hiperparâmetros

Agora que temos nosso conjunto de dados pronto, precisamos determinar a métrica de avaliação adequada para determinar os hiperparâmetros ótimos. Pelo discutido na Seção 6.1.4, a métrica de avaliação adequada aqui é a métrica $\boldsymbol{F_{0.5}}$, pois a classe positiva é mais importante que a classe negativa; e o *prejuízo* gerado pelos falsos positivos é maior que o prejuízo gerado pelos falsos negativos, pois uma instância classificada como pulsar deve passar por verificações posteriores que requerem de trabalho adicional para verificar se realmente são, portanto classificar uma (RFI) como pulsar (falso positivo) gera *mais prejuízo* que classificar um pulsar como (RFI) (falso negativo).

Assim, fazendo 5×20 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de avaliação $F_{0.5}$) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 8.

#	Modelo	$\boldsymbol{hp_1}$	$\boldsymbol{hp_2}$	$\boldsymbol{hp_3}$	$\boldsymbol{F_{0.5}}$	acurácia	precisão	recall
1	RF	gini	$H = 5$	$n_e = 500$	**0.9103**	**0.9783**	0.9371	0.8177
2	k-NN	$k = 30$	-	-	0.9102	0.9767	0.9460	0.7913
3	LR	$C = 10^3$	l_2	liblinear	0.9085	0.9782	0.9340	0.8198
4	LSVC	$C = 10^2$	l_2	$i_s = 1$	0.9077	0.9769	0.9402	0.7987
5	$\mathcal{L}_k^{\Phi_p}$	$p = 3$	$k = 1$	$H = 2$	0.9049	0.9738	**0.9544**	0.7502
6	DT	entropy	$H = 4$	best	0.8993	0.9772	0.9213	**0.8228**

Tabela 8 – Melhores representantes de cada modelo no Dataset 2

As pontuações médias, com relação à métrica $F_{0.5}$, dos modelos mais promissores no conjunto de treinamento; *sugerem* uma leve superioridade dos modelos RF e k-NN, os quais

podem ser de desempenho equivalente, sobre os restantes modelos. As pontuações médias também sugerem que o desempenho dos modelos LR, LSVC e $\mathcal{L}_k^{\Phi_p}$ pode ser equivalente e levemente melhor do que o desempenho do modelo DT.

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos e testamos os melhores representantes de cada modelo nos conjuntos de treinamento e teste obtidos no final do passo 1, obtendo assim os resultados da Tabela 9.

#	Modelo	hp_1	hp_2	hp_3	$F_{0.5}$	acurácia	precisão	recall
1	LSVC	$C = 10^2$	l_2	$i_s = 1$	**0.9227**	0.9794	0.9548	0.8132
2	RF	gini	$H = 5$	$n_e = 500$	0.9218	0.9797	0.9512	0.8205
3	LR	$C = 10^3$	l_2	liblinear	0.9209	**0.9801**	0.9476	**0.8278**
4	$\mathcal{L}_k^{\Phi_p}$	$p = 3$	$k = 1$	$H = 2$	0.9150	0.9762	**0.9591**	0.7729
5	k-NN	$k = 30$	-	-	0.9128	0.9767	0.9512	0.7857
6	DT	entropy	$H = 4$	best	0.9109	0.9787	0.9356	0.8242

Tabela 9 – Avaliação final dos representantes de cada modelo no Dataset 2

Os resultados da Tabela 9, parecem *discordar* com os resultados médios no conjunto de treinamento da Tabela 8, agora a situação parece ser diferente: os dados *sugerem* que o desempenho dos modelos LSVC, RF e LR, pode ser equivalente e levemente melhor do que o desempenho dos modelos $\mathcal{L}_k^{\Phi_p}$, k-NN e DT, modelos que por sua vez também podem ter desempenho equivalente.

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 9 e a Tabela 10 com os resultados do *teste de significância* para cada par de modelos.

5. Conclusões finais para o Dataset 2

Da Figura 9 e da Tabela 10, vemos que o comportamento aparentemente contraditório dos passos 2 e 3 está dentro da variabilidade nas pontuações que podemos esperar em modelos que possuem uma alta probabilidade de ser *praticamente equivalentes*. Nesse conjunto de dados, vemos que as maiores probabilidades de ser equivalentes são para os modelos LSVC e LR com probabilidade 1.0, RF e k-NN com probabilidade 0.9835 e o par RF, LR com probabilidade 0.9872. Por último, com relação ao novo modelo, $\mathcal{L}_k^{\Phi_p}$ e o k-NN tem uma probabilidade 0.9661 de ser praticamente equivalentes e em geral, a probabilidade de ser equivalente com o modelo $\mathcal{L}_k^{\Phi_p}$ é superior a 0.92, salvo no caso do modelo DT, pois aí a probabilidade chega apenas a 0.6490.

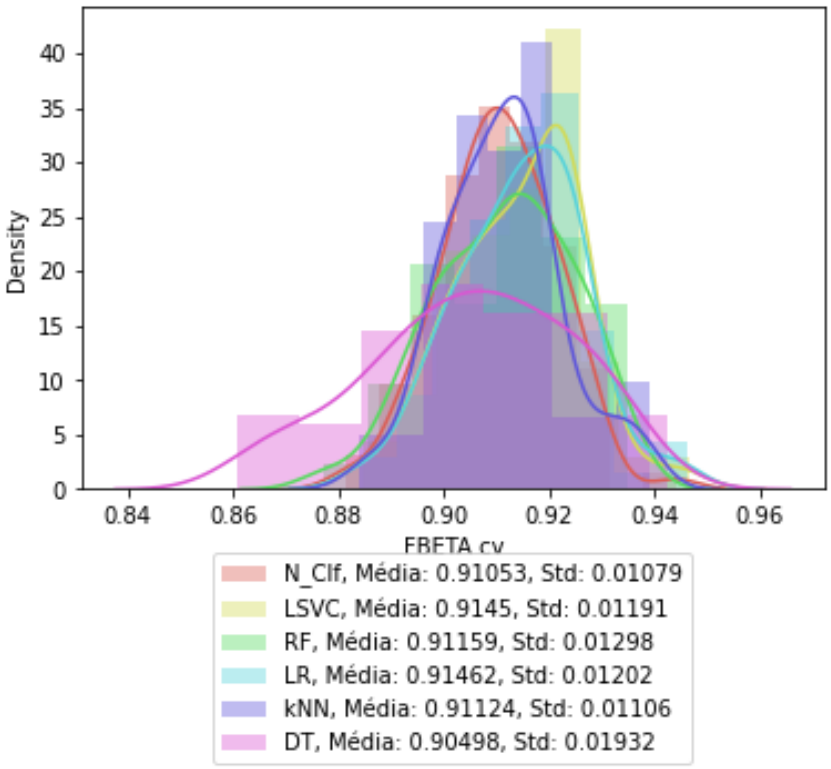

Figura 9 – Distribuição CV Dataset 2

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	LSVC	0.0585	0.0002	0.9413
$\mathcal{L}_k^{\Phi_p}$	RF	0.0367	0.0137	0.9497
$\mathcal{L}_k^{\Phi_p}$	LR	0.0759	0.0004	0.9237
$\mathcal{L}_k^{\Phi_p}$	k-NN	0.0229	0.0109	0.9661
$\mathcal{L}_k^{\Phi_p}$	DT	0.0418	0.3092	0.6490
LSVC	RF	0.0002	0.0248	0.9750
LSVC	LR	0.0	0.0	1.0
LSVC	k-NN	0.0004	0.0404	0.9592
LSVC	DT	0.0074	0.4759	0.5167
RF	LR	0.0128	0.0	0.9872
RF	k-NN	0.0064	0.0101	0.9835
RF	DT	0.0139	0.3250	0.6611
LR	k-NN	0.0001	0.0309	0.9690
LR	DT	0.0052	0.4809	0.5139
k-NN	DT	0.0224	0.3206	0.6570

Tabela 10 – Probabilidades Dataset 2

6.2.3 Dataset 3: MAGIC Gamma Telescope

O conjunto de dados *MAGIC Gamma Telescope* foi gerado mediante simulações de Monte Carlo para simular o registro de partículas *gamma* de alta energia em um telescópio gamma Cherenkov atmosférico baseado no solo usando técnicas de imagem.

As classes do conjunto de dados são `g`, que representa os eventos *gamma*, e `h` (hadron), que representa os eventos marcados como ruído atmosférico de fundo. Por razões técnicas, o número de eventos do tipo `h` é subestimado, pois nos dados reais, a classe `h` representa a maioria dos eventos. Também, a acurácia como métrica de avaliação da classificação não é significativa para esses dados, pois classificar um evento de ruído de fundo (`h`) como sinal (`g`), é pior do que classificar um evento sinal como ruído de fundo.

O conjunto de dados completo e a descrição acima, foram obtidos do site *UCI Machine Learning Repository*[8].

1. Pré-processamento de dados

A Tabela 11, resume as informações básicas do conjunto de dados, sem o vetor de rótulos, que vamos utilizar no que segue. Como o dataset não possui dados *não válidos*, podemos utilizar ele sem a necessidade de fazer pré-processamentos de limpeza.

Informações básicas	
Nº de instâncias	19020
Dimensão dos dados	10
Possui dados não válidos?	Não
Tipo de dados	`float64`

Tabela 11 – Informações básicas do Dataset 3

Agora com relação ao vetor de rótulos, a classe positiva é a classe `g`, que representa o evento de interesse e que codificamos com o rótulo 1, e a classe negativa é `h`, codificada com rótulo 0, representando o ruído de fundo. A distribuição de classes é mostrada na Figura 10.

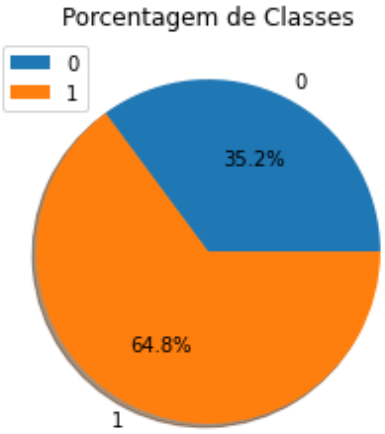

Figura 10 – Classes Dataset 3

O número de observações da classe minoritária, a classe 0, é aproximadamente o 54.23% do número de observações da classe majoritária ($N_m/N_M \approx 0.5423$), portanto,

[8] O conjunto de dados pode ser baixado desde o seguinte link: https://doi.org/10.24432/C52C8B.

segundo a convenção da Seção 6.1.4, o conjunto de dados assim conformado não precisa de técnicas de balanceamento de classes.

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

Número de instâncias	
Conjunto de treinamento	12680
Conjunto de teste	6340

2. Otimização de hiperparâmetros

Agora que temos nosso conjunto de dados pronto, precisamos determinar a métrica de avaliação adequada para determinar os hiperparâmetros ótimos. Pelo discutido na Seção 6.1.4, a métrica de avaliação adequada aqui é a métrica $\boldsymbol{F_{0.5}}$, pois a classe positiva é mais importante que a classe negativa; e como apontado no preâmbulo, classificar um ruído de fundo (hadron) como *sinal* (falso positivo) gera *mais prejuízo* que classificar um sinal como ruído de fundo (falso negativo).

Assim, fazendo 5×20 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de avaliação $F_{0.5}$) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 12.

#	**Modelo**	$\boldsymbol{hp_1}$	$\boldsymbol{hp_2}$	$\boldsymbol{hp_3}$	$\boldsymbol{F_{0.5}}$	**acurácia**	**precisão**	**recall**
1	RF	gini	$H = 12$	$n_e = 500$	**0.8830**	**0.8733**	**0.8678**	**0.9492**
2	DT	entropy	$H = 9$	best	0.8638	0.8450	0.8499	0.9243
3	k-NN	$k = 4$	-	-	0.8609	0.8222	0.8591	0.8682
4	LSVC	$C = 10^2$	l_2	$i_s = 1$	0.8223	0.7894	0.8079	0.8882
5	LR	$C = 10^2$	l_2	liblinear	0.8216	0.7926	0.8043	0.8989
6	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 1$	$H = 6$	0.8190	0.7899	0.8011	0.8993

Tabela 12 – Melhores representantes de cada modelo no Dataset 3

As pontuações médias, com relação à métrica $F_{0.5}$, dos modelos mais promissores no conjunto de treinamento; *sugerem* uma clara superioridade do modelo RF por sobre os modelos restantes. As pontuações também sugerem que o desempenho dos modelos DT e k-NN pode ser equivalente e melhor que o desempenho dos modelos LSVC, LR e $\mathcal{L}_k^{\Phi_p}$, os quais por sua vez também podem ter desempenho equivalente.

