UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CIÊNCIA DA COMPUTAÇÃO

Alek Fröhlich

Fundamentos de Aprendizado de Máquina: Análise Funcional voltada a Métodos de Kernel

Florianópolis
2022

Alek Fröhlich

Fundamentos de Aprendizado de Máquina: Análise Funcional voltada a Métodos de Kernel

Trabalho de Conclusão de Curso submetido ao Curso de Graduação em Ciência da Computação do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para obtenção do título de Bacharel em Ciência da Computação.
Orientador: Prof. Danilo Royer, Dr.
Coorientador: Prof. Mauro Roisenberg, Dr.

Florianópolis
2022

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

Fröhlich , Alek
Fundamentos de Aprendizado de Máquina : Análise Funcional voltada a Métodos de Kernel / Alek Fröhlich ; orientador, Danilo Royer, coorientador, Mauro Roisenberg, 2022.
107 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2022.

Inclui referências.

1. Ciências da Computação. 2. Espaços de Hilbert de Reprodução. 3. Análise Funcional. 4. Aprendizado de Máquina. 5. Métodos de Kernel. I. Royer, Danilo. II. Roisenberg, Mauro. III. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. IV. Título.

Alek Fröhlich

Fundamentos de Aprendizado de Máquina:
Análise Funcional voltada a Métodos de Kernel

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo curso de Graduação em Ciência da Computação.

Florianópolis, 23 de Março de 2022.

Prof. Jean Martina, Dr.
Coordenador do Curso

Banca Examinadora:

Prof. Danilo Royer, Dr.
Orientador
Universidade Federal de Santa Catarina

Prof. Mauro Roisenberg, Dr.
Coorientador
Universidade Federal de Santa Catarina

Prof. Luciano Bedin, Dr.
Avaliador
Universidade Federal de Santa Catarina

Dedico este trabalho ao Guto e à Lu.

AGRADECIMENTOS

Gostaria de agradecer a meus pais por me darem o suporte necessário para seguir minhas ambições e meu interesse relativamente tardio por áreas teóricas da Computação e da Matemática. Gostaria de agradecer ao meu orientador Danilo e a meu coorientador Mauro pela enorme liberdade que me deram ao longo do desenvolvimento deste trabalho. Gostaria também de agradecer as pessoas com quem tive contato no Labsec, pois foi uma época bem divertida e produtiva da minha graduação na Computação. Em particular, gostaria de agradecer ao Jean, ao Custódio e ao Zatta.

Por fim, gostaria de agradecer aos meus amigos Leo, Mosi e Jota, por me ensinarem aspectos muito importantes da vida. Em especial, gostaria de agradecer ao Mosi por despertar minha curiosidade, ao Leo por repetidas vezes me ensinar como ser um ser humano melhor e ao Jota por me inspirar a curtir a vida. Obviamente, há várias outras pessoas que mereciam ser citadas aqui. A todas essas pessoas: obrigado por tudo!

"Wir müssen wissen – wir werden wissen."
(HILBERT, 1930)

RESUMO

O uso de kernels é um dos principais paradigmas de Aprendizado de Máquina. Métodos de Kernel estão naturalmente associado a espaços de Hilbert de Reprodução (EHR) e Análise Funcional por meio do Teorema de Moore-Aronszajn. Neste trabalho, apresentamos os elementos iniciais da Análise Funcional e então os usamos para investigar a natureza desses espaços de funções. Ao final, abordamos um método de kernel e provamos sua corretude.

Palavras-chave: Espaços de Hilbert de Reprodução. Análise Funcional. Aprendizado de Máquina. Métodos de Kernel.

ABSTRACT

Kernel Methods are one of the main paradigms in Machine Learning. The symmetric positive-definite kernels employed by these methods are naturally associated with Reproducing Kernel Hilbert Spaces via the celebrated Moore-Aronszajn Theorem. In this work, we present elementary concepts in Functional Analysis so that we can explore the underlying nature of these function spaces. By the end, we illustrate the Kernel Methodology via a simple kernel method.

Key-words: Reproducing Kernel Hilbert Spaces. Functional Analysis. Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

O Aprendizado de Máquina é um dos campos mais aquecidos da ciência (UNION, 2022). De fato, é difícil encontrar uma área do conhecimento humano que não tenha sido afetada pela revolução da Inteligência Artificial. Dentro do contexto de Aprendizado de Máquina, os Métodos de Kernel vem sendo aplicados em diversas áreas e em diversos problemas. Para listar alguns, Métodos de Kernel vem sendo empregados no problema de classificação de proteínas (LESLIE; ESKIN; NOBLE, 2001), na identificação de doenças na Tireoide (SHANKAR et al., 2020) e na classificação de aplicações multimídia (MORENO; HO; VASCONCELOS, 2003).

A troca de representação de conjuntos de dados é uma técnica utilizada no Aprendizado de Máquina. Por certo, é muito comum que Métodos de Redução de Dimensionalidade e Seleção de Atributos como Análise de Componentes Principais (Principal Component Analysis - PCA em inglês) sejam aplicados em uma etapa de pré-processamento em fluxos de trabalho de Aprendizado de Máquina (AM) (NAVOT, 2006). Para alguns algoritmos conhecidos como Métodos de Kernel, há disponível uma técnica de troca de representação poderosa: o Truque do Kernel. Para estes algoritmos, a única informação relevante do conjunto de dados é o valor do produto interno entre pares de vetores de treino. Isto é, a quantidade: $\langle x_i, x_j \rangle$. Por consequência, qualquer transformação $\phi : X \to H$ aplicada sobre o conjunto de dados só aparecerá em expressões da forma: $\langle \phi(x_i), \phi(x_j) \rangle$. É um fato interessante que conseguimos encontrar funções da forma $K(x,y) = \langle \phi(x), \phi(y) \rangle$, onde ϕ e H estão implícitos, que são úteis para a resolução de problemas de Aprendizado de Máquina. Um exemplo é o Kernel Gaussiano, que é altamente usado em conexão com algoritmos como Kernel-SVM e Kernel-PCA:

$$K(x,y) = e^{-\gamma \|x-y\|^2}.$$

Como esses algoritmos dependem do uso do produto interno, sabemos que a imagem H de ϕ é um espaço com produto interno (ou pré-espaço de Hilbert, como nos referiremos no texto). Isso motiva a pergunta: qual seria a natureza desses espaços? É uma consequência do Teorema de Moore-Aronszajn, que H é um espaço de Hilbert de Reprodução (EHR), cuja definição pode ser vista a seguir:

Definição 1.1. Dizemos que um espaço de Hilbert H é um espaço de Hilbert de Reprodução se for um espaço de funções $f : X \to \mathbb{R}$ e todo funcional avaliação $T_x : H \to \mathbb{R}$ dado por $T_X(f) = f(x)$ for contínuo.

Esses espaços foram primeiramente estudados no contexto de Análise Harmônica por Stanisław Zaremba (ZAREMBA, 1907) e, simultaneamente, no contexto de Equações Integrais por James Mercer (MERCER, 1909). Eventualmente, o assunto foi sistematicamente desenvolvido por Nachman Aronszajn e Stefan Bergman (ARONSZAJN, 1950). É possível se dizer que a relação de kernels de reprodução com Aprendizado de Máquina começou com a descoberta

de que o Perceptron não era capaz de classificar conjuntos de dados não linearmente separáveis (MINSKY, 1969). De fato, o Kernel-Perceptron foi o primeiro algoritmo de classificação a ser adaptado para o uso de kernels (AIZERMAN, 1964). Dentre os vários usuários de kernels, o algoritmo mais proeminente é o Kernel-SVM, introduzido por Vapnik na década de 90 (BOSER; GUYON; VAPNIK, 1992). De forma similar ao desenvolvimento sistemático da teoria de kernels de reprodução, demorou um pouco para que os Métodos de Kernels fossem desenvolvidos sistematicamente. Neste caso, a figura que mais se destaca é Bernhard Schölkopf, autor de uma das principais referências na área (SCHLKOPF, 2018).

1.1 OBJETIVOS

O objetivo principal deste trabalho é esclarecer a natureza dos espaços de Hilbert de Reprodução e sua conexão com os Métodos de Kernel. Ao longo do texto, vamos desenvolver ferramentas que nos permitirão compreender melhor as consequências da definição altamente técnica dos EHR. A saber, vamos discutir noções alternativas de base para um espaço vetorial, veremos como noções analíticas de convergência e continuidade se manifestam na teoria dos espaços normados de dimensão infinita e abordaremos ortogonalidade em espaços de Hilbert. De maneira específica, este trabalho almeja:

- Apresentar elementos de Análise Funcional relevantes para a análise dos Métodos de Kernel.
- Motivar o uso de Métodos de Kernel e do Truque do Kernel através do problema de Classificação Binária.
- Adaptar o Perceptron para o uso de kernels. Isso incluí a demonstração de sua corretude em conexão ao uso de kernels.
- Introduzir uma caracterização equivalente para uma função ser um kernel em termos de funções simétricas e positivo-definidas.
- Apresentar de forma rigorosa alguns exemplos de kernels importantes para a prática de Aprendizado de Máquina.
- Demonstrar o Teorema de Moore-Aronszajn e estabelecer a relevância dos espaços de Hilbert de Reprodução para a teoria dos Métodos de Kernel.

1.2 ORGANIZAÇÃO DO TEXTO

O texto está organizado da seguinte maneira: no Capítulo 2, introduzimos espaços normados e discutimos as peculiaridades encontradas em espaços vetoriais de dimensão infinita. Em particular, vemos que o conceito usual de base é (em grande parte) insatisfatório em

dimensão infinita. Discutimos também como noções analíticas de convergência e topologia se fazem necessárias para o estudo adequado desses espaços. Finalmente, abordamos operadores e funcionais sobre esses espaços.

O Capítulo 3 aborda espaços de Hilbert, que são uma classe especial de espaços normados. A teoria dos espaços de Hilbert é uma das áreas mais belas da Análise Funcional, pois grande parte dos resultados são intuitivos e de fácil compreensão. De fato, é possível se dizer que os espaços de Hilbert são a melhor generalização que se conhece do típico $\mathbb{R}^n$ ou $\mathbb{C}^n$. Neste texto, nos concentraremos em espaços vetoriais reais, pois são os espaços relevantes para o Aprendizado de Máquina.

No Capítulo 4, introduzimos brevemente o Aprendizado Supervisionado e, então, o Perceptron Binário. Incluímos também uma demonstração de sua corretude com respeito a qualquer espaço de Hilbert. Por fim, munidos do conhecimento dos capítulos anteriores, abarcamos espaços de Hilbert de Reprodução e o Teorema de Moore-Aronszajn. Terminamos o texto com uma conclusão que reflete sobre o que foi abordado e discute continuações naturais deste trabalho.

2 ESPAÇOS NORMADOS

> Mathematical objects are determined by, and understood by, the network of relationships they enjoy with all the other objects of their species.
>
> *Barry Mazur*

Neste capítulo, abordamos espaços normados[1]. Esses espaços nada mais são do que espaços vetoriais com a noção adicional de norma (tamanho) de vetor. Essa discussão é relevante pois todo espaço de Hilbert é também um espaço normado. Assim, conseguimos primeiro nos concentrar em aspectos relacionados à norma. Outra razão que nos motiva a apresentar espaços normados, ao invés de seguir diretamente para espaços de Hilbert, é que espaços normados são a linguagem "natural" da Análise Funcional[2]. A relação entre os diversos tipos de espaços estudados em Análise Funcional é ilustrada na Figura 1. Dito isso, a partir de agora, pensaremos em Análise Funcional como o estudo de espaços normados e operadores entre esses espaços[3]. Vamos começar relembrando a definição de espaço vetorial.

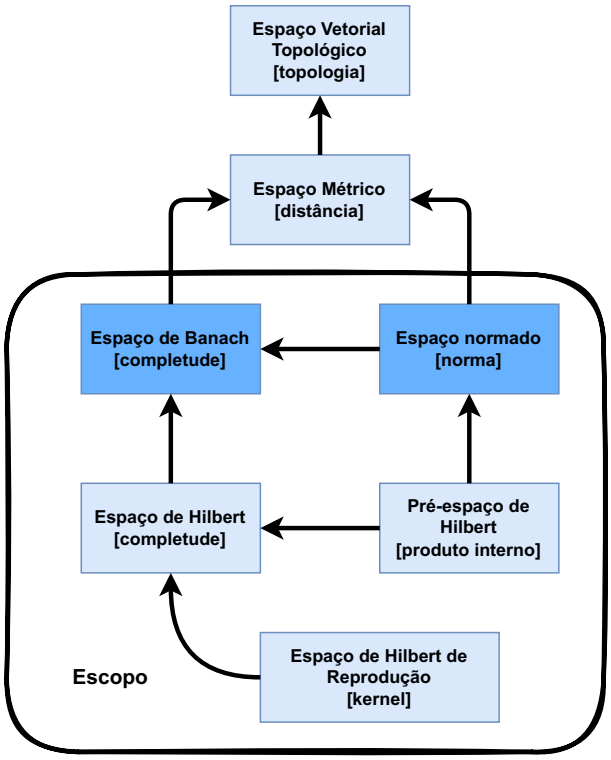


Figura 1 – Relação "é um tipo de" entre diferentes tipos de espaços estudados em Análise Funcional.

[1] O texto pressupõe familiaridade com Álgebra Linear e Análise Real.

[2] Pelo menos em sua versão elementar. Em abordagens mais aprofundadas, é comum se deparar com o conceito de espaço vetorial topológico, que é uma generalização dos espaços normados.

[3] Em Análise Funcional, é comum se referir a transformações lineares como operadores.

Definição 2.1. Dado X um conjunto. Se tivermos duas operações[4] $+ : X \times X \to X$ e $\cdot : \mathbb{R} \times X \to X$ definidas de forma que as seguintes regras sejam satisfeitas, então $V = (X, +, \cdot)$ é um espaço vetorial (real).

1. **Comutatividade da soma:** $\forall u, v : u + v = v + u$.
2. **Associatividade da soma:** $\forall u, v, w : (u + v) + w = u + (v + w)$.
3. **Identidade da soma:** Existe um vetor nulo 0 tal que $\forall u : 0 + u = u$.
4. **Inversa da soma:** Para todo vetor u, existe um inverso $-u$, tal que $u + (-u) = 0$.
5. **Identidade do produto:** Existe um escalar neutro 1 tal que $\forall u : 1 \cdot u = u$.
6. **Compatibilidade:** A ordem que multiplicamos dois escalares não importa: podemos primeiro multiplicá-los em $\mathbb{R}$ e depois usar o resultado para escalar o vetor, ou podemos fazer duas multiplicações por escalar em sequência; o resultado será o mesmo. Formalmente, $\forall \lambda, \mu, u : (\lambda \mu) \cdot u = \lambda \cdot (\mu \cdot u)$.
7. **Distributividade 1:** $\forall \lambda, u, v : \lambda \cdot (u + v) = \lambda \cdot u + \lambda \cdot v$.
8. **Distributividade 2:** $\forall \lambda, \mu, u : (\lambda + \mu) \cdot u = \lambda \cdot u + \mu \cdot u$.

Os axiomas acima capturam a essência do que um conjunto necessita satisfazer para que possamos pensar nos seus elementos como vetores. Como veremos a seguir, os espaços vetoriais de interesse para Análise Funcional são tipicamente espaços de sequências e de funções, o que nos leva imediatamente para questão: como pensar em sequências e funções como vetores? Vamos começar com sequências. Seja $\mathbb{R}^{\mathbb{N}}$ o conjunto de todas as sequências de números reais. Note que $\mathbb{R}^{\mathbb{N}}$ é um espaço vetorial com soma e produto coordenada a coordenada. Podemos pensar nos elementos de $\mathbb{R}^{\mathbb{N}}$ como tuplas com infinitos índices, um para cada número natural; e.g., $x = (x_1, x_2, ...,)$. Essa visão é inclusive reforçada pela notação $\mathbb{R}^{\mathbb{N}}$. Assim, é evidente que sequências são vetores, a única diferença é que há infinitos eixos nesse sistema de coordenadas. Analogamente, podemos pensar no espaço vetorial das funções de $\mathbb{R}$ em $\mathbb{R}$, o $F(\mathbb{R}, \mathbb{R})$ como sendo o espaço das tuplas indexadas pelos números reais, o $\mathbb{R}^{\mathbb{R}}$. Assim, temos um índice para cada real: $(x_\alpha)_{\alpha \in \mathbb{R}}$ e o sistema de coordenadas possui incontáveis eixos. Seguindo essa maneira de pensar, os espaços de funções que veremos são todos subespaços de $\mathbb{R}^{\mathbb{R}}$. Note que podemos falar de vetores em qualquer número de coordenadas (há um espaço vetorial para cada cardinal[5]). Com efeito, o espaço vetorial $\mathbb{R}^{\kappa}$, em que κ é um cardinal, recebe

[4] Formalmente, não há diferença entre operação e função. Usamos a primeira notação por vantagens óbvias: imagine só usar $+(u, v)$ ao invés de $u + v$.

[5] Aprendemos na escola que um número cardinal serve para medir quantidades absolutas, ou seja, tamanhos de conjuntos. Os primeiros números cardinais são os naturais: $1, 2, ...$; como é possível construír uma hierarquia de conjuntos infinitos, um maior que o outro, é preciso estender os números naturais para que possamos representar o tamanho de conjuntos arbitrariamente grandes. Esses números recebem o nome de cardinais. Para mais informações, consulte (HALMOS, 2011).

o nome de espaço vetorial livre. O nome "livre" está associado ao fato de que não precisamos de nenhuma informação extra para construir esses espaços, basta determinar o tamanho da base.

É fato que em Álgebra Linear todo espaço vetorial possui uma base no seguinte sentido:

Definição 2.2. Seja V um espaço vetorial e E um conjunto de vetores linearmente independentes em V. Dizemos que E é uma base (de Hamel[6]) quando, para qualquer vetor v em V, há uma decomposição de v como combinação linear dos vetores de E:

$$v = x_1e_1 + x_2e_2 + ... + x_ne_n, \quad x_i \in \mathbb{R}, e_i \in E.$$

A prova desse resultado é trivial para espaços de dimensão finita. Com efeito, relembremos da seguinte definição:

Definição 2.3. Dizemos que um espaço vetorial V é de dimensão finita quando possui uma base (de Hamel) com um número finito de elementos.

O processo de determinação da dimensão de um espaço de dimensão finita tende a ser bem simples. De fato, considere $\mathbb{R}^n$. É imediata a intuição de que os vetores $\{(1,0,...,0)$, $(0,1,...,0)$, ..., $(0,0,...,1)\}$ formam uma base. Basta assim verificar as condições formais. Em um típico espaço de dimensão infinita (há excessões, é interessante voltar para esta questão após ver o espaço ℓ^0), a "base canônica" tende a não ser uma base (de Hamel). De fato, pense no espaço $\mathbb{R}^{\mathbb{N}}$ discutido anteriormente. A "base canônica" deste espaço é composta pelas sequências que são 1 na cordenada n e 0 em todas as outras; a extensão natural da base de $\mathbb{R}^n$. Todavia, não é possível representar os vetores de $\mathbb{R}^{\mathbb{N}}$ como combinações lineares finitas de vetores dessa "base" (o conjunto resultante seria ℓ^0). Precisaríamos de séries para isto. Como sabemos, a introdução de séries necessita de alguma noção de convergência. Isso e outras questões, como continuidade de operadores, faz com que a Análise Funcional seja um mix[7] de Álgebra com Análise (e Topologia). Dito isso, a existência de bases de Hamel para espaços vetoriais quaisquer passa a ser um fato não trivial, a ponto de necessitarmos de um argumento que nos convença de sua veracidade. O argumento que faremos precisará de um resultado fundamental de Teoria dos Conjuntos: o Lema de Zorn.

Lema de Zorn

> The Axiom of Choice is obviously true, the well-ordering principle obviously false, and who can tell about Zorn's lemma?
>
> Jerry Bonna

[6] Introduziremos outras noções de base. Assim, passaremos a nos referir às bases usuais como bases de Hamel.

[7] Note como essa situação se reflete na nomenclatura: Álgebra Linear e Análise Funcional. Ambas as áreas estudam espaços vetoriais e transformações lineares, mas os métodos empregados são bem diferentes.

Para apreciar o uso do Lema de Zorn, é importante entendermos o contexto em que está inserido. Desde o início, os matemáticos sempre estiveram preocupados com a validação de seus métodos e teorias. Isto é, preocupados com a pergunta "como garantir que o que estou fazendo está certo?". Os detalhes de como essa questão vem sendo abordada ao longo dos anos variam, mas uma coisa é invariante: o uso do Método Axiomático.

O Método Axiomático consiste em especificar uma coleção (tipicamente pequena) de premissas/axiomas e então investigar as consequências desses axiomas. Demorou um pouco para os matemáticos perceberem que a unificação da Matemática sobre um único[8] conjunto de axiomas é um grande auxílio à produtividade. De fato, até a revolução das teorias de conjunto do século XX, era comum que os diversos objetos da Matemática fossem tratados de maneira isolada. Isso não impedia totalmente a interação entre as diversas partes da Matemática, mas tornava grande parte das tentativas vagas e imprecisas.

Como é possível ver pela definição dada para espaços vetoriais, a linguagem unificadora da Matemática é a Teoria dos Conjuntos (tipicamente, a versão conhecida como Zermelo-Frankel-Choice, ZFC). Essa estruturação da Matemática em torno de uma teoria de conjuntos é muito similar a estruturação de ambientes de programação em torno de um sistema operacional. É possível viver longe da linguagem de baixo nível, e muitos matemáticos fazem isso. Contudo, às vezes é necessário fazer o uso de primitivas. É exatamente esse o caso com o Lema de Zorn. Vejamos as definições e o lema a seguir:

Definição 2.4. Dizemos que uma relação $\leq$ sobre um conjunto X é uma ordem parcial se possui as seguintes propriedades:

1. **Reflexividade:** $\forall x : x \leq x$.
2. **Anti-Simetria:** $\forall x, y : x \leq y$ e $y \leq x$ juntos implicam em $x = y$.
3. **Transitividade:** $\forall x, y, z : x \leq y$ e $y \leq z$ juntos implicam em $x \leq z$.

Observação. Note que a relação de inclusão $\subseteq$ sempre define uma relação de ordem parcial sobre uma coleção de conjuntos. Veja um exemplo na Figura 2.

Definição 2.5. Seja $\leq$ uma relação de ordem parcial sobre um conjunto X e seja L um subconjunto de X. Dizemos que L é uma cadeia se cada dois elementos de L são comparáveis. Isto é, se $\forall x, y \in L : x \leq y$ ou $y \leq x$.

Lema 2.1. (Lema de Zorn) Seja $\mathfrak{F}$ um conjunto parcialmente ordenado pela relação $\leq$. Se toda cadeia possui uma cota superior em $\mathfrak{F}$, então $\mathfrak{F}$ possui (pelo menos) um elemento maximal x_0 tal que não há outro x_1 em $\mathfrak{F}$ com $x_0 \leq x_1$.

[8] Atualmente, teorias de conjuntos não são mais os únicos fundamentos disponíveis para Matemática. De fato, áreas da Matemática com alto caráter algébrico tendem a se basear em Teoria das Categorias e áreas teóricas da Computação tendem a se basear em Teoria dos Tipos.

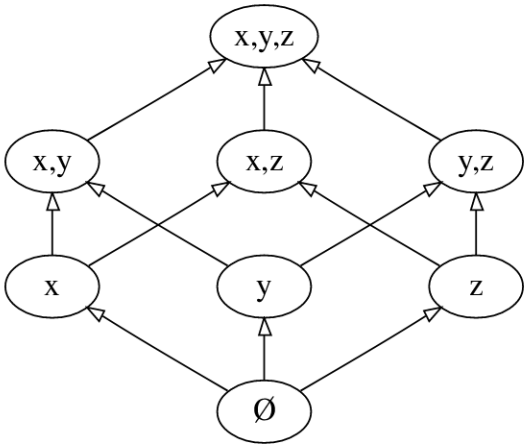


Figura 2 – Ordem parcial sobre o conjunto potência de $\{x,y,z\}$. Fonte: I, KSmrq, distribuida pela licença CC BY-SA 3.0.

Teorema 2.1. Todo espaço vetorial admite uma base de Hamel.

Demonstração. Seja V um espaço vetorial. Uma base em V nada mais é do que um subconjunto maximal de V cujos vetores são linearmente independentes. Assim, basta definir $\mathfrak{F}$ como sendo a coleção de subconjuntos L.I. em V, $\leq$ como $\subseteq$ e verificar que toda cadeia possui uma cota superior. Com efeito, seja $(L_i)_{i\in I}$ uma cadeia e tome $L = \bigcup_{i\in I} L_i$. Veja que os vetores de L continuam sendo linearmente independentes. De fato, se $a_1v_1 + a_2v_2 + ... + a_nv_n = 0$ e $v_1 \in L_1, v_2 \in L_2, ..., v_n \in L_n$, basta ordenar os L_j e escolher o maior de todos, vamos denotá-lo por L_k. Note que todos os vetores da combinação linear pertencem a L_k. Assim, a única escolha de escalares que pode gerar um vetor nulo é $a_1 = a_2 = ... = a_n = 0$. Podemos concluir que L pertence a $\mathfrak{F}$ e que, portanto, toda cadeia em $\mathfrak{F}$ possui uma cota superior também em $\mathfrak{F}$. Seja M um conjunto maximal de $\mathfrak{F}$, vamos verificar que M também é gerador. Com efeito, suponha que não o seja, então é possível pegar um vetor v_0 fora do span de M. Isso nos diz que $M' = M \cup \{v_0\}$ é um conjunto L.I., mas isso é um absurdo, pois implicaria $M' > M$. Assim, não há tal v_0 e M é uma base para V. □

Observação. É interessante notar que essa estratégia para mostrar a existência de uma base de Hamel já se faz necessária nos exemplos mais simples de espaços de dimensão infinita. De fato, não é possível construir uma base explícita para $\mathbb{R}^{\mathbb{N}}$. É necessário usar o Lema de Zorn. A situação é mais estranha do que parece, como pode ser visto a seguir:

Extra 1. Como em $\mathbb{R}^n$ há bases de Hamel canônicas compostas por tuplas de zeros e uns, gostaríamos de pensar que haveria uma base de Hamel para $\mathbb{R}^{\mathbb{N}}$ composta somente por sequências de zeros e uns. Contudo, não há tal base. É possível se convencer disso através de uma contradição; seria possível encontrá-la?

Voltaremos às questões analíticas mais à frente. Por ora, vejamos alguns exemplos.

2.1 EXEMPLOS DE ESPAÇOS VETORIAIS

Exemplo 2.1. $\mathbb{R}^n$. Esse espaço tem dimensão n e possui como base os vetores canônicos $\{(1,0,...,0),(0,1,0,...,0),...,(0,0,...,1)\}$.

Exemplo 2.2. $P_n(\mathbb{R})$, o conjunto dos polinômios de grau até n. Esse espaço tem dimensão $n+1$ e possui como base os polinômios $\{1,x,x^2,...,x^n\}$.

Exemplo 2.3. Em geral, soluções para equações lineares homogêneas tendem a dar origem a espaços vetoriais. Vejamos os seguintes exemplos:

1. Seja A uma matriz $n \times n$ e x um vetor $n \times 1$, então as soluções do sistema de equações $Ax = 0$ formam um subespaço de $\mathbb{R}^n$. Note que esse espaço é o núcleo da transformação linear correspondente à matriz.

2. Seja $\mathcal{F}$ o conjunto das funções que satisfazem a equação diferencial $a_0 f + a_1 f' + ... + a_n f^{(n)} = 0$, então $\mathcal{F}$ é um espaço vetorial. Novamente, esse conjunto é o núcleo de uma transformação linear; a saber, $Df \mapsto a_0 f + a_1 f' + ... + a_n f^{(n)}$.

Exemplo 2.4. Vejamos alguns espaços vetoriais de sequências. Os seguintes conjuntos são todos espaços vetoriais se considerarmos que a sequência $x+y$ é dada pela soma coordenada a coordenada: $x+y = (x_1+y_1, x_2+y_2, ...)$ e que a sequência $\lambda \cdot x$ é dada por $(\lambda x_1, \lambda x_2, ...)$.

Nome	Descrição
ℓ^0	Espaço das sequências que só são diferentes de zero em finitas posições.
ℓ^p	Espaços das sequências tais que $\sum_{i=1}^{\infty} \|x_i\|^p < \infty, 1 \leq p < \infty$.
ℓ^∞	Espaço das sequências limitadas.
c	Espaço das sequências convergentes.
c_0	Espaço das sequências que convergem para zero.

Exemplo 2.5. Agora, vejamos os principais exemplos de espaços vetoriais de funções. Fixemos X um conjunto (pensemos em $\mathbb{R}$ ou $[a,b]$). Os seguintes conjuntos são todos espaços vetorias se considerarmos que a função[9] $f+g : X \to \mathbb{R}$ é dada pontualmente por $(f+g)(x) = f(x)+g(x)$ e que a função λf é dada pontualmente por $(\lambda \cdot f)(x) = \lambda f(x)$.

[9] Note que a soma de dois vetores deve ser um vetor, logo, a soma de duas funções é também uma função. Idem para o produto por escalar.

Nome	Descrição
$C(X)$	Espaço das funções contínuas da forma $f : X \to \mathbb{R}$.
$C^k(X)$	Espaço das funções cuja k-ésima derivada é contínua.
$C^\infty(X)$	Espaço das funções que são infinitamente diferenciáveis.
$C^\omega(X)$	Espaço das funções analíticas.
$R[a,b]$	Espaço das funções Riemann integráveis.
$L^p(X)$	Quoc. do espaço das funções mensuráveis tais que $(\int_X \lvert f(t)\rvert^p dt)^{\frac{1}{p}} < \infty$, $1 \leq p < \infty$.
$B(X)$	Espaço das funções limitadas.
$F(X,E)$	Espaço das funções da forma $f : X \to E$, em que E é um espaço vetorial qualquer.

Há uma observação a ser feita sobre a tabela acima a respeito de L^p. Um aluno de Computação provavelmente não está familiarizado com funções mensuráveis. Tendo isso em mente, motivemos brevemente a ideia. Na teoria de integração à Riemann (que é vista nos cursos de Cálculo), exigimos que as funções integráveis sejam aquelas não muito descontínuas (formalmente, o conjunto de descontinuidades precisa ter medida nula). É possível desenvolver uma outra teoria, a teoria de Lebesgue, que concorda com a teoria de Riemann no sentido que toda função integrável à Riemann também é integrável à Lebesgue e o valor de suas integrais é o mesmo, mas que é mais potente no sentido de permitir considerar a integral de várias funções que não são Riemann integráveis. Por exemplo, a função indicadora dos racionais $1_\mathbb{Q}(x) = 1$ se $x \in \mathbb{Q}$ e 0 se $x \notin \mathbb{Q}$ não é Riemann integrável, mas possui integral de Lebesgue, cujo valor é zero. A condição imposta sobre as funções integráveis nessa teoria é que sejam mensuráveis[10]. Naturalmente, a soma de funções mensuráveis é mensurável. O mesmo vale para o produto por escalar e várias outras coisas como o limite pontual. Assim, o conjunto das funções mensuráveis (pense Lebesgue integráveis) é um espaço vetorial. Gostaríamos, por motivos que ficarão claros quando introduzirmos o conceito de norma, que o único vetor que possuisse norma nula fosse o vetor nulo. Como vimos com $1_\mathbb{Q}$, isso não é o caso. Portanto, o real espaço L^p é composto por classes de equivalências de funções[11].

Exemplo 2.6. Continuando na linha de P_m, que já é um espaço de funções, temos vários outros exemplos de espaços que também possuem dimensão finita. De fato, basta tomar um subconjunto finito de vetores L.I. de um espaço de funções de dimensão infinita. Alguns exemplos, em ordem ascendente de complexidade são:

1. Em $C(\mathbb{R})$, considere o subespaço de dimensão dois gerado por sen e e^x.
2. Em $C(\mathbb{R})$, as potências naturais de e^x formam um conjunto L.I.: $e^x, e^{2x}, ..., e^{nx}$.
3. Em $C[-\pi,\pi]$, fixe $n \in \mathbb{N}$, então as seguintes $2n$ funções formam uma base para os polinômios trigonométricos de grau até n: $\operatorname{sen} t, \operatorname{sen} 2t, ..., \operatorname{sen} nt, \cos t, \cos 2t, ..., \cos nt$.

[10] Para mais detalhes, consulte a referência (AXLER, 2020).

[11] É interessante notar que as funções contínuas possuem representantes canônicos. Então, de certa forma, é possível dizer que L^p é um espaço de funções e não um espaço de classe de equivalências de funções.

Extra 2. Quanto maior um espaço vetorial, mais difícil é (estatisticamente) para que um conjunto aleatório de n vetores seja linearmente dependente. De fato, observemos que, se $f_1, f_2, ..., f_n$ são linearmente dependentes, então a seguinte transformação linear sempre possui núcleo não trivial[a], independente da escolha de $x_1, x_2, ..., x_n$:

$$\begin{bmatrix} f_1(x_1) & f_2(x_1) & ... & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & ... & f_n(x_2) \\ ... & ... & ... & ... \\ f_1(x_n) & f_2(x_n) & ... & f_n(x_n) \end{bmatrix}$$

Use isso para verificar que algum conjunto finito de funções é linearmente independente.

[a] O que significa que não é injetiva, não é sobrejetiva e que possuí determinante zero.

2.2 INTRODUZINDO A NORMA

A estrutura analítica que usamos para introduzir noções de convergência, continuidade e topologia é a norma. Nesta seção, definimos o conceito de norma e damos exemplos de como usá-lo em combinação com alguns dos espaços já abordados.

Definição 2.6. Seja V um espaço vetorial, uma norma $\|\cdot\|$ em V é uma função de V em $\mathbb{R}$ que obedece as seguintes propriedades. O par $X = (V, \|\cdot\|)$ recebe o nome de espaço normado.

1. **Separar pontos:** $\|x\| = 0$ se, e somente se, $x = 0$.
2. **Homogeniedade:** $\forall \lambda, x : \|\lambda \cdot x\| = |\lambda| \, \|x\|$.
3. **Desigualdade triangular:** $\forall x, y : \|x + y\| \leq \|x\| + \|y\|$.

Exemplo 2.7. Podemos dotar $\mathbb{R}^n$ da norma 1: se $x = (x_1, x_2, ..., x_n)$, então $\|x\|_1 = \sum_{i=1}^{n} |x_i|$. Note que, se $n = 1$, então $\|\cdot\|_1$ é simplesmente o valor absoluto $|\cdot|$. Alternativamente, poderíamos equipar $\mathbb{R}^n$ com a norma 2: $\|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{1/2}$. De maneira geral, para todo $1 \leq p \leq \infty$, podemos dotar $\mathbb{R}^n$ da norma p: $\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$. Para $p = \infty$, a norma é dada por $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$. Vamos denotar esses espaços por ℓ_n^p.

Observação. Como veremos no próximo exemplo, é possível definir as normas p nos espaços ℓ^p. Isso faz com que seja possível pensar nos elementos de ℓ_n^p como sequências da forma $(x_1, x_2, ..., x_n, 0, 0, ...)$. De fato, a norma de ambas as representações seria a mesma. Assim, é possível reutilizar a prova de que $\|\cdot\|_p$ é uma norma em ℓ^p para justificar que também é uma norma em ℓ_n^p.

Exemplo 2.8. Podemos estender a norma p do exemplo anterior para espaços vetoriais de dimensão infinita. Consideremos o espaço vetorial $\mathbb{R}^{\mathbb{N}}$, então $\|x\|_p = (\sum_{i=1}^{\infty} |x_i|^p)^{1/p}$. Notemos,

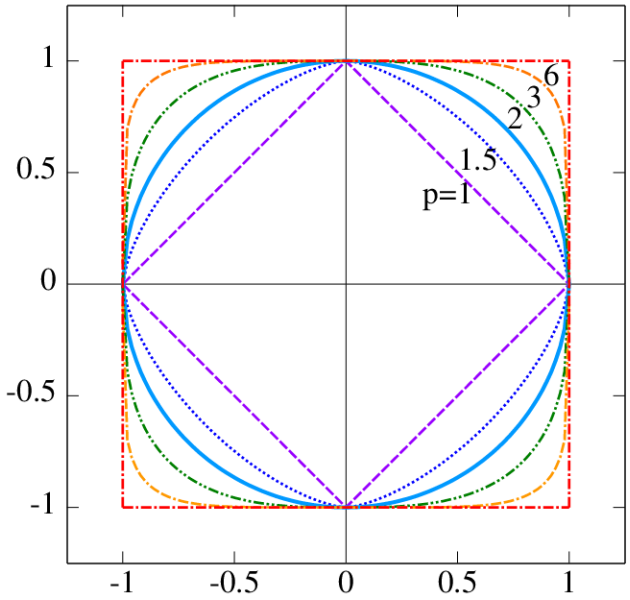


Figura 3 – Conjunto $B_1 = \{x \in \ell_2^p : \|x\|_p = 1\}$ para alguns valores de p. Fonte: Quartl, distribuida pela licença CC BY-SA 3.0.

entretanto, que nem sempre é o caso que $|x_1|^p + |x_2|^p + \ldots$ converge e, portanto, precisamos nos restringir ao subconjunto em que isso seja verdade. Este subconjunto é um subespaço de $\mathbb{R}^{\mathbb{N}}$ e denotamos ele por ℓ^p.

Observação. Quando introduzimos o espaço ℓ^p, nós não demonstramos que ℓ^p é de fato um subespaço vetorial de $\mathbb{R}^{\mathbb{N}}$. Esta justificativa não é trivial e está altamente ligada a $\|\cdot\|_p$ ser uma norma. Assim, resolvemos ambas as questões de maneira conjunta.

Demonstração. Seja ℓ^p o subconjunto de $\mathbb{R}^{\mathbb{N}}$ composto pelas sequências p-somáveis, vamos demonstrar que esse conjunto é um subespaço de $\mathbb{R}^{\mathbb{N}}$. Com efeito, este conjunto não é vazio pois $0 \in \ell^p$. Definimos sobre este conjunto a função $\|x\|_p = (\sum_{i=1}^{\infty} |x_i|^p)^{1/p}$. Note que $\|x\|_p = 0$ implica que $x = 0$, e vice-versa. Também é possível ver que $\|\lambda \cdot x\| = |\lambda| \, \|x\|$, pois $\|\lambda \cdot x\| = (\lim_{n\to\infty} R_n)^{\frac{1}{p}}$, em que $R_n = \sum_{i=1}^{n} |\lambda x|^p = |\lambda|^p S_n$, $S_n = \sum_{i=1}^{n} |x_i|^p$, e $\lim_{n\to\infty} R_n = |\lambda^p| \lim_{n\to\infty} S_n$. Assim, ℓ^p é fechado pela multiplicação por escalar. Resta mostrar que $x + y \in \ell^p$ quando x e y pertencem a ℓ^p. Vemos isso através da prova que $\|x+y\|_p \leq \|x\|_p + \|y\|_p$. Note que a norma p não necessariamente está definida sobre esse vetor, mas isso é fácil de resolver: primeiro se mostra a desigualdade, depois se conclui que ℓ^p é um espaço vetorial, então se diz que $\|x+y\|_p \leq \|x\|_p + \|y\|_p$; a parte realmente importante é a demonstração da desigualdade. Para fazê-la, precisamos da seguinte desigualdade auxiliar (desigualdade de Hölder[12]):

$$\sum_{i=1}^{\infty} |x_i y_i| \leq \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^{\infty} |y_i|^q \right)^{\frac{1}{q}}, \quad \frac{1}{p} + \frac{1}{q} = 1. \tag{2.1}$$

Agora, observemos que a desigualdade vale para quaisquer sequências x e y, o que significa que ambos os lados podem ser infinito. Dito isso,

[12] Uma demonstração desse resultado pode ser encontrada na seção 1.2 da referência (KREYSZIG, 1989).

$$\sum_{i=1}^{\infty}|x_i+y_i|^p=\sum_{i=1}^{\infty}|x_i+y_i||x_i+y_i|^{p-1}\leq\sum_{i=1}^{\infty}|x_i||x_i+y_i|^{p-1}+\sum_{i=1}^{\infty}|y_i||x_i+y_i|^{p-1}.$$

Podemos então aplicar a desigualdade de Hölder para obter:

$$\sum_{i=1}^{\infty}|x_i+y_i|^p\leq\|x\|_p\left(\sum_{i=1}^{\infty}|x_i+y_i|^{(p-1)q}\right)^{\frac{1}{q}}+\|y\|_p\left(\sum_{i=1}^{\infty}|x_i+y_i|^{(p-1)q}\right)^{\frac{1}{q}}.$$

Note que $(p-1)q=1$. Usando isso e fatorando, temos:

$$\sum_{i=1}^{\infty}|x_i+y_i|^p\leq(\|x\|_p+\|y\|_p)\left(\sum_{i=1}^{\infty}|x_i+y_i|^p\right)^{\frac{1}{q}}.$$

Basta reagrupar a série para obter a desigualdade desejada:

$$\left(\sum_{i=1}^{\infty}|x_i+y_i|^p\right)^{\frac{1}{p}}=\|x+y\|_p\leq\|x\|_p+\|y\|_p.$$

□

Exemplo 2.9. O espaço das sequências limitadas ℓ^∞ com a norma do máximo dada por $\|x\|_\infty=\sup_{i\in\mathbb{N}}|x_i|$ é um espaço normado.

Demonstração. Começamos verificando que ℓ^∞ é um subespaço de $\mathbb{R}^\mathbb{N}$ e, portanto, é um espaço vetorial. Para isto, consideremos duas sequências limitadas x,y e um escalar real λ. Sabemos que existem constantes reais a,b tais que $|x_i|\leq a$ e $|y_i|\leq b$ para todo índice i. Percebe-se que $\lambda\cdot x$ e $x+y$ também são sequências limitadas, pois $|\lambda x_i|=|\lambda||x_i|\leq|\lambda|a$ e $|x_i+y_i|\leq|x_i|+|y_i|\leq a+b$. Como $0\in\ell^\infty$, podemos concluir que ℓ^∞ é um subespaço de $\mathbb{R}^\mathbb{N}$. Agora verificamos os axiomas da norma. Claramente, $\|x\|_\infty=0\iff x=0$. Adicionalmente, $\|\lambda\cdot x\|=\sup_{i\in\mathbb{N}}|\lambda x_i|=\sup_{i\in\mathbb{N}}|\lambda||x_i|=|\lambda|\,\|x\|_\infty$. Por fim, $|x_i+y_i|\leq|x_i|+|y_i|\leq\|x\|_\infty+\|y\|_\infty$, o que nos diz que $\|x\|_\infty+\|y\|_\infty$ é uma cota superior para o conjunto $\{|x_i+y_i|\}_{i\in\mathbb{N}}$. Pela definição do supremo, sabemos que $\|x+y\|_\infty$ é a menor cota superior daquele conjunto. Assim, podemos concluir que $\|x+y\|_\infty\leq\|x\|_\infty+\|y\|_\infty$ e que ℓ^∞ é um espaço normado. □

Exemplo 2.10. Seja $C[a,b]$ o espaço vetorial das funções contínuas com domínio em $[a,b]$, então $\|f\|_\infty=\sup\{|f(x)|:x\in[a,b]\}$ é uma norma[13] e $f_n\to f$ com respeito à $\|\cdot\|_\infty$ se, e somente se, as f_n convergem uniformemente para f.

Demonstração. Note que $\|f\|_\infty=0$ sse $f=0$. Além disso, lembrando que $\sup rA=r\sup A$ para $r\geq0$, temos $\|rf\|=\sup\{|r||f(x)|:x\in[a,b]\}=|r|\,\|f\|_\infty$. Por fim, temos $\|f+g\|_\infty\leq\|f\|_\infty+\|g\|_\infty$ uma vez que, para todo $x\in[a,b]$, $|f(x)+g(x)|\leq|f(x)|+|g(x)|\leq\sup\{|f(x)|:$

[13] $\|\cdot\|_\infty$ está bem definida pois f contínua sobre $[a,b]$ implica que f é limitada.

$x \in [a,b]\} + \sup\{|g(x)| : x \in [a,b]\}$ (o supremo é menor ou igual a toda cota superior). A demonstração de que $f_n \xrightarrow{\|\cdot\|_\infty} f \iff f_n \xrightarrow{\textbf{unif.}} f$ é deixada como sugestão de exercício para o leitor. □

Exemplo 2.11. Também em $C[a,b]$, podemos definir a norma p da seguinte maneira: $\|f\|_p = (\int_a^b |f|^p)^{1/p}$.

Demonstração. Ao contrário de ℓ^p, onde precisamos mostrar que o conjunto era de fato um espaço vetorial, já sabemos que $C[a,b]$ forma um espaço vetorial, então basta verificar as propriedades da norma. De fato, se $\|f\|_p = 0$, temos que uma função não negativa e continua possui integral zero, isso implica que a função é nula[14]. Além disso, $\|0\|_p$ claramente é zero. Para a segunda propriedade, veja que $\|\lambda \cdot f\|_p = (\int_a^b |\lambda f(t)|^p dt)^{\frac{1}{p}} = (\int_a^b |\lambda|^p |f(t)|^p dt)^{\frac{1}{p}}$. Usando a linearidade da integral, é possível concluir que $\|\lambda \cdot f\|_p = |\lambda| \|f\|_p$. Falta agora mostrar que $\|f+g\|_p \leq \|f\|_p + \|g\|_p$. Para isto, usamos outra versão da desigualdade de Hölder, desta vez para funções:

$$\int_a^b |f(t)||g(t)|dt \leq \left(\int_a^b |f(t)|^p dt\right)^{\frac{1}{p}} \left(\int_a^b |g(t)|^q dt\right)^{\frac{1}{q}}. \tag{2.2}$$

O resto da demonstração segue de maneira análoga a ℓ^p. □

2.3 CONVERGÊNCIA E CONTINUIDADE EM ESPAÇOS NORMADOS

Em cálculo, dizemos que uma série é convergente se a sequência de suas somas parciais P_n converge para algum número. Ou seja, se, para todo ε existe N tal que, para todo $n \geq N$, temos $|Pn - \sum_{i=1}^{\infty} x_i e_i| < \varepsilon$. Crucial para esta definição foi a existência de uma função distância $d(x,y) = |x-y|$; em espaços normados, também temos uma escolha de função distância[15]. De fato, basta definir

$$d(x,y) = \|x-y\|.$$

Observação. Note que d assim definida é simétrica e satisfaz a seguinte desigualdade:

$$\forall x,y,z : d(x,y) \leq d(x,z) + d(z,y).$$

Essa desigualdade, que recebe o nome de desigualdade triangular, nos diz que é sempre mais rápido ir de um ponto a outro diretamente, sem passar por um terceiro ponto intermediário.

Com essa função distância, podemos dar as seguintes definições:

[14] Se não fosse nula, a função deveria ser positiva em algum intervalo por conta da continuidade. Isso contradiria o fato de que a integral da função é nula.

[15] Funções que satisfazem propriedades similares as de $|\cdot|$ recebem o nome de métrica. Assim, passaremos a nos referir à $d(x,y) = \|x-y\|$ como sendo a métrica induzida pela norma.

Definição 2.7. Dizemos que uma sequência $(x_n)_{n\in\mathbb{N}}$ é convergente em um espaço normado X se há um vetor $x \in X$ satisfazendo a seguinte condição:

$$\forall \varepsilon > 0 : \exists N > 0 : \forall n \geq N : \|x - x_n\| < \varepsilon.$$

Denotamos a convergência de x_n para x por $x_n \to x$.

Observação. Essa definição é fundamental. De fato, para muitos conceitos de Análise, há definições em termos de sequências. Por exemplo, veremos a seguir que funções contínuas entre dois espaços normados X e Y podem ser caracterizadas como aquelas tais que $f(x_n) \to f(x)$ sempre que $x_n \to x$.

Definição 2.8. Dizemos que uma série $\sum_{i=1}^{\infty} x_i e_i$, na qual $x_i \in \mathbb{R}$ e $e_i \in X$, é convergente se há um vetor $\sum_{i=1}^{\infty} x_i e_i \in X$ tal que a sequência das somas parciais $P_n = \sum_{i=1}^{n} x_i e_i$ converge para $\sum_{i=1}^{\infty} x_i e_i$.

Definição 2.9. Dizemos que uma função $f : X \to Y$ entre espaços normados X e Y é contínua no ponto x se satisfaz a seguinte condição:

$$\forall \varepsilon > 0 : \exists \delta > 0 : \forall y : \|x - y\| < \delta \Rightarrow \|f(x) - f(y)\| < \varepsilon. \tag{2.3}$$

Uma função é contínua se for contínua em todos os pontos.

Observação. Salvo casos que possam gerar ambiguidade, usaremos $\|x\|$ e $\|f(x)\|$ ao invés de $\|x\|_X$ e $\|f(x)\|_Y$.

Definição 2.10. Seja X um espaço normado e $A \subseteq X$ um subconjunto de X. Dizemos que A é limitado se há $M > 0$ tal que:

$$\forall a \in A : \|a\| \leq M.$$

Proposição 2.1. As propriedades básicas que estamos acostumados ao lidar com limites de sequências reais continuam valendo em espaços normados. Com efeito, sejam $(x_n)_{n\in\mathbb{N}}$ e $(y_n)_{n\in\mathbb{N}}$ sequências em um espaço normado X e λ um escalar real.

1. Se $x_n \to x$ e $y_n \to y$, então $x_n + y_n \to x + y$.
2. Se $x_n \to x$, então $\lambda \cdot x_n \to \lambda \cdot x$.
3. O limite é único; se $x_n \to x_1$ e $x_n \to x_2$, então $x_1 = x_2$.
4. Uma sequência convergente é limitada.

Demonstração. As provas de 1 e 2 são análogas ao caso real. Para 3, note que $\|x_1 - x_2\| = \|x_1 - x_n + x_n - x_2\| \leq \|x_1 - x_n\| + \|x_n - x_2\|$. O lado direito da desigualdade vai para zero quando $n \to \infty$. Isso nos diz que $\|x - y\| < \varepsilon$, para todo ε, o que, por sua vez, nos permite concluir que $\|x_1 - x_2\| = 0$ e $x_1 = x_2$. Para 4, veja que a convergência de $(x_n)_{n\in\mathbb{N}}$ nos diz que há $N > 0$ tal

que $\|x-x_n\| < 1$ para todo $n \geq N$. Assim, sabemos que $\|x_n\| \leq \|x_n - x\| + \|x\| < \|x\| + 1$ para todo $n \geq N$. Definindo $M = \max\{\|x_1\|, \|x_2\|, ..., \|x_{N-1}\|, \|x\| + 1\}$, temos $\|x_n\| \leq M$ para todo n. □

Proposição 2.2. (Caracterização Sequencial) Uma função $f : X \to Y$ é contínua se, e somente se, para toda sequência $(x_n)_{n\in\mathbb{N}}$ com $x_n \to x$ em X, tem-se $f(x_n) \to f(x)$.

Demonstração. Seja $f : X \to Y$ uma função contínua e $(x_n)_{n\in\mathbb{N}}$ uma sequência convergente, então, se escolhermos N de forma que $\|x_n - x\| < \delta$ para todo $n \geq N$, em que δ está associado à ε na continuidade de f sobre x, então temos que $\|f(x_n) - f(x)\| < \varepsilon$ para todo $n \geq N$. Isso nos mostra que $f(x_n) \to f(x)$.

A outra direção da prova é menos construtiva. Suponha que f não é contínua, então, em particular, f não é contínua sobre algum ponto x. Isso se traduz em $\exists\varepsilon > 0 : \forall\delta > 0 : \exists x' : \|x - x'\| < \delta$ e $\|f(x) - f(x')\| \geq \varepsilon$. Escolhendo $\delta = 1/n$ para cada n, podemos montar uma sequência $(x_n)_{n\in\mathbb{N}}$ que converge para x mas cuja imagem nunca fica mais próxima do que ε. Podemos concluir que $f(x_n) \not\to f(x)$, o que nos dá a contrapositiva. □

Corolário 2.1. Funções contínuas se comportam bem com respeito a operações algébricas. De fato, sejam $f : X \to Y$ e $g : X \to Y$ funções contínuas entre espaços normados X e Y e seja λ um escalar real, então

1. $f + g$ é contínua.
2. $\lambda \cdot f$ é contínua.

Demonstração. Seja $(x_n)_{n\in\mathbb{N}}$ uma sequência tal que $x_n \to x$. Pelo teoremas anteriores, temos que $(f+g)(x_n) = f(x_n) + g(x_n) \to f(x) + g(x) = (f+g)(x)$ e $(\lambda \cdot f)(x_n) = \lambda f(x_n) \to \lambda f(x) = (\lambda \cdot f)(x)$, o que nos permite concluir que $f + g$ e $\lambda \cdot f$ são ambas contínuas. □

Proposição 2.3. (Desigualdade triangular reversa) Seja X um espaço normado, então, para quaisquer x e y em X, temos que

$$| \|x\| - \|y\| | \leq \|x - y\|.$$

Demonstração. Pela desigualdade triangular, temos $\|x\| \leq \|x - y\| + \|y\|$, o que nos dá $\|x\| - \|y\| \leq \|x - y\|$. Adicionalmente, $\|y\| \leq \|x - y\| + \|x\|$ e $-\|x\| + \|y\| \leq \|x - y\|$. Combinando essas duas desigualdades, podemos concluir que $| \|x\| - \|y\| | \leq \|x - y\|$. □

Corolário 2.2. A norma é uma função uniformemente contínua. Isto é,

$$\forall\varepsilon > 0 : \exists\delta > 0 : \forall x, y : \|x - y\| < \delta \to | \|x\| - \|y\| | < \varepsilon. \tag{2.4}$$

Podemos concluir que vetores próximos possuem normas próximas.

Demonstração. Dado $\varepsilon > 0$, basta selecionar $\delta = \varepsilon$. □

É comum nos depararmos com a seguinte situação: temos uma sequência $(x_n)_{n\in\mathbb{N}}$ que parece convergir, pois seus elementos ficam cada vez mais próximos uns dos outros, mas que não possui um candidato para ser seu limite. Por exemplo, considere a seguinte série:

$$\sum_{i=1}^{\infty}(-1)^n\frac{1}{n^2}$$

É fácil mostrar que ela converge absolutamente, mas seu limite não é tão claro. Outra situação comum é quando obtemos os x_i de forma não construtiva, como pode ser vista na prova da Proposição 2.5. Isso justifica a seguinte definição:

Definição 2.11. Dizemos que uma sequência $(x_n)_{n\in\mathbb{N}}$ é de Cauchy quando satisfaz a seguinte condição:

$$\forall\varepsilon>0:\exists N>0:\forall n,m\geq N:\|x_n-x_m\|<\varepsilon.$$

Proposição 2.4. Toda sequência convergente é de Cauchy e toda sequência de Cauchy é limitada.

Demonstração. Seja $(x_n)_{n\in\mathbb{N}}$ uma sequência tal que $x_n\to x$. Mostraremos que $(x_n)_{n\in\mathbb{N}}$ é de Cauchy. Com efeito, note que $\|x_n-x_m\|\leq\|x_n-x\|+\|x-x_m\|$ para todo n e m. Como $(x_n)_{n\in\mathbb{N}}$ é convergente, sabemos que, dado $\varepsilon>0$, há $N>0$ tal que $\forall n,m\geq N:\|x_n-x\|<\varepsilon/2$. Assim, $\|x_n-x_m\|<\varepsilon$ desde que deixemos $n,m\geq N$ e $(x_n)_{n\in\mathbb{N}}$ é de Cauchy.

Agora, mostramos que sequências de Cauchy são limitadas. De fato, seja $(x_n)_{n\in\mathbb{N}}$ uma sequência de Cauchy, então há $N>0$ tal que $\forall n,m\geq N:\|x_n-x_m\|<1$. Em especial, temos $\forall n\geq N:\|x_N-x_n\|<1$. Isso nos permite concluir que $\|x_n\|\leq\|x_n-x_N\|+\|x_N\|<\|x\|_N+1$ sempre que $n\geq N$. Se deixarmos $M=\max\{\|x_1\|,\|x_2\|,...,\|x_{N-1},\|x_N\|+1\|$, então $\|x_n\|\leq M$ para todo n e $(x_n)_{n\in\mathbb{N}}$ é limitada. □

Teorema 2.2. Em $(\mathbb{R},|\cdot|)$, Uma sequência é convergente se, e somente se, for de Cauchy.

Demonstração. É possível encontrar uma prova desse teorema na seção 2.6 da referência (ABBOTT, 2015). □

Nem sempre isso é verdade para um espaço normado, o que justifica a seguinte definição:

Definição 2.12. Dizemos que um espaço normado X é de Banach quando toda sequência de Cauchy em X é convergente. Além disso, dizemos que um subconjunto E de X é completo se toda sequência de Cauchy composta por elementos de E converge para um elemento de E.

Exemplo 2.12. $\mathbb{R}^n$ é completo com respeito à qualquer norma. De fato, veremos mais à frente que todo espaço normado de dimensão finita é de Banach.