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos os melhores representantes de cada modelo no conjunto de treinamento *completo* para logo avaliar o desempenho

no conjunto de teste, obtendo assim os resultados da Tabela 13, os quais concordam com os resultados médios no conjunto de treinamento da Tabela 12, sugerindo o mesmo comportamento relativo entre modelos descrito no passo 2.

#	Modelo	hp_1	hp_2	hp_3	$F_{0.5}$	acurácia	precisão	recall
1	RF	gini	$H = 12$	$n_e = 500$	**0.8834**	**0.8730**	**0.8688**	**0.9472**
2	DT	entropy	$H = 9$	best	0.8714	0.8536	0.8585	0.9270
3	k-NN	$k = 4$	-	-	0.8642	0.8292	0.8606	0.8789
4	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 1$	$H = 6$	0.8219	0.7923	0.8052	0.8966
5	LR	$C = 10^2$	l_2	liblinear	0.8167	0.7863	0.7990	0.8956
6	LSVC	$C = 10^2$	l_2	$i_s = 1$	0.7969	0.7740	0.7692	0.9307

Tabela 13 – Avaliação final dos representantes de cada modelo no Dataset 3

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 11.

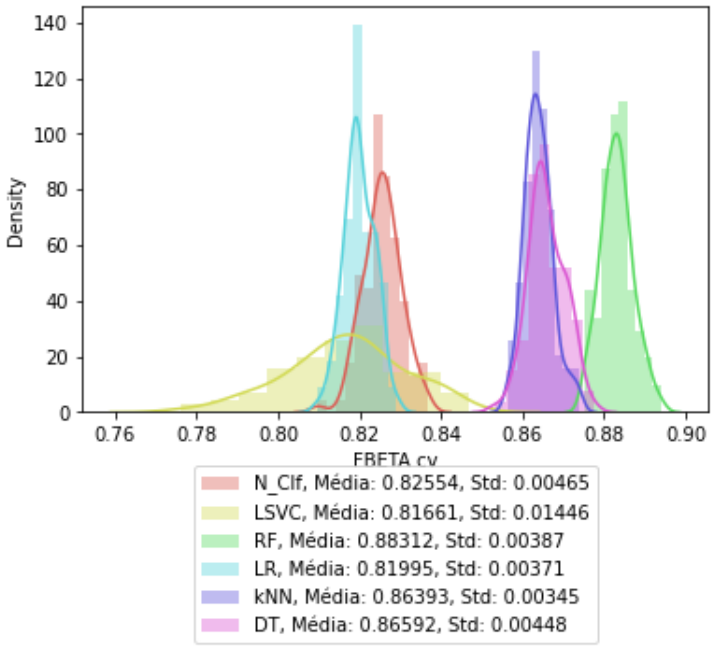

Figura 11 – Distribuição CV Dataset 3

Finalmente, na Tabela 14 temos os resultados do *teste de significância* para cada par de modelos.

5. Conclusões finais para o Dataset 3

Da Figura 11 e da Tabela 14, vemos que o modelo RF possui uma alta probabilidade, maior que 0.999, de ter um desempenho melhor que o resto dos modelos. Também

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	LSVC	0.0057	0.4425	0.5518
$\mathcal{L}_k^{\Phi_p}$	RF	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	LR	0.0	0.0433	0.9567
$\mathcal{L}_k^{\Phi_p}$	k-NN	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	DT	1.0	0.0	0.0
LSVC	RF	1.0	0.0	0.0
LSVC	LR	0.1741	0.0307	0.7952
LSVC	k-NN	1.0	0.0	0.0
LSVC	DT	1.0	0.0	0.0
RF	LR	0.0	1.0	0.0
RF	k-NN	0.0	1.0	0.0
RF	DT	0.0	0.9994	0.0006
LR	k-NN	1.0	0.0	0.0
LR	DT	1.0	0.0	0.0
k-NN	DT	0.0005	0.0	0.9995

Tabela 14 – Probabilidades Dataset 3

vemos que os modelos $\mathcal{L}_k^{\Phi_p}$ e LR, possuem uma alta probabilidade, igual a 0.9567, de ter desempenho equivalente no Dataset, assim como também é o caso do par k-NN e DT, que possuem uma probablidade de ter desempenho equivalente igual a 0.9995.

6.2.4 Dataset 4: MiniBooNE particle identification

Este conjunto de dados é tomado do experimento *MiniBooNE* e é usado para distinguir *neutrinos de elétrons* (sinal) de *neutrinos de múons* (ruído de fundo).

O conjunto de dados e a descrição acima, foram obtidos do site *UCI Machine Learning Repository*[9] e mais detalhes sobre o experimento MiniBooNE podem ser consultados em [3].

1. Pré-processamento de dados

A Tabela 15, resume as informações básicas do conjunto de dados, sem o vetor de rótulos, que vamos utilizar no que segue. Como o dataset não possui dados *não válidos*, podemos utilizar ele sem a necessidade de fazer pré-processamentos de limpeza.

Agora com relação ao vetor de rótulos, a classe positiva é a classe 1, que representa o evento de interesse (sinal), e a classe negativa é 0, representando o ruído de fundo. A distribuição de classes é mostrada na Figura 12.

O número de observações da classe minoritária, a classe 1, é aproximadamente o 39% do número de observações da classe majoritária ($N_m/N_M \approx 0.39$), portanto, segundo

[9] O conjunto de dados pode ser baixado desde o seguinte link: https://doi.org/10.24432/C5QC87.

Informações básicas	
Nº de instâncias	130064
Dimensão dos dados	50
Possui dados não válidos?	Não
Tipo de dados	`float64`

Tabela 15 – Informações básicas do Dataset 4

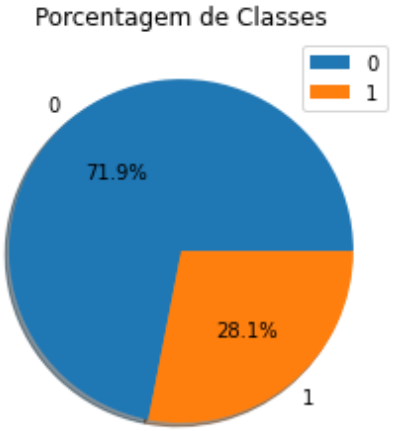

Figura 12 – Classes Dataset 4

a convenção da Seção 6.1.4, o conjunto de dados assim conformado não precisa receber técnicas de balanceamento de classes.

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

Número de instâncias	
Conjunto de treinamento	86709
Conjunto de teste	43355

2. Otimização de hiperparâmetros

Agora que temos nosso conjunto de dados pronto, precisamos determinar a métrica de avaliação adequada para determinar os hiperparâmetros ótimos. Pelo discutido na Seção 6.1.4, a métrica de avaliação adequada aqui é a métrica $\boldsymbol{F_{0.5}}$, pois a classe positiva é mais importante que a classe negativa; e o *prejuízo* gerado pelos falsos positivos é maior que o prejuízo gerado pelos falsos negativos, pois uma instância classificada como sinal deve passar por verificações posteriores que requerem de trabalho adicional, portanto classificar um ruído de fundo como sinal (falso positivo) gera *mais prejuízo* que classificar um sinal como ruído de fundo (falso negativo).

Assim, fazendo 5×20 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de avaliação $F_{0.5}$) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 16.

#	Modelo	hp_1	hp_2	hp_3	$F_{0.5}$	acurácia	precisão	recall
1	RF	entropy	$H = 11$	$n_e = 500$	**0.8775**	**0.9298**	**0.8799**	**0.8683**
2	DT	gini	$H = 9$	best	0.8335	0.9074	0.8319	0.8398
3	LR	$C = 10^3$	l_2	liblinear	0.8266	0.8919	0.8488	0.7482
4	k-NN	$k = 8$	-	-	0.8027	0.8832	0.8159	0.7540
5	LSVC	$C = 10$	l_2	$i_s = 0.1$	0.7975	0.8693	0.8427	0.6567
6	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 5$	$H = 9$	0.7327	0.8469	0.7428	0.6951

Tabela 16 – Melhores representantes de cada modelo no Dataset 4

As pontuações médias, com relação à métrica $F_{0.5}$, dos modelos mais promissores no conjunto de treinamento; sugerem uma clara superioridade do modelo RF por sobre os modelos restantes. As pontuações também sugerem que o desempenho dos modelos DT e LR pode ser equivalente e melhor que o desempenho dos modelos k-NN e LSVC, os que por sua vez podem ter desempenho equivalente e melhor do que o modelo $\mathcal{L}_k^{\Phi_p}$, que é o modelo com pior desempenho nesse dataset.

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos os melhores representantes de cada modelo no conjunto de treinamento *completo* para logo avaliar o desempenho no conjunto de teste, obtendo assim os resultados da Tabela 17, os quais concordam com os resultados médios no conjunto de treinamento da Tabela 16, sugerindo o mesmo comportamento relativo entre modelos descrito no passo 2.

#	Modelo	hp_1	hp_2	hp_3	$F_{0.5}$	acurácia	precisão	recall
1	RF	entropy	$H = 11$	$n_e = 500$	**0.8733**	**0.9282**	**0.8745**	**0.8686**
2	DT	gini	$H = 9$	best	0.8309	0.907	0.8264	0.8494
3	LR	$C = 10^3$	l_2	liblinear	0.8243	0.8918	0.8439	0.7539
4	k-NN	$k = 8$	-	-	0.8018	0.8832	0.8139	0.7570
5	LSVC	$C = 10$	l_2	$i_s = 0.1$	0.7943	0.8677	0.8396	0.6534
6	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 5$	$H = 9$	0.7297	0.8461	0.7370	0.7022

Tabela 17 – Avaliação final dos representantes de cada modelo no Dataset 4

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 13.

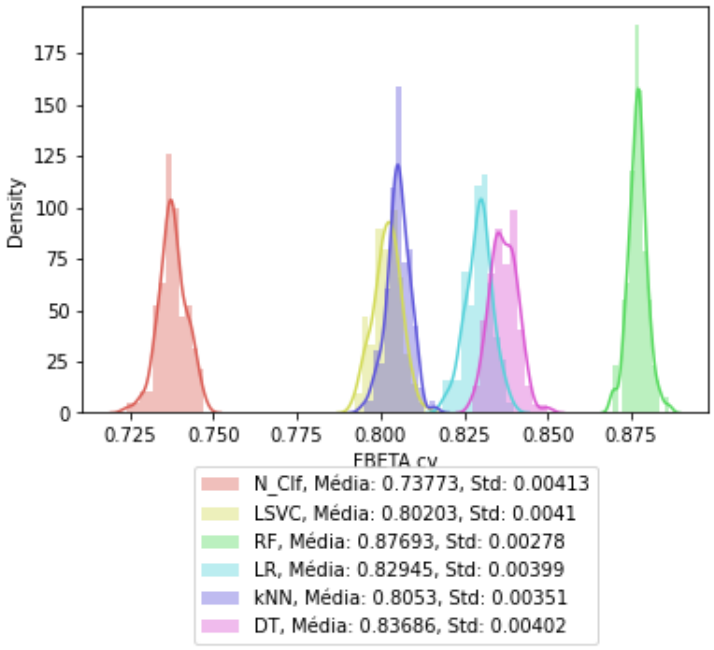

Figura 13 – Distribuição CV Dataset 4

Finalmente, na Tabela 18 temos os resultados do *teste de significância* para cada par de modelos.