Demonstração. Provas das completudes de ℓ^2_n, ℓ^p,ℓ^∞ e $(C[a,b],\|\cdot\|_\infty)$ podem ser encontradas na seção 1.5 da referência (KREYSZIG, 1989). □

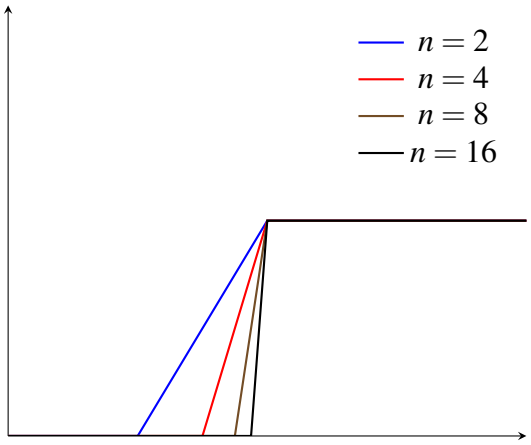


Figura 4 – Sequência das f_n no Exemplo 2.16.

Exemplo 2.13. ℓ^p é completo.

Exemplo 2.14. ℓ^∞ é completo.

Exemplo 2.15. $C[a,b]$ com a norma $p = \infty$ é completo.

Exemplo 2.16. $C[a,b]$ com a norma $1 \leq p < \infty$ é incompleto. É possível completar esse espaço usando o Teorema 2.3. É interessante notar que o completamento desse espaço é isomorfo (e isométrico) ao espaço L^p introduzido anteriormente.

Demonstração. Para simplificar a prova, iremos assumir que $[a,b] = [0,2]$. A prova para o caso geral é análoga. De fato, considere a sequência de Cauchy:

$$f_n(x) = \begin{cases} 0, & \text{se } 0 \leq x \leq 1 - 1/n, \\ n(x - 1 + 1/n), & \text{se } 1 - 1/n < x \leq 1, \\ 1, & \text{se } x > 1. \end{cases}$$

Veja que a diferença entre f_n e f_m, se $n, m \geq N$, pode ser superestimada pela área $1/N$. Se tomarmos N grande o suficente, teremos f_n e f_m arbitrariamente perto uma da outra, o que significa que a sequência de fato é Cauchy. Veremos que as f_n não podem convergir, pois caso convergissem, convergiriam para uma função descontínua. Prosseguimos por contradição. Seja f o limite das f_n, então, temos $\|f_n - f\|_p^p = \int_0^2 |f_n - f|^p < \varepsilon$ para n grande o suficente. Note que podemos separar essa integral em três partes: $\int_0^{1-1/n} |f_n - f|^p$, que simplifica para $\int_0^{1-1/n} |f|^p$; $\int_{1-1/n}^1 |f_n - f|^p$, que não simplifica; e $\int_1^2 |f_n - f|^p$, que simplifica para $\int_0^2 |1 - f|^p$. Como as três integrais são não negativas (note o módulo), temos, em particular, que a última fica arbitrariamente pequena; mas a última não depende em n. Portanto, obtemos $\int_0^2 |1 - f|^p = 0$ e $f = 1$ no intervalo $[1,2]$. Mostramos agora que $f(x) = 0$ para todo x em $[0,1)$, o que implica que f é descontínua em $x = 1$. Com efeito, suponha que $f(x) \neq 0$ para algum $x \in [0,1)$. Isso nos diz que $|f|^p(x) > 0$. Através da continuidade de $|f|^p$, podemos também inferir que há um intervalo $(x - \delta, x + \delta) \subseteq (0,1)$ em que a função fica inteiramente acima do eixo x. Isto é um problema

para nossa hipótese de que as f_n convergem para f, pois a integral de 0 até $1-1/n$ deveria ficar arbitrariamente pequena e acabamos de mostrar que, se pegarmos n tal que $1-1/n > x+\delta$, ela fica pelo menos igual à área dessa região positiva. Logo, $f = 0$ no intervalo $[0,1)$ e f é descontínua. Isto também nos leva a uma contradição, uma vez que f deveria pertencer a $C[0,2]$. Somos então forçados a abandonar nossa hipótese que as f_n tem limite; $C[0,2]$ com a norma p não é completo. □

Definição 2.13. Dizemos que uma série $\sum_{i=1}^{\infty} v_i$ de vetores em um espaço normado converge absolutamente se $\sum_{i=1}^{\infty} \|v_i\|$ converge em $\mathbb{R}$.

Proposição 2.5. Em um espaço de Banach, toda série que converge absolutamente é convergente.

Demonstração. Seja $\sum_{i=1}^{\infty} v_i$ uma série que converge absolutamente, então $\sum_{i=1}^{\infty} \|v_i\|$ converge e a cauda $\sum_{i=m}^{\infty} \|v_i\|$ tende a zero quando m tende a infinito. Vamos mostrar que a sequência $P_n = \sum_{i=1}^{n} v_i$ das somas parciais é de Cauchy. Sem perda de generalidade, suponha que $n > m$, então:

$$\|P_n - P_m\| = \left\| \sum_{i=m+1}^{n} v_i \right\| \leq \sum_{i=m+1}^{n} \|v_i\| \leq \sum_{i=m}^{\infty} \|v_i\| .$$

Note que a diferença $\|P_n - P_m\|$ é majorada pela cauda que vai pra zero quando m vai para infinito. Podemos concluir que P_n é de Cauchy e, portanto, que converge. □

É surpreendente o fato de que todo espaço normado incompleto "vive" dentro de um espaço maior que é completo. Antes disso, vamos relembrar a noção de isometria:

Definição 2.14. Dizemos que uma transformação linear $T : X \to Y$ entre espaços normados X e Y é uma isometria se preserva distâncias:

$$\forall x : \|Tx\| = \|x\| . \tag{2.5}$$

Observação. Isometrias são sempre injetivas, pois $Tx = Ty$ implica $T(x-y) = 0$, que, por sua vez, implica em $\|x-y\| = 0$ e $x = y$.

Teorema 2.3. Seja $X = (V, \|\cdot\|)$ um espaço normado, então existe um espaço de Banach $\tilde{X} = (\tilde{V}, \|\tilde{\cdot}\|)$ que possui um subespaço W isométrico à X. Esse W é denso[16] em $\tilde{X}$.

Demonstração. A prova desse teorema não é difícil, mas é entediante. Geralmente, primeiro se prova um resultado análogo no contexto mais geral de espaços métricos (tuplas (X,d) nas quais d é uma função distância) e então o teorema é estabelecido como corolário. O presente teorema, por sua vez, é usado para provar que todo pré-espaço de Hilbert "vive" dentro de um espaço maior que é de Hilbert. Pode-se encontrar provas para esses três teoremas nas seções 1.6, 2.3 e 3.2 da referência (KREYSZIG, 1989). A última versão será necessária para provar o teorema de Moore-Aronszajn na seção de Kernels e Aprendizado. □

[16] Densidade será definida na Seção 2.5.

2.4 BASES DE SCHAUDER

Nesta seção introduzimos a noção de base de Schauder. Após introduzirmos a noção de separabilidade na seção seguinte, discutiremos por que nem todo espaço normado admite uma base de Schauder. De fato, em um certo sentido, um espaço normado precisa ser pequeno para admitir uma base de Schauder.

Definição 2.15. Seja X um espaço normado, dizemos que uma sequência $(e_n)_{n\in\mathbb{N}}$ de elementos de X é uma base de Schauder para X se, para todo $x \in X$, há uma única sequência de escalares $(x_n)_{n\in\mathbb{N}}$ tal que:

$$x = \sum_{i=1}^{\infty} x_i e_i. \tag{2.6}$$

Exemplo 2.17. A "base canônica" de ℓ^p é uma base de Schauder.

Demonstração. Seja $x = (x_1, x_2, ...)$ e $P_n = \sum_{i=1}^{\infty} x_i e_i$, então $\|x - P_n\|_p = (\sum_{i=n+1}^{\infty} |x_i|^p)^{\frac{1}{p}}$. Note que o lado direito vai para zero quando n vai para infinito pois a série $\sum_{i=1}^{\infty} |x_i|^p$ é convergente. Assim, temos que todo elemento pode ser escrito como uma expansão em termos dos vetores da "base canônica". Resta mostrar que essa expansão é única. Para isso, note que qualquer mudança nos coeficientes x_i implica em um limite x diferente. Assim, não há como haver duas representações para x e a "base canônica" é uma base de Schauder. □

Teorema 2.4. Seja X um espaço normado, $(e_n)_{n\in\mathbb{N}}$ uma base de Schauder para X, $(x_n)_{n\in\mathbb{N}}$ e $(y_n)_{n\in\mathbb{N}}$ as coordenadas de dois vetores x e y em X e λ um escalar, então:

1. $x + y = \sum_{i=1}^{\infty} (x_i + y_i) e_i$,
2. $\lambda \cdot x = \sum_{i=1}^{\infty} (\lambda x_i) e_i$.

Demonstração. Para 1, temos que $x + y = \sum_{i=1}^{\infty} x_i e_i + \sum_{i=1}^{\infty} y_i e_i = \lim_{n\to\infty} P_n + \lim_{n\to\infty} Q_n$, em que P_n e Q_n são, respectivamente, as sequências das somas parciais de x e y. Pela Proposição 2.1, podemos fatorar o limite para obter que $x + y = \sum_{i=1}^{\infty} (x_i + y_i) e_i$. A demonstração do item 2 se dá de maneira análoga. □

Proposição 2.6. É sempre possível normalizar uma base de Schauder. Isto é, se $(v_n)_{n\in\mathbb{N}}$ é uma base de Schauder para um espaço normado X, então $(e_n)_{n\in\mathbb{N}}$ dada por $e_n = v_n / \|v_n\|$ também é uma base de Schauder para X.

Demonstração. Todo vetor $x \in X$ possui uma única representação como $x = \sum_{i=1}^{\infty} x_i v_i$. Podemos rearranjar os termos para achar uma representação na base normalizada. De fato, $x = \sum_{i=1}^{\infty} (x_i \|v_i\|) e_i$. Se houvesse um vetor com duas representações $\sum_{i=1}^{\infty} y_i e_i$ e $\sum_{i=1}^{\infty} z_i e_i$ com respeito à base $(e_n)_{n\in\mathbb{N}}$, poderíamos reescrever e_i como $v_i / \|v_i\|$ e então obteríamos uma contradição (haveria duas representações para um vetor na base $(v_n)_{n\in\mathbb{N}}$). Podemos concluir que $(e_n)_{n\in\mathbb{N}}$ também é uma base de Schauder para X. □

Observação. Análogo ao caso de dimensão finita - em que tinhamos vetores expressos da forma $x = x_1e_1 + x_2e_2 + ... + x_ne_n$ e podíamos expandir o valor de uma transformação linear em x como sendo a combinação linear de seus valores na base: $f(x) = \sum_{i=1}^{n} x_i f(e_i)$ - em espaços normados com base de Schauder podemos expandir transformações lineares contínuas como séries: $f(x) = \sum_{i=1}^{\infty} x_i f(e_i)$.

2.5 INTERLÚDIO TOPOLÓGICO

Nesta seção, discutiremos propriedades que conjuntos de pontos em um espaço normado podem ou não possuir. Ao contrário de Álgebra Linear[17], nos depararemos mais frequentemente com conjuntos "estranhos" para os quais não possuimos grande intuição. Topologia - a área da matemática que estuda espaços abstratos de pontos e suas propriedades - nos ajudará a raciocinar nesses casos. De fato, considere as seguintes definições:

Definição 2.16. Seja X um espaço normado e x um elemento de X, definimos os seguintes conjuntos:

1. **A bola aberta de raio r:** $B_r(x) = \{y \in X : \|x-y\| < r\}$.
2. **A bola fechada de raio r:** $\overline{B}_r = \{y \in X : \|x-y\| \leq r\}$.
3. **A esfera de raio r:** $S_r(x) = \{y \in X : \|x-y\| = r\}$.

Definição 2.17. Seja A um subconjunto de um espaço normado X, dizemos que $a \in A$ é um ponto no interior de A se for possível encontrar um $r > 0$ tal que $B_r(a) \subseteq A$. Denotamos por $\text{Int}A$ o subconjunto de A composto pelos pontos interiores de A.

Definição 2.18. Seja $A \subseteq X$ um subconjunto de um espaço normado X, dizemos que $x \in X$ é um ponto limite de A se toda bola centrada em x contém algum ponto de A. Isto é:

$$\forall r > 0 : B_r(x) \cap A \neq \emptyset.$$

O subconjunto de X composto pelos pontos limites de A é denotado por $\bar{A}$.

Definição 2.19. Dizemos que um conjunto é aberto se não contiver nenhum ponto além de seu interior e dizemos que um conjunto é fechado se contiver todos seus pontos limites. Ou seja, A é aberto se $\text{Int}A = A$ e é fechado se $\bar{A} = A$.

Proposição 2.7. (Caracterização Sequencial) Um subconjunto $A \subseteq X$ é fechado se, e somente se, toda sequência $(x_n)_{n\in\mathbb{N}}$ em A que converge, converge para um elemento de A.

Demonstração. Seja A um subconjunto fechado de X e seja $(x_n)_{n\in\mathbb{N}}$ uma sequência em A tal que $x_n \to x \in X$. Note que a definição de convergência nos diz que toda bola $B_\varepsilon(x)$ centrada em

[17] Conjunto de Cantor, conjunto de Vitali, etc.

x contém infinitos pontos da sequência $(x_n)_{n\in\mathbb{N}}$. Como ε é arbitrário e a sequência é composta de pontos de A, podemos concluir que x é um ponto limite e, portanto, que pertence à A.

Suponha agora que toda sequência de elementos em A que converge, converge para algum elemento em A e que x é um ponto limite de A. Pela definição de ponto limite, sabemos que toda bola $B_\varepsilon(x)$ contém algum ponto de A. Assim, se montarmos uma sequência de bolas $B_1, B_2, ...$ todas centradas em x e com raios dados pela sequência $(\frac{1}{n})_{n\in\mathbb{N}}$, teremos como extrair uma sequência $(x_n)_{n\in\mathbb{N}}$ de pontos de A onde $\|x-x_n\| < \frac{1}{n}$. Claramente, $x_n \to x$. Pela suposição original, podemos concluir que $x \in X$ e que A contém todos seus pontos limites. □

Corolário 2.3. Um subconjunto de um espaço de Banach é fechado se, e somente se, for completo.

Demonstração. Seja Y um subconjunto fechado de um espaço de Banach X e seja $(x_n)_{n\in\mathbb{N}}$ uma sequência de Cauchy em Y. Como X é de Banach, sabemos que $x_n \to x$ para algum $x \in X$. Por Y ser fechado, sabemos também que $x \in Y$. Assim, toda sequência de Cauchy em Y converge para algum elemento de Y e podemos dizer que Y é completo.

Seja Y um subconjunto completo de X e seja $(x_n)_{n\in\mathbb{N}}$ uma sequência em Y tal que $x_n \to x$ para algum $x \in X$. Note que, em particular, $(x_n)_{n\in\mathbb{N}}$ é de Cauchy. Como Y é completo, temos que $x \in Y$, o que nos permite concluir que Y é fechado. □

Exemplo 2.18. A bola aberta é aberta e a bola fechada é fechada.

Exemplo 2.19. O conjunto $\{1/n : n \in \mathbb{N}\}$ não é fechado pois não contém o ponto limite 0.

Observação. O interior $\mathrm{Int}A$ de um conjunto A é o maior subconjunto de A que é aberto. De maneira análoga, o fecho $\bar{A}$ de um conjunto A é o menor subconjunto de X que contém A e é fechado.

Definição 2.20. Dizemos que um conjunto $D \subseteq X$ é denso em X se $\overline{D} = X$. Alternativamente, D é denso se, para todo $x \in X$ e todo $\varepsilon > 0$, houver um vetor $d \in D$ tal que $\|d-x\| < \varepsilon$.

Exemplo 2.20. $\mathbb{Q}$ é denso em $\mathbb{R}$ e $\mathbb{Q}^n$ é denso em ℓ_n^p.

Demonstração. Uma demonstração de que $\mathbb{Q}$ é denso em $\mathbb{R}$ pode ser encontrada na seção 1.4 da referência (ABBOTT, 2015). Para a densidade de $\mathbb{Q}^n$ em ℓ_n^p, seja $x = (x_1, x_2, ..., x_n)$ um vetor de ℓ_n^p e $\varepsilon > 0$ um número real positivo, então, pela densidade de $\mathbb{Q}$ em $\mathbb{R}$, conseguimos encontrar um vetor com coordenadas racionais $r = (r_1, r_2, ..., r_n)$ também em ℓ_n^p tal que $|x_i - r_i|^p < \varepsilon^p/n$ para $i = 1, 2, ..., n$. Assim, $\|x-r\|_p = (\sum_{i=1}^n |x_i - r_i|^p)^{\frac{1}{p}} < (\sum_{i=1}^n \varepsilon^p/n)^{\frac{1}{p}} = \varepsilon$. Assim, concluímos que todo vetor $x \in \ell_n^p$ pode ser arbitrariamente bem aproximado por vetores em $\mathbb{Q}^n$. Portanto, $\mathbb{Q}^n$ é denso em ℓ_n^p. □

Observação. Note que todos os conjuntos acima são enumeráveis. Isso motiva a seguinte definição.

Definição 2.21. Dizemos que um espaço normado X é separável se há um subconjunto $D \subseteq X$ denso e enumerável.

Exemplo 2.21. ℓ_n^p e ℓ^p são separáveis, enquanto ℓ^∞ não é.

Demonstração. Mostramos que ℓ^∞ não é separável, as outras provas podem ser encontradas na seção 1.3 da referência (KREYSZIG, 1989). Com efeito, considere o conjunto E composto pelas sequências de zeros e uns. Esse conjunto é não enumerável e a distância entre cada dois pontos distintos em E é 1. Assim, se D é um conjunto denso, D precisa conter um elemento distinto para cada bola $B_{1/2}(e)$ centrada sobre uma sequência $e \in E$. Isso força D a ser não enumerável, uma vez que o conjunto de todas as sequências de zeros e uns é não enumerável. □

Observação. Todo espaço normado que possui uma base de Schauder é separável. De fato, é possível replicar o argumento de ℓ^p: seja $(e_n)_{n\in\mathbb{N}}$ uma base de Schauder para X, então o conjunto das combinações lineares racionais $(r_1e_1 + r_2e_2 + ... + r_ne_n)$ é um subconjunto denso e enumerável de X. Com isso, podemos concluir que ℓ^∞ não possui base de Schauder.

Uma terceira propriedade interessante que um conjunto pode possuir é compacidade. Vejamos a definição a seguir:

Definição 2.22. Dizemos que um subconjunto K em X é compacto se toda sequência $(x_n)_{n\in\mathbb{N}}$ em K admite uma subsequência convergente (com limite em K).

Proposição 2.8. Seja K um conjunto compacto em X, então K é fechado e é limitado.

Demonstração. Começamos mostrando que K é fechado. Faremos isso usando a caracterização sequencial de conjuntos fechados. Com efeito, seja $(x_n)_{n\in\mathbb{N}}$ uma sequência de elementos de K tal que $x_n \to x$ para algum $x \in X$. Como K é compacto, $(x_n)_{n\in\mathbb{N}}$ tem uma subsequência convergente $(x_{n'})_{n'\in\mathbb{N}'}$ com $x_{n'} \to x' \in K$. Note que $x' = x$, pois todos os termos da sequência $(x_n)_{n\in\mathbb{N}}$ eventualmente ficam arbitrariamente próximos de x, incluindo os pontos $x_{n'}$ da subsequência que converge pra x'. Assim, $x \in K$ e K é fechado.

Agora, mostramos que K é limitado. Para isto, buscaremos encontrar uma contradição a partir da hipótese de que não é limitado. De fato, se K não fosse limitado, então haveria uma sequência $(x_n)_{n\in\mathbb{N}}$ tal que $\|x_n\| \geq n$. Essa sequência teria de possuir uma subsequência convergente por conta da compacidade de K. Entretanto, pela Proposição 2.1, isso implicaria que os termos dessa subsequência seriam limitados. Isso é um absurdo, uma vez que toda subsequência cresce indeterminadamente. □

Observação. Em alguns espaços normados vale a recíproca. Isto é, todo conjunto fechado e limitado é compacto. Chamamos essa propriedade de Heine-Borel e é um fato crucial para a análise dos espaços normados que $\mathbb{R}$ possui a propriedade de Heine-Borel[18].

[18] Há uma prova deste resultado na seção 3.3 da referência (ABBOTT, 2015).

Exemplo 2.22. Em ℓ_n^1, a esfera S_1 de raio unitário é compacta.

Teorema 2.5. Seja $f : X \to Y$ uma função contínua entre espaços normados X e Y e K um subconjunto compacto de X, então $f(K)$ é compacto em Y.

Demonstração. Seja $(y_n)_{n\in\mathbb{N}}$ uma sequência de elementos em $f(K)$, podemos escrever $y_n = f(x_n)$ para algum $x_n \in K$. Se considerarmos a sequência $(x_n)_{n\in\mathbb{N}}$, então, pela compacidade de K, $(x_n)_{n\in\mathbb{N}}$ deve possuir alguma subsequência convergente $(x_{n'})_{n'\in\mathbb{N}'}$ com $x_{n'} \to x' \in K$. Pela caracterização sequêncial de funções contínuas, temos que $f(x_{n'}) \to f(x') \in f(K)$. Assim, $(y_n)_{n\in\mathbb{N}}$ possui uma subsequência convergente em $f(K)$ e $f(K)$ também é compacto. □

Corolário 2.4. Seja $f : K \to \mathbb{R}$ uma função real contínua definida num subconjunto compacto de um espaço normado X, então existem x_M e x_m em K tal que $f(x_M)$ é o valor máximo atingido por f e $f(x_m)$ é o valor mínimo atingido por f.

Demonstração. Pelo resultado anterior, $f(K)$ é um subconjunto compacto de $\mathbb{R}$. Isso significa que a imagem é um conjunto fechado e limitado. Como é limitado, possuí ínfimo e supremo e, como é fechado, o ínfimo e o supremo são elementos de $f(K)$. Assim, existe um valor máximo $f(x_M)$ e um valor mínimo $f(x_m)$. □

Figura 5 – Primeiras etapas na construção da Poeira de Cantor. Fonte: Domínio Público.

Extra 3. Um subconjunto muito interessante do intervalo $[0,1]$ é a Poeira de Cantor. De fato, comece com o intervalo $C_1 = [0,1]$, depois remova o terço do meio para obter $C_2 = [0,\frac{1}{3}] \cup [\frac{2}{3},1]$. Continue repetindo esse processo, cada vez removendo o terço do meio de cada subintervalo de C_n. Isso define uma sequência $(C_n)_{n\in\mathbb{N}}$ de subconjuntos de $[0,1]$ que ficam cada vez menores. Veja esse processo na Figura 5. É possível falar do limite desta sequência como sendo a intersecção de todos os seus elementos:

$$C = \bigcap_{n=1}^{\infty} C_n.$$

Este C é a Poeira de Cantor. Verifique:

1. C não é vazio e não é enumerável.
2. C é compacto.

2.6 NORMAS EQUIVALENTES

O quão importante é a escolha da norma para um espaço vetorial? Vimos que todas as definições analíticas: topologia, convergência e continuidade dependem diretamente da norma. De fato, só é possível determinar quais sequências são convergentes e quais funções são contínuas uma vez que se tenha fixado uma norma. Assim, é natural buscar uma condição a partir da qual duas normas produzam as mesmas funções contínuas, sequências convergentes e topologia. Nesse espírito, vejamos a seguinte definição:

Definição 2.23. Dado um espaço vetorial V, dizemos que duas normas $\|\cdot\|$ e $\|\cdot\|'$ para V são equivalentes quando há números reais positivos a, b tais que, para todo $x \in V$

$$\|x\| \leq a \|x\|',$$

$$\|x\|' \leq b \|x\|.$$

Observação. Se definirmos $\|\cdot\| \cong \|\cdot\|'$ como "$\|\cdot\|$ é equivalente a $\|\cdot\|'$", então $\cong$ forma uma relação de equivalência sobre o conjunto de todas as normas sobre V.

Exemplo 2.23. Em $\mathbb{R}^n$, temos que $\|\cdot\|_1 \cong \|\cdot\|_\infty$. De fato, $\|x\|_1 = \sum_{i=1}^n |x_i| \leq \sum_{i=1}^n \|x\|_\infty = n\|x\|_\infty$ e $\|x\|_\infty = \max |x_i| \leq \sum_{i=1}^n |x_i| = \|x\|_1$. Logo $a = n$ e $b = 1$.

Exemplo 2.24. Em $\mathbb{R}^n$, temos que $\|\cdot\|_2 \cong \|\cdot\|_\infty$. De fato, $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2} \leq \sqrt{\sum_{i=1}^n \|x\|_\infty^2} = \sqrt{n}\|x\|_\infty$ e $\|x\|_\infty = \max |x_i| \leq \sqrt{\sum_{i=1}^n |x_i|^2} = \|x\|_1$. Logo $a = \sqrt{n}$ e $b = 1$.

Observação. Por transitividade, temos que $\|\cdot\|_1 \cong \|\cdot\|_2$.

Teorema 2.6. Quando duas normas são equivalentes, toda estrutura analítica é indistinguível. De fato, seja V um espaço vetorial e $\|\cdot\|$ e $\|\cdot\|'$ duas normas equivalentes sobre V, então:

1. Se uma sequência $(x_n)_{n\in\mathbb{N}}$ de vetores de V converge para $x \in V$ com respeito à norma $\|\cdot\|$, então também converge com respeito à $\|\cdot\|'$ e vice-versa.

2. Se uma sequência $(x_n)_{n\in\mathbb{N}}$ de vetores de V é Cauchy com respeito à norma $\|\cdot\|$, então também o é com respeito à $\|\cdot\|'$ e vice-versa.

3. Seja $f : V \to W$ uma função contínua com respeito à $\|\cdot\|_V$ e $\|\cdot\|_W$ e sejam $\|\cdot\|_V \cong \|\cdot\|_V'$, $\|\cdot\|_W \cong \|\cdot\|_W'$. Assim, T é contínua com respeito à $\|\cdot\|_V'$ e $\|\cdot\|_W'$.

4. Se $A \subseteq V$ é aberto com respeito à norma $\|\cdot\|$, então também é aberto com respeito à norma $\|\cdot\|'$. Idem para fechados e compactos.

5. Se V é um espaço de Banach com respeito à $\|\cdot\|$, então também o é com respeito à norma $\|\cdot\|'$, e vice-versa.

Demonstração. Para 1, suponha que $x_n \to x$ com respeito à $\|\cdot\|$. Em particular, tendo fixado $\varepsilon > 0$, há $N > 0$ tal que $\forall n \geq N : \|x - x_n\| < \varepsilon/b$, em que b é a constante dada pela equivalência de norma. Com isso, podemos concluir que $\|x - x_n\|' \leq b\|x - x_n\| < \varepsilon$ sempre que $n \geq N$. Portanto, $x_n \to x$ com respeito à norma $\|\cdot\|'$. Note que o caso em que $x_n \to x$ com respeito à $\|\cdot\|'$ é análogo. A prova para 2 é muito similar a que acabamos de ver e, portanto, a omitimos.

Para 3, seja x um ponto no domínio de f, demonstraremos que f é contínua sobre x com respeito às normas $\|\cdot\|'_V$ e $\|\cdot\|'_W$. Com efeito, tendo fixado $\varepsilon > 0$, sabemos que há $\delta > 0$ tal que $\|f(x) - f(y)\|_W < \varepsilon/b_W$ sempre que $\|x - y\|_V < \delta$, em que b_W está associada à equivalência entre $\|\cdot\|_W$ e $\|\cdot\|'_W$. Assim, se deixarmos $\delta' = \delta/a_V$, em que a_V está associada à equivalência entre $\|\cdot\|_V$ e $\|\cdot\|'_V$, então $\|x - y\|_V < \delta$ sempre que $\|x - y\|'_V < \delta'$. Com isso, temos δ' que faz com que $\forall y : \|x - y\|'_V < \delta' \to \|f(x) - f(y)\|'_W \leq b_W\|f(x) - f(y)\|_W < \varepsilon$. Como as escolhas de ε e x foram arbitrárias, podemos concluir que f é contínua com respeito à $\|\cdot\|'_V$ e $\|\cdot\|'_W$.

Para 4, suponha que $A \subseteq X$ é um conjunto aberto, então, para todo $x \in A$, temos r tal que $B_r(x) \subseteq A$. Gostaríamos de mostrar que há r' para cada x que faz com que $B'_{r'}(x) \subseteq A$. Para isto, veja que, se definirmos $r' = r/a$, onde a está relacionada com a equivalência de $\|\cdot\|$ e $\|\cdot\|'$, então $y \in B'_{r'}(x)$ implica em $\|x - y\|' < r/a$, que por sua vez, implica em $\|x - y\| \leq a\|x - y\|' < r$. Podemos concluir que $y \in B_r(a) \subseteq A$. Assim, $B'_{r'}(x) \subseteq A$ e A também é aberto com respeito à $\|\cdot\|'$. O caso de fechados e compactos é imediato se considerarmos o item 1 desse teorema em conjunto com as caracterizações sequenciais que demos para fechados e compactos na seção passada.

Para 5, suponha que V é de Banach com respeito à $\|\cdot\|$ e que $(x_n)_{n\in\mathbb{N}}$ é uma sequência de Cauchy com respeito à $\|\cdot\|'$. Pelo item 2, $(x_n)_{n\in\mathbb{N}}$ também é de Cauchy com respeito à $\|\cdot\|$. Como $(V, \|\cdot\|)$ é de Banach, sabemos que $x_n \to x$ com respeito à $\|\cdot\|$ para algum $x \in V$. Assim, utilizando o item 1, podemos concluir que $(x_n)_{n\in\mathbb{N}}$ é convergente em $(V, \|\cdot\|')$ e que esse espaço é de Banach. □

Observação. Em $C[a,b]$, as normas $1 \leq p < \infty$ não são equivalentes à norma $p = \infty$, pois os espaços resultantes das primeiras são incompletos, enquanto $(C[a,b], \|\cdot\|_\infty)$ é completo.

2.7 ESPAÇOS NORMADOS DE DIMENSÃO FINITA

Espaços normados de dimensão finita são de interesse pois aparecem frequentemente, às vezes como núcleo/imagem de uma transformação linear, às vezes como parte de argumentos, etc. Como veremos nesta seção, espaços normados de dimensão finita possuem muitas propriedades interessantes; e.g., todos são de Banach e possuem a propriedade de Heine-Borel. Além disso, os operadores lineares são todos contínuos. Em dimensão infinita, a situação não é tão simples: é comum se deparar com espaços incompletos, nenhum espaço possui a propriedade de Heine-Borel e há exemplos de transformações lineares importantes que são descontínuas. Vejamos agora um resultado que parece ser simples, mas será fonte principal dos resultados

mencionados para dimensão finita.

Lema 2.2. Sejam $e_1, e_2, ..., e_n$ vetores linearmente independentes em um espaço normado X, então, existe $b > 0$ tal que, para quaisquer $x_1, x_2, ..., x_n \in \mathbb{R}$, tem-se

$$\sum_{i=1}^{n} |x_i| \leq b \left\| \sum_{i=1}^{n} x_i e_i \right\|. \tag{2.7}$$

Observação. Este lema nos diz duas coisas: combinações lineares com coeficientes grandes produzem vetores grandes; e, se definirmos $\|x\|_* = \sum_{i=1}^{n} |x_i|$, onde $x = \sum_{i=1}^{n} x_i e_i$ e $\{e_1, e_2, ..., e_n\}$ é uma base, então $\|x\|_* \leq b\|x\|$. Isso é grande parte do caminho para mostrar que todas as normas sobre X, quando X tem dimensão finita, são equivalentes entre si.

Demonstração. Vamos reformular um pouco o problema. Note que uma solução para a desigualdade existe se, e somente se, existe uma solução para seguinte desigualdade:

$$1 \leq b \left\| \sum_{i=1}^{n} y_i e_i \right\|, \quad \sum_{i=1}^{n} |y_i| = 1. \tag{2.8}$$

A inexistência de solução para essa desigualdade implicaria que é possível construir combinações lineares com coeficientes de mesmo tamanho, mas com norma arbitrariamente pequena. Isso não é possível, pois podemos encontrar coeficientes $(y_1, y_2, ..., y_n)$ que produzem um vetor de norma mínima. De fato, considere o conjunto compacto $S = \{y \in \mathbb{R}^n : \|y\|_1 = 1\}$. Podemos definir $F : S \to \mathbb{R}$ tal que

$$F(y) = \left\| \sum_{i=1}^{n} y_i e_i \right\|.$$

F assim definida é contínua. Com efeito, pela desigualdade triangular reversa, temos:

$$|F(y) - F(z)| = \left| \left\| \sum_{i=1}^{n} y_i e_i \right\| - \left\| \sum_{i=1}^{n} z_i e_i \right\| \right| \leq \left\| \sum_{i=1}^{n} y_i e_i - \sum_{i=1}^{n} z_i e_i \right\| = \left\| \sum_{i=1}^{n} (y_i - z_i) e_i \right\|.$$

Podemos aplicar a desigualdade triangular tradicional para obter:

$$|F(y) - F(z)| \leq \sum_{i=1}^{n} |y_i - z_i| \, \|e_i\|.$$

Majorando pelo máximo $M = \max \|e_i\|$:

$$|F(y) - F(z)| \leq M \sum_{i=1}^{n} |y_i - z_i|.$$

Observamos que a soma da direita é simplesmente $\|y - z\|_1$. Isso mostra que F é contínua. Como F é contínua, o Corolário 2.4 nos diz que há[19] $y_0 \in S$ com $0 < F(y_0) \leq F(y)$ para todo $y \in Y$. Por sua vez, isso nos permite concluir que há solução para a desigualdade 2.8. □

[19] $F(y_0) > 0$ pois o vetor nulo, único com norma nula, não pertence à S.

Teorema 2.7. Seja $\|\cdot\|$ uma norma qualquer em X espaço normado de dimensão finita n. Se definirmos $\|x\|_* = \sum_{i=1}^n |x_i|$, onde $x = x_1e_1 + x_2e_2 + ... + x_ne_n$, $x_i \in \mathbb{R}$ e $\{e_1, e_2, ..., e_n\}$ é uma base de X, então $\|\cdot\|_*$ é uma norma. Além disso, $\|\cdot\|_*$ é equivalente à $\|\cdot\|$. Assim, só há uma classe de equivalência para as normas sobre X.

Demonstração. A verificação de que $\|\cdot\|_*$ de fato é uma norma é simples e fica como sugestão de exercício para o leitor. Assim, nos concentramos em mostrar que existe $a > 0$ tal que $\|x\| \leq a\|x\|_*$, para todo x. Com efeito, se deixarmos $a = \max \|e_i\|$, então $\|x\| \leq a\|x\|_*$. Por fim, sabemos pelo Teorema 2.2 que há $b > 0$ tal que $\|x\|_* \leq b\|x\|$ para todo x. Podemos concluir que $\|\cdot\|$ e $\|\cdot\|_*$ são equivalentes. □

Corolário 2.5. Todo espaço normado de dimensão finita é de Banach.

Demonstração. Seja X um espaço normado de dimensão finita e seja $\{e_1, e_2, ..., e_n\}$ uma base de X. Utilizando o resultado anterior, mostraremos que X é completo com relação à $\|\cdot\|_*$. Como $\|\cdot\| \cong \|\cdot\|_*$, pelo Teorema 2.6, teremos que X também é completo com relação à $\|\cdot\|$.

Com efeito, considere uma sequência de Cauchy arbitrária $(x_k)_{k\in\mathbb{N}}$ em X. Temos que $\forall \varepsilon > 0 : \exists N > 0 : \forall k, l \geq N : \sum_{i=1}^n |x_i^k - x_i^l| = \|x_k - x_l\|_* < \varepsilon$. Observe que, ao fixar i, obtemos uma sequência de Cauchy $(x_i^k)_{k\in\mathbb{N}}$ em $\mathbb{R}$. Como $\mathbb{R}$ é completo, temos $x_i^k \to b_i$ para algum $b_i \in \mathbb{R}$. Defina $b = b_1e_1 + b_2e_2 + ... + b_ne_n$. Logo, $\|x_k - b\|_* = \sum_{i=1}^n |x_i^k - b_i|$. Ao deixarmos $N = \max N_1, N_2, ..., N_n$ onde N_i é tal que $\forall k \geq N_i : |x_i^k - b_i| < \varepsilon/n$, obtemos $\|x_k - b\|_* < \varepsilon$ quando $k \geq N$ e, portanto, $x_k \to b$ em X. Toda sequência de Cauchy em $(X, \|\cdot\|_*)$ converge. Consequentemente, X é completo com respeito à sua norma original. □

Corolário 2.6. Seja X um espaço normado, então todo subespaço de X que tem dimensão finita é fechado.

Demonstração. Seja Y um subespaço de X com dimensão finita. Se restringirmos a norma de X para Y, temos que $Z = (Y, \|\cdot\|_{|_Y})$ é um espaço normado. Pelo teorema anterior, sabemos que Y é completo. Adicionalmente, pelo Corolário 2.3, temos que Y é fechado. □

Teorema 2.8. Sejam X e Y espaços normados, onde X tem dimensão finita, e seja $T : X \to Y$ um operador linear entre X e Y. Então, T é contínuo.

Demonstração. Este resultado segue do último lema se usarmos a norma $\|\cdot\|_*$. Realmente, fixe uma base $\{e_1, e_2, ..., e_n\}$ para X e tome $x = x_1e_1 + x_2e_2 + ... + x_ne_n \in X$. Veja que $\|Tx\| = \|T(\sum_{i=1}^n x_ie_i)\| = \|\sum_{i=1}^n x_iTe_i\| \leq \sum_{i=1}^n |x_i| \|Te_i\|$. Ao definir $k = \max\{\|Te_i\|\}$, podemos ver que $\sum_{i=1}^n |x_i| \|Te_i\| \leq k\|x\|_*$. Veremos na seção seguinte que todo operador limitado (isto é, tal que há $k > 0$ com $\|Tx\| \leq k\|x\|$ para todo x) é contínuo. Assim, T é contínuo com respeito às normas $\|\cdot\|_Y$ e $\|\cdot\|_*$. Concluímos relembrando que $\|\cdot\|_*$ é equivalente à $\|\cdot\|_X$ e que isto implica na continuidade de T com respeito à $\|\cdot\|_X$. □

Teorema 2.9. Todos os espaços normados de dimensão finita gozam da propriedade de Heine-Borel. Isto é, os conjuntos compactos são exatamente os conjuntos fechados e limitados.

Demonstração. Seja $K \subseteq X$ um espaço normado de dimensão finita. Se K é compacto, então K é fechado e limitado. Mostremos então que, em X, vale a recíproca. Com efeito, seja K um subconjunto fechado e limitado de X e seja $(x_k)_{k\in\mathbb{N}}$ uma sequência em X. Adotamos novamente a norma $\|\cdot\|_*$. Como K é limitado, há $M \in \mathbb{R}$ tal que $\forall k : \|x_k\|_* \leq M$; ou seja, $\sum_{i=1}^n |x_i^k| \leq M$ e $|x_i^k|$ com i fixo são sequências limitadas em $\mathbb{R}$. Pelo teorema de Bolzano Weierstrass[20], há uma subsequência convergente de $(x_1^k)_{k\in\mathbb{N}}$, a qual denotamos por $(x_1^{k'})$. Aplicando novamente o teorema, obtemos uma subsequência de $(x_2^{k'})_{k'\in\mathbb{N}}$ que é convergente, a qual denotamos por $(x_2^{k''})_{k''\in\mathbb{N}}$. Prosseguindo assim, obtemos um conjunto de índices N' de forma que as sequências $(x_i^L)_{L\in N'}$ convergem. De fato, seja c_i o limite da sequência $(x_i^L)_{L\in N'}$, então, se definirmos $c = \sum_{i=1}^n c_i e_i$, temos que $\sum_{i=1}^n |x_i^{k'} - c_i| \to 0$ e $x_{k'} \to c$. Finalizamos a prova observando que, como K é fechado, $c \in K$. □

Lema 2.3. (Lema de Riesz) Sejam Y e Z subespaços de um espaço normado X tais que Y é fechado e está propriamente contido em Z, então para cada número real r no intervalo $(0,1)$ há $z \in Z$ tal que

$$\|z\| = 1, \quad \|z - y\| \geq r, \forall y \in Y. \tag{2.9}$$

Demonstração. Uma prova deste teorema pode ser encontrada na seção 2.5 da referência (KREYSZIG, 1989). □

Observação. O que este lema nos diz é que podemos achar um vetor $z \in Z - Y$ quase perpendicular ao subespaço Y.

Corolário 2.7. A bola unitária fechada é compacta em um espaço normado se, e somente se, o espaço possuir dimensão finita. Assim, podemos concluir que nenhum espaço de dimensão infinita possui a propriedade de Heine-Borel.

Demonstração. Suponha, por meio de contradição, que há um espaço normado X com dimensão infinita cuja bola unitária é compacta. Vamos construir uma sequência de vetores $(x_n)_{n\in\mathbb{N}}$ na bola que não possui subsequência convergente. De fato, começamos com um vetor qualquer x_1 com norma unitária e então aplicamos o Lema de Riesz em $Y = \text{Span}\{x_1\}$ e $Z = X$ para obter $x_2 \in X$ tal que $\|x_2 - y\| \geq 1/2$, para qualquer $y \in Y$. Em particular, temos que $\|x_2 - x_1\| \geq 1/2$. Podemos repetir esse passo tomando $Y = \text{Span}\{x_1, x_2\}$ para obter $x_3 \in X$ com $\|x_3 - y\| \geq 1/2$ para todo $y \in Y$. Novamente, isso nos dá que $\|x_3 - x_1\|$ e $\|x_3 - x_2\|$ são ambas pelo menos $1/2$. Prosseguindo dessa maneira, é possível construir uma sequência de vetores $(x_n)_{n\in\mathbb{N}}$ tais que $\|x_i - x_j\| \geq 1/2$, para quaisquer índices i e j. Pela nossa suposição original, $(x_n)_{n\in\mathbb{N}}$ possui uma subsequência convergente. Assim, é possível encontrar $N > 0$ a partir do qual todos os pontos da subsequência estão a uma distância menor do que $1/2$ um do outro. Isso é um absurdo. Portanto, podemos concluir que a bola unitária não é um compacto quando a dimensão de X é infinita. □

[20] Este teorema é apresentado como Teorema 2.5.5 na referência (ABBOTT, 2015).

Extra 4. Como a condição de Heine-Borel falha em espaços normados de dimensão infinita, somos motivados a investigar o que está faltando nesses espaços para que conjuntos fechados e limitados sejam compactos. A resposta para essa pergunta varia de acordo com o espaço. Em $C[a,b]$ com a norma do supremo $\|\cdot\|_\infty$, o que falta é a condição de equicontinuidade. Dizemos que um conjunto $A \subseteq C[a,b]$ é uniformemente equicontinuo se:

$$\forall \varepsilon > 0 : \exists \delta > 0 : \forall x, y, f : |x-y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon.$$

Assim, em $(C[a,b], \|\cdot\|_\infty)$, um conjunto é compacto se, e somente se, for fechado, limitado e uniformemente equicontinuo. Esse resultado recebe o nome de Teorema de Arzelà-Ascoli.

2.8 OPERADORES E FUNCIONAIS

Operadores e funcionais tendem a ser a parte mais interessante tanto de Álgebra Linear quanto de Análise Funcional. Nesta seção, relembraremos a definição de operadores e funcionais, veremos alguns exemplos e introduziremos o conceito de operador/funcional limitado.

Definição 2.24. Sejam X e Y espaços normados, então um operador linear é uma função $T : X \to Y$ satisfazendo:

1. $\forall x, y : T(x+y) = Tx + Ty$.
2. $\forall \lambda, x : T(\lambda \cdot x) = \lambda \cdot Tx$.

Se adicionalmente houver uma constante $k > 0$ tal que

$$\forall x : \|Tx\| \leq k \|x\|$$

dizemos que operador é limitado.

Definição 2.25. Um funcional linear $f : X \to Y$ é um operador linear cujo contradomínio é $\mathbb{R}$. Dizemos que f é limitado se há uma constante $k > 0$ tal que

$$\forall x : |f(x)| \leq k \|x\|.$$

Lema 2.4. Seja $T : X \to Y$ um operador linear entre espaços normados X e Y, então, são equivalentes:

1. T é uniformemente contínuo.
2. T é contínuo.
3. T é contínuo na origem.

4. T é limitado.

Demonstração. É imediato que $1 \to 2$ e $2 \to 3$. Agora, suponha que T é contínuo na origem. Se fixarmos $\varepsilon = 1$ na definição de continuidade, temos $\exists \delta : \forall x : \|x\| < \delta \to \|Tx\| < 1$. Seja agora x um elemento qualquer de X, podemos normalizar x para que sua norma fique menor que δ: $\|\delta/(2\|x\|)x\| = \delta/(2\|x\|)\|x\| = \delta/2 < \delta$. Com isso, $\|T(\delta/(2\|x\|)x)\| < 1$ vira $\|Tx\| < 2/\delta\|x\|$, que mostra que T é limitado.

Para $4 \to 1$, considere $\delta = \varepsilon/k$ na definição de continuidade uniforme, então, para quaisquer x e y com $\|x-y\| < \delta$, temos $\|Tx - Ty\| = \|T(x-y)\| \leq k\|x-y\| < \varepsilon$. □

Definição 2.26. Definimos a norma de um operador limitado como sendo

$$\|T\| = \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|} = \sup_{\|x\|=1} \|Tx\|. \tag{2.10}$$

Observação. Podemos pensar na quantidade $\frac{\|Tx\|}{\|x\|}$ como representando o fator de escala de T sobre x. É possível ver que o fator de escala é o mesmo para qualquer vetor (exceto o vetor nulo) no espaço gerado por x. De fato, $\frac{\|T(\lambda \cdot x)\|}{\|\lambda \cdot x\|} = \frac{|\lambda|\|Tx\|}{|\lambda|\|x\|} = \frac{\|Tx\|}{\|x\|}$. Por esta razão, só é necessário considerar os vetores unitários na hora de calcular a norma de T.

Exemplo 2.25. O operador nulo entre dois espaços normados $0 : X \to Y$ dado por $T(x) = 0$ é limitado e $\|T\| = 0$.

Exemplo 2.26. O operador identidade $I : X \to X$ dado por $I(x) = x$ é limitado e possuí norma $\|I\| = 1$.

Exemplo 2.27. Seja $\{x_1, x_2, ..., x_m\}$ uma base para X e $\{y_1, y_2, ..., y_m\}$ uma base para Y, em que X e Y são espaços normados de dimensão finita, então toda transformação linear $T : X \to Y$ pode ser representada por uma matriz M em que $m_{i,j}$ é a i-ésima coordenada de $T(x_j)$ com respeito à base fixada de Y.

Exemplo 2.28. O operador derivada $D : C^1[0,1] \to C[0,1], D : f \mapsto f'$ não é limitado. De fato, se considerarmos a norma do máximo $\|f\|_\infty = \max_{x \in [0,1]} |f(x)|$, então todo elemento da sequência $f_n = \text{sen}(2\pi nx)$ tem norma unitária, porém, suas respectivas derivadas são arbitrariamente grandes: $\|f_n'\|_\infty = \|2\pi n cos(2\pi nx)\|_\infty = 2\pi n$.

Exemplo 2.29. O operador de Volterra (integral indefinida) $V : C[0,1] \to C^1[0,1]$ dado por $V(f)(y) = \int_0^y f(t)dt$ é limitado. Com efeito, note que $\|V(f)\|_\infty = \max_{y \in [0,1]} |\int_0^y f(t)dt| \leq \max_{y \in [0,1]} \int_0^y |f(t)|dt \leq \max_{y \in [0,1]} \int_0^y \|f\|_\infty dt = \|f\|_\infty$. Isso mostra que V é limitado e que $\|V\| \leq 1$. É possível atingir $\frac{\|V(g)\|_\infty}{\|g\|_\infty} = 1$ com a função constante $g = 1$. Assim, concluimos que $\|V\| = 1$.

Exemplo 2.30. Em um espaço normado de funções X, é sempre possível definir o funcional avaliação como $T_x(f) = f(x)$. T_x é linear pois $T_x(f+g) = (f+g)(x) = f(x) + g(x) = T_x(f) + T_x(g)$ e $T_x(\lambda \cdot f) = (\lambda \cdot f)(x) = \lambda f(x) = \lambda T_x(f)$. A continuidade de T_x depende do espaço X.

Com efeito, suponha que X é $C[a,b]$ com a norma do máximo, então $|T_x(f)| = |f(x)| \leq \|f\|_\infty$ e T_x é limitado. Alternativamente, suponha que X é $C[0,1]$ com a norma $p = 2$. É possível criar uma sequência de funções cuja norma 2 sempre é 1, mas $|f_n(0)|$ cresce intermitentemente. Com efeito, seja

$$f_n(x) = \begin{cases} n - \frac{n^3}{3}x, & \text{se } x \leq \frac{3}{n^2}, \\ 0, & \text{se } x > \frac{3}{n^2}. \end{cases}$$

Então $f_n(0) = n$, mas

$$\|f_n\|_2^2 = \int_0^{\frac{3}{n^2}} \left(n - \frac{n^3}{3}x\right)^2 dx = \int_0^{\frac{3}{n^2}} n^2 - \frac{2n^4x}{3} + \frac{n^6}{9}x^2 dx = 1.$$

Assim, T_0 é ilimitado e, portanto, descontínuo.

Exemplo 2.31. O funcional integral definida $A(f) : C[a,b] \to \mathbb{R}$ dado por $A(f) = \int_a^b f(t)dt$ é limitado. De fato, $\|A(f)\|_\infty = |\int_a^b f(t)dt| \leq \int_a^b |f(t)|dt \leq \int_a^b \|f\|_\infty dt = (b-a)\|f\|_\infty$. Assim, $\|A\| \leq (b-a)$. Novamente, podemos usar $f = 1$ para concluir que $\|A\| = b - a$.

Proposição 2.9. O núcleo de um operador limitado é fechado, mas sua imagem não necessáriamente o é.

Demonstração. Seja $T : X \to Y$ um operador limitado e $(x_n)_{n\in\mathbb{N}}$ uma sequência de vetores no núcleo de T com limite x. Como T é limitado, temos que $0 = T(x_n) \to T(x)$. Isso força $T(x) = 0$ e $x \in \operatorname{Ker} T$.

Considere agora o operador $\lambda : \ell^1 \to \ell^1$ dado por $\lambda(x)(j) = (1/jx_j)$. Isto é, o operador que manda $(x_1, x_2, ...)$ para $(x_1, 1/2x_2, ...)$. Claramente, λ é linear e limitado. Contudo, sua imagem não é fechada. De fato, considere a sequência a seguir:

$$s_1 = (1, 0, 0, ...)$$
$$s_2 = (1, \frac{1}{2}, 0, ...)$$
$$...$$

A sequência formada pelas imagens $\lambda(s_n)$ converge para $(\frac{1}{n^2})$, que não pertence à imagem do operador λ. Assim, podemos concluir que $\lambda(\ell^1)$ não é um subespaço fechado de ℓ^1. □

Teorema 2.10. (Hahn-Banach) Seja $f : Z \to \mathbb{R}$ um funcional linear limitado definido num subespaço Z de um espaço normado X, então, existe uma extensão $\tilde{f}$ do funcional f para X tal que

$$\|f\| = \left\|\tilde{f}\right\|.$$

Demonstração. Uma prova detalhada deste teorema pode ser encontrada na seção 4.2 da referência (KREYSZIG, 1989).

□

3 ESPAÇOS DE HILBERT E ANALOGIAS GEOMÉTRICAS

> The notion of a Hilbert space sheds light on so much of modern mathematics, from number theory to quantum mechanics, that if you do not know at least the rudiments of Hilbert space theory then you cannot claim to be a well-educated mathematician.
>
> *Timothy Gowers*

Espaços de Hilbert são espaços vetoriais equipados com três noções adicionais: distância (advinda da métrica), tamanho de vetor (advinda da norma) e ortogonalidade (advinda do produto interno). Além disso, espaços de Hilbert são completos, o que facilita o uso de técnicas da Análise. Por fim, esses espaços possuem uma geometria muito similar à Euclidiana, o que faz com que muitos teoreomas clássicos - como o Teorema de Pitágoras - possuam análogos para espaços de Hilbert. Isso auxilia na resolução de problemas. Desta forma, esse capítulo objetiva apontar para o fato de que quando se estiver trabalhando com um espaço de Hilbert, é aconselhavel usar intuição geométrica. A Figura 6 mostra os espaços que serão estudados neste capítulo.

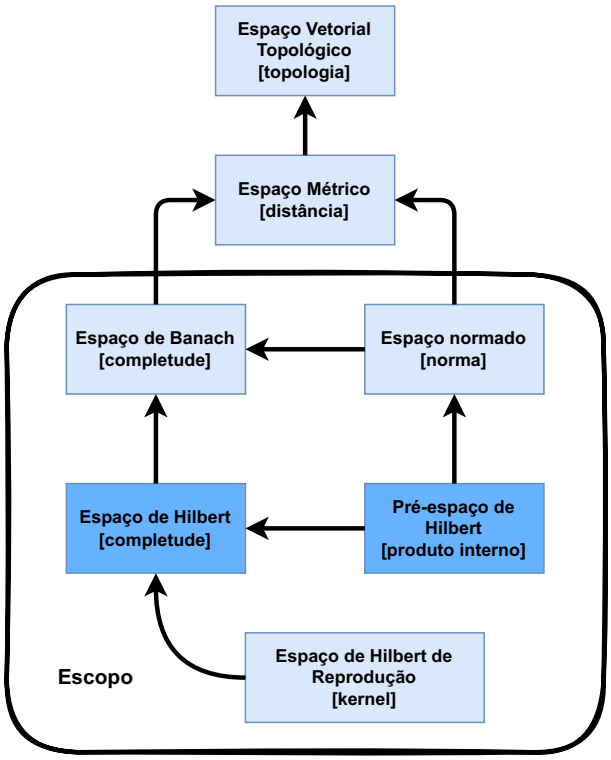


Figura 6 – Relação “é um tipo de” entre diferentes tipos de espaços estudados em Análise Funcional.

Definição 3.1. Seja $V = (X, +, \cdot)$ um espaço vetorial e seja $\langle \cdot, \cdot, \rangle : X \times X \to \mathbb{R}$ uma função que satisfaça as seguintes propriedades:

1. **Positivo-definida:** $\forall x : \langle x,x \rangle \geq 0$ e $\langle x,x \rangle = 0 \iff x = 0$.
2. **Simetria:** $\forall x, y : \langle x,y \rangle = \langle y,x \rangle$.
3. **Linearidade:** $\forall x, y, z, \lambda : \langle x + \lambda \cdot z, y \rangle = \langle x,y \rangle + \lambda \langle z,y \rangle$.

Então $P = (V, \langle \cdot, \cdot \rangle)$ é um pré-espaço de Hilbert. É sempre possível definir uma norma em um pré-espaço de Hilbert. De fato, basta definir

$$\|x\| = \sqrt{\langle x,x \rangle}. \tag{3.1}$$

A prova deste fato depende da desigualdade de Cauchy-Schwarz, que será vista em uma seção posterior. Por fim, caso P seja um espaço de Banach com respeito à $\|\cdot\|$, então o chamamos de espaço de Hilbert. Neste caso, usamos a letra H ao invés de P.

Definição 3.2. Dizemos que dois vetores x e y em um pré-espaço de Hilbert P são ortogonais quando $\langle x,y \rangle = 0$. Denotamos ortogonalidade por $x \perp y$. Similarmente, dizemos que x é ortogonal a um subconjunto $Y \subseteq P$, $x \perp Y$, se $x \perp y$ para todo $y \in Y$.

3.1 EXEMPLOS

Nesta seção, abordaremos três exemplos: $\boldsymbol{\ell}^2_n$, $\boldsymbol{\ell}^2$ e $\mathbf{L}^2[0,1]$. Na seção sobre kernels e aprendizado de máquina, serão abordados mais alguns exemplos.

Exemplo 3.1. Seja $V = \boldsymbol{\ell}^2_n$ e $\langle x,y \rangle = \sum_{i=1}^n x_i y_i$ o produto escalar usual, então, $H = (V, \langle \cdot, \cdot \rangle)$ é um espaço de Hilbert.

Demonstração. Sejam $x = (x_1, x_2, ..., x_n), y = (y_1, y_2, ..., y_n)$ e $z = (z_1, z_2, ..., z_n)$ vetores em V e λ um escalar, começamos observando que $\langle x,x \rangle = \sum_{i=1}^n x_i^2 \geq 0$ e que essa soma é zero se, e somente se, $x = 0$. Adicionalmente, pela comutatividade da multiplicação, temos que $\langle x,y \rangle = \langle y,x \rangle$. Por fim, $\langle x + \lambda \cdot z, y \rangle = \sum_{i=1}^n (x_i + \lambda z_i) y_i = \sum_{i=1}^n x_i y_i + \lambda \sum_{i=1}^n z_i y_i = \langle x,y \rangle + \lambda \langle z,y \rangle$. Portanto, H é um pré-espaço de Hilbert. Agora, observe que $\|x\| = \sqrt{\langle x,x \rangle} = \|x\|_2$. Vimos na Seção 2.3 que $\mathbb{R}^n$ munido desta norma é completo. Assim, H é um espaço de Hilbert. □

Exemplo 3.2. Seja $V = \boldsymbol{\ell}^2$ o espaço das sequências quadrado-somáveis e seja $\langle x,y \rangle = \sum_{i=1}^\infty x_i y_i$, então, $H = (V, \langle \cdot, \cdot \rangle)$ é um espaço de Hilbert.