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	LSVC	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	RF	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	LR	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	k-NN	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	DT	1.0	0.0	0.0
LSVC	RF	1.0	0.0	0.0
LSVC	LR	1.0	0.0	0.0
LSVC	k-NN	0.0006	0.0	0.9994
LSVC	DT	1.0	0.0	0.0
RF	LR	0.0	1.0	0.0
RF	k-NN	0.0	1.0	0.0
RF	DT	0.0	1.0	0.0
LR	k-NN	0.0	1.0	0.0
LR	DT	0.1472	0.0	0.8528
k-NN	DT	1.0	0.0	0.0

Tabela 18 – Probabilidades Dataset 4

5. Conclusões finais para o Dataset 4

A informação da Figura 13 e da Tabela 18, confirma a tendência vista inicialmente, e com alta probabilidade, o modelo $\mathcal{L}_k^{\Phi_p}$ possui um desempenho pior que o resto dos

modelos para esse Dataset. Também os modelos LSVC e k-NN, tem uma probabilidade igual a 0.9994 de terem desempenho equivalente, e os modelos LR e DT, mostram um grau de proximidade na gráfica mas com uma probabilidade de serem equivalentes no Dataset igual a 0.8528.

6.2.5 Dataset 5: Higgs Boson Machine Learning Challenge

O seguinte Dataset, foi construído a partir de simulações oficiais do experimento *ATLAS*. O simulador tem duas partes. Na primeira, colisões próton-próton aleatórias são simuladas com base no conhecimento que acumulado em física de partículas. Ele reproduz as explosões microscópicas aleatórias resultantes das colisões próton-próton. Na segunda parte, as partículas resultantes são rastreadas através de um modelo virtual do detector. O processo produz eventos simulados com propriedades que imitam as propriedades estatísticas dos eventos reais com informações adicionais sobre o que aconteceu durante a colisão, antes que as partículas sejam medidas no detector.

O conjunto de dados completo e a descrição acima, foram obtidos do site *Kaggle*[10], e mais detalhes sobre o conjunto de dados podem ser encontrados em [2].

1. Pré-processamento de dados

Esse dataset possui dados com o valor `-999.0`, valor que segundo a informação dada com o próprio dataset, representa um dado que deve ser tratado como `NaN`.

O conjunto de dados, originalmente possui 30 features, das quais *sete* têm pouco mais de 70% dos valores igual a `-999.0`, *três* features com pouco mais de 39% desses valores e *uma* feature com pouco mais de 15% de valores iguais a `-999.0`, logo vamos eliminar as colunas com mais de 39% de valores `-999.0`, e na coluna com pouco mais de 15% desses valores, eles serão imputados pela média do resto dos valores na coluna.

Feito isso, a Tabela 19 resume as informações básicas do conjunto de dados, sem o vetor de rótulos, que vamos utilizar no que segue.

Informações básicas	
Nº de instâncias	818238
Dimensão dos dados	20
Possui dados não válidos?	Não
Tipo de dados	`int`, `float64`

Tabela 19 – Informações básicas do Dataset 5

Com relação ao vetor de rótulos, a classe alvo é `s` (sinal), que será codificada por 1, e a classe negativa `b` (background), que será codificada por 0. A Figura 14 mostra a porcentagem de cada classe no dataset. O número de observações da classe minoritária é

[10] O conjunto de dados pode ser baixado desde o seguinte link: https://www.kaggle.com/competitions/higgs-boson/data.

aproximadamente o $51,90\%$ do número de observações da classe majoritária ($N_m/N_M \approx 0.51897$), portanto, segundo a convenção da Seção 6.1.4, o dataset não tem problemas de desbalance de classes.

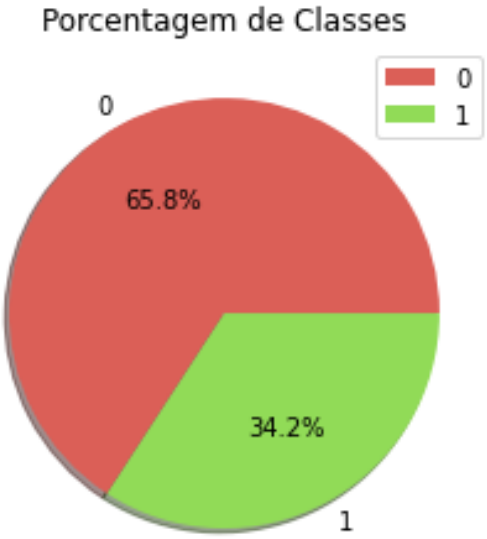

Figura 14 – Classes Dataset 5

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

Número de instâncias	
Conjunto de treinamento	545492
Conjunto de teste	272746

2. Otimização de hiperparâmetros

Antes de começar, aclaramos que para esse conjunto de dados, o algoritmo k-NN *não participará* do processo de comparação devido aos excessivos tempos de cálculo do modelo sobre conjuntos de dados de essa envergadura.

Agora precisamos determinar a métrica de avaliação adequada para determinar os hiperparâmetros ótimos. Pelo discutido na Seção 6.1.4, a métrica de avaliação adequada aqui é a métrica $\boldsymbol{F_{0.5}}$, pois a classe positiva é mais importante que a classe negativa, e como é comum na deteção de partículas, cada evento classificado como sinal, deve passar por verificações posteriores, logo, supomos que o *prejuízo* gerado pelos falsos positivos é maior do que o gerado pelos falsos negativos.

Assim, fazendo 5×2 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de

avaliação $F_{0.5}$) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 20, onde vemos uma clara diferença no desempenho dos modelos, sendo o modelo RF o de melhor métrica e desempenho geral, seguido do modelo DT, e o algoritmo $\mathcal{L}_k^{\Phi_p}$ é o pior colocado no ranking. Também vemos que os algoritmos LR e LSVC parecem ter um desempenho muito próximo ou praticamente equivalente.

#	Modelo	hp_1	hp_2	hp_3	$F_{0.5}$	acurácia	precisão	recall
1	RF	gini	$H = 10$	$n_e = 10^2$	**0.7606**	**0.8281**	**0.7798**	0.6926
2	DT	entropy	$H = 10$	best	0.7459	0.8210	0.7591	**0.6976**
3	LR	$C = 10^3$	l_2	liblinear	0.6192	0.7404	0.6482	0.5252
4	LSVC	$C = 10$	l_2	$i_s = 1$	0.6142	0.7376	0.6475	0.5092
5	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 5$	$H = 3$	0.5831	0.7182	0.6012	0.5204

Tabela 20 – Melhores representantes de cada modelo no Dataset 5

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos os melhores representantes de cada modelo no conjunto de treinamento *completo* para logo avaliar o desempenho no conjunto de teste, obtendo assim os resultados da Tabela 21, os quais concordam com os resultados médios no conjunto de treinamento da Tabela 20, sugerindo o mesmo comportamento relativo entre modelos descrito no passo 2.

#	Modelo	hp_1	hp_2	hp_3	$F_{0.5}$	acurácia	precisão	recall
1	RF	gini	$H = 10$	$n_e = 10^2$	**0.7572**	**0.8262**	**0.7759**	0.6906
2	DT	entropy	$H = 10$	best	0.7415	0.81978	0.7503	**0.7081**
3	LR	$C = 10^3$	l_2	liblinear	0.6168	0.7390	0.6458	0.5230
4	LSVC	$C = 10$	l_2	$i_s = 1$	0.6114	0.7361	0.6446	0.5071
5	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 5$	$H = 3$	0.5829	0.7182	0.6015	0.5187

Tabela 21 – Avaliação final dos representantes de cada modelo no Dataset 5

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 15 e na Tabela 22 temos os resultados do *teste de significância* para cada par de modelos.

5. Conclusões finais para o Dataset 5

Da Tabela 22 e da Figura 15, vemos que o comportamento observado desde o início se mantém e o modelo com melhor desempenho no Dataset é o RF seguido do

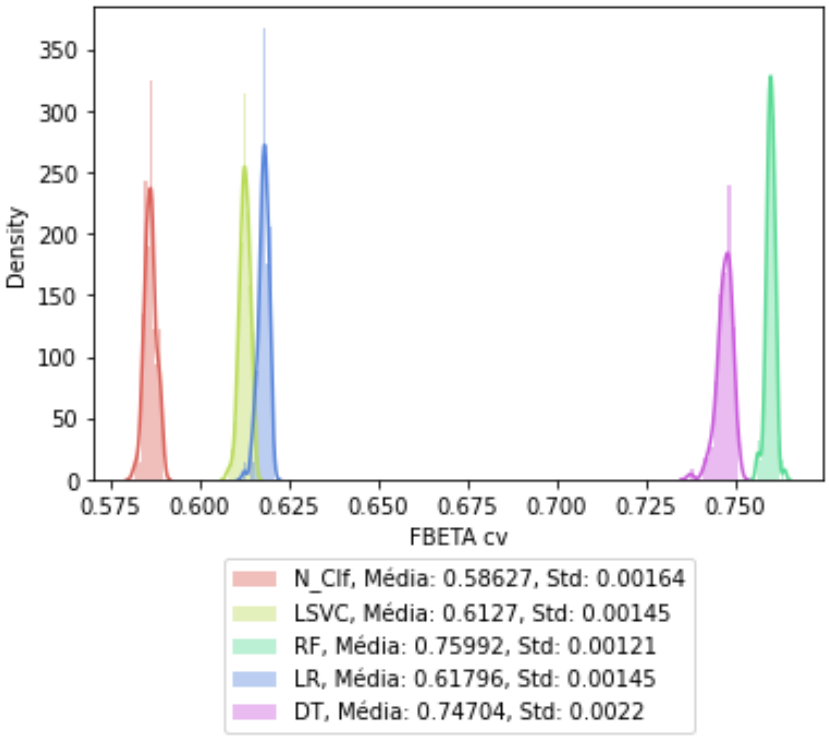

Figura 15 – Distribbuição CV Dataset 5

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	RF	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	LSVC	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	LR	1.0	0.0	0.0
$\mathcal{L}_k^{\Phi_p}$	DT	1.0	0.0	0.0
LSVC	RF	1.0	0.0	0.0
LSVC	LR	0.0	0.0	1.0
LSVC	DT	1.0	0.0	0.0
RF	LR	0.0	1.0	0.0
RF	DT	0.0	0.9988	0.0012
LR	DT	1.0	0.0	0.0

Tabela 22 – Probabilidades Dataset 5

DT com uma probabilidade do RF ter melhor desempenho do que o DT igual a 0.9988. Também confirmamos que o modelo $\mathcal{L}_k^{\Phi_p}$ possui o pior desempenho no Dataset, com probabilidade 1.0, e que os modelos LSVC e LR são de desempenho equivalente no Dataset, com probabilidade 1.0.

6.2.6 Dataset 6: Human Activity Recognition from Continuous Ambient Sensor Data

O seguinte *dataset*, representa dados coletados com sensores ambientais em *residências inteligentes* com residentes voluntários. Sensores de movimento PIR ambiente,

sensores de porta/temperatura e sensores de interruptor de luz são colocados em aqueles locais da casa do voluntário que estejam relacionados com atividades específicas da vida diária que se deseja capturar, e onde os dados são coletados continuamente enquanto os residentes executam suas rotinas normais.

O problema de classificação é prever a atividade que está ocorrendo na casa inteligente e sendo observada pelos sensores ambientais.

O conjunto de dados completo e a descrição acima, foram obtidos do site *UCI Machine Learning Repository*[11], e mais detalhes sobre o conjunto de dados podem ser encontrados em [19, 20].

1. Pré-processamento de dados

O conjunto de dados é composto pelos registros coletados de trinta casas inteligentes, casas que podem ser de vários modelos e que possuem diferentes configurações, portanto, para os testes numéricos utilizaremos os dados de apenas uma das casas. A pasta cujos arquivos possuem o maior número de observações, é a pasta `csh113`, e portanto, vamos extrair o nosso conjunto de dados desse diretório: o arquivo `csh113.ann.features.csv`. A Figura 16, é uma imagem tomada diretamente do dataset, e mostra o modelo e a distribuição dos sensores da casa `csh113`.