Demonstração. Constatamos na Seção 2.3 que $(V, \|\cdot\|)$ é um espaço de Banach. Assim, basta verificar que $\langle \cdot, \cdot \rangle$ está bem definida e que satisfaz as propriedades da definição 3.1. De fato, note que $xy \leq \frac{1}{2}(x^2 + y^2)$, o que faz com que $\sum_{i=1}^\infty x_i y_i \leq \frac{1}{2}(\sum_{i=1}^\infty x_i^2 + \sum_{i=1}^\infty y_i^2) < \infty$, mostrando que $\langle x,y \rangle$ está bem definida. Os axiomas de produto interno podem ser verificados através de propriedades básicas de manipulação de séries. □

Exemplo 3.3. Seja $V = C[a,b]$ o espaço vetorial das funções contínuas e seja $\langle f,g\rangle$ dada por $\int_a^b f(t)g(t)dt$, então, $P = (H, \langle\cdot,\cdot\rangle)$ é um pré-espaço de Hilbert.

Demonstração. A verificação dos axiomas de produto interno se dá através do uso de propriedades básicas da integral de Riemman/Lebesgue. □

Observação. Como consequência do contra-exemplo 2.16, P não é um espaço de Hilbert. Entretanto, todo pré-espaço de Hilbert é isomorfo a um subespaço W de um espaço de Hilbert H, como será visto na Seção 3.3.

Extra 5. Em pré-espaços de Hilbert de dimensão finita, cada produto interno dá origem a uma matriz, como pode ser visto a seguir:

$$\langle\cdot,\cdot\rangle \Rightarrow I = \begin{bmatrix} \langle e_1,e_1\rangle & \langle e_1,e_2\rangle & \ldots & \langle e_1,e_n\rangle \\ \langle e_2,e_1\rangle & \langle e_2,e_2\rangle & \ldots & \langle e_2,e_n\rangle \\ \ldots & \ldots & \ldots & \ldots \\ \langle e_n,e_1\rangle & \langle e_n,e_2\rangle & \ldots & \langle e_n,e_n\rangle \end{bmatrix}$$

Observamos que, por conta da linearidade, dada uma base $e_1, e_2, ..., e_n$, o produto interno é totalmente caracterizado por seu valor nos elementos da base: $\langle e_i, e_j\rangle$. Assim,

- As entradas de I não são arbitrárias. Note, por exemplo, que I é simétrica.
- Qual outra condição devemos impor sobre uma matriz $n \times n$ simétrica para que defina um produto interno?

Extra 6. Dado um espaço vetorial V, é sempre possível definir um produto interno sobre V?

3.2 DESIGUALDADES DE CAUCHY-SCHWARZ E TRIANGULAR

Em $\mathbb{R}^n$, sabemos que, dado um vetor v, podemos decompor qualquer vetor u como $u = w + \lambda \cdot v$, onde $w \perp v$. Também é possível fazer isso em pré-espaços de Hilbert.

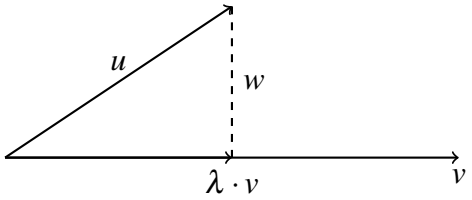


Figura 7 – Decomposição ortogonal de u em termos de v e w.

Teorema 3.1. Seja P um pré-espaço de Hilbert e sejam x e y vetores de P, então, x pode ser escrito como $x = w + \lambda \cdot y$, onde $w \perp y$.

Demonstração. Com efeito, sejam x e y vetores de P. Se $y = 0$, então defina $w = x$ e escolha qualquer escalar λ. Se $y \neq 0$, defina $\lambda = \frac{\langle x,y\rangle}{\langle y,y\rangle}$ e $w = x - \lambda \cdot y$, então $\langle w,y\rangle = \langle x - \lambda \cdot y, y\rangle = \langle x,y\rangle - \lambda \langle y,y\rangle = 0$. Assim, $w \perp y$. □

Desigualdade de Cauchy-Schwarz

Agora veremos uma das desigualdades mais importantes da Matemática (STEELE, 2004).

Teorema 3.2. Sejam x e y vetores num pré-espaço de Hilbert P, então,

$$|\langle x,y\rangle| \leq \|x\| \, \|y\| . \tag{3.2}$$

Além disso, temos igualdade se, e somente se, um dos vetores é múltiplo do outro.

Demonstração. Começamos observando que, se $y = 0$, então a desigualdade é imediata. Se y não é zero, então podemos usar o resultado anterior para decompor x como $x = w + \lambda \cdot y$, onde $w \perp y$. Com isso, temos:

$$\|w + \lambda \cdot y\|^2 = \langle w,w\rangle + 2\langle w, \lambda \cdot y\rangle + \langle \lambda \cdot y, \lambda \cdot y\rangle = \|w\|^2 + \lambda^2 \|y\|^2 \geq \lambda^2 \|y\|^2 .$$

Expandindo λ, obtemos:

$$\|x\|^2 \geq \frac{\langle x,y\rangle^2}{\|y\|^2}.$$

Multiplicando por $\|y\|^2$ e extraindo a raiz quadrada, obtemos a desigualdade desejada. Note que temos igualdade se, e somente se, $\|w\| = 0$, o que, por sua vez, nos permite concluir que temos igualdade se, e somente se, x é múltiplo de y. □

Desigualdade Triangular

Teorema 3.3. Seja P um pré-espaço de Hilbert e sejam x e y vetores de P, então, $\|x+y\| \leq \|x\| + \|y\|$.

Demonstração. Sejam x e y vetores em um pré-espaço de Hilbert P, então $\|x+y\|^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle = \|x\|^2 + 2\langle x,y\rangle + \|y\|^2$. Aplicando Cauchy-Schwarz, obtemos $2\langle x,y\rangle \leq 2\|x\|\,\|y\|$. Agora podemos fatorar o lado direito como $(\|x\| + \|y\|)^2$. Extraindo raizes quadradas de ambos os lados, obtemos a desigualdade desejada. □

Proposição 3.1. Se definirmos $\|x\| = \sqrt{\langle x,y\rangle}$, então $\|\cdot\|$ é uma norma.

Demonstração. De fato,

1. Se $\|x\| = 0$ para algum vetor x, então $\langle x,x\rangle = 0$, o que implica em $x = 0$. Além disso, se $x = 0$, então $\|x\| = 0$.
2. Seja λ um escalar, então $\|\lambda \cdot x\| = \sqrt{\langle \lambda \cdot x, \lambda \cdot x\rangle} = \sqrt{\lambda^2 \langle x,x\rangle} = |\lambda| \sqrt{\langle x,x\rangle} = |\lambda| \, \|x\|$.
3. Pela proposição anterior, se x e y são vetores, então $\|x+y\| \leq \|x\| + \|y\|$.

□

3.3 O COMPLETAMENTO DE UM PRÉ-ESPAÇO DE HILBERT

Proposição 3.2. Seja P um pré-espaço de Hilbert e $(x_n)_{n\in\mathbb{N}}$ e $(y_n)_{n\in\mathbb{N}}$ sequências convergentes em P tais que $x_n \to x$ e $y_n \to y$, então temos $\langle x_n, y_n\rangle \to \langle x,y\rangle$.

Demonstração. Podemos escrever $|\langle x_n,y_n\rangle - \langle x,y\rangle|$ como $|\langle x_n,y_n\rangle - \langle x,y_n\rangle + \langle x,y_n\rangle - \langle x,y\rangle|$. Assim, temos:

$$|\langle x_n,y_n\rangle - \langle x,y\rangle| \leq |\langle x_n,y_n\rangle - \langle x,y_n\rangle| + |\langle x,y_n\rangle - \langle x,y\rangle|.$$

Simplificando, usando a linearidade do produto interno:

$$|\langle x_n,y_n\rangle - \langle x,y\rangle| \leq |\langle x_n - x,y_n\rangle| + |\langle x,y_n - y\rangle|.$$

Agora, aplicamos a desigualdade de Cauchy-Schwarz para obter:

$$|\langle x_n,y_n\rangle - \langle x,y\rangle| \leq \|x_n - x\| \, \|y_n\| + \|x\| \, \|y_n - y\| \, .$$

Como $(y_n)_{n\in\mathbb{N}}$ converge, sabemos pela Proposição 2.1 que $\{\|y_n\|\}_{n\in\mathbb{N}}$ é limitado superiormente. Portanto, sabemos que há um número real M tal que $\|y_n\| \leq M$, para todo n. Assim,

$$|\langle x_n,y_n\rangle - \langle x,y\rangle| \leq \|x_n - x\| M + \|x\| \, \|y_n - y\| \, .$$

Essa expressão claramente tende a zero quando $n \to \infty$. Assim, podemos concluir que $\langle \cdot,\cdot\rangle$ é uma função contínua. □

Teorema 3.4. Seja $P = (V, \langle \cdot,\cdot\rangle)$ um pré-espaço de Hilbert, então existe um espaço de Hilbert $H = (\tilde{V}, \langle \cdot,\tilde{\cdot}\rangle)$ que possui um subespaço W isomorfo[1] à P. Esse W é denso em H.

3.4 ALGUNS RESULTADOS GEOMÉTRICOS

Teorema de Pitágoras

Comecemos com algo simples: o Teorema de Pitágoras. Dado um triângulo retângulo com lados a,b,c, temos que $a^2 + b^2 = c^2$. Por outro lado, se um triângulo tem lados a,b,c

[1] Similarmente ao isomorfismo que preservava distâncias no enunciado do Teorema 2.3, este isomorfismo preserva o produto interno: $\langle Tx,\tilde{Ty}\rangle = \langle x,y\rangle$.

satisfazendo a equação acima, então é um triângulo retângulo. Será que isso vale para espaços de Hilbert?

Teorema 3.5. Seja P um pré-espaço de Hilbert, então, $\|x+y\|^2 = \|x\|^2 + \|y\|^2$ se, e somente se, x e y são vetores ortogonais.

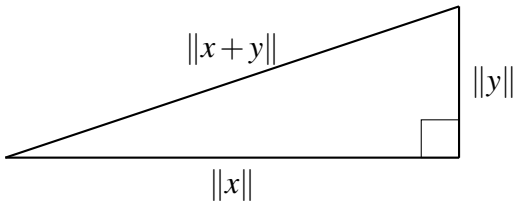


Figura 8 – Teorema de Pitágoras, versão para pré-espaços de Hilbert.

Demonstração. Se $x \perp y$, então $\langle x,y\rangle = 0$. Assim, $\|x+y\|^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + 2\langle x,y\rangle + \langle y\rangle = \|x\|^2 + \|y\|^2$. Para a volta, note que $\|x+y\|^2 = \|x\|^2 + \|y\|^2$ implica em $\langle x,y\rangle = 0$ e $x \perp y$. □

Regra do Paralelogramo

Outro resultado elementar de geometria, que possui um análogo em espaços de Hilbert, é a regra do paralelogramo. Sejam x e y vetores no plano e $\|\cdot\|_2$ a norma usual de $\mathbb{R}^2$, a identidade pode ser expressa da seguinte forma:

$$2\|x\|_2^2 + 2\|y\|_2^2 = \|x+y\|_2^2 + \|x-y\|_2^2. \tag{3.3}$$

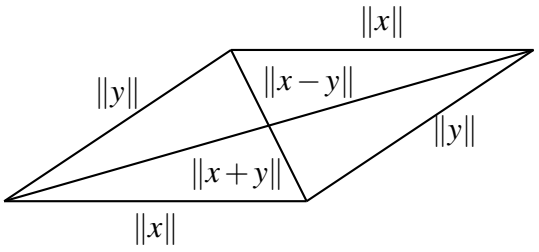


Figura 9 – Regra do Paralelogramo, versão para pré-espaços de Hilbert.

Em termos de espaços de Hilbert, onde a norma $\|\cdot\|$ não necessariamente é a norma-2 do plano, temos:

Teorema 3.6. Seja P um pré-espaço de Hilbert e sejam x e y vetores em P, então,

$$2\|x\|^2 + 2\|y\|^2 = \|x+y\|^2 + \|x-y\|^2. \tag{3.4}$$

Demonstração. Sejam x e y vetores em p, então $\|x+y\|^2 + \|x-y\|^2 = \langle x+y, x+y\rangle + \langle x-y, x-y\rangle = 2\langle x,x\rangle + 2\langle x,y\rangle - 2\langle x,y\rangle + 2\langle y,y\rangle = 2\|x\|^2 + 2\|y\|^2$. □

Extra 7. Um fato interessante é que, se um espaço normado X é tal que sua norma satisfaz a Regra do Paralelogramo, então X é na verdade um pré-espaço de Hilbert e seu produto interno é dado por:

$$\langle x,y\rangle = \frac{\|x+y\|^2 - \|x-y\|^2}{4}.$$

Isso serve como um bom teste para detectar se uma norma advém de um produto interno. De fato, podemos verificar que ℓ^p é um pré-espaço de Hilbert se, e somente se, $p=2$ pois, se pegarmos e_1 e e_2 da base de Schauder canônica de ℓ^p, então:

$$2\|e_1\|_p^2 + 2\|e_2\|_p^2 = \|e_1+e_2\|_p^2 + \|e_1-e_2\|_p^2,$$

simplifica para:

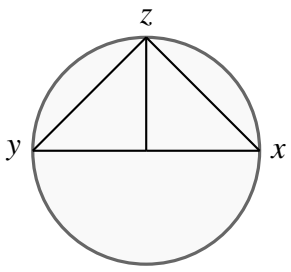
$$4 = 2(2)^{\frac{2}{p}},$$

que, por sua vez, simplifica para:

$$2 = 2^{\frac{2}{p}}.$$

Claramente, a última igualdade só se dá no caso $p=2$.

Extra 8. Na introdução deste capítulo, foi dito que vários resultados da Geometria Euclidiana possuem análogos na teoria dos pré-espaços de Hilbert. Tente encontrar algum desses análogos. A seguir, damos um exemplo: o Teorema de Thales na sua versão para pré-espaços de Hilbert.



Exemplo: Teorema de Thales. Sejam x,y dois vetores tais que $y=-x$. Insira esses dois vetores em um círculo de raio $\|x\|$ e escolha um vetor z qualquer que possua $\|z\| = \|x\|$, então, $x-z \perp y-z$, ou seja, o triângulo formado pelos pontos x,y,z é retângulo.

3.5 CONJUNTOS ORTONORMAIS E BASES DE HILBERT

Definição 3.3. Dizemos que um conjunto de vetores E em um pré-espaço de Hilbert é ortonormal se, para cada dois vetores $e,e' \in E$, tivermos:

$$\langle e, e' \rangle = \begin{cases} 0, & \text{se } e \neq e', \\ 1, & \text{se } e = e'. \end{cases}$$

Assumimos que o leitor esteja familiarizado com o conceito de base ortonormal em espaços vetoriais de dimensão finita. Desta forma, sugerimos a análise da seguinte lista de afirmações[2]. Fixe $e_1, e_2, ..., e_n$ vetores ortonormais em P, então,

1. **É fácil computar normas:** Sejam $\alpha_1, \alpha_2, ..., \alpha_n$ escalares, então $\|\sum_{i=1}^{n} \alpha_i e_i\|^2 = \sum_{i=1}^{n} \alpha_i^2$.
2. **É fácil computar coeficientes:** Seja $x = \sum_{i=1}^{n} \alpha_i e_i$ um vetor em $\text{Span}\{e_1, e_2, ..., e_n\}$, então $\alpha_i = \langle x, e_i \rangle$.
3. **Dependência Linear:** Os vetores $e_1, e_2, ..., e_n$ são linearmente independentes.
4. **Gram-Schmidt:** Todo conjunto de vetores linearmente independentes pode ser ortonormalizado; isto é, dado um conjunto $V = \{v_1, v_2, ..., v_n\}$ de vetores linearmente independentes, podemos obter um conjunto $W = \{w_1, w_2, ..., w_n\}$ de vetores ortonormais tais que $\text{Span}\, V = \text{Span}\, W$.

Para se convencer da praticalidade de bases ortonormais, sugerimos computar as coordenadas de $(4,5,6,7)$ com respeito à base canônica de $\mathbb{R}^4$ e depois computar as coordenadas do mesmo vetor com respeito à base: $(1,9,3,1), (-2,5,8,7), (3,3,5,7), (6,4,2,1)$. Em espaços vetoriais de dimensão finita, bases ortonormais não apenas facilitam a representação de vetores, como também viabilizam a solução de problemas de otimização da forma:

$$\min_{y \in Y} \quad \|x - y\|. \tag{3.5}$$

Onde, dado um vetor $x \in H$ e um subespaço $Y \subseteq H$, busca-se encontrar um vetor $y \in Y$ que melhor aproxime x. Antes de mostrarmos a existência e unicidade de tal y, precisamos primeiro considerar algumas definições.

Definição 3.4. Seja P um pré-espaço de Hilbert e $M \subseteq P$ um subconjunto de P, então definimos o complemento ortogonal de M como:

$$M^{\perp} := \{x \in P : \forall m \in M : \langle x, m \rangle = 0\}. \tag{3.6}$$

Observação. $M^{\perp}$ é um subespaço de P. De fato, $0 \in M^{\perp}$ e, se x e y pertencem a $M^{\perp}$ e λ é um escalar, então $\langle x + \lambda \cdot y, m \rangle = \langle x, m \rangle + \lambda \langle y, m \rangle = 0$ e $x + \lambda \cdot y \in M^{\perp}$.

Definição 3.5. Sejam V e W subespaços vetoriais de um pré-espaço de Hilbert P, definimos a soma de V e W como:

$$V + W := \{v + w : v \in V, w \in W\}. \tag{3.7}$$

[2] O capítulo 6 da referência (AXLER, 2015) aborda esses resultados de maneira detalhada.

Caso $V \cap W = \{0\}$, denotamos $V + W$ por $V \oplus W$, que recebe o nome de soma direta.

Observação. $V + W$ é um subespaço de P. Com efeito, $0 \in V + W$ e, se $x = v_1 + w_1$ e $y = v_2 + w_2$ pertencem a $V + W$ e λ é um escalar, então $x + \lambda \cdot y = (v_1 + \lambda \cdot v_2) + (w_1 + \lambda \cdot w_2) \in V + W$.

A propriedade mais interessante de somas diretas é que podemos escrever todo vetor de $V \oplus W$ de maneira única como a soma de um vetor de V com um vetor de W. Exploraremos essa propriedade em conjunto com o teorema a seguir.

Teorema 3.7. Seja P um pré-espaço de Hilbert e Y um subespaço de P com dimensão finita, então,

$$P = Y \oplus Y^{\perp}. \tag{3.8}$$

Demonstração. Por conta de Gram-Schmidt, sabemos da existência de uma base ortonormal para Y. Tendo fixado uma base ortonormal $e_1, e_2, ..., e_n$ de Y, considere um vetor $x \in P$ e defina $y = \sum_{i=1}^{n} \langle x, e_i \rangle e_i$ e $z = x - y$. Assim, $x = y + z$. Agora, basta verificar que $z \perp Y$. Com efeito, para cada vetor da base e_j, temos $\langle z, e_j \rangle = \langle x, e_j \rangle - \sum_{i=1}^{n} \langle x, e_i \rangle \langle e_i, e_j \rangle = \langle x, e_j \rangle - \langle x, e_j \rangle = 0$. Adicionalmente, temos que $Y \cap Y^{\perp} = \{0\}$ pois o único vetor que é ortogonal a si mesmo é 0. □

Observação. Chegamos à seguinte conclusão: dado um subespaço Y de dimensão finita, todo vetor de P pode ser escrito de maneira única como a soma de um vetor de Y com um vetor de $Y^{\perp}$. Já havíamos observado uma decomposição similar no caso em que $\dim Y = 1$. De fato, w é ortogonal à reta gerada por y no Teorema 3.1. Assim, podemos definir o operador projeção, que leva cada vetor de P em sua componente no subespaço Y.

Definição 3.6. Seja P um pré-espaço de Hilbert, Y um subespaço de P com dimensão finita e seja $x = y + z$ um vetor de P tal que $y \in Y, z \in Y^{\perp}$, então, a projeção de x em Y é dada por

$$P_Y(x) = y. \tag{3.9}$$

Proposição 3.3. Seja P como na definição anterior, então:

1. P_Y é um operador linear de P em P.
2. $\operatorname{Ker} P_Y = Y^{\perp}$.
3. $\operatorname{Im} P_Y = Y$.
4. $P_Y^2 = P_Y$.
5. $\|P_Y(x)\| \leq \|x\|$.

Demonstração. Para 1, seja $x_1 = y_1 + z_1$ e $x_2 = y_2 + z_2$, em que $y_1, y_2 \in Y$ e $z_1, z_2 \in Y^{\perp}$. Assim, se λ é um escalar, temos $P_Y(x_1 + \lambda \cdot x_2) = y_1 + \lambda \cdot y_2 = P_Y(x_1) + \lambda \cdot P_Y(x_2)$ e P_Y é linear.

Para 2, note que $P(x) = P_Y(y+z) = y$ é igual a zero se, e somente se $y = 0$ e $x \in Y^{\perp}$. Assim, temos que $\text{Ker}\, P_Y = Y^{\perp}$. A demonstração de 3 é análoga.

Para 4, observe que $P_Y(P_Y(y+z)) = P_Y(y+0) = y = P_Y(y+z)$.

Para 5, notemos que $y \perp z$, o que nos dá $\|x\|^2 = \|y+z\|^2 = \|y\|^2 + \|z\|^2 \leq \|y\|^2 = \|P_Y(x)\|^2$. Extraíndo raizes quadradas, obtemos a desigualdade desejada. □

Observação. Se P_Y é um operador projeção, $\{e_1, e_2, ..., e_n\}$ é uma base ortonormal de Y e $x = y+z$, onde $y \in Y$ e $z \in Y^{\perp}$, então $\langle x, e_i \rangle = \langle y+z, e_i \rangle = \langle y, e_i \rangle = \langle P_Y(x), e_i \rangle$ para $i = 1, 2, ..., n$. Assim, conseguimos computar as coordenadas da projeção de x em Y através de:

$$P_Y(x) = \sum_{i=1}^{n} \langle x, e_i \rangle e_i. \tag{3.10}$$

Teorema 3.8. Seja P um pré-espaço de Hilbert, x um vetor de P e Y um subespaço de P com dimensão finita, então:

$$\forall y \in Y : \|x - P_Y(x)\| \leq \|x - y\|. \tag{3.11}$$

Demonstração. Escreva $x = P_Y(x) + z$, em que $z \in Y^{\perp}$, então $\|x-y\|^2 = \|(P_Y(x) - y) + z\|^2 = \|P_Y(x) - y\|^2 + \|z\|^2 \geq \|z\|^2 = \|x - P_Y(x)\|^2$. Extraindo raizes quadradas, obtemos a desigualdade desejada. □

Extra 9. Encontre o polinômio de grau até 3 que melhor aproxima $\text{sen}\, x$ com respeito à norma $\|f\| = \sqrt{\int_{-\pi}^{+\pi} f(x)^2 dx}$.

- Agora encontre o polinômio de grau até 5. Os coeficientes de e_1, e_2 e e_3 não mudaram. Por que?
- Compares aproximações obtidas com a série de Taylor para $\text{sen}\, x$ truncada nos graus 3 e 5. Qual é a melhor aproximação, a série truncada ou a projeção ortogonal?
- Observe que, em certo sentido, a aproximação pela série de Taylor é local, enquanto a aproximação pela projeção ortogonal é global.

Generalizando

Até agora, trabalhamos com finitos vetores ortonormais e assumimos que Y tinha dimensão finita. A seguir, abordaremos sequências de vetores ortonormais e estenderemos os resultados para Y fechado. A inexistência de uma base de Hamel ortonormal para espaços de

Hilbert de dimensão infinita nos levará a discutir novas noções de base ortonormal. Em particular, desenvolveremos o conceito de base de Hilbert e veremos como bases de Hilbert possuem muitas das propriedades interessantes que bases de Hamel ortonormais possuem em dimensão finita.

Definição 3.7. Seja H um espaço de Hilbert e $B \subseteq H$ um conjunto de vetores ortonormais em H, dizemos que B é uma base de Hilbert[3] se esgotar todos os eixos de H no seguinte sentido.

$$\forall h \in H : [\forall b \in B : h \perp b] \rightarrow h = 0 \tag{3.12}$$

Observação. Note que a base de Schauder canônica para ℓ^2 que vimos no capítulo passado claramente é uma base de Hilbert. Isso se repete em outros espaços. De fato, toda base de Hilbert em um espaço cuja dimensão é enumerável é uma base de Schauder cujos vetores são ortonormais. A situação ficará mais evidente depois que introduzirmos as propriedades algébricas das bases de Hilbert. Por enquanto, nos concentramos em mostrar que há um conceito de dimensão para estas bases. Com efeito, considere a seguinte proposição.

Proposição 3.4. Todas as bases de Hilbert de um dado espaço de Hilbert possuem a mesma cardinalidade. Chamamos esta cardinalidade de dimensão ortogonal de H.

Demonstração. A prova deste teorema envolve técnicas fora do escopo deste trabalho. O leitor interessado pode encontrar uma prova na seguinte referência (MARTINI, 2012). □

Agora conseguimos ver porque não há bases de Hamel ortonormais para espaços de Hilbert com dimensão infinita. De fato, note que conjuntos ortonormais em ℓ^2 são no máximo enumeráveis, enquanto que, pelo próximo teorema, todas as bases de Hamel para ℓ^2 são não enumeráveis. Essa situação se repete em outros espaços de Hilbert interessantes, como $\mathbf{L}^2[a,b]$.

Teorema 3.9. Nenhuma base num espaço de Banach de dimensão infinita pode ser enumerável.

Demonstração. Esse resultado é uma consequência do Teorema de Categoria de Baire, que não abordamos nesse trabalho. O leitor interessado pode encontrar uma demonstração na referência (MINTU, 2012). □

Extra 10. É possível adaptar a prova de que todo espaço vetorial possui uma base de Hamel para provar que todo espaço de Hilbert possui uma base de Hilbert.

Revisitando $\boldsymbol{\ell^2}$

Se considerarmos a base de Schauder canônica para $\boldsymbol{\ell^2}$, podemos verificar que $(e_n)_{n\in\mathbb{N}}$ possui algumas propriedades desejáveis de uma base:

[3] Alternativamente, conjunto ortonormal total.

1. Todo vetor $x \in \boldsymbol{\ell}^2$ pode ser expresso de forma única como uma série da forma $\sum_{i=1}^{\infty} a_i e_i$. De fato, basta deixar $a_i = \langle x, e_i \rangle$, como fizemos no caso de dimensão finita.

2. A norma de um vetor $x \in \boldsymbol{\ell}^2$ é dada por $\|x\|^2 = \sum_{i=1}^{\infty} \langle x, e_i \rangle^2$, análogo ao caso finito.

3. O produto interno de dois vetores $x, y \in \boldsymbol{\ell}^2$ é dado por $\langle x, y \rangle = \sum_{i=1}^{\infty} \langle x, e_i \rangle \langle y, e_i \rangle$, também análogo ao caso finito.

4. $(e_n)_{n\in\mathbb{N}}$ "esgota" todas os eixos de $\boldsymbol{\ell}^2$ no seguinte sentido. Se há um vetor $h \perp e_n$ para todo n, ou seja, uma nova direção, então $h = 0$.

Gostaríamos de investigar até que ponto essa ideia nos leva. Para isto, começamos considerando espaços de Hilbert com dimensão ortogonal enumerável.

Bases de Hilbert enumeráveis

Primeiro, verificamos que séries da forma $\sum_{i=1}^{\infty} \langle x, e_i \rangle e_i$ sempre convergem quando $(e_n)_{n\in\mathbb{N}}$ é uma sequência ortonormal. Este fato é consequência da desigualdade de Bessel, vista a seguir.

Teorema 3.10. Seja $(e_n)_{n\in\mathbb{N}}$ uma sequência ortonormal em um pré-espaço de Hilbert P, então, para todo $x \in P$,

$$\sum_{i=1}^{\infty} |\langle x, e_i \rangle|^2 \leq \|x\|^2 . \tag{3.13}$$

Demonstração. Podemos reciclar resultados de dimensão finita. De fato, seja $Y = \text{Span}(e_n)_{n\leq N}$, então $\sum_{i=1}^{N} |\langle x, e_i \rangle|^2 = \|P_Y(x)\|^2 \leq \|x\|^2$. Como a série em questão é monótona e limitada pela desigualdade anterior, temos que ela deve convergir. □

Seja $x \in H$ um vetor em um espaço de Hilbert H, com esta última desigualdade e o seguinte teorema, conseguimos mostrar que $y = \sum_{i=1}^{\infty} \langle x, e_i \rangle e_i$ converge.

Teorema 3.11. Dada uma sequência ortonormal $(e_n)_{n\in\mathbb{N}}$ em um espaço de Hilbert H, então as seguintes proposições são verdadeiras.

1. $\sum_{i=1}^{\infty} x_i e_i$ converge se, e somente se $(x_n)_{n\in\mathbb{N}} \in \ell^2$.