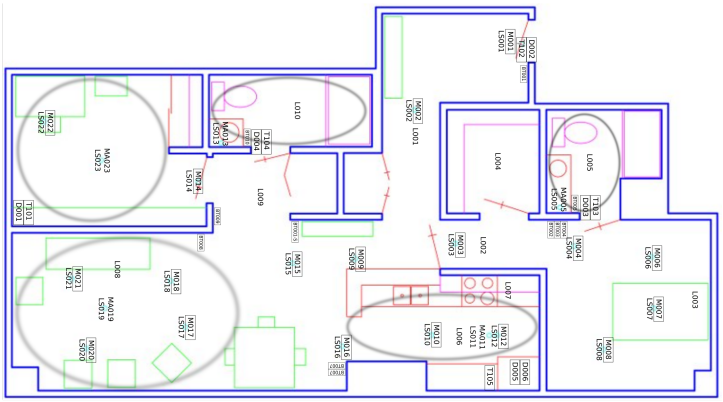

Figura 16 – Modelo da casa csh113. Imagem tomada do dataset.

A Tabela 23, resume as informações básicas do conjunto de dados, sem o vetor de rótulos, que vamos utilizar no que segue. Como o dataset não possui dados *não válidos* nem *instâncias duplicadas*, podemos utilizar ele sem a necessidade de fazer pré-processamentos de limpeza.

Agora vamos fazer uma rápida análise do vetor de rótulos, cujas classes são as atividades realizadas pelos residentes da casa. Em total, são *trinta e cinco* atividades ou

[11] O conjunto de dados pode ser baixado desde o seguinte link: https://doi.org/10.24432/C5D60P.

Informações básicas	
Nº de instâncias	3190818
Dimensão dos dados	36
Possui dados não válidos?	Não
Tipo de dados	`float64`

Tabela 23 – Informações básicas do Dataset 6

classes que listamos na Tabela 24, junto com a porcentagem que cada classe representa do total.

Classe e porcentagem		**Classe e porcentagem**	
`Other_Activity`	28.693%	`Watch_TV`	0.906%
`Personal_Hygiene`	12.378%	`Work_At_Table`	0.781%
`Work_On_Computer`	8.308%	`Enter_Home`	0.779%
`Cook_Breakfast`	7.779%	`Leave_Home`	0.757%
`Entertain_Guests`	6.932%	`Bed_Toilet_Transition`	0.478%
`Groom`	5.793%	`Evening_Meds`	0.371%
`Sleep`	3.293%	`Morning_Meds`	0.346%
`Toilet`	3.156%	`Cook_Lunch`	0.206%
`Wash_Dishes`	3.117%	`Work`	0.158%
`Bathe`	2.614%	`Step_Out`	0.094%
`Wash_Breakfast_Dishes`	2.199%	`Wash_Lunch_Dishes`	0.062%
`Cook`	2.183%	`Cook_Dinner`	0.055%
`Phone`	1.981%	`Sleep_Out_Of_Bed`	0.036%
`Dress`	1.695%	`Eat`	0.017%
`Drink`	1.482%	`Eat_Lunch`	0.012%
`Relax`	1.371%	`Wash_Dinner_Dishes`	0.009%
`Read`	1.010%	`Eat_Dinner`	0.004%
`Eat_Breakfast`	0.947%	-	-

Tabela 24 – Atividades ou classes do Dataset 6

Para reduzir o problema para um de classificação binária, e para facilitar o trabalho dos classificadores, vamos *escolher* uma das classes mais numerosas para identificar, assim escolhendo `Work_On_Computer` como atividade alvo, codificamos o vetor de rótulos associando o rótulo 1 à atividade `Work_On_Computer` e associando o rótulo 0 para as atividades restantes. Dessa forma, o nosso *novo* vetor de rótulos, possui a distribuição de classes da Figura 17.

O número de observações da classe minoritária é aproximadamente o 9.06% do número de observações da classe majoritária ($N_m/N_M = 0.0906$), portanto, segundo a convenção da Seção 6.1.4, é recomendado *avaliar* o uso de técnicas de balanceamento de classes.

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do

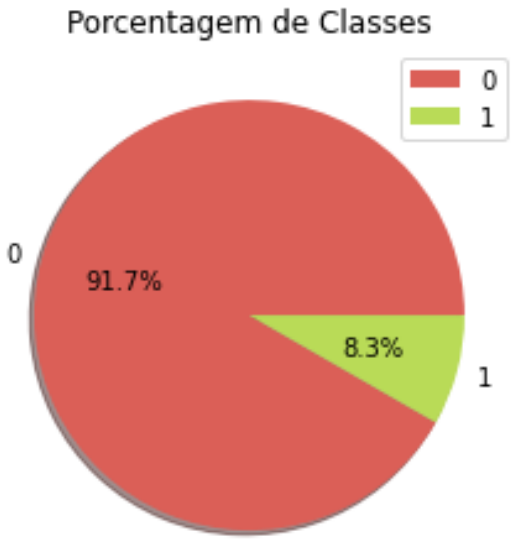

Figura 17 – Classes Dataset 6

módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

Número de instâncias	
Conjunto de treinamento	2127212
Conjunto de teste	1063606

2. Otimização de hiperparâmetros

Antes de começar, aclaramos que para esse conjunto de dados, o algoritmo k-NN *não participará* do processo de comparação devido aos excessivos tempos de cálculo do modelo sobre conjuntos de dados de essa envergadura.

Agora que temos nosso conjunto de dados pronto, precisamos determinar a métrica de avaliação adequada para determinar os hiperparâmetros ótimos. Pelo discutido na Seção 6.1.4, a métrica de avaliação adequada aqui é a métrica $\boldsymbol{F_1}$, pois a classe positiva é mais importante que a classe negativa, e se não dispomos de alguma restrição adicional, podemos supor que o *prejuízo* gerado pelos falsos positivos e falsos negativos é *equivalente*.

Provas preliminares feitas com os classificadores concorrentes na comparação, mostram que técnicas de balance de classes não melhoram a métrica $\boldsymbol{F_1}$ de forma significativa (diferença inferior ao 1%), assim, para poupar esforço computacional, não faremos pré-processamentos de balance de classes para esse conjunto de dados.

Assim, fazendo 5×2 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de avaliação F_1) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 25.

#	Modelo	hp_1	hp_2	hp_3	F_1	acurácia	precisão	recall
1	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 1$	$H = 10$	**0.9117**	**0.9850**	**0.8910**	0.9334
2	RF	gini	$H = 10$	$n_e = 10^2$	0.8646	0.9755	0.7987	**0.9424**
3	DT	gini	$H = 10$	best	0.8640	0.9757	0.8066	0.9303
4	LR	$C = 10^3$	l_2	liblinear	0.8432	0.9713	0.7718	0.9292
5	LSVC	$C = 10$	l_2	$i_s = 0.1$	0.8429	0.9711	0.7677	0.9345

Tabela 25 – Melhores representantes de cada modelo no Dataset 6

As pontuações médias, com relação à métrica F_1, dos modelos mais promissores no conjunto de treinamento; mostram uma diferença clara entre o modelo $\mathcal{L}_k^{\Phi_p}$ e os modelos restantes. Os resultados sugerem que para esse problema de classificação binária, o modelo $\mathcal{L}_k^{\Phi_p}$ é melhor que os restantes, que os modelos DT e RF são praticamente equivalentes e com desempenho melhor do que o dos modelos LR e LSVC, os que por sua vez também parecem ser praticamente equivalentes.

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos os melhores representantes de cada modelo no conjunto de treinamento *completo* para logo avaliar o desempenho no conjunto de teste, obtendo assim os resultados da Tabela 26, os quais concordam com os resultados médios no conjunto de treinamento da Tabela 25, sugerindo o mesmo comportamento relativo entre modelos descrito no passo 2.

#	Modelo	hp_1	hp_2	hp_3	F_1	acurácia	precisão	recall
1	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 1$	$H = 10$	**0.9204**	**0.9865**	**0.9024**	0.9390
2	RF	gini	$H = 10$	$n_e = 10^2$	0.8642	0.9754	0.7984	**0.9418**
3	DT	gini	$H = 10$	best	0.8628	0.9754	0.8050	0.9296
4	LR	$C = 10^3$	l_2	liblinear	0.8434	0.9713	0.7723	0.9288
5	LSVC	$C = 10$	l_2	$i_s = 0.1$	0.8429	0.9711	0.7683	0.9337

Tabela 26 – Avaliação final dos representantes de cada modelo no Dataset 6

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 18 e na Tabela 27 temos os resultados do *teste de significância* para cada par de modelos.

5. Conclusões finais para o Dataset 6

Da Tabela 27 e da Figura 18, vemos que o modelo $\mathcal{L}_k^{\Phi_p}$ tem um desempenho marcadamente superior ao resto de classificadores com probabilidade 1.0. Também os

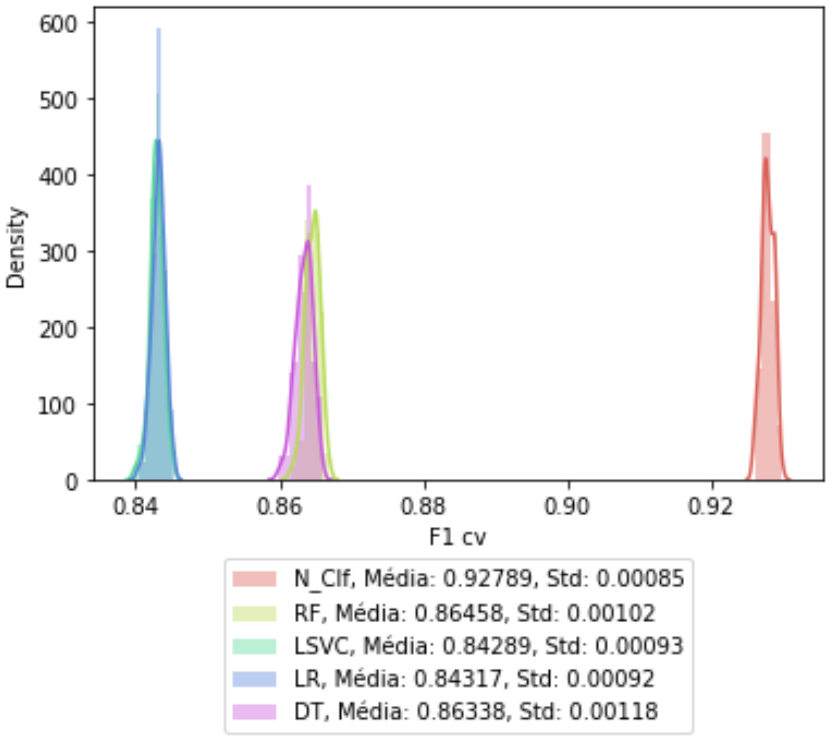

Figura 18 – Distribbuição CV Dataset 6

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	RF	0.0	1.0	0.0
$\mathcal{L}_k^{\Phi_p}$	LSVC	0.0	1.0	0.0
$\mathcal{L}_k^{\Phi_p}$	LR	0.0	1.0	0.0
$\mathcal{L}_k^{\Phi_p}$	DT	0.0	1.0	0.0
RF	LSVC	0.0	1.0	0.0
RF	LR	0.0	1.0	0.0
RF	DT	0.0	0.0	1.0
LSVC	LR	0.0	0.0	1.0
LSVC	DT	1.0	0.0	0.0
LR	DT	1.0	0.0	0.0

Tabela 27 – Probabilidades Dataset 6

resultados confirmam a tendência inicial de que os modelos RF e DT são de desempenho equivalente no Dataset com probabilidade 1.0, situação que se repete com os modelos LSVC e LR e com os modelos RF e DT.

6.2.7 Dataset 7: WISDM Parte I: Smartphone and Smartwatch Activity and Biometrics Dataset

Esse Dataset está composto por series temporais de leituras de sensores (acelerômetro e do giroscópio), coletados desde smartphones e smartwatches enquanto 51 sujeitos de teste executam 18 atividades durante 3 minutos cada um.

O problema de classificação é prever a atividade que está sendo executada pelo usuário do aparelho usando as leituras do acelerômetro e/ou giroscópio.

O conjunto de dados completo e a descrição acima, foram obtidos do site *UCI Machine Learning Repository*[12], e mais detalhes sobre o conjunto de dados podem ser encontrados em [60].

1. Pré-processamento de dados

O conjunto de dados está dividido em duas pastas, uma para os dados coletados desde os smartphones e outra com os dados coletados desde os smartwatches. O Dataset 7 constará apenas dos dados do acelerômetro do smartphone. O conjunto de dados é composto de 51 arquivos, um por cada sujeito de teste, os quais possuem 5 colunas: a primeira coluna indica o tempo no qual foi registrada a leitura, mais as três componentes da leitura do acelerômetro e a última é a atividade realizada ou classe. Para extraer informação da coluna temporal, existem diversos métodos que transformam o conjunto de dados em conjuntos de dimensão maior e com menor número de instâncias, [60]. Aqui apenas será retirada a coluna temporal, obtendo um Dataset com apenas 3 features. A Tabela 28, resume as informações básicas do conjunto de dados, sem o vetor de rótulos, que vamos utilizar no que segue. Como o dataset não possui dados *não válidos*, podemos utilizar ele sem a necessidade de fazer pré-processamentos de limpeza.

Informações básicas	
Nº de instâncias	4804403
Dimensão dos dados	3
Possui dados não válidos?	Não
Tipo de dados	`float64`

Tabela 28 – Informações básicas do Dataset 7

Agora vamos fazer uma rápida análise do vetor de rótulos, que possue 18 classes ou atividades. Pela forma de obter as amostras, todas as classes tem aproximadamente a mesma proporção, isto é, cada classe representa aproximadamente o 5.5% da amostra. A Tabela 29, mostra a lista das atividades ou classes.

Para reduzir o problema para um de classificação binária, vamos *escolher* uma das classes para identificar, assim escolhendo `Typing` como atividade alvo, codificamos o vetor de rótulos associando o rótulo 1 à atividade `Typing` e associando o rótulo 0 para as atividades restantes. Dessa forma, o nosso *novo* vetor de rótulos, possui a distribuição de classes da Figura 19.

O número de observações da classe minoritária é aproximadamente o 5.4% do número de observações da classe majoritária ($N_m/N_M = 0.054$), portanto, segundo a

[12] O conjunto de dados pode ser baixado desde o seguinte link: `https://doi.org/10.24432/C5HK59`.

Walking	Jogging
Stairs	Sitting
Standing	Typing
Eating Soap	Brushing Teeth
Eating Chips	Eating Pasta
Drinking from Cup	Eating Sandwich
Kicking (Soccer Ball)	Playing Catch
Dribblinlg	Writing
Clapping	Folding Clothes

Tabela 29 – Multiclasses Dataset 7

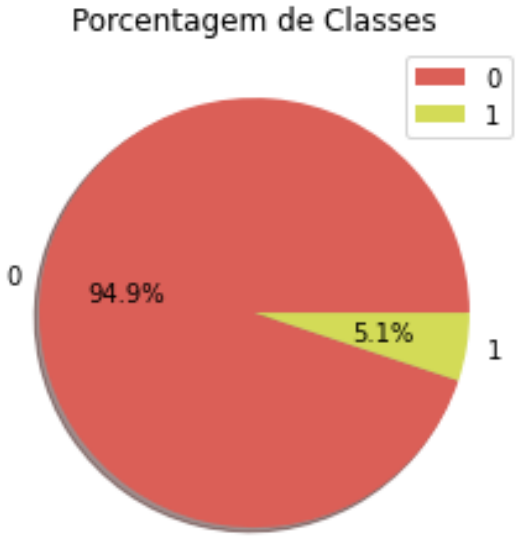

Figura 19 – Classes Dataset 7

convenção da Seção 6.1.4, é recomendado considerar o uso alguma técnica de balance de classes.

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

Número de instâncias	
Conjunto de treinamento	3202935
Conjunto de teste	1601468

2. Otimização de hiperparâmetros

Agora que temos nosso conjunto de dados pronto, precisamos determinar a métrica de avaliação adequada para determinar os hiperparâmetros ótimos. Pelo discutido na Seção 6.1.4, a métrica de avaliação adequada aqui é a métrica $\boldsymbol{F_1}$, pois a classe positiva é

mais importante que a classe negativa, e se não dispomos de alguma restrição adicional, podemos supor que o *prejuízo* gerado pelos falsos positivos e falsos negativos é *equivalente*.

Provas preliminares feitas com os classificadores concorrentes na comparação, mostram que técnicas de balance de classes não melhoram a métrica $\boldsymbol{F_1}$ de forma significativa (diferença inferior ao 1%), assim, para poupar esforço computacional, não faremos pré-processamentos de balance de classes para esse conjunto de dados.

Assim, fazendo 5×2 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de avaliação F_1) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 30.

#	Modelo	hp_1	hp_2	hp_3	F_1	acurácia	precisão	recall
1	k-NN	$k = 7$	-	-	**0.9063**	**0.9906**	0.9228	**0.8904**
2	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 3$	$H = 12$	0.8945	0.9893	0.9048	0.8845
3	RF	gini	$H = 12$	$n_e = 10^2$	0.7924	0.9820	**0.9690**	0.6702
4	DT	gini	$H = 12$	best	0.7841	0.9795	0.8509	0.7277
5	LR	$C = 0.01$	l_2	liblinear	0.0	0.9487	0.0	0.0
6	LSVC	$C = 0.01$	l_2	$i_s = 1$	0.0	0.9487	0.0	0.0

Tabela 30 – Melhores representantes de cada modelo no Dataset 7

As pontuações médias, com relação à métrica F_1, dos modelos mais promissores no conjunto de treinamento; mostram uma diferença clara entre o par de modelos k-NN, $\mathcal{L}_k^{\Phi_p}$ e os modelos restantes. Os resultados também sugerem que os modelos RF e DT são praticamente equivalentes e com desempenho melhor do que o dos modelos LR e LSVC, os que por sua vez parecem ser praticamente equivalentes com o classificador que a cada amostra associa o rótulo da classe majoritária, classificador que não possui nenhum poder preditivo.

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos os melhores representantes de cada modelo no conjunto de treinamento *completo* para logo avaliar o desempenho no conjunto de teste, obtendo assim os resultados da Tabela 31, os quais concordam com os resultados médios no conjunto de treinamento da Tabela 30, sugerindo o mesmo comportamento relativo entre modelos descrito no passo 2.

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 20.

#	Modelo	hp_1	hp_2	hp_3	F_1	acurácia	precisão	recall
1	k-NN	$k = 7$	-	-	**0.9068**	**0.9906**	0.9238	**0.8904**
2	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 3$	$H = 12$	0.8970	0.9896	0.9071	0.8870
3	RF	gini	$H = 12$	$n_e = 10^2$	0.7902	0.9818	**0.9684**	0.6674
4	DT	gini	$H = 12$	best	0.7912	0.9799	0.8449	0.7439
5	LR	$C = 0.01$	l_2	liblinear	0.0	0.9487	0.0	0.0
6	LSVC	$C = 0.01$	l_2	$i_s = 1$	0.0	0.9487	0.0	0.0

Tabela 31 – Avaliação melhores representantes de cada modelo no Dataset 7

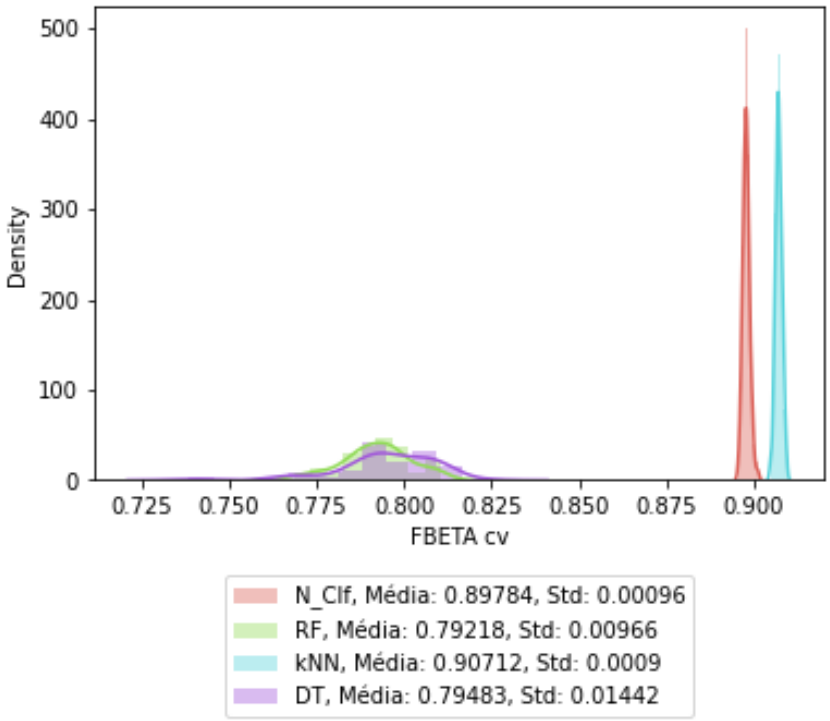

Figura 20 – Distribbuição CV Dataset 7

Finalmente, na Tabela 32 temos os resultados do *teste de significância* para cada par de modelos, onde os modelos LR e LSVC não participam por não possuir poder preditivo.

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	RF	0.0	1.0	0.0
$\mathcal{L}_k^{\Phi_p}$	k-NN	0.0192	0.0	0.9808
$\mathcal{L}_k^{\Phi_p}$	DT	0.0	1.0	0.0
RF	k-NN	1.0	0.0	0.0
RF	DT	0.2160	0.0887	0.6953
k-NN	DT	0.0	1.0	0.0

Tabela 32 – Probabilidades Dataset 7

5. Conclusões finais para o Dataset 7

Da Tabela 32 e da Figura 20, vemos que os modelos k-NN e $\mathcal{L}_k^{\Phi_p}$ possuem uma alta probabilidade de ser praticamente equivalentes, igual a 0.9808, e eles têm um desempenho marcadamente superior ao resto de classificadores com probabilidade 1.0. Também os resultados confirmam a tendência inicial de que os modelos LR e LSVC não possuem poder preditivo para esse Dataset e com relação aos modelos RF e DT, a probabilidade de eles serem equivalente é apenas de 0.6953, baixa demais para poder assumir equivalência no desempenho, veredito diferente ao obtido apenas com os passos anteriores sem o teste de significância.

6.2.8 Dataset 8: WISDM Parte II: Smartphone and Smartwatch Activity and Biometrics Dataset

Aqui vamos considerar o mesmo conjunto de dados do qual foi extraído o Dataset 7, mas dessa vez vamos considerar o conjunto de leituras do acelerômetro do *Smartwatch*.

1. Pré-processamento de dados

As atividades ou classes são as mesmas da Tabela 29 com o mesmo porcentagem aproximado de 5.5% para cada classe. O formato e frequência de coleta de dados é o mesmo tanto para o Smartphone como para o Smartwatch, portanto procederemos da mesma forma como no Dataset 7 e aqui também será extraída a primeira coluna com a variável temporal. A Tabela 33, resume as informações básicas do Dataset 8.

Informações básicas	
Nº de instâncias	3777046
Dimensão dos dados	3
Possui dados não válidos?	Não
Tipo de dados	`float64`

Tabela 33 – Informações básicas do Dataset 8

Igual que antes, para reduzir o problema para um de classificação binária, vamos *escolher* uma das classes para identificar, assim escolhendo `Sitting` como atividade alvo, codificamos o vetor de rótulos associando o rótulo 1 à atividade `Sitting` e associando o rótulo 0 para as atividades restantes. Dessa forma, o nosso *novo* vetor de rótulos, possui a distribuição de classes da Figura 21.

O número de observações da classe minoritária é aproximadamente o 5.98% do número de observações da classe majoritária ($N_m/N_M = 0.0598$), portanto, segundo a convenção da Seção 6.1.4, é recomendado *avaliar* o uso alguma técnica de balance de classes.