2. Se $y = \sum_{i=1}^{\infty} y_i e_i$, então $y_i = \langle y, e_i \rangle$. Isto implica na unicidade da representação.

Demonstração. Defina $S_N = \sum_{n=1}^{N} x_n e_n$ e $R_N = \sum_{n=1}^{N} x_n^2$. Note que, para todo $N > M$, temos $\|S_M - S_N\|^2 = \left\|\sum_{n=M+1}^{N} x_n e_n\right\|^2 = \sum_{n=M+1}^{N} x_n^2 = R_N - R_M = |R_N - R_M|$. Assim, S_N é Cauchy se, e somente se, R_N também o é. Como ambos os espaços são completos, temos que S_N converge em H se, e somente se, R_N converge em $\mathbb{R}$. Por fim, observe que R_N convergente implica que $(x_n)_{n\in\mathbb{N}} \in \ell^2$.

Para o segundo item, usamos a continuidade do produto interno. Com efeito, Seja $y = \sum_{i=1}^{\infty} y_i e_i$, então $\langle y, e_j \rangle = \langle \sum_{i=1}^{\infty} y_i e_i, e_j \rangle = \sum_{i=1}^{\infty} y_i \langle e_i, e_j \rangle$. Observe que todas as parcelas serão zero, com exceção de $i = j$. Com isso, temos a igualdade desejada. □

Observação. No caso de bases de Hilbert, todos os vetores podem ser expressos como séries da forma $x = \sum_{i=1} x_i e_i$. Realmente, seja $(e_n)_{n\in\mathbb{N}}$ uma base de Hilbert para o espaço H, então $y = \sum_{i=1}^{\infty} \langle x, e_i \rangle e_i$ está bem definida e $x - y \perp (e_n)_{n\in\mathbb{N}}$. Usando a definição de bases de Hilbert, vemos que $x - y = 0$ e, portanto, que $x = y$. Além disso, quando estamos falando de bases de Hilbert, a desigualdade de Bessel se torna uma igualdade e produtos internos podem ser computados como o produto escalar de ℓ^2: $\langle x, y \rangle = \langle \sum_{i=1}^{\infty} x_i e_i, \sum_{j=1}^{\infty} y_j e_j \rangle = \sum_{i=1}^{\infty} \sum_{i=1}^{\infty} x_i y_j \langle e_i, e_j \rangle = \sum_{i=1}^{\infty} x_i y_i$.

Vamos agora resolver o problema de otimização.

Teorema 3.12. Seja Y um subespaço fechado de um espaço de Hilbert H, então $H = Y \oplus Y^{\perp}$.

Demonstração. Como Y é fechado, temos que também é um espaço de Hilbert. Assim, pelo Extra 10 sabemos que há uma base de Hilbert $(e_n)_{n\in\mathbb{N}}$ para Y. Definindo $y = \sum_{i=1}^{\infty} \langle x, e_i \rangle e_i$, temos que $z = x - y \perp e_n$, para todo n. Note que isso implica que $z \in Y^{\perp}$. Assim, $x = y + z$ e $H = Y + Y^{\perp}$. Novamente a soma é direta pois o único vetor ortogonal a si mesmo é o vetor nulo. □

Como anteriormente, definimos o operador projeção, que continua com as mesmas propriedades. Em particular, temos que $\langle P_Y(x), e_i \rangle = \langle x, e_i \rangle$, possibilitando o cálculo das coordenadas do vetor y:

$$P_Y(x) = \sum_{i=1}^{\infty} \langle x, e_i \rangle e_i. \tag{3.14}$$

Teorema 3.13. Seja Y um subespaço fechado de um espaço de Hilbert H, então

$$\forall y \in Y : \|x - P_Y(x)\| \leq \|x - y\|. \tag{3.15}$$

Demonstração. A demonstração é análoga ao caso de dimensão finita. □

É possível ir além

Há espaços de Hilbert com dimensão ortogonal não enumerável? Sim, existem vários. Provavelmente, o mais famoso é o espaço das funções quase-periódicas. Mesmo assim, sua importância é pequena se comparada aos espaços com dimensão enumerável (WEAVER, 2016). A seguir, construimos um exemplo acessível.

Exemplo 3.4. Seja $\mathfrak{F}(\mathbb{R}, \mathbb{R})$ o espaço vetorial das funções de $\mathbb{R}$ em $\mathbb{R}$. Defina $E = \text{Span}(e_\lambda)$, onde $e_\lambda(x) = 1$ se $x = \lambda$, 0 caso contrário, então, podemos definir uma função bilinear $\langle \cdot, \cdot \rangle$ especificando seus valores nos vetores da base $\langle e_\lambda, e_\mu \rangle$ de tal forma que os e_λ sejam ortonormais.

Note que essa função satisfaz os axiomas de produto interno. Assim $(E, \langle\cdot,\cdot\rangle)$ é um pré-espaço de Hilbert. Podemos completá-lo para obtermos um espaço de Hilbert em que os e_λ formam um conjunto ortonormal. Como todo conjunto ortonormal está contido em uma base de Hilbert, podemos inferir que a dimensão ortogonal deste espaço recém construído é não enumerável.

A existência desses espaços nos força a refletir sobre o problema da base. Será que nossa ideia também funciona quando a base de Hilbert $(e_\lambda)_{\lambda\in\Lambda}$ não é enumerável? Surpeendentemente, sim. E são poucos os ajustes que precisam ser feitos. Comecemos com o problema mais aparente: como lidar com expressões da forma $\sum x_\lambda e_\lambda$. O teorema a seguir nos dá um caminho.

Teorema 3.14. Seja B um conjunto ortonormal em um espaço de Hilbert, então, para todo $x \in H$, o seguinte conjunto sempre é enumerável.

$$\{\lambda \in \Lambda : |\langle x, e_\lambda\rangle| > 0\}. \tag{3.16}$$

Demonstração. Note que o conjunto acima pode ser reescrito como a união dos $A_n = \{\lambda \in \Lambda : |\langle x, e_\lambda\rangle| > 1/n\}$. Vamos mostrar que cada um desses conjuntos é finito, isso será o suficiente para mostrar que o conjunto acima é enumerável. Com efeito, suponha que A_N é infinito para algum N. Seja $(e_n)_{n\in\mathbb{N}}$ uma sequência ortonormal de vetores em B tal que $|\langle x, e_n\rangle| > 1/N$, para todo n. Note que $\sum_{n=1}^{\infty} |\langle x, e_n\rangle|^2$ não converge, contrariando a desigualdade de Bessel. Assim, temos uma contradição e A_n é finito para todo n. □

Seja $x \in H$ um vetor em um espaço de Hilbert qualquer, então, pelo teorema anterior, podemos ver que x só possui componentes em no máximo contáveis eixos. Assim, tendo fixada uma base $(e_\lambda)_{\lambda\in\Lambda}$, podemos definir $y = \sum\langle x, e_\lambda\rangle e_\lambda = \sum_{i=1}^{\infty}\langle x, e_i\rangle e_i$, onde os e_i são tais que $\langle x, e_i\rangle \neq 0$. Sabemos que y está bem definido pois $(e_i)_{i\in\mathbb{N}}$ é uma sequência ortonormal. Além disso, podemos ver que $x - y \perp e_\lambda$. Como $(e_\lambda)_{\lambda\in\Lambda}$ é uma base, temos que $x = y$. Assim, todo vetor de H pode ser expresso de forma única como combinação linear dos vetores da base. Isso é o suficiente para fazer sentido de expressões da forma $\sum x_\lambda e_\lambda$; como só precisamos de contáveis vetores da base para cada x, convencionamos que só há contáveis x_λ diferentes de zero na expressão anterior. Observe as seguintes igualdades:

1. Se $x = \sum x_\lambda e_\lambda$, então $\|x\|^2 = \sum x_\lambda^2$.
2. Se além de x, tivermos $y = \sum y_\lambda e_\lambda$, então $\langle x, y\rangle = \sum x_\lambda y_\lambda$.

A nossa solução para o problema de otimização pode ser adaptada para estas novas bases. De fato, considere os seguintes teoremas:

Teorema 3.15. Seja Y um subespaço fechado de um espaço de Hilbert H, então $H = Y \oplus Y^\perp$.

Demonstração. Sejam $x \in H$ um vetor de H e Y um subespaço fechado de H. Pelo Extra 10, Y admite uma base de Hilbert $(e_\lambda)_{\lambda\in\Lambda}$. Dado $x \in H$, podemos definir $y = \sum\langle x, e_\lambda\rangle e_\lambda$. Na

discussão anterior, vimos que $z = x - y \perp e_\lambda$ para todo $\lambda \in \Lambda$. Neste caso, não necessariamente temos que $z = 0$, mas podemos concluir que $z \perp Y$, uma vez que todo vetor de Y é uma série da forma $\sum a_\lambda e_\lambda$ e $\langle z, \sum a_\lambda e_\lambda \rangle = \sum a_\lambda \langle z, e_\lambda \rangle = 0$. A soma é direta pois 0 é o único vetor ortogonal a si mesmo. □

Teorema 3.16. Seja Y um subespaço fechado de um espaço de Hilbert H, então:

$$\forall y \in Y : \|x - P_Y(x)\| \leq \|x - y\|. \tag{3.17}$$

Demonstração. A prova é análoga ao caso de dimensão finita. □

3.6 FUNCIONAIS EM ESPAÇOS DE HILBERT

Nesta seção, veremos exemplos de funcionais definidos sobre espaços de Hilbert e abordaremos o importante Teorema de Representação de Riesz, que permite expressar funcionais $f : H \to \mathbb{R}$ em termos de produtos internos com vetores fixos. Isto é, dado um funcional f, há um vetor $z \in H$, tal que $f(x) = \langle x, z \rangle$ para todo $x \in H$.

Exemplo 3.5. Se $H = \ell^2_n$, então $\varphi(x_1, x_2, ..., x_n) = x_1$ é um funcional. Adicionalmente, se $H = \ell^2_3$, $\psi(x, y, z) = x + y + z$ também é um funcional. De forma geral, se $f : \mathbb{R}^n \to \mathbb{R}$ for um funcional e $v_1, v_2, ..., v_n$ for uma base, então $f(x) = f(\sum_{i=1}^n x_i v_i) = \sum_{i=1}^n x_i f(v_i) = \langle x, z \rangle$, onde $z = (f(v_1), f(v_2), ..., f(v_n))$. Note que essa representação também é possível para outros espaços vetoriais de dimensão finita. Como veremos a frente, essa representação ainda será possível em espaços de dimensão infinita, onde será de crucial importância.

Exemplo 3.6. De forma geral, se H é um espaço de Hilbert, então $f(x) = \langle x, z \rangle$, onde $z \in H$ é fixo, sempre será um funcional. Esse funcional será limitado e sua norma será $\|f\| = \|z\|$. Com efeito, $|f(x)| = \langle x, z \rangle \leq \|z\| \|x\|$, o que mostra que $\|f\| \leq \|z\|$. Adicionalmente, temos $\frac{|f(z)|}{\|z\|} = \frac{\|z\|^2}{\|z\|} = \|z\|$, o que nos permite concluir $\|f\| \geq \|z\|$. Assim, $\|f\| = \|z\|$ e f é limitado.

Observação. Em $H = \ell^2$, funcionais assumem a forma de séries: $f(x) = \sum_{i=1}^\infty x_i z_i$, onde $z \in \ell^2$ é uma sequência fixa. Enquanto que em $L^2[a, b]$, funcionais assumem a forma de integrais: $f(x) = \int_a^b x(t) z(t) dt$.

Exemplo 3.7. Em espaços de funções[4], é sempre possível definir o funcional avaliação $L_x(f) = f(x)$.

Teorema 3.17. (Teorema de Representação de Riesz) Seja H um espaço de Hilbert e $f : H \to \mathbb{R}$ um funcional limitado, então existe um único vetor $z \in H$ tal que:

$$f(x) = \langle x, z \rangle.$$

Além disso, $\|f\| = \|z\|$.

[4] Em certo sentido, este exemplo é idêntico ao primeiro; em ambos os casos, estamos pegando um vetor e estamos projetando uma de suas coordenadas.

Demonstração. Seja $f : H \to \mathbb{R}$ um funcional sobre H. Se $f = 0$, então $z = 0$ satisfaz $f(x) = \langle x, z\rangle$ para todo x. Se $f \neq 0$, então $\operatorname{Ker} f \neq H$. Como $\operatorname{Ker} f$ é fechado, sabemos que $H = \operatorname{Ker} f \oplus (\operatorname{Ker} f)^{\perp}$ e $(\operatorname{Ker} f)^{\perp} \neq \emptyset$. Seja z_0 algum vetor de norma unitária em $(\operatorname{Ker} f)^{\perp}$. A motivação de escolher z_0 nesse espaço é a seguinte: se $f(x) = \langle x, z\rangle$ para algum z, então $f(x) = 0 \iff x \perp y$, o que mostra que $z \perp \operatorname{Ker} f$. Para continuarmos nossa análise, fixemos $x \in H$. Assim, definindo $v = f(x)z_0 - f(z_0)x$, temos $f(v) = 0$. Como $z_0 \perp \operatorname{Ker} f$, podemos também concluir que $\langle f(x)z_0 - f(z_0)x, z_0\rangle = 0$. Por sua vez, isso nos permite concluir $f(x)\langle z_0, z_0\rangle - \langle x, f(z_0)\cdot z_0\rangle = 0$. Por fim, $f(x) = \langle x, f(z_0)\cdot z_0\rangle$.

Agora mostramos a unicidade de z. De fato, sejam z_1 e z_2 tais que $f(x) = \langle x, z_1\rangle = \langle x, z_2\rangle$ para todo x. Em particular, se escolhermos $x = z_1 - z_2$, obtemos $\langle z_1 - z_2, z_1 - z_2\rangle = 0$. Isso nos permite concluir que $z_1 = z_2$. Finalmente, $\|f\| = \|z\|$ pela mesma razão do Exemplo 3.6. □

4 KERNELS E APRENDIZADO

> I heard the reiteration of the following claim: Complex theories do not work; simple algorithms do. I would like to demonstrate that in the area of science a good old principle is valid: Nothing is more practical than a good theory.
>
> Vladimir Vapnik

Neste último capítulo, iremos introduzir rapidamente o problema de aprendizado supervisionado para então discutirmos um algoritmo de classificação binária: o Perceptron. Será possível perceber, durante a demonstração da corretude de tal algoritmo, que nada o impede de funcionar em um espaço de Hilbert qualquer. Isso, aliado ao problema de classificação de datasets não linearmente separáveis, será motivação para abordarmos transformações não lineares (feature maps) e Métodos de Kernel. Por fim, discutiremos a relação de kernels com espaços de Hilbert de Reprodução (EHR) e daremos exemplos de kernels relevantes para a prática de Aprendizado de Máquina. A figura abaixo ilustra a relação dos EHR com os outros espaços estudados anteriormente.

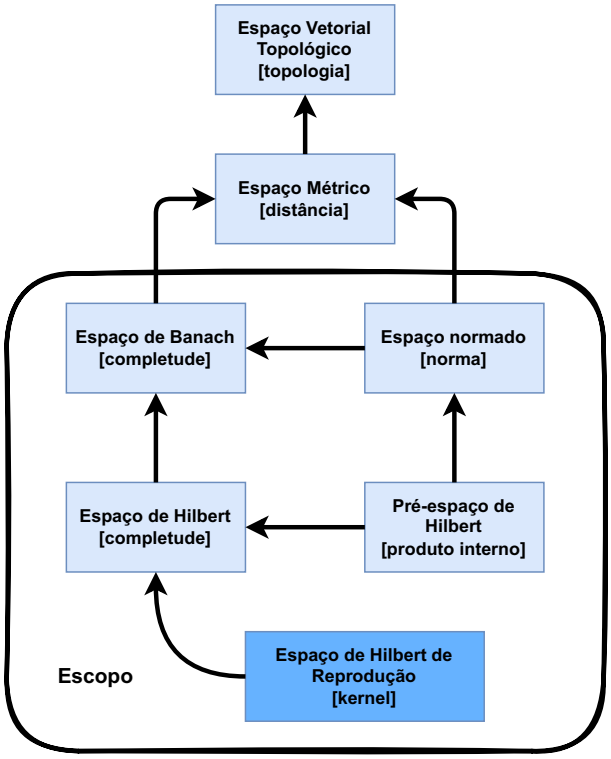


Figura 10 – Relação “é um tipo de” entre diferentes tipos de espaços estudados em Análise Funcional.

4.1 APRENDIZADO SUPERVISIONADO

O aprendizado supervisionado é vital para várias áreas da ciência, pois permite a extração semi-automatizada de padrões estatísticos em conjuntos potencialmente grandes de dados. Em sua versão mais clássica, quando temos um conjunto de pontos (x_1, y_1),(x_2, y_2),...,(x_m, y_m) e queremos aprender uma hipótese $h : X \to Y$, ele costuma figurar como meio de correlacionar um conjunto de características X com um conjunto Y. Tipicamente, $X = \mathbb{R}^d$, enquanto Y às vezes é um conjunto discreto de rótulos e às vezes é algum subconjunto de $\mathbb{R}$. Quando Y é discreto, o problema recebe o nome de classificação e, quando Y é contínuo, o problema recebe o nome de regressão. Vejamos um exemplo simples de classificação. Suponha que X é um subconjunto de $\mathbb{R}$ e Y é $\{\text{Maligno}, \text{Benigno}\}$. Tendo visto um conjunto de treino, podemos obter uma h que nos possibilite prever se futuros tumores são benignos ou malignos. Agora, vejamos um exemplo de regressão: suponha que $X = \mathbb{R}^2$ seja tal que cada x_i represente o par (área da casa em m^2, número de quartos). Poderíamos colocar $Y = \mathbb{R}$ e então teríamos um contexto que poderíamos usar para formalizar o problema de previsão de preço de casa tendo informações de área e de número de quartos.

Esses dois exemplos têm uma característica em comum, enquanto o próximo exemplo não terá essa característica (tente advinhar o que está faltando). Considere os pontos (x_i, y_i) como apresentados no gráfico abaixo. O objetivo do aprendizado é encontrar uma função $h : X \to Y$. Como poderíamos proceder?

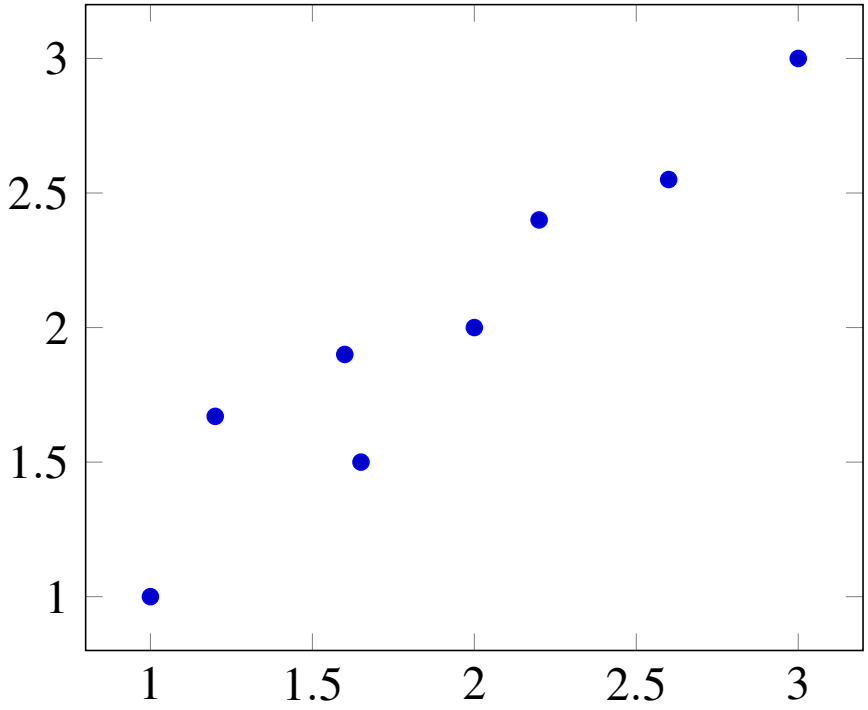


Figura 11 – Conjunto de dados sem contexto.

Claramente, esse exemplo está descontextualizado. O que representam X e Y? Vejamos as consequências de proceder sem termos essa noção. Para começar, note que há infinitas

funções de $\mathbb{R}$ para $\mathbb{R}$. Dois exemplos arbitrários são mostrados a seguir. Como poderíamos decidir qual das duas funções regressoras é melhor?

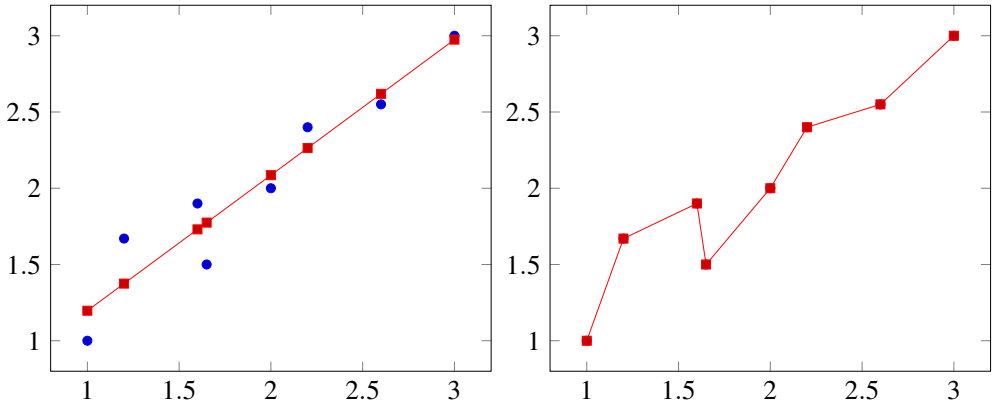


Figura 12 – Duas funções regressoras.

Poderíamos seguir critérios subjetivos como: a da esquerda é mais simples (Navalha de Occam) ou a da direita é mais bonita. Entretanto, a questão que devemos observar é a seguinte: sem sabermos nada *a priori*, isto é, sem que façamos alguma suposição sob a relação de X com Y, é impossível traçar hipóteses científicas; qualquer forma de preferência, na ausência de premissas, seria puramente arbitrária. Chamaremos estas suposições de viés indutivo[1]. Assim, sem alguma forma de viés indutivo, é impossível aprender. Poderíamos supor, por exemplo, que a relação que estamos aproximando é linear. Neste caso, a regressão linear da esquerda seria a melhor solução. A busca por viéses bem fundamentados é parte fundamental da Teoria Estatística do Aprendizado.

Além do viés indutivo, temos também que considerar o problema da representatividade dos $(x_1, y_1), ..., (x_m, y_m)$ usados anteriormente para encontrar a hipótese h. Uma analogia agora cairá bem. Suponha o cenário em que um aluno se prepara para uma prova de cálculo I (limites, derivadas e integrais). Há uma vasta coleção de textos sobre cálculo e há uma coleção ainda maior de problemas que poderiam ser abordados num curso de cálculo. Em quais questões e seções o aluno deveria concentrar suas energias? Bom, se é a P1, que não inclui integrais, não seria inteligente se ele estudasse justamente o capítulo sobre integrais (embora seja o mais divertido). Ele provavelmente se sairia melhor na prova caso se dedicasse ao capítulo de limites; ou seja, durante o ato do aprendizado, é importante priorizar as partes de X que serão de mais importância para o teste. De maneira análoga, é de se esperar que o teste seja escolhido de forma a contemplar as partes de X que serão mais úteis para o eventual uso da hipótese h. No caso do aluno estudando para a prova, é de se esperar que as questões sejam escolhidas de

[1] Uma palavrinha sobre a origem dessa nomenclatura: o racicínio indutivo se distingue do dedutivo (aquele que os matemáticos usam em suas provas), pois aceita argumentos em que a veracidade da conclusão não é determinada pela veracidade das premissas (a conclusão portanto é provável e não necessária); um exemplo de indução que todos nós fizemos quando éramos crianças é o seguinte: "toda vez que eu solto a bola, ela cai; então, dá próxima vez que eu soltar, ela também vai cair". Note que exemplos desse tipo abarcam as ciências naturais no problema filosófico: como determinar se uma indução é válida? A continuação dessa discussão pode ser encontrada na Encíclopedia de Stanford sobre Filosofia (HENDERSON, 2020).

forma a maximizar a aplicabilidade das ferramentas do cálculo para o aluno depois que este tenha concluído a disciplina.

Até agora, temos duas conclusões importantes: é necessário fazer suposições para aprender e é preciso escolher o conjunto de treino e testes de maneira inteligente para que possamos gerar hipóteses que generalizem bem. Note que não discutimos como é possível encontrar tais hipóteses. Com efeito, há várias abordagens possíveis, mas não teremos espaço para abordá-las nesse trabalho[2]. O que faremos é investigar uma solução específica para o problema de classificação. Veja a discussão a seguir.

4.2 PERCEPTRON

Suponha que $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ é um conjunto de dados onde cada x_i pertence a um espaço de entrada $X \subset \mathbb{R}^d$ e cada y_i é ± 1 (rótulos negativos e positivos). O Perceptron é um algoritmo de classificação e, portanto, seu objetivo é achar uma função de decisão que permita classificar qualquer ponto $x \in X$, inclusive aqueles não vistos no conjunto de treinamento. No caso do Perceptron binário, essa função de decisão define duas regiões no espaço: H_+ e H_-, em que $H_+ \subset X$ é composta por todos os pontos de X que são classificados como $+1$ e $H_- \subset X$ é composta por todos os pontos que são classificados como -1. Como veremos em sequência, essas regiões são precisamente o que fica acima e abaixo de um hiperplano orientado (no caso $d = 2$, uma reta), como pode ser visto na figura abaixo.

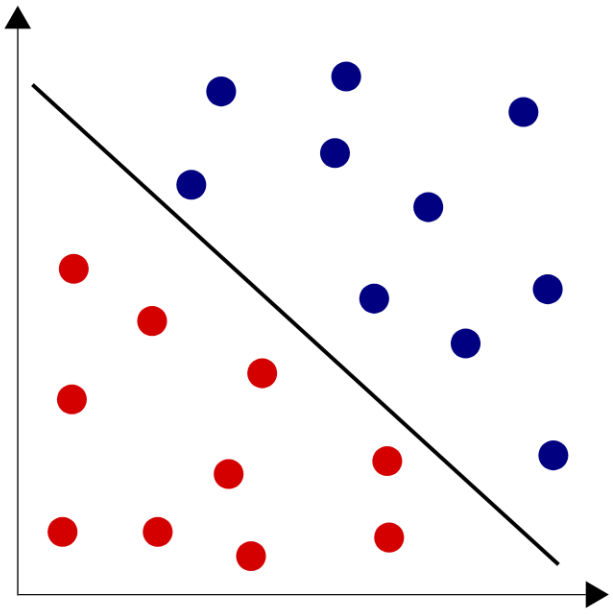


Figura 13 – Exemplo de hiperplano orientado separando duas classes de pontos. Fonte: Mekeor, distribuída pela licença CC BY-SA 3.0.

[2] A referência (DEISENROTH, 2020) abarca as principais abordagens.

4.3 FUNÇÃO DE DECISÃO

A função de decisão do Perceptron é da forma:

$$f(x) = \langle w, x \rangle + b,$$

onde w é um vetor de $\mathbb{R}^d$, $b \in \mathbb{R}$ e $\langle \cdot, \cdot \rangle$ é o produto interno usual. Ou seja, estamos trabalhando com ℓ_n^2. Generalizaremos essa situação para espaços de Hilbert quaisquer quando estivermos trabalhando com kernels. Por ora, note que $f(x) = 0$ define um hiperplano ortogonal ao vetor w. Adicionalmente, $f(x) > 0$ define a região positiva (acima) do hiperplano e $f(x) < 0$ define a região negativa (abaixo) do hiperplano. O objetivo do Perceptron então é encontrar w e b tais que todos os pontos x_i tais que $y_i = +1$ fiquem na região positiva e todos os pontos x_j tais que $y_j = -1$ fiquem na região negativa. Podemos condensar essa condição como:

$$\forall i : y_i f(x_i) > 0.$$

Observação. Assumiremos em toda nossa análise que o conjunto de treino é linearmente separável; i.e., que existem w e b que satisfazem a equação acima e que é possível encontrá-los de forma a deixar uma pequena margem ao redor do hiperplano sem que nenhum ponto de treino invada essa margem. Podemos formalizar essa condição como:

$$\forall i : y_i f(x_i) > r,$$

para algum $r > 0$.