Antes de ir na busca dos hiperparâmetros ótimos, vamos dividir o conjunto de dados em conjuntos de treinamento e teste, utilizando a função `train_test_split` do

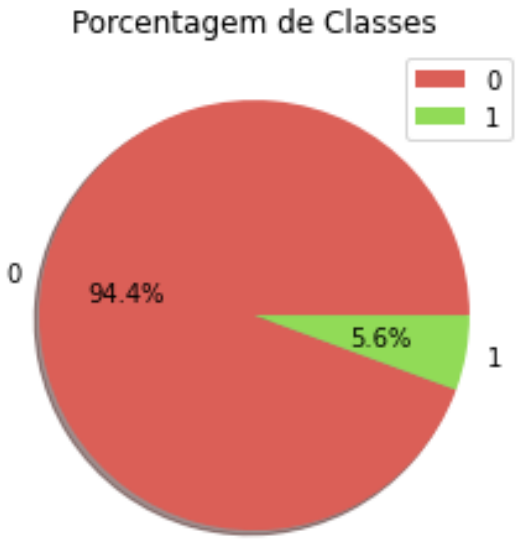

Figura 21 – Classes Dataset 8

módulo `sklearn`. Mantendo a proporção entre classes, reservamos 2/3 do dataset para o conjunto de *treinamento* e o 1/3 restante para o conjunto de *teste*, obtendo:

Número de instâncias	
Conjunto de treinamento	2518030
Conjunto de teste	1259016

2. Otimização de hiperparâmetros

Procedendo de igual maneira que no Dataset 7, temos que a métrica de avaliação também será $\boldsymbol{F_1}$ e depois de fazer provas preliminares, as técnicas de balance de classes aplicadas não melhoram a métrica $\boldsymbol{F_1}$ de forma significativa (diferença inferior ao 1%), assim, não faremos pré-processamentos de balance de classes para esse conjunto de dados.

Logo, fazendo 5×2 validações cruzadas da forma especificada na Seção 6.1.1; a média para cada métrica de avaliação nos melhores representantes (respeito da métrica de avaliação F_1) de cada modelo ao longo dessas validações, é apresentada no *ranking* da Tabela 34.

#	Modelo	$\boldsymbol{hp_1}$	$\boldsymbol{hp_2}$	$\boldsymbol{hp_3}$	$\boldsymbol{F_1}$	**acurácia**	**precisão**	**recall**
1	k-NN	$k = 7$	-	-	**0.7705**	**0.9764**	0.8534	**0.7022**
2	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 5$	$H = 12$	0.7574	0.9749	0.8316	0.6954
3	DT	gini	$H = 12$	best	0.6920	0.9694	0.7993	0.6106
4	RF	gini	$H = 12$	$n_e = 10^2$	0.6711	0.9703	**0.8920**	0.5379
5	LSVC	$C = 10^3$	l_2	$i_s = 1$	0.0108	0.8560	0.0057	0.1
6	LR	$C = 10$	l_2	liblinear	0.0001	0.9428	0.0021	0.0

Tabela 34 – Melhores representantes de cada modelo no Dataset 8

As pontuações médias, com relação à métrica F_1, dos modelos mais promissores no conjunto de treinamento; mostram uma diferença clara entre os modelos. Os primeiros lugares são para os modelos k-NN e $\mathcal{L}_k^{\Phi_p}$, seguidos pelos modelos DT e RF, todos melhores do que os modelos LSVC e LR, que por sua vez parecem ser praticamente equivalentes com o classificador que a cada amostra associa o rótulo da classe majoritária, classificador que não possui nenhum poder preditivo.

3. Avaliação dos melhores representantes

Seguindo com a estratégia de comparação, agora treinamos os melhores representantes de cada modelo no conjunto de treinamento *completo* para logo avaliar o desempenho no conjunto de teste, obtendo assim os resultados da Tabela 35, os quais concordam com os resultados médios no conjunto de treinamento da Tabela 34, sugerindo o mesmo comportamento relativo entre modelos descrito no passo 2.

#	Modelo	hp_1	hp_2	hp_3	F_1	acurácia	precisão	recall
1	k-NN	$k = 7$	-	-	**0.7694**	**0.9764**	0.8554	**0.6992**
2	$\mathcal{L}_k^{\Phi_p}$	$p = 2$	$k = 5$	$H = 12$	0.7586	0.9751	0.8360	0.6943
3	DT	gini	$H = 12$	best	0.6872	0.9697	0.8209	0.5909
4	RF	gini	$H = 12$	$n_e = 10^2$	0.6697	0.9702	**0.8912**	0.5363
5	LSVC	$C = 10^3$	l_2	$i_s = 1$	0.0	0.9436	0.0	0.0
6	LR	$C = 10$	l_2	liblinear	0.0001	0.9428	0.0020	0.0

Tabela 35 – Avaliação melhores representantes de cada modelo no Dataset 8

4. Teste de significância estatística

Agora, para realizar o teste de significância estatística, vamos considerar o conjunto de dados *completo* para fazer 5×20 validações cruzadas, obtendo o gráfico de frequências da Figura 22 e na Tabela 36 temos os resultados do *teste de significância* para cada par de modelos, onde os modelos LR e LSVC não participam por não possuir poder preditivo.

Modelo 1	Modelo 2	1 Pior que 2	1 Melhor que 2	1 Equivalente com 2.
$\mathcal{L}_k^{\Phi_p}$	RF	0.0	1.0	0.0
$\mathcal{L}_k^{\Phi_p}$	k-NN	0.2856	0.0	0.7144
$\mathcal{L}_k^{\Phi_p}$	DT	0.0	1.0	0.0
RF	k-NN	1.0	0.0	0.0
RF	DT	1.0	0.0	0.0
k-NN	DT	0.0	1.0	0.0

Tabela 36 – Probabilidades Dataset 8

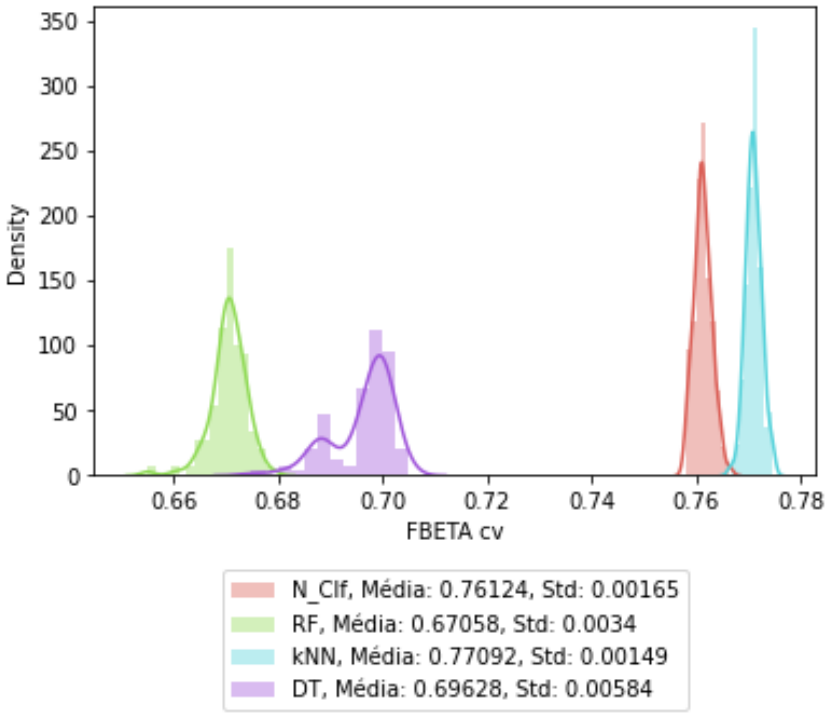

Figura 22 – Distribuição CV Dataset 8

5. Conclusões finais para o Dataset 8

Da Tabela 36 e da Figura 22, vemos que os modelos k-NN e $\mathcal{L}_k^{\Phi_p}$ possuem uma probabilidade de ser praticamente equivalentes igual a 0.7144, baixa demais para assumir equivalencia prática entre esses modelos, mas eles têm um desempenho marcadamente superior ao resto de classificadores com probabilidade 1.0. Também os resultados confirmam a tendência inicial de que os modelos LR e LSVC não possuem poder preditivo para esse Dataset.

6.2.9 Tempo versus tamanho da amostra

Finalizamos o capítulo com uma pequena análise do tempo de cálculo do algoritmo $\mathcal{L}_k^{\Phi_p}$, comparado com o tempo de cálculo dos modelos clássicos. Para esse fim, vamos medir o tempo de treinamento/teste de cada modelo em função do tamanho da amostra sem utilizar nenhuma técnica de aceleração, como por exemplo as opções de cálculo paralelo nas implementações dos modelos clássicos da biblioteca `sklearn`. Para detalhes sobre os hiperparâmetros utilizados em cada modelo, consultar o Apêndice A.2.

O conjunto de dados utilizados será uma amostra de 250000 instâncias do Dataset 5, e para obter curvas mais suaves, para cada subamostra $d_n \in (\mathbb{R}^{20} \times \{0,1\})^n$, com $n = 5000, 10000, \ldots, 250000$, serão feitos 5 processos de treinamento/teste, considerando 2/3 da subamostra para treinamento e o 1/3 restante para teste; obtendo o tempo médio para cada modelo, gerando assim a Figura 23, onde o novo classificador $\mathcal{L}_k^{\Phi_p}$ é denotado por **NC**. Da Figura 23, vemos que para dados de dimensão $d = 20$, para amostras de tamanho

a partir de $n = 60000$, o tempo relativo entre modelos se mantém estável e marcando uma clara diferença entre o modelo k-NN e os demais. Também vemos, que mesmo sendo um algortimo baseado em distâncias, o modelo $\mathcal{L}_k^{\Phi_p}$ atinge tempos de cálculo comparáveis com os modelos mais rápidos, representando assim uma alternativa a ser considerada para conjuntos de dados de porte médio/grande.

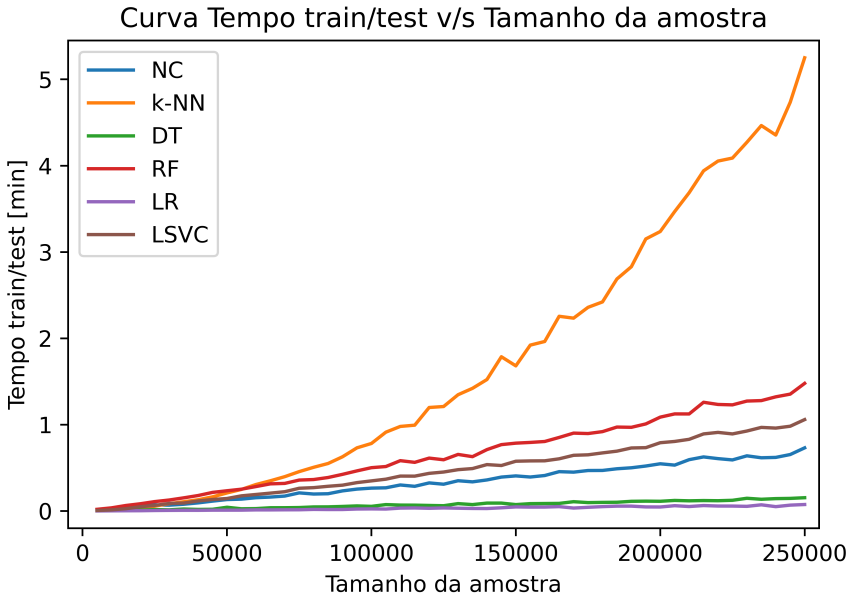

Figura 23 – Tempos de execução

Finalmente, na Tabela 37, temos uma lista das modestas características do notebook utilizado no desenvolvimento do protótipo da implementação do algoritmo $\mathcal{L}_k^{\Phi_p}$ e nos tempos de cálculo da Figura 23.

Características técnicas	
Fabricante	Samsung
Modelo	RF411
Ano	2011
Processador	Intel Core i5-2410
Memória RAM	8 GB DDR3-SDRAM
Disco SSD	120 GB
Sistema Operacional	Windows 7 Professional

Tabela 37 – Características técnicas do notebook utilizado para o protótipo

7 CONCLUSÕES E TRABALHOS FUTUROS

Nesse último capítulo, escrevemos algumas linhas sobre o fruto do presente trabalho em função dos objetivos traçados inicialmente, e também sobre as limitações próprias da pesquisa e as linhas de trabalho futuras.

Primeiro lembremos o objetivo principal dessa pesquisa:

"Desenvolver novos algoritmos de aprendizagem supervisionada não paramétrica no espaço euclidiano de dimensão finita, utilizando para tomar a decisão, a particular estrutura métrica dos números p-ádicos em lugar da estrutura métrica dos números reais".

Os objetivos secundários, que listamos a continuação, em seu conjunto procuram concretizar o objetivo principal da pesquisa, portanto, é mediante eles que escreveremos as conclusões do trabalho.

Dado que *aparentemente* a área à qual pertence essa pesquisa é absolutamente nova, as únicas ferramentas disponíveis para abordar a situação são a *teoria da aprendizagem estatística* e a análise não-arquimediana junto com a *análise p-ádica*. Assim, o primeiro passo foi examinar de perto a teoria de aprendizagem estatística em contextos mais gerais, (espaços métricos), processo que gerou os conteúdos do Capítulo 2, onde em particular foi visto com detalhes o conceito de *consistência universal* e também o principal resultado que permite *transportar* o problema de aprendizagem para outros espaços: a técnica de *redução boreliana de dimensionalidade*, obra de um dos orientadores [53]. Já no Capítulo 3, são analisados em detalhe os principais conceitos e resultados da análise não-arquimediana e da análise p-ádica[1], processo que permitiu compreender com clareza a particular estrutura topológica do espaço *ultramétrico* $(\mathbb{Q}_p, |\cdot|_p)$, e em particular, a estrutura de árvore associada ao subespaço $(\mathbb{Z}_p, |\cdot|_p)$.

Graças aos conhecimentos sobre a estrutura topológica dos *inteiros* p-ádicos adquiridos no estágio anterior, no seguinte estágio da investigação, foi utilizada essa estrutura para construir o primeiro algoritmo de classificação binária *não paramétrico* que trabalha com números p-ádicos, algoritmo que logo é estendido de maneira natural ao espaço ultramétrico $d \in \mathbb{N}$ dimensional $(\mathbb{Z}_p^d, \|\cdot\|_p)$. No Capítulo 4, além de ser registrada com detalhes dita construção, é também feita a construção de uma função injetora e boreliana, que denotamos por $\Phi_p : \mathbb{R}_+^d \to \mathbb{Q}_p^d$, que será a responsável de transportar nossos dados do domínio euclidiano para o *mundo p-ádico*, e assim poder usufruir das vantagens da estrutura de árvore do classificador p-ádico, que denotamos por $\mathcal{L}_{(k,p)}$, gerando assim, mediante a composição de uma aplicação com uma regra de aprendizagem; o classificador no espaço euclidiano que usa a estrutura métrica p-ádica para fazer a classificação: o classificador

[1] Os tópicos sobre análise não-arquimediana e números p-ádicos, em geral não são muito conhecidos pela comunidade científica, e até mesmo pela comunidade matemática em geral.

$\mathcal{L}_k^{\Phi_p} := \mathcal{L}_{(k,p)}^{\Phi_p}$. Utilizando mais uma vez a composição de uma regra de aprendizagem com uma aplicação, podem ser definidos os algoritmos de classificação p-ádicos, $\mathcal{L}_{(k,p)}^{T}$ e os algortimos de classificação no espaço euclidiano $\mathcal{L}_{(k,p)}^{T\circ\Phi_p}$, onde $T : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$ é uma aplicação linear (T contínua $\Rightarrow$ T boreliana) bijetora que de forma análoga ao caso real, pode ser representada por uma matriz de inteiros p-ádicos.

No seguinte estágio, que foi registrado no Capítulo 5, depois de obter uma expressão matemática para o algoritmo $\mathcal{L}_{(k,p)}$, expressão que permite definir uma regra de aprendizagem em qualquer espaço métrico e que denotamos por ${}^{+}k$-NN; é feito um estudo de consistência de dita regra, obtendo como resultado que a nova regra é universalmente consistente em uma classe de espaços métricos que contém o espaço $(\mathbb{Z}_p^d, \|\cdot\|_p)$ como um dos seus membros: a classe dos espaços métricos de dimensão σ-finita no sentido de *Nagata*. Esse resultado combinado com a técnica de redução boreliana de dimensionalidade, *implica* a consistência universal das regras $\mathcal{L}_{(k,p)}^{T\circ\Phi_p}$ no espaço euclidiano, e das regras $\mathcal{L}_{(k,p)}^{T}$ no espaço p-ádico, para $T : \mathbb{Z}_p^d \to \mathbb{Z}_p^d$ transformação linear bijetora. Somado ao anterior, também obtemos um resultado sobre a consistência da regra de aprendizagem dada pelo voto majoritário de classificadores binários definidos por regras de aprendizagem que pertencem a uma família consistente, resultado que permite estabelecer a consistência das *Florestas p-ádicas Aleatórias*.

Finalmente, constatada a consistência universal dos novos algoritmos desenvolvidos, no Capítulo 6, testes numéricos são realizados com o novo classificador $\mathcal{L}_k^{\Phi_p}$. Em total, foram realizados testes ao longo de 8 conjuntos de dados. O desempenho do modelo $\mathcal{L}_k^{\Phi_p}$ varia de conjunto a conjunto, mas de modo geral, nos casos onde o desempenho do novo classificador é fraco em comparação ao resto, seu desempenho não se afasta significativamente do desempenho dos modelos concorrentes, e em outros casos, o desempenho do novo modelo fica nos primeiros lugares do ranking, com performance comparável à do modelo k-NN, representando uma nova alternativa a ser considerada para resolver problemas de classificação binária.

Para finalizar, vamos listar alguns trabalhos para serem realizados com os algoritmos num futuro próximo.

1. Implementar de forma *profissional* a versão *multiclasse* do modelo $\mathcal{L}_k^{\Phi_p}$, comparando seu desempenho em variados conjuntos de dados com o desempenho de modelos clássicos, incluindo as redes neurais.

2. Implementação das *florestas p-ádicas aleatórias* do Capítulo 5, para a qual é necessária a implementação computacional das operações aritméticas dos números p-ádicos e também experimentar com a escolha aleatória das coordenadas ou varíaveis do conjunto de dados para construir as árvores, de maneira similar a como é feito na Floresta Aleatória.

3. Abordar o *procedimento de atualização dinâmica* ou *dynamic update procedure*, para adicionar novos pontos de dados na árvore p-ádica, e fazer a árvore crescer para acomodá-los e assim evitar construir uma nova árvore para incluir os novos pontos.

4. Estudar a influência da dimensão dos dados no desempenho do classificador $\mathcal{L}_k^{\Phi_p}$ desde o ponto de vista teórico e prático.

5. No âmbito da comparação de modelos e considerando uma família de conjuntos de dados, para cada um desses conjuntos, determinar qual é o subconjunto de instâncias onde cada modelo da comparação erra e determinar se existe correlação desses subconjuntos com as instâncias onde erra o classificador $\mathcal{L}_k^{\Phi_p}$. O caso de uma baixa correlação, sugere que o classificador $\mathcal{L}_k^{\Phi_p}$ pode errar de uma maneira diferente que os modelos clássicos. Essa informação pode ser muito valiosa na prática, em particular para construir classificadores do tipo ensemble cujo desempenho seja melhor que o desempenho de cada classificador individualmente. Fazer a mesma análise utilizando conjuntos de dados *simulados* com diferentes fronteiras entre as classes (linear, não linear, etc.), para tentar dilucidar os padrões que o classificador $\mathcal{L}_k^{\Phi_p}$ consegue capturar melhor.

6. Ainda no âmbito da comparação de modelos, estudar quais são as características dos conjuntos de dados onde o desempenho do classificador $\mathcal{L}_k^{\Phi_p}$ é melhor que o desempenho dos outros modelos.

7. Estudar desde o ponto de vista teórico como escolher o valor de $k \in \mathbb{N}$ para obter o melhor desempenho do classificador $\mathcal{L}_k^{\Phi_p}$ para um determinado conjunto de dados. Um estudo de características similares no contexto do classificador k-NN, é feito em [39].

Finalmente, esperamos converter esse trabalho num artigo científico no futuro próximo.

REFERÊNCIAS

[1] Steven Abney. *Semisupervised learning for computational linguistics*. CRC Press, 2007.

[2] Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balazs Kegl e David Rousseau. "Learning to discover: the higgs boson machine learning challenge". *In*: *URL http://higgsml. lal. in2p3. fr/documentation* 9 (2014).

[3] A. A. Aguilar-Arevalo *et al.* "Significant Excess of Electronlike Events in the MiniBooNE Short-Baseline Neutrino Experiment". *In*: *Physical Review Letters* 121.22 (nov. de 2018).

[4] Patrice Assouad e Thierry Quentin de Gromard. "Recouvrements, derivation des mesures et dimensions". *In*: *Revista matemática iberoamericana* 22.3 (2006), pp. 893–953.

[5] Alessio Benavoli, Giorgio Corani, Janez Demšar e Marco Zaffalon. "Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis". *In*: *The Journal of Machine Learning Research* 18.1 (2017), pp. 2653–2688.

[6] A. S. Besicovitch. "A general form of the covering principle and relative differentiation of additive functions". *In*: *Mathematical Proceedings of the Cambridge Philosophical Society* 41.2 (1945), pp. 103–110.

[7] Gérard Biau, Luc Devroye e Gäbor Lugosi. "Consistency of random forests and other averaging classifiers." *In*: *Journal of Machine Learning Research* 9.9 (2008).

[8] Gérard Biau e Erwan Scornet. "A random forest guided tour". *In*: *Test* 25.2 (2016), pp. 197–227.

[9] Patrick Billingsley. *Probability and measure*. 2rd ed. Wiley series in probability and mathematical statistics. Wiley, 1986.

[10] Christopher M Bishop e Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[11] Patrick Erik Bradley. "Mumford dendrograms". *In*: *The Computer Journal* 53.4 (2010), pp. 393–404.

[12] Patrick Erik Bradley. "On p-adic classification". *In*: *p-Adic Numbers, Ultrametric Analysis, and Applications* 1 (2009), pp. 271–285.

[13] L Breiman, J Friedman, R Olshen e C Stone. "Classification and Regression Trees. Belmont, Ca: Wadsworth and Brooks". *In*: *Statistics/Probability Series* (1984).

[14] Leo Breiman. "Random forests". *In*: *Machine learning* 45.1 (2001), pp. 5–32.

[15] Jason Brownlee. *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning*. Machine Learning Mastery, 2020.

[16] Xavier Caruso. "Computations with p-adic numbers". *In*: *Journées Nationales de Calcul Formel*. Les cours du CIRM. (2018).

[17] Frédéric Cérou e Arnaud Guyader. "Nearest neighbor classification in infinite dimension". *In*: *ESAIM: Probability and Statistics* 10 (2006), pp. 340–355.

[18] Benoıt Collins, Sushma Kumari e Vladimir G Pestov. "Universal consistency of the k-nn rule in metric spaces and nagata dimension". *In*: *ESAIM: Probability and Statistics* 24 (2020), pp. 914–934.

[19] Diane J Cook, Aaron S Crandall, Brian L Thomas e Narayanan C Krishnan. "CASAS: A smart home in a box". *In*: *Computer* 46.7 (2012), pp. 62–69.

[20] Diane J Cook, Narayanan C Krishnan e Parisa Rashidi. "Activity discovery and activity recognition: A new partnership". *In*: *IEEE transactions on cybernetics* 43.3 (2013), pp. 820–828.

[21] Corinna Cortes e Vladimir Vapnik. "Support-vector networks". *In*: *Machine learning* 20 (1995), pp. 273–297.

[22] T. Cover e P. Hart. "Nearest neighbor pattern classification". *In*: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27.

[23] Luc Devroye. "On the almost everywhere convergence of nonparametric regression function estimates". *In*: *The Annals of Statistics* (1981), pp. 1310–1319.

[24] Luc Devroye, László Györfi e Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Vol. 31. Springer Science & Business Media, 1996.

[25] Branko Dragovich e A Yu Dragovich. "A p-adic model of DNA sequence and genetic code". *In*: *P-Adic Numbers, Ultrametric Analysis, and Applications* 1 (2009), pp. 34–41.