4.4 O ALGORITMO

Algoritmo 1 Perceptron Binário

$w \leftarrow 0$
$b \leftarrow 0$
$R \leftarrow \max_{1 \leq i \leq n} ||x_i||$
while houver erros de classificação **do**
  for $i = 1$ to n **do**
    if $y_i(\langle w, x_i \rangle + b) \leq 0$ **then**
      $w \leftarrow w + y_i x_i$
      $b \leftarrow b + y_i R^2$
    end if
  end for
end while
return (w, b)

O algoritmo é básico e parece funcionar bem em exemplos simples, mas como ter certeza que ele sempre termina com um hiperplano que corretamente classifique o conjunto de

dados? Para isso, precisamos demonstrar a corretude do algoritmo. Note que, se conseguirmos garantir que ele sempre para (dado um conjunto de dados linearmente separável), então também conseguimos garantir que ele sempre encontra um hiperplano válido. Isso se dá pois o algoritmo não para enquanto há erros de classificação. Como veremos a seguir, é possível encontrar um limite superior para o número de erros t que o algoritmo comete. Esse limite dependerá de duas métricas importantes: a margem e a escala do conjunto de dados. Definimos esses conceitos antes de partirmos para o teorema principal.

Definição 4.1. Seja x_i um vetor de treino e (w,b) os parâmetros que definem um hiperplano, então a margem funcional entre x_i e o hiperplano é dada por

$$\gamma_i = y_i(\langle w, x_i \rangle + b).$$

Se $\|w\| = 1$, então γ_i representa a distância entre o hiperplano e o ponto de treino. Quando isso acontece, chamamos γ_i de margem geométrica do vetor x_i. Se tomarmos o mínimo entre as margens dos exemplos, temos a margem funcional (geométrica, respectivamente) do hiperplano, como pode ser visto a seguir:

$$\gamma = \min_{1 \leq i \leq n} \gamma_i.$$

Observação. Note que a escolha de representantes (w,b) para um dado hiperplano é um problema mal posto no sentido de que há vários (w,b) correspondendo a um mesmo hiperplano. De fato, observamos que $\langle w, x \rangle + b = 0$ se, e somente se, $\langle \lambda \cdot w, x \rangle + \lambda b = 0$, para qualquer escalar λ.

Abordaremos agora a prova de corretude do Perceptron. A demonstração supõe que o conjunto de dados é separável pela origem. Isto é, que há um hiperplano da forma $(w,0)$ que classifica corretamente o conjunto de dados. A razão para esta simplificação é que assim não é necessário supor que[3] $X = \ell_n^2$. De fato, seria possível implementar este algoritmo sobre qualquer espaço de Hilbert e é exatamente isso que faremos em conexão com o Truque do Kernel, como veremos mais à frente.

Teorema 4.1. Seja $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ um conjunto de treino linearmente separável pela origem tal que haja w^* que corretamente o classifique. Isto é, tal que $\forall i : y_i \langle w^*, x_i \rangle \geq \gamma$ para alguma margem γ. Sob estas condições, o número de erros t que o algoritmo comete antes de terminar é limitado superiormente da seguinte forma:

$$t \leq \left(\frac{R}{\gamma} \right)^2.$$

Demonstração. A prova que segue é uma adaptação da prova do Teorema 2.3 da referência (CRISTIANINI, 2000). Deixe o algoritmo rodar sobre o dado conjunto de treinamento. Sua

[3] Muito provavelmente há uma prova elegante para o caso geral, mas não a encontrei.

execução define uma sequência de vetores w_t, onde w_t é o vetor de peso atualizado logo após o t-ésimo erro e $w_0 = 0$. Por definição, temos que:

$$w_{t+1} = w_t + y_i x_i.$$

onde i vem do vetor de treino que foi classificado erroneamente. Usando essa definição recursiva, obtemos:

$$\langle w_t, w^* \rangle = \langle w_{t-1}, w^* \rangle + y_i \langle x_i, w^* \rangle.$$

Como w^* classifica x_i corretamente, temos:

$$\langle w_{t-1}, w^* \rangle + y_i \langle x_i, w^* \rangle \geq \langle w_{t-1}, w^* \rangle + \gamma.$$

Iterando até chegar em w_0, ficamos com

$$\langle w_t, w^* \rangle \geq t\gamma.$$

Agora, nós buscamos uma desigualdade na outra direção; conseguimos isso limitando $\|w_t\|$:

$$||w_t||^2 = ||w_{t-1}||^2 + 2\langle w_{t-1}, y_i x_i \rangle + ||x_i||^2.$$

Como w_{t-1} errou a classe de x_i, nós sabemos que a expressão do meio é menor ou igual a zero, o que permite a simplificação a seguir:

$$||w_t||^2 \leq ||w_{t-1}||^2 + ||x_i||^2 \leq ||w_{t-1}||^2 + R^2.$$

Aplicando isso várias vezes e depois tomando raízes quadradas, obtemos:

$$||w_t|| \leq \sqrt{t}R.$$

Agora podemos encadear as desigualdades:

$$t\gamma \leq \langle w_t, w^* \rangle \leq ||w_t||||w^*|| \leq \sqrt{t}R.$$

Elevando ao quadrado e cancelando t, obtemos:

$$t\gamma^2 \leq R^2.$$

Que finalmente nos leva a desigualdade desejada:

$$t \leq \frac{R^2}{\gamma}.$$

□

4.5 TRANSFORMAÇÕES NÃO LINEARES

Como vimos na seção passada, o Perceptron Binário funciona sobre qualquer espaço de Hilbert. Isso se torna interessante quando consideramos o seguinte problema: muitos conjuntos de dados relevantes sobre os quais gostaríamos de aplicar o Perceptron não são linearmente separáveis. A Figura 14 ilustra este cenário. Vemos que não há escolha de (w,b) que seja capaz de separar esse conjunto de dados. Uma solução possível é mudar a representação dos dados. De fato, poderíamos buscar uma função $\phi : X \to H$, onde H é outro espaço de Hilbert de tal forma que $\phi(X)$ fosse linearmente separável.

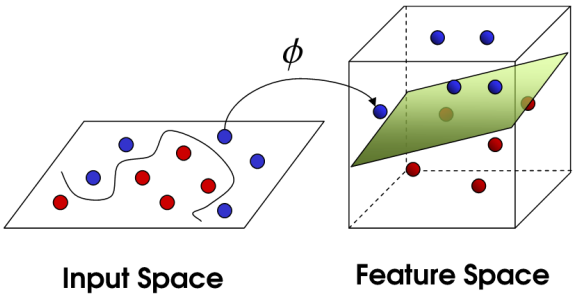


Figura 14 – ϕ faz com que o conjunto de dados se torne linearmente separável. Fonte: adaptado de (VOCATURO; PERNA; ZUMPANO, 2019).

Há dois problemas com essa abordagem. O primeiro e mais sério é o seguinte: como encontrar uma transformação que torne um dado conjunto de treino linearente separável? Dependendo da dimensão[4] e do conjunto de dados, essa pergunta pode ser fácil de responder, mas, em geral, não é. Outro problema que também deve ser considerado é a eficácia de computar $\{\phi(x_1), \phi(x_2,), ..., \phi(x_n)\}$. Dependendo de ϕ, esse cálculo pode ser proibitivo. Esses problemas sugerem a pergunta: não seria possível encontrar uma ϕ ou uma coleção parametrizada de $(\phi_i)_{i\in I}$ que seja altamente versátil e que sirva para vários conjuntos de dados encontrados na prática? A resposta para essa pergunta é sim: não só é possível, como conseguimos também pular a etapa em que ϕ é computada. De fato, vamos agora introduzir a ideia que viabiliza esse atalho: o Truque do Kernel.

4.6 TRUQUE DO KERNEL

Para podermos falar do Truque do Kernel, precisamos primeiro fazer algumas modificações no Perceptron. Note que o vetor w encontrado pelo Algoritmo 1 é combinação linear (com escalares inteiros) dos vetores de treino $x_1, x_2, ..., x_n$. De fato, inicializamos w como zero e depois apenas o atualizamos somando $\pm x_i$. Assim, podemos escrever $w = \sum_{i=1}^{n} \alpha_i x_i$, em que $\alpha_i \in \mathbb{Z}$. Essa observação possibilita uma reformulação do Perceptron, chamada de ver-

[4] Nos casos em que $d \leq 3$, claramente temos mais intuição visual.

são dual do Perceptron, onde o objetivo de aprendizado é encontrar $\alpha_1, \alpha_2, ..., \alpha_n$ e b tais que $(\sum_{i=1}^{n} \alpha_i x_i, b)$ produza um hiperplano separante. Com efeito, considere o seguinte algoritmo.

Algoritmo 2 Perceptron Binário (Dual)

```
α ← 0
b ← 0
R ← max_{1≤i≤n} ||x_i||
while houver erros de classificação do
    for i = 1 to n do
        if y_i(Σ_{j=1}^{n} α_j⟨x_j, x_i⟩ + b) ≤ 0 then
            α_i ← α_i + y_i
            b ← b + y_i R^2
        end if
    end for
end while
return (α, b)
```

Observação. Os vetores de treino x_i só aparecem no algoritmo acima na forma de produtos internos $\langle x_j, x_i \rangle$. Assim, caso usemos uma transformação não linear $\phi : X \to H$, ela também só aparecerá na forma $\langle \phi(x_j), \phi(x_i) \rangle$. Essa observação motiva a seguinte definição.

Definição 4.2. Seja X um espaço de entrada e $K : X \times X \to \mathbb{R}$ uma função da forma

$$K(x,y) = \langle \phi(x), \phi(y) \rangle_H$$

para algum espaço de Hilbert H e transformação não linear $\phi : X \to \mathbb{R}$, então dizemos que K é um kernel.

Definição 4.3. Uma função $K : X \times X \to \mathbb{R}$ é dita positivo-definida se, para cada escolha de $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$, tem-se:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0. \tag{4.1}$$

Proposição 4.1. Todo kernel é uma função simétrica e positivo-definida.

Demonstração. Seja $K : X \times X \to \mathbb{R}$ um Kernel, então $K(x,y) = \langle \phi(x), \phi(y) \rangle = \langle \phi(y), \phi(x) \rangle = K(y,x)$ para cada x e y em X por conta da simetria do produto interno. Além disso, dados $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$, temos $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle$. Usando a linearidade do produto interno, podemos ver que $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) = \langle z, z \rangle$, onde $z = \sum_{i=1}^{n} c_i \phi(x_i)$. Note que essa última expressão é o produto interno de um vetor com ele mesmo e, portanto, deve ser maior ou igual a zero. □

O Truque do Kernel consiste em buscar por kernels fáceis de computar cuja transformação não linear implícita seja versátil e consiga separar vários conjunto de dados encontrados na prática. Esses kernels de fato existem e discutiremos alguns deles na próxima seção. Por

ora, vejamos um exemplo simples de uma função que é um kernel fácil de computar e que é útil em alguns cenários.

Exemplo 4.1. Seja $X = \mathbb{R}^2$, então $K(x,y) = \langle x,y\rangle^2$ é um Kernel. Com efeito, sejam $x = (x_1,x_2)$ e $y = (y_1,y_2)$ vetores em $\mathbb{R}^2$, então $K(x,y) = \langle x,y\rangle^2 = (x_1y_1 + x_2y_2)^2 = x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2$. Note que, se definirmos $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, então $K(x,y) = \langle \phi(x), \phi(y)\rangle_3$, onde $\langle\cdot,\cdot\rangle_3$ é o produto escalar usual de $\mathbb{R}^3$.

Observação. É possível generalizar o exemplo anterior em algumas direções. Na sua forma mais geral, o kernel $K(x,y) = (\langle x,y\rangle_d + c)^n$, em que o espaço de entrada é $\mathbb{R}^d$ e $c \geq 0$, recebe o nome de Kernel Polinomial de grau n.

4.7 KERNELS E ESPAÇOS DE HILBERT DE REPRODUÇÃO

Nesta seção, abordaremos uma caracterização interessante de kernels em termos de funções simétricas e positivo-definidas. Além disso, discutiremos como, a partir de kernels conhecidos, podemos construir novos kernels. Esses resultados nos possibilitarão introduzir o Kernel Gaussiano, que é um dos kernels mais utilizados por ser bem versátil e fácil de configurar (WIJ, 2015). A caracterização de kernels mencionada anteriormente envolve a construção de um espaço de Hilbert de Reprodução, cuja definição relembramos a seguir:

Definição 4.4. Dizemos que um espaço de Hilbert H é um espaço de Hilbert de Reprodução se for um espaço de funções $f : X \to \mathbb{R}$ e todo funcional avaliação $T_x : H \to \mathbb{R}$ dado por $T_X(f) = f(x)$ for contínuo.

Observação. Vimos no Capítulo 2 que $P = C[a,b]$ com a norma $p = 2$ é um pré-espaço de Hilbert e que todo subespaço de P de dimensão finita é de Hilbert. Isso motiva vários exemplos, como pode ser visto a seguir.

Exemplo 4.2. Seja Y um subespaço de dimensão finita de $C[a,b]$, então $H = (Y, \langle\cdot,\cdot\rangle_{Y\times Y})$ é um espaço de Hilbert. Por construção, H é um espaço de Hilbert de funções, o que significa que podemos definir os funcionais avaliação sobre H. Adicionalmente, pelo Teorema 2.8, temos que todos os funcionais avaliação sobre H são contínuos. Isso nos permite concluir que todo subespaço Y de $C[a,b]$ é um EHR.

Apenas através de sua definição, não é possível ver qual a relevância de espaços de Hilbert de Reprodução para o estudo de kernels. Para isto, precisamos da seguinte proposição.

Proposição 4.2. Seja H um EHR com funções definidas sobre um conjunto X, então é possível construir um kernel $K : X \times X \to \mathbb{R}$ tal que:

1. $\forall x : K_x := K(x,\_) \in H$.
2. $\forall x : \forall f : X \to \mathbb{R} \in H : f(x) = \langle f, K_x\rangle$.

Demonstração. Seja $T_x : H \to \mathbb{R}$ o funcional avaliação para o ponto x, então, pelo Teorema de Representação de Riesz, há uma única função $K_x \in H$ tal que $T_x(f) = \langle f, K_x \rangle$. Isso nos permite definir uma transformação não linear $\phi(x) = K_x$, que, por sua vez, nos permite definir um kernel $K(x,y) = \langle \phi(x), \phi(y) \rangle = \langle K_x, K_y \rangle$. Como K_y é uma função de H, temos $K_x(y) = T_y(K_x) = \langle K_x, K_y \rangle = K(x, \_)$. Assim, as funções K_x obtidas pelo Teorema de Riesz satisfazem ambas as condições do enunciado. □

Observação. Com a proposição anterior, fica claro a razão por trás do nome espaço de Hilbert de Reprodução. De fato, todo EHR H vem equipado com um kernel que reproduz as funções de H. É possível introzudir o conceito de EHR diretamente através dessa condição, entretanto, assim não seria possível dar exemplos de maneira tão rápida como fizemos.

Um fato muito interessante é que a recíproca da Proposição 4.2 também vale. Isto é, dado um kernel $K : X \times X \to \mathbb{R}$ definido sobre um conjunto X, é possível construir um único espaço de Hilbert de Reprodução H_K cujo kernel associado é exatamente K. Assim, podemos concluir que todo kernel está unicamente associado a um EHR H_K e vice-versa. Esse fato elucida a natureza dos kernels. De fato, com esse resultado, sabemos que todo método de kernel opera sobre um EHR cujos elementos são combinações lineares[5] das funções K_x induzidas pelo kernel. Vejamos o seguinte teorema.

Teorema 4.2. (Moore-Aronszajn) Seja $K : X \times X \to \mathbb{R}$ uma função simétrica e positivo-definida sobre um conjunto X, então existe um único espaço de Hilbert de Reprodução H_K cujo kernel associado é K.

Demonstração. Começamos definindo $K_x = K(x, \_)$ para cada $x \in X$ e construindo o espaço $H_0 = \text{Span}\{K_x\}$. Vamos definir um produto interno sobre H_0. Para isto, sejam f, g e h funções em H_0 e λ um escalar. Podemos escrever $f = \sum_{i=1}^{n} a_i K_{x_i}, g = \sum_{j=1}^{m} b_j K_{y_j}$ e $h = \sum_{i=1}^{n} c_i K_{x_i}$, então

$$\langle f, g \rangle_0 = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(x_i, y_j),$$

define um produto interno. De fato, primeiro note que $\langle \cdot, \cdot \rangle_0$ está bem definida, pois

$$\langle f, g \rangle_0 = \sum_{i=1}^{n} a_i g(x_i) = \sum_{j=1}^{m} b_j f(y_j).$$

Além disso, conseguimos ver que $\langle \cdot, \cdot \rangle_0$ é bilinear:

$$\langle f + \lambda \cdot h, g \rangle_0 = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + \lambda c_i) b_j K(x_i, y_j) = \langle f, g \rangle_0 + \lambda \langle h, g \rangle_0.$$

Essa função também claramente é simétrica e $\langle f, f \rangle_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$ por conta de K ser positivo-definida. Nos resta mostrar que $\langle f, f \rangle_0 = 0 \Rightarrow f = 0$. Para isto,

[5] Na verdade, limites de sequências de combinações lineares.

precisamos do fato que a desigualdade de Cauchy-Schwarz também vale para formas bilineares simétricas e positivo-definidas (a função $\langle \cdot, \cdot \rangle_0$ satisfaz essas condições). Assim, temos:

$$|f(x)| = |\langle f, K_x \rangle_0| \leq \|f\|_0 \|K_x\|_0 = 0.$$

Portanto, $\langle f, f \rangle_0 = 0 \Rightarrow f = 0$ e $\langle \cdot, \cdot \rangle_0$ é um produto interno. Podemos concluir que $P = (H_0, \langle \cdot, \cdot \rangle_0)$ é um pré-espaço de Hilbert. Veja que temos quase todas as condições do teorema satisfeitas: P é um pré-espaço de Hilbert que possui um kernel de reprodução K. O que falta é a completude de P. Para isto, precisaremos trabalhar um pouco mais.

Começamos notando que toda sequência de Cauchy $(f_n)_{n \in N}$ em H_0 admite limite pontual. De fato, veja que:

$$|f_n(x) - f_m(x)| = |\langle f_n - f_m, K_x \rangle_0| \leq \|f_n - f_m\|_0 \|K_x\|.$$

Assim, se fixarmos x, $(f_n(x))_{n \in \mathbb{N}}$ é uma sequência de Cauchy em $\mathbb{R}$. Pela completude de $\mathbb{R}$, sabemos que $f_n(x) \to f(x)$ para algum número $f(x) \in \mathbb{R}$. Isso nos permite definir o limite pontual de toda sequência de Cauchy:

$$f(x) = \lim_{n \to \infty} f_n(x).$$

Iremos denotar por H o conjunto de todos os limites de sequências de Cauchy em H_0. Claramente, $H_0 \subseteq H$, uma vez que, para toda função $f \in H$, podemos montar a sequência de Cauchy $(f)_{n \in \mathbb{N}}$ cujo limite pontual é a própria f. Assim definido, H também é um espaço vetorial. De fato, $0 \in H$ e é possível ver que, se $(f_n)_{n \in \mathbb{N}}$ e $(g_n)_{n \in \mathbb{N}}$ são sequências de Cauchy com limites pontuais f e g e λ é um escalar, então $(f_n + \lambda \cdot g_n)_{n \in \mathbb{N}}$ é uma sequência de Cauchy cujo limite pontual é $f + \lambda \cdot g$. Vamos agora definir um produto interno sobre H de forma que H se torne um espaço de Hilbert de Reprodução. Com efeito, sejam f e g funções em H, então $f = \lim_{n \to \infty} f_n$ e $g = \lim_{n \to \infty} g_n$, em que $(f_n)_{n \in \mathbb{N}}$ e $(g_n)_{n \in \mathbb{N}}$ são sequências de Cauchy em H_0. Podemos definir um produto interno $\langle \cdot, \cdot \rangle$ sobre H da seguinte maneira:

$$\langle f, g \rangle = \lim_{n \to \infty} \langle f_n, g_n \rangle_0.$$

Precisamos mostrar que este limite existe e que não depende da escolha das sequências $(f_n)_{n \in \mathbb{N}}$ e $(g_n)_{n \in \mathbb{N}}$. Começamos mostrando que o limite existe. Faremos isso mostrando que $(\langle f_n, g_n \rangle_0)_{n \in \mathbb{N}}$ é uma sequência de Cauchy em $\mathbb{R}$. Com efeito,

$$|\langle f_n, g_n \rangle_0 - \langle f_m, g_m \rangle_0| = |\langle f_n, g_n \rangle_0 - \langle f_n, g_m \rangle_0 + \langle f_n, g_m \rangle_0 - \langle f_m, g_m \rangle_0|,$$

que podemos agupar para obter:

$$|\langle f_n, g_n \rangle_0 - \langle f_m, g_m \rangle_0| = |\langle f_n, g_n - g_m \rangle_0 - \langle f_n - f_m, g_m \rangle_0|.$$

Aplicando a desigualdade triangular e depois Cauchy-Schwarz, obtemos:

$$|\langle f_n, g_n\rangle_0 - \langle f_m, g_m\rangle_0| \leq \|f_n\|_0 \|g_n - g_m\|_0 + \|f_n - f_m\|_0 \|g_m\|_0 .$$

Como as sequências $(f_n)_{n\in\mathbb{N}}$ e $(g_n)_{n\in N}$ são de Cauchy, sabemos que são limitadas. Sejam M_1 e M_2 números reais tais que $\|f_n\|_0 \leq M_1$ e $\|g_n\|_0 \leq M_2$ para todo n. Com isso,

$$|\langle f_n, g_n\rangle_0 - \langle f_m, g_m\rangle_0| \leq M_1 \|g_n - g_m\|_0 + \|f_n - f_m\|_0 M_2$$

e a sequência em questão é Cauchy. Portanto, $\lim_{n\to\infty}\langle f_n, g_n\rangle_0$ existe. Vamos agora mostrar que este limite independe da escolha das sequências $(f_n)_{n\in\mathbb{N}}$ e $(g_n)_{n\in\mathbb{N}}$. Para isto, sejam $(f'_n)_{n\in\mathbb{N}}$ e $(g'_n)_{n\in\mathbb{N}}$ outras sequências de Cauchy em H_0 tais que $f_n \to f$ e $g_n \to g$ pontualmente, então $|\langle f_n, g_n\rangle_0 - \langle f'_n, g'_n\rangle_0| \to 0$ por um argumento análogo ao que demos para mostrar que $\lim_{n\to\infty}\langle f_n, g_n\rangle_0$ existe. Portanto, $\langle\cdot,\cdot\rangle$ está bem definido. Não é difícil verificar que tal função é simétrica e bilinear usando as propriedades de $\langle\cdot,\cdot\rangle_0$. Note que, para toda f e $g \in H_0$, temos:

$$\langle f, g\rangle = \langle f, g\rangle_0,$$

uma vez que podemos escolher as sequências $(f)_{n\in\mathbb{N}}$ e $(g)_{n\in\mathbb{N}}$ para computar o produto interno em H. Isso nos permite concluir que $\langle 0, 0\rangle = 0$. Por fim, veja que K ainda é um kernel de reprodução para H, uma vez que $K_x \in H_0 \subseteq H$ para todo x e $\langle f, K_x\rangle = \lim_{n\to\infty}\langle f_n, K_x\rangle_0 = f_n(x) = f(x)$. Assim, se $f \in H$ é tal que $\langle f, f\rangle = 0$, podemos inferir que:

$$|f(x)| = |\langle f, K_x\rangle| \leq \|f\| \|K_x\| = 0,$$

e $f = 0$. Portanto, H é um pré-espaço de Hilbert com um kernel de reprodução. Veremos agora que H é completo e, portanto, um EHR com kernel associado K. Precisaremos do fato que H_0 é denso em H. De fato, seja $f \in H$, então há uma sequência de funções $(f_n)_{n\in\mathbb{N}}$ em H_0 tal que $f_n \to f$ pontualmente. É possível mostrar[6] que isso implica em $\|f - f_n\| \to 0$, o que nos permite concluir que H_0 é denso em f.

Por fim, seja $(f_n)_{n\in\mathbb{N}}$ uma sequência de Cauchy em H. Pela densidade de H_0 em H, podemos montar outra sequência $(f'_n)_{n\in\mathbb{N}}$, desta vez de elementos em H_0, tal que $\|f_n - f'_n\| \to 0$. Isso nos permite mostrar que $(f'_n)_{n\in\mathbb{N}}$ é de Cauchy:

$$\left\|f'_n - f'_m\right\| \leq \left\|f'_n - f_n\right\| + \|f_n - f_m\| + \left\|f_m - f'_m\right\| .$$

Dado $\varepsilon > 0$, claramente conseguimos escolher N grande o suficiente tal que $\|f'_n - f_n\| < \varepsilon/3, \|f_n - f_m\| < \varepsilon/3$ e $\|f_m - f'_m\| < \varepsilon/3$ para todo $m, n \geq N$. Assim, $(f'_n)_{n\in\mathbb{N}}$ é de Cauchy, o que significa que possui limite pontual em H. Seja $f = \lim_{n\to\infty} f_n$, então

$$\|f - f_n\| \leq \left\|f - f'_n\right\| + \left\|f'_n - f_n\right\| .$$

O lado direito claramente vai para zero quando $n \to \infty$, o que nos permite concluir que H é completo. Agora, nos voltamos a questão da unicidade. Seja G outro EHR com kernel de

[6] Uma prova similar a este fato pode ser encontrada na referência (VERT, 2020).

reprodução K, então $H_0 \subset G$. Não é difícil verificar que G também deve conter o completamento H de H_0. Basta então verificar que toda função de G também está em H. Para isto, seja $f \in G$, como H é um subespaço fechado de G, podemos usar a decomposição ortogonal que vimos no Capítulo 3 para escrever $G = H \oplus H^{\perp}$. Assim, há uma único decomposição de f como $f = g + h$, em que $g \in H$ e $h \in H^{\perp}$. Como K é o kernel de reprodução tanto de H quanto de G, temos:

$$f(x) = \langle f, K_x \rangle_G = \langle g, K_x \rangle_G + \langle h, K_x \rangle_G.$$

h é ortogonal à K_x, o que nos permite concluir que $f(x) = \langle g, K_x \rangle_G = \langle g, K_x \rangle_H = g(x)$. Assim, $f \in H$ e $G = H$, como queríamos demonstrar. □

Observação. Pelo teorema anterior, podemos concluir que os kernels são exatamente as funções simétricas e positivo-definidas.

Teorema 4.3. Sejam K_1 e K_2 kernels definidos sobre o conjunto $X \subseteq \mathbb{R}^d$, $\lambda \in \mathbb{R}_+, f : X \to \mathbb{R}$ uma função, $\phi : X \to \mathbb{R}^e$ uma transformação não linear e $K_3 : \mathbb{R}^e \times \mathbb{R}^e \to \mathbb{R}$ um kernel definido sobre $\mathbb{R}^e \times \mathbb{R}^e$, então as seguintes funções também são kernels:

1. $K(x,y) = K_1(x,y) + K_2(x,y)$,
2. $K(x,y) = \lambda K_1(x,y)$,
3. $K(x,y) = K_1(x,y)K_2(x,y)$,
4. $K(x,y) = f(x)f(y)$,
5. $K(x,y) = K_3(\phi(x), \phi(y))$.

Demonstração. Fixe $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$. As funções dos itens 1,2,3 e 5 são todas simétricas por conta da simetria de K_1, K_2 e K_3, enquanto a função do item 4 é simétrica por conta da comutatividade da multiplicação. Agora, vamos mostrar que cada uma dessas funções é positivo-definida. Para 1, observe que:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (K_1(x_i,x_j) + K_2(x_i,x_j)) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_1(x_i,x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_2(x_i,x_j) \geq 0,$$

uma vez que K_1 e K_2 são ambas positivo-definidas. Assim, K é um kernel. O segundo item é análogo. Para o terceiro item, é necessário usar alguns fatos relacionados ao produto de Schur de duas matrizes[7]. O quarto item é imediato se fizermos a fatoração certa:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j f(x_i) f(x_j) = (\sum_{i=1}^{n} c_i f(x_i))^2 \geq 0.$$

Por fim, o item 5 já foi explorado na Proposição 4.1. □

[7] Há uma prova do presente teorema na seção 3.3.2 da referência (CRISTIANINI, 2000).

Observação. Note que o conjunto dos kernels sobre um conjunto X é quase um espaço vetorial, a condição limitante sendo que os escalares λ devem ser não negativos.

Exemplo 4.3. Usando o teorema anterior, conseguimos mostrar que $K(x,y) = (\langle x,y\rangle_d + c)^n$, em que $\langle\cdot,\cdot\rangle_d$ é o produto interno usual de $\mathbb{R}^d$ e $c \geq 0$, é um Kernel. Com efeito, vamos mostrar que $K_1(x,y) = \langle x,y\rangle_d + c$ é um kernel e então usaremos o fato que $K(x,y) = K_1(x,y)^n$ para concluir que este é um kernel. Seja $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ a transformação não linear dada por $\phi(x) = (x_1, x_2, ..., x_n, \sqrt{c})$, então $K_1(x,y) = \langle \phi(x), \phi(y)\rangle_{d+1}$, o que nos permite concluir que K_1 e K são kernels.