[26] Hubert Haoyang Duan. *Applying Supervised Learning Algorithms and a New Feature Selection Method to Predict Coronary Artery Disease*. (2014). arXiv: 1402.0459 [cs.LG].

[27] Alberto Fernández, Salvador Garcıa, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk e Francisco Herrera. *Learning from imbalanced data sets*. Vol. 10. Springer, 2018.

[28] Gerald B Folland. *Real analysis: modern techniques and their applications*. Vol. 40. John Wiley & Sons, 1999.

[29] Fernando Quadros Gouvêa. *Primeiros passos p-ádicos*. IMPA, 1989.

[30] Alfred Haar. "Der Massbegriff in der Theorie der kontinuierlichen Gruppen". *In*: *Annals of mathematics* (1933), pp. 147–169.

[31] Paul R. Halmos. *Measure Theory*. Graduate Texts in Mathematics, 18. Springer, 1974.

[32] Trevor Hastie, Robert Tibshirani, Jerome H Friedman e Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[33] Haibo He e Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[34] Israel N Herstein. *Abstract algebra*. John Wiley & Sons, 1996.

[35] L. P. Kaelbling, M. L. Littman e A. W. Moore. *Reinforcement Learning: A Survey*. (1996). arXiv: cs/9605103 [cs.AI].

[36] Svetlana Katok. *p-adic Analysis Compared with Real*. Vol. 37. American Mathematical Soc., 2007.

[37] Alexander Kechris. *Classical descriptive set theory*. Vol. 156. Springer-Verlag, 1995.

[38] Andrei Y Khrennikov e Marcus Nilsson. *P-adic deterministic and random dynamics*. Vol. 574. Springer Science & Business Media, 2004.

[39] Samory Kpotufe. *Escaping the curse of dimensionality with a tree-based regressor*. (2009). arXiv: 0902.3453 [stat.ML].

[40] Sushma Kumari. *Topics in Random Matrices and Statistical Machine Learning*. (2018). arXiv: 1807.09419 [stat.ML].

[41] Sushma Kumari e Vladimir G. Pestov. *Universal consistency of the k-NN rule in metric spaces and Nagata dimension. II*. (2023). arXiv: 2305.17282 [cs.LG].

[42] Henri Leon Lebesgue. *Leçons sur l'intégration et la recherche des fonctions primitives professées au Collège de France*. 1ª ed. Cambridge Library Collection - Mathematics. Cambridge University Press, 2009.

[43] R Lyon. *Htru2, 2016,[online] Available: https://figshare. com/articles/dataset*. Rel. técn. HTRU2/3080389/1.

[44] Robert J Lyon, BW Stappers, Sally Cooper, John Martin Brooke e Joshua D Knowles. "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach". *In*: *Monthly Notices of the Royal Astronomical Society* 459.1 (2016), pp. 1104–1123.

[45] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[46] Claude Nadeau e Yoshua Bengio. "Inference for the generalization error". *In*: *Advances in neural information processing systems* 12 (1999).

[47] Jun-Iti Nagata. "Open problems left in my wake of research". *In*: *Topology and its Applications* 146 (2005), pp. 5–13.

[48] Jun-iti Nagata. "On a special metric and dimension". *In*: *Fundamenta Mathematicae* 2.55 (1964), pp. 181–194.

[49] Phillip A Ostrand. "A conjecture of J. Nagata on dimension and metrization". *In*: (1965).

[50] Vladimir Pestov. "A universally consistent learning rule with a universally monotone error". *In*: *Journal of Machine Learning Research* 23.157 (2022), pp. 1–27.

[51] Vladimir G. Pestov. *A learning problem whose consistency is equivalent to the non-existence of real-valued measurable cardinals*. (2020). arXiv: 2005.01886 [cs.LG].

[52] Vladimir G. Pestov. *Elementos da Teoria de Aprendizagem de Máquina Supervisionada*. Editora do IMPA, 2019.

[53] Vladimir G. Pestov. "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?" *In*: *Computers and Mathematics with Applications* 65.10 (2013), pp. 1427–1437.

[54] David Preiss. "Dimension of metrics and differentiation of measures". *In*: *General topology and its relations to modern analysis and algebra, V (Prague, 1981)* 3 (1983), pp. 565–568.

[55] Halsey Lawrence Royden e Patrick Fitzpatrick. *Real analysis*. Vol. 32. Macmillan New York, 1988.

[56] Wilhelmus Hendricus Schikhof. *Ultrametric calculus: An introduction to p-adic analysis*. Vol. 4. Cambridge University Press, 2007.

[57] Shai Shalev-Shwartz e Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[58] Marina Sokolova e Guy Lapalme. "A systematic analysis of performance measures for classification tasks". *In*: *Information processing & management* 45.4 (2009), pp. 427–437.

[59] Charles J Stone. "Consistent nonparametric regression". *In*: *The annals of statistics* (1977), pp. 595–620.

[60] Gary M Weiss. "Wisdm smartphone and smartwatch activity and biometrics dataset". *In*: *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set* 7 (2019), pp. 133190–133202.

APÊNDICE A – RESULTADOS AUXILIARES E DETALHES TÉCNICOS DA IMPLEMENTAÇÃO COMPUTACIONAL

A.1 RESULTADOS AUXILIARES

Proposição A.1.1. *Sejam (X, τ_X) e (Y, τ_Y), espaços topológicos e suponha que τ_Y possua uma base enumerável U. Então $\mathscr{B}_Y = \sigma(U)$ e $f : X \to Y$ é Borel mensurável, se e somente se, $f^{-1}(O) \in \mathscr{B}_X$, para todo $O \in U$.*

Demonstração. Como U é base enumerável de τ_Y, temos que $\tau_Y \subset \sigma(U)$, de onde $\mathscr{B}_Y \subset \sigma(U)$. A outra inclusão é imediata, logo $\mathscr{B}_Y = \sigma(U)$.

($\Rightarrow$) É imediato, pois $U \subset \mathscr{B}_Y$.

($\Leftarrow$) A família $\Gamma = \{E \subset Y : f^{-1}(E) \in \mathscr{B}_X\}$ é uma sigma álgebra de subconjuntos de Y que contem a base U, logo $\mathscr{B}_Y = \sigma(U) \subset \Gamma$ e assim, para todo $E \in \mathscr{B}_Y$, $f^{-1}(E) \in \mathscr{B}_X$, i.é, f é Borel mensurável. ■

Proposição A.1.2. *Sejam $(X_i, \tau_i), i \in [n]$, espaços topológicos e considere $X = \prod_{i=1}^n X_i$ o espaço produto, (X, τ_X) com τ_X a topologia produto, i.é, $\tau_X = \tau(\mathcal{G})$, onde*

$$\mathcal{G} := \{\pi_i^{-1}(O_i) : O_i \in \tau_i, i \in [n]\},$$

então

$$\otimes_{i=1}^n \mathscr{B}_{X_i} \subset \mathscr{B}_X.$$

Ainda, se cada $X_i, i \in [n]$ possui uma base enumerável para τ_i, então

$$\otimes_{i=1}^n \mathscr{B}_{X_i} = \mathscr{B}_X.$$

Corolário A.1.0.1. *Se $(X_i, \rho_i), i \in [n]$ são espaços métricos separáveis, então*

$$\otimes_{i=1}^n \mathscr{B}_{X_i} = \mathscr{B}_X,$$

onde $\mathscr{B}_{X_i} = \sigma(\tau_i)$ e τ_i é a topologia induzida por ρ_i, para todo $i \in [n]$.

Demonstração. Por um resultado clássico de análise, temos que para $i \in [n]$, (X_i, ρ_i) é separável, se e somente se, τ_i possui uma base enumerável. O resultado segue da Proposição A.1.2. ■

Proposição A.1.3. *Sejam (X, σ_X), $(Y_\alpha, \sigma_\alpha)$, $\alpha \in A$ e (Y, σ_Y), espaços mensuráveis, onde $Y = \prod_{\alpha \in A} Y_\alpha$ e $\sigma_Y = \otimes_{\alpha \in A} \sigma_\alpha$. Se $\pi_\alpha : Y \to Y_\alpha$, $\alpha \in A$ são as projeções canônicas, então $f : X \to Y$ é (σ_X, σ_Y)-mensurável, se e somente se, $f_\alpha := \pi_\alpha \circ f$ é $(\sigma_X, \sigma_\alpha)$-mensurável, $\forall \alpha \in A$.*

Demonstração. ($\Rightarrow$) Se f é (σ_X, σ_Y)-mensurável, então f_α é a composição de funções mensuráveis. ($\Leftarrow$) A Proposição A.1.1 também é valida si trocamos $\mathscr{B}_X$ por σ_X e $\mathscr{B}_Y$ por

$\sigma(U)$ com $U \subset 2^Y$, nesse caso, f é $(\sigma_X, \sigma(U))$-mensurável, se e somente se, $f^{-1}(E) \in \sigma_X$, para todo $E \in U$. Logo, se toda f_α é mensurável, temos, para todo $\alpha \in A$ e para todo $E_\alpha \in \sigma_\alpha$, $f^{-1}(\pi_\alpha^{-1}(E_\alpha)) = f_\alpha^{-1}(E_\alpha) \in \sigma_X$, onde $\sigma_Y = \otimes_{\alpha \in A}\sigma_\alpha = \sigma(\mathcal{C})$ e $\mathcal{C} = \{\pi_\alpha^{-1}(E_\alpha) : E_\alpha \in \sigma_\alpha, \alpha \in A\}$, logo, f é (σ_X, σ_Y)-mensurável. ■

A.2 DADOS TÉCNICOS SOBRE A IMPLEMENTAÇÃO

Esse apêndice, contém uma lista com as bibliotecas de Python 3 utilizadas na implementação do algoritmo $\mathcal{L}_k^{\Phi_p}$ e uma lista de hiperparâmetros dos modelos clássicos utilizados na comparação do Capítulo 6, assim como os valores dos hiperparâmetros usados em cada modelo para comparar os tempos de execução da Seção 6.2.9.

Na Tabela 38, temos os módulos ou bibliotecas utilizados na implementação do novo algortimo e para exibir resultados.

Módulo Python	**Versão**
`matplotlib`	`3.4.3`
`numba`	`0.53.1`
`numpy`	`1.21.1`
`pandas`	`1.3.1`
`pyarrow`	`8.0.0`
`scikit-learn`	`1.0.2`

Tabela 38 – Módulos Python utilizados na implementação

Na Tabela 39, temos uma lista dos hiperparâmetros que foram considerados para cada modelo no processo de otimização do Capítulo 6. Foram selecionados apenas os hiperparâmetros mais relevantes, deixando os outros hiperparâmetros próprios da implementação dos modelos em seus valores por defeito. Os valores possíveis considerados para cada hiperparâmetro dependem do tamanho de cada Dataset e foram escolhidos de tal forma de não gerar tempos de cálculo excessivos.

Modelo	**Hiperparâmetros**
$\mathcal{L}_k^{\Phi_p}$	k,p,H
k-NN	k
LSVC	C, `intercept_scaling`,`penalty`
LR	C,`solver`, `penalty`
DT	`max_depth`, `criterion`,`splitter`
RF	`max_depth`, `n_estimators`,`criterion`

Tabela 39 – Hiperparâmetros dos modelos clássicos

Finalmente, na Tabela 40, temos os valores dos hiperparâmetros utilizados em cada modelo na medição dos tempos de cálculo sobre uma subamostra do Dataset 5 da Seção 6.2.9.

Modelo	**Hiperparâmetros**
$\mathcal{L}_k^{\Phi_p}$	$k = 1$,$p = 2$,$H = 10$
k-NN	$k = 1$
LSVC	$C = 10$, `intercept_scaling`=1,`penalty`=$l2$
LR	$C = 1000$,`solver` = $liblinear$, `penalty` = $l2$
DT	`max_depth` = 10, `criterion`=$entropy$,`splitter` = $best$
RF	`max_depth` = 10, `n_estimators` = 100,`criterion`=$gini$

Tabela 40 – Valores dos hiperparâmetros para medir os tempos de cálculo