Corolário 4.1. Seja $K_1 : X \times X \to R$ um kernel sobre $X \times X$ e p um polinômio com coeficientes não negativos, então $K(x,y) = p(K_1)(x,y)$ é um kernel.

Demonstração. A demonstração é análoga ao exemplo anterior. □

Teorema 4.4. Seja $(K_n)_{n\in\mathbb{N}}$ uma sequência de kernels sobre $X \times X$ tal que, para todo x e y em X, $\lim_{n\to\infty} K_n(x,y)$ exista, então podemos definir o limite pontual dos K_n:

$$K(x,y) = \lim_{n\to\infty} K_n(x,y).$$

Assim definida, K é um kernel.

Demonstração. Sejam $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$, então $K(x,y) = \lim_{n\to\infty} K_n(x,y) = \lim_{n\to\infty} K_n(y,x) = K(y,x)$, uma vez que toda função K_n é simétrica. Adicionalmente, temos que:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \lim_{n\to\infty} K_n(x_i,x_j).$$

Podemos tirar o limite para fora da última expressão, o que nos dá:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) = \lim_{n\to\infty}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_n(x_i,x_j).$$

Como cada K_n é positivo-definida, temos que todo elemento da sequência a_n dada por $a_n = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_n(x_i,x_j)$ é maior ou igual a zero. Portanto,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) = \lim_{n\to\infty} a_n \geq 0.$$

Podemos concluir que K é um kernel. □

Exemplo 4.4. O Kernel Gaussiano é dado por $K(x,y) = K(x,y) = e^{-\gamma\|x-y\|^2}$, onde $\gamma = \frac{1}{2\sigma^2}$. Vamos verificar que é mesmo um kernel. De fato, note que podemos reescrevê-lo como $K(x,y) = e^{-\gamma\|x\|^2} e^{-\gamma\|y\|^2} e^{2\gamma\langle x,y\rangle}$. Dessa forma, se definirmos $f(x) = e^{-\gamma\|x\|^2}$, então $K_1(x,y) = f(x)f(y)$, o produto dos dois primeiros termos da reformulação de K, é um kernel pelo item 4 do Teorema 4.3. Pelo item 3 do mesmo teorema, sabemos que basta verificar que $K_2(x,y) = e^{2\gamma\langle x,y\rangle}$ também é um kernel para que possamos concluir que K é um kernel. Usando a expansão de e^x em série

de potência, vejamos que $K_2(x,y) = \sum_{n=0}^{\infty} \frac{K_3(x,y)}{n!}$, onde $K_3(x,y) = 2\gamma\langle x,y\rangle$ é um kernel. Pelos Corolário 4.1 e Teorema 4.4, somos capazes de concluir que K_2 e, portanto, K são kernels.

Observação. Usando o Kernel Gaussiano, é sempre possível separar perfeitamente um conjunto de dados finito. Para mais informações, consulte a referência (PAUL, 2015).

4.8 CONSEQUÊNCIAS DA ABORDAGEM POR EHR

A abordagem de Métodos de Kernel por espaços de Hilbert de Reprodução abre espaço para alguns resultados interessantes. Entre eles, se destaca o Teorema do Representante, cujo enunciado pode ser visto a seguir.

Teorema 4.5. (Teorema do Representante) Sejam $K : X \times X \to \mathbb{R}$ um kernel com EHR associado H_K, $(x_1,y_1),(x_2,y_2),...,(x_n,y_n) \in X \times \mathbb{R}$ um conjunto de dados, $g : [0,\infty) \to \mathbb{R}$ uma função estritamente crescente e $E : (X \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$ uma função de erro, então podemos definir o seguinte funcional de erro regularizado sobre H_K:

$$f \mapsto E((x_1,y_1,f(x_1)),(x_2,y_2,f(x_2)),...,(x_n,y_n,f(x_n))) + g(\|f\|).$$

Dessa forma, qualquer função f^* que atinja o valor mínimo deste funcional admite uma representação da forma:

$$f^*(x) = \sum_{i=1}^{n} \alpha_i K(x,x_i).$$

Demonstração. Uma demonstração para este teorema pode ser encontrada na referência (SCHÖLKOPF; HERBRICH; SMOLA, 2001). □

Observação. O processo de treinamento de alguns métodos de kernel (como o Kernel-SVM) pode ser formulado em termos de problemas de otimização; isto é, em termos de encontrar um valor mínimo para um funcional de erro:

$$f^* = \underset{f \in H_K}{\operatorname{argmin}} \{E((x_1,y_1,f(x_1)),(x_2,y_2,f(x_2)),...,(x_n,y_n,f(x_n))) + g(\|f\|)\}.$$

Nestes casos, o Teorema do Representante nos diz que só é necessário buscar a solução f^* em um subespaço de dimensão finita: $\operatorname{Span}\{K_{x_1},K_{x_2},...,K_{x_n}\}$. Isso reduz um problema de otimização de dimensão potencialmente infinita (como no caso do Kernel Gaussiano) para um problema de dimensão finita, possibilitando implementações numéricas como programas de computador.

5 CONCLUSÕES

Neste trabalho, abordamos espaços normados e espaços de Hilbert. Discutimos as razões pelas quais noções analíticas de convergência e topologia se fazem necessárias para o estudo satisfatório de espaços vetoriais de dimensão infinita. Vimos também como a geometria dos espaços de Hilbert possibilita uma teoria simples e harmoniosa. Por fim, usamos os primeiros dois capítulos como base para desenvolvermos uma melhor intuição sobre os espaços de Hilbert de Reprodução onde operam os Métodos de Kernel.

Há várias extensões naturais desse trabalho. De fato, seria possível abordar outros métodos de kernel, como o Kernel-PCA e Kernel-SVM. Em outra direção, seria possível discutir mais a fundo sobre alguns dos kernels, como o Gaussiano e o Laplaciano. Seria interessante construir os EHR associados a esses kernels e então utilizar métodos da Análise Funcional para revelar sua estrutura. Valiosas também seriam as adições com respeito às propriedades estatísticas do Kernel Gaussiano. Essas questões serão abordadas em trabalhos futuros.

REFERÊNCIAS

ABBOTT, S. **Understanding analysis**. New York: Springer, 2015. ISBN 978-1493927111.

AIZERMAN, M. A. Theoretical foundations of the potential function method in pattern recognition learning. **Automation and Remote Control**, v. 25, p. 821–837, 1964.

ARONSZAJN, N. Theory of reproducing kernels. **Transactions of the American Mathematical Society**, p. 337–404, 1950.

AXLER, S. **Linear Algebra Done Right**. [S.l.]: Springer, 2015. ISBN 978-3319110790.

AXLER, S. **Measure, Integration & Real Analysis**. Springer International Publishing, 2020. Disponível em: https://doi.org/10.1007/978-3-030-33143-6.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the Fifth Annual Workshop on Computational Learning Theory**. New York, NY, USA: Association for Computing Machinery, 1992. (COLT '92), p. 144–152. ISBN 089791497X. Disponível em: https://doi.org/10.1145/130385.130401.

CRISTIANINI, N. **An introduction to support vector machines and other kernel-based learning methods**. Cambridge New York: Cambridge University Press, 2000. ISBN 9780511801389.

DEISENROTH, M. **Mathematics for machine learning**. Cambridge, United Kingdom New York, NY: Cambridge University Press, 2020. ISBN 110845514X.

HALMOS, P. **Naive Set Theory**. Mansfield Centre, Conn: Martino Pub, 2011. ISBN 1614271313.

HENDERSON, L. **The Problem of Induction**. 2020. Disponível em: https://plato.stanford.edu/archives/spr2020/entries/induction-problem.

HILBERT, D. **Hilbert's radio address**. 1930. David Hilbert's radio address. Disponível em: https://www.maa.org/press/periodicals/convergence/david-hilberts-radio-address-english-translation.

KREYSZIG, E. **Introductory Functional Analysis with Applications**. Wiley, 1989. (Wiley Classics Library). ISBN 9780471504597. Disponível em: https://books.google.com.br/books?id=nZmpQgAACAAJ.

LESLIE, C.; ESKIN, E.; NOBLE, W. S. The spectrum kernel: A string kernel for svm protein classification. In: **Biocomputing 2002**. [S.l.]: World Scientific, 2001. p. 564–575.

MARTINI. **Showing the basis of a Hilbert Space have the same cardinality**. 2012. Mathematics Stack Exchange. Disponível em: https://math.stackexchange.com/q/232182.

MERCER, J. Functions of positive and negative type and their connection with the theory of integral equations. **Philosophical Transactions of the Royal Society**, p. 441–458, 1909.

MINSKY, M. **Perceptrons; an introduction to computational geometry**. Cambridge, Mass: MIT Press, 1969. ISBN 9780262130431.

MINTU. **Let X be an infinite dimensional Banach space. Prove that every Hamel basis of X is uncountable.** 2012. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/217516 (version: 2019-07-10). Disponível em: https://math.stackexchange.com/q/217516.

MORENO, P.; HO, P.; VASCONCELOS, N. A kullback-leibler divergence based kernel for svm classification in multimedia applications. **Advances in neural information processing systems**, v. 16, 2003.

NAVOT, A. **On the role of feature selection in machine learning**. Tese (Doutorado) — Citeseer, 2006.

PAUL. **How can SVM find a feature space where linear separation is always possible?** 2015. Disponível em: https://stats.stackexchange.com/a/168346/340026.

SCHLKOPF, B. **LEARNING WITH KERNELS : support vector machines, regularization, optimization, and beyond**. Place of publication not identified: MIT Press, 2018. ISBN 9780262536578.

SCHÖLKOPF, B.; HERBRICH, R.; SMOLA, A. J. A generalized representer theorem. In: SPRINGER. **International conference on computational learning theory**. [S.l.], 2001. p. 416–426.

SHANKAR, K. et al. Optimal feature-based multi-kernel svm approach for thyroid disease classification. **The journal of supercomputing**, Springer, v. 76, n. 2, p. 1128–1143, 2020.

STEELE, J. **The Cauchy-Schwarz master class : an introduction to the art of mathematical inequalities**. Cambridge, UK New York: Cambridge University Press, 2004. ISBN 978-0521837750.

UNION, E. **A European approach to artificial intelligence**. 2022.

VERT, J.-P. **Aronszajn's theorem**. 2020. Disponível em: https://members.cbio.mines-paristech.fr/ jvert/svn/kernelcourse/notes/aronszajn.pdf.

VOCATURO, E.; PERNA, D.; ZUMPANO, E. Machine learning techniques for automated melanoma detection. In: **2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)**. [S.l.: s.n.], 2019. p. 2310–2317.

WEAVER, N. **Nonseparable Hilbert spaces**. 2016. MathOverflow. URL:https://mathoverflow.net/q/230435 (version: 2016-02-07). Disponível em: https://mathoverflow.net/q/230435.

WIJ. **What makes the Gaussian kernel so magical for PCA, and also in general?** 2015. Cross Validated. URL:https://stats.stackexchange.com/q/133037 (version: 2015-01-11). Disponível em: https://stats.stackexchange.com/q/133037.

ZAREMBA, S. L'équation biharmonique et une class remarquable de functions fondamentales harmoniques. **Bulletin International de l'Acadmie des Sciences de Cracovie**, p. 147–196, 1907.

APÊNDICE A – ARTIGO SBC

Fundamentos de Aprendizado de Máquina: Análise Funcional voltada a Métodos de Kernel

Alek Fröhlich[1]

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina
Caixa Postal 476 – Florianópolis – SC – Brazil

alek.frohlich@grad.ufsc.br

Abstract. *Kernel Methods are one of the main paradigms in Machine Learning. The symmetric positive-definite kernels employed by these methods are naturally associated with Reproducing Kernel Hilbert Spaces (RKHS) via the celebrated Moore-Aronszajn Theorem. In this work, we explore this connection to rigorously justify the correctness of the kPerceptron classification algorithm. Furthermore, we indicate how this connection may be used to justify the kernel methodology as a whole.*

Resumo. *A abordagem por Métodos de Kernel é um dos principais paradigmas em Aprendizado de Máquina. Os kernels positivo-definidos utilizados por esses métodos estão naturalmente associados a espaços de Hilbert de Reprodução (EHR) por meio do teorema de Moore-Aronszajn. Neste trabalho, exploramos essa conexão para justificar a corretude do algoritmo de classificação binária kPerceptron. Além disso, indicamos como é possível estender essas considerações para justificar a abordagem por kernels como um todo.*

1. Introdução

A troca de representação de conjuntos de dados é uma técnica utilizada no Aprendizado de Máquina. Por certo, é muito comum que Métodos de Redução de Dimensionalidade e Seleção de Atributos como Análise de Componentes Principais (Principal Component Analysis - PCA em inglês) sejam aplicados em uma etapa de pré-processamento em fluxos de trabalho de Aprendizado de Máquina (AM) [Navot 2006]. Para alguns algoritmos conhecidos como Métodos de Kernel, há disponível uma técnica de troca de representação poderosa: o Truque do Kernel. Para estes algoritmos, a única informação relevante do conjunto de dados é o valor do produto interno entre pares de vetores de treino. Isto é, a quantidade: $\langle x_i, x_j \rangle$. Por consequência, qualquer transformação $\phi : X \to H$ aplicada sobre o conjunto de dados só aparecerá em expressões da forma: $\langle \phi(x_i), \phi(x_j) \rangle$. É um fato interessante que conseguimos encontrar funções da forma $K(x, y) = \langle \phi(x), \phi(y) \rangle$, onde ϕ e H estão implícitos, que são úteis para a resolução de problemas de Aprendizado de Máquina. Um exemplo é o Kernel Gaussiano, que é altamente usado em conexão com algoritmos como kSVM e kPCA:

$$K(x, y) = e^{-\gamma \|x-y\|^2}.$$

Como esses algoritmos dependem do uso do produto interno, sabemos que a imagem H de ϕ é um espaço com produto interno. Isso motiva a pergunta: qual seria a

natureza desses espaços? É uma consequência do Teorema de Moore-Aronszajn, que H é um espaço de Hilbert de Reprodução (EHR), cuja definição pode ser vista a seguir:

Definição 1.1. Dizemos que um espaço de Hilbert H é um espaço de Hilbert de Reprodução se for um espaço de funções $f : X \to \mathbb{R}$ e todo funcional avaliação $T_x : H \to \mathbb{R}$ dado por $T_X(f) = f(x)$ for contínuo.

Esses espaços foram primeiramente estudados no contexto de Análise Harmônica por Stanisław Zaremba [Zaremba 1907] e, simultaneamente, no contexto de Equações Integrais por James Mercer [Mercer 1909]. Eventualmente, o assunto foi sistematicamente desenvolvido por Nachman Aronszajn e Stefan Bergman [Aronszajn 1950]. É possível se dizer que a relação de kernels de reprodução com Aprendizado de Máquina começou com a descoberta de que o Perceptron não era capaz de classificar conjuntos de dados não linearmente separáveis [Minsky 1969]. De fato, o Kernel-Perceptron foi o primeiro algoritmo de classificação a ser adaptado para o uso de kernels [Aizerman 1964]. Dentre os vários usuários de kernels, o algoritmo mais proeminente é o kSVM, introduzido por Vapnik na década de 90 [Boser et al. 1992]. De forma similar ao desenvolvimento sistemático da teoria de kernels de reprodução, demorou um pouco para que os Métodos de Kernels fossem desenvolvidos sistematicamente. Neste caso, a figura que mais se destaca é Bernhard Schölkopf, autor de uma das principais referências na área [SCHLKOPF 2018].

Neste trabalho, generalizamos a corretude do Perceptron Binário com relação a qualquer espaço de entrada, desde que seja um espaço com produto interno. Em seguida, desenvolvemos um pouco da teoria de kernels a ponto de conseguir justificar que os principais kernels da literatura são de fato kernels. Por fim, demonstramos o teorema de Moore-Aronszajn, que esclarece a conexão de Métodos de Kernel com espaços de Hilbert de Reprodução.

O resto deste artigo está organizado da seguinte maneira: na seção 2, motivamos o Truque do Kernel por meio do problema de classificação binária e do kPerceptron. Na seção 3, desenvolvemos uma caracterização equivalente para funções serem kernels, isso nos permitirá construir uma rápida teoria de kernels, o que, por sua vez, permitirá concluir que funções como o Kernel Gaussiano e os kernels polinomiais são de fato kernels. Na seção 4, apresentamos o teorema de Moore-Aronszajn e ilustramos como essa conexão com EHR pode ser usada para justificar a corretude de outros Métodos de Kernel como o kSVM. Por fim, apresentamos conclusões e possíveis trabalhos futuros na seção 5.

2. Preliminares: Classificação Binária e o Truque do Kernel

Suponha que $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ é um conjunto de dados onde cada x_i pertence a um espaço de entrada $X \subset \mathbb{R}^d$ e cada y_i é ± 1 (rótulos negativos e positivos). O Perceptron é um algoritmo de classificação e, portanto, seu objetivo é achar uma função de decisão que permita classificar qualquer ponto $x \in X$, inclusive aqueles não vistos no conjunto de treinamento. No caso do Perceptron binário, essa função de decisão define duas regiões no espaço: H_+ e H_-, em que $H_+ \subset X$ é composta por todos os pontos de X que são classificados como $+1$ e $H_- \subset X$ é composta por todos os pontos que são classificados como -1. Como veremos em sequência, essas regiões são precisamente o que fica acima e abaixo de um hiperplano orientado (no caso $d = 2$, uma reta), como pode ser visto na figura abaixo.

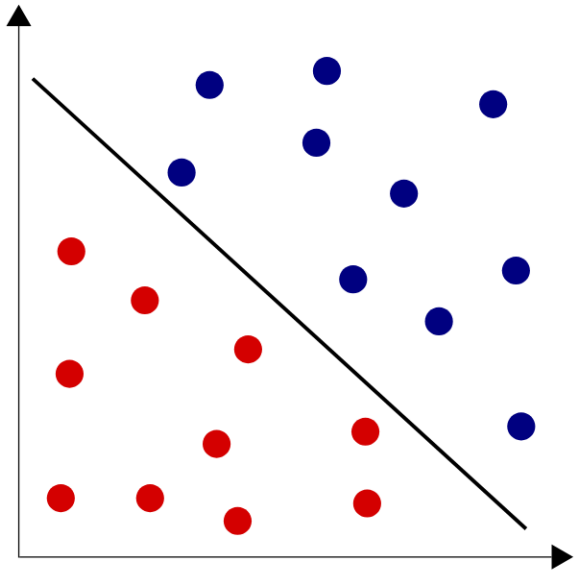


Figura 1. Exemplo de hiperplano orientado separando duas classes de pontos. Fonte: Mekeor, distribuída pela licença CC BY-SA 3.0.

2.1. Função de Decisão

A função de decisão do Perceptron é da forma:

$$f(x) = \langle w, x \rangle + b,$$

onde w é um vetor de $\mathbb{R}^d$, $b \in \mathbb{R}$ e $\langle \cdot, \cdot \rangle$ é o produto interno usual. Ou seja, estamos trabalhando com ℓ_n^2. Generalizaremos essa situação para espaços de Hilbert quaisquer quando estivermos trabalhando com kernels. Por ora, note que $f(x) = 0$ define um hiperplano ortogonal ao vetor w. Adicionalmente, $f(x) > 0$ define a região positiva (acima) do hiperplano e $f(x) < 0$ define a região negativa (abaixo) do hiperplano. O objetivo do Perceptron então é encontrar w e b tais que todos os pontos x_i tais que $y_i = +1$ fiquem na região positiva e todos os pontos x_j tais que $y_j = -1$ fiquem na região negativa. Podemos condensar essa condição como:

$$\forall i : y_i f(x_i) > 0.$$

Observação. Assumiremos em toda nossa análise que o conjunto de treino é linearmente separável; i.e., que existem w e b que satisfazem a equação acima e que é possível encontrá-los de forma a deixar uma pequena margem ao redor do hiperplano sem que nenhum ponto de treino invada essa margem. Podemos formalizar essa condição como:

$$\forall i : y_i f(x_i) > r,$$

para algum $r > 0$.

2.2. O Algoritmo

O algoritmo é básico e parece funcionar bem em exemplos simples, mas como ter certeza que ele sempre termina com um hiperplano que corretamente classifique o conjunto de

Algoritmo 1 Perceptron Binário

$w \leftarrow 0$
$b \leftarrow 0$
$R \leftarrow \max_{1 \leq i \leq n} ||x_i||$
while houver erros de classificação **do**
  for $i = 1$ to n **do**
    if $y_i(\langle w, x_i \rangle + b) \leq 0$ **then**
      $w \leftarrow w + y_i x_i$
      $b \leftarrow b + y_i R^2$
    end if
  end for
end while
return (w, b)

dados? Para isso, precisamos demonstrar a corretude do algoritmo. Note que, se conseguirmos garantir que ele sempre para (dado um conjunto de dados linearmente separável), então também conseguimos garantir que ele sempre encontra um hiperplano válido. Isso se dá pois o algoritmo não para enquanto há erros de classificação. Como veremos a seguir, é possível encontrar um limite superior para o número de erros t que o algoritmo comete. Esse limite dependerá de duas métricas importantes: a margem e a escala do conjunto de dados. Definimos esses conceitos antes de partirmos para o teorema principal.

Definição 2.1. Seja x_i um vetor de treino e (w, b) os parâmetros que definem um hiperplano, então a margem funcional entre x_i e o hiperplano é dada por

$$\gamma_i = y_i(\langle w, x_i \rangle + b).$$

Se $\|w\| = 1$, então γ_i representa a distância entre o hiperplano e o ponto de treino. Quando isso acontece, chamamos γ_i de margem geométrica do vetor x_i. Se tomarmos o mínimo entre as margens dos exemplos, temos a margem funcional (geométrica, respectivamente) do hiperplano, como pode ser visto a seguir:

$$\gamma = \min_{1 \leq i \leq n} \gamma_i.$$

Observação. Note que a escolha de representantes (w, b) para um dado hiperplano é um problema mal posto no sentido de que há vários (w, b) correspondendo a um mesmo hiperplano. De fato, observamos que $\langle w, x \rangle + b = 0$ se, e somente se, $\langle \lambda \cdot w, x \rangle + \lambda b = 0$, para qualquer escalar λ.

Abordaremos agora a prova de corretude do Perceptron. A demonstração supõe que o conjunto de dados é separável pela origem. Isto é, que há um hiperplano da forma $(w, 0)$ que classifica corretamente o conjunto de dados. A razão para esta simplificação é que assim não é necessário supor que $X = \ell_n^2$. De fato, seria possível implementar este algoritmo sobre qualquer espaço com produto interno e é exatamente isso que faremos em conexão com o Truque do Kernel, como veremos mais à frente.

Teorema 2.1. Seja $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ um conjunto de treino linearmente separável pela origem tal que haja w^* que corretamente o classifique. Isto é, tal que

$\forall i : y_i \langle w^*, x_i \rangle \geq \gamma$ para alguma margem γ. Sob estas condições, o número de erros t que o algoritmo comete antes de terminar é limitado superiormente da seguinte forma:

$$t \leq \left(\frac{R}{\gamma} \right)^2.$$

Demonstração. A prova que segue é uma adaptação da prova do Teorema 2.3 da referência [Cristianini 2000]. Deixe o algoritmo rodar sobre o dado conjunto de treinamento. Sua execução define uma sequência de vetores w_t, onde w_t é o vetor de peso atualizado logo após o t-ésimo erro e $w_0 = 0$. Por definição, temos que:

$$w_{t+1} = w_t + y_i x_i.$$

onde i vem do vetor de treino que foi classificado erroneamente. Usando essa definição recursiva, obtemos:

$$\langle w_t, w^* \rangle = \langle w_{t-1}, w^* \rangle + y_i \langle x_i, w^* \rangle.$$

Como w^* classifica x_i corretamente, temos:

$$\langle w_{t-1}, w^* \rangle + y_i \langle x_i, w^* \rangle \geq \langle w_{t-1}, w^* \rangle + \gamma.$$

Iterando até chegar em w_0, ficamos com

$$\langle w_t, w^* \rangle \geq t\gamma.$$

Agora, nós buscamos uma desigualdade na outra direção; conseguimos isso limitando $\|w_t\|$:

$$||w_t||^2 = ||w_{t-1}||^2 + 2\langle w_{t-1}, y_i x_i \rangle + ||x_i||^2.$$

Como w_{t-1} errou a classe de x_i, nós sabemos que a expressão do meio é menor ou igual a zero, o que permite a simplificação a seguir:

$$||w_t||^2 \leq ||w_{t-1}||^2 + ||x_i||^2 \leq ||w_{t-1}||^2 + R^2.$$

Aplicando isso várias vezes e depois tomando raízes quadradas, obtemos:

$$||w_t|| \leq \sqrt{t} R.$$

Agora podemos encadear as desigualdades:

$$t\gamma \leq \langle w_t, w^* \rangle \leq ||w_t|| ||w^*|| \leq \sqrt{t} R.$$

Elevando ao quadrado e cancelando t, obtemos:

$$t\gamma^2 \leq R^2.$$

Que finalmente nos leva a desigualdade desejada:

$$t \leq \frac{R^2}{\gamma}.$$

□

2.3. Transformações não lineares

Como vimos na seção passada, o Perceptron Binário funciona sobre qualquer espaço com produto interno. Isso se torna interessante quando consideramos o seguinte problema: muitos conjuntos de dados relevantes sobre os quais gostaríamos de aplicar o Perceptron não são linearmente separáveis. A Figura 2 ilustra este cenário. Vemos que não há escolha de (w, b) que seja capaz de separar esse conjunto de dados. Uma solução possível é mudar a representação dos dados. De fato, poderíamos buscar uma função $\phi : X \to H$, onde H é outro espaço com produto interno de tal forma que $\phi(X)$ fosse linearmente separável.

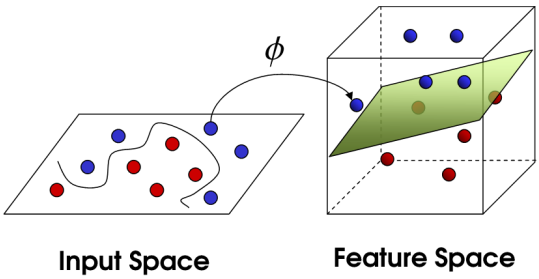


Figura 2. ϕ faz com que o conjunto de dados se torne linearmente separável. Fonte: adaptado de [Vocaturo et al. 2019].

Há dois problemas com essa abordagem. O primeiro e mais sério é o seguinte: como encontrar uma transformação que torne um dado conjunto de treino linearente separável? Dependendo da dimensão[1] e do conjunto de dados, essa pergunta pode ser fácil de responder, mas, em geral, não é. Outro problema que também deve ser considerado é a eficácia de computar $\{\phi(x_1), \phi(x_2,), ..., \phi(x_n)\}$. Dependendo de ϕ, esse cálculo pode ser proibitivo. Esses problemas sugerem a pergunta: não seria possível encontrar uma ϕ ou uma coleção parametrizada de $(\phi_i)_{i \in I}$ que seja altamente versátil e que sirva para vários conjuntos de dados encontrados na prática? A resposta para essa pergunta é sim: não só é possível, como conseguimos também pular a etapa em que ϕ é computada. De fato, vamos agora introduzir a ideia que viabiliza esse atalho: o Truque do Kernel.

2.4. Truque do Kernel

Para podermos falar do Truque do Kernel, precisamos primeiro fazer algumas modificações no Perceptron. Note que o vetor w encontrado pelo Algoritmo 1 é

[1]Nos casos em que $d \leq 3$, claramente temos mais intuição visual.

combinação linear (com escalares inteiros) dos vetores de treino $x_1, x_2, ..., x_n$. De fato, inicializamos w como zero e depois apenas o atualizamos somando $\pm x_i$. Assim, podemos escrever $w = \sum_{i=1}^{n} \alpha_i x_i$, em que $\alpha_i \in \mathbb{Z}$. Essa observação possibilita uma reformulação do Perceptron, chamada de versão dual do Perceptron, onde o objetivo de aprendizado é encontrar $\alpha_1, \alpha_2, ..., \alpha_n$ e b tais que $(\sum_{i=1}^{n} \alpha_i x_i, b)$ produza um hiperplano separante. Com efeito, considere o seguinte algoritmo:

Algoritmo 2 Perceptron Binário (Dual)

```
α ← 0
b ← 0
R ← max_{1≤i≤n} ||x_i||
while houver erros de classificação do
    for i = 1 to n do
        if y_i(Σ_{j=1}^{n} α_j ⟨x_j, x_i⟩ + b) ≤ 0 then
            α_i ← α_i + y_i
            b ← b + y_i R^2
        end if
    end for
end while
return (α, b)
```

Observação. Os vetores de treino x_i só aparecem no algoritmo acima na forma de produtos internos $\langle x_j, x_i \rangle$. Assim, caso usemos uma transformação não linear $\phi : X \to H$, ela também só aparecerá na forma $\langle \phi(x_j), \phi(x_i) \rangle$. Essa observação motiva a seguinte definição.

Definição 2.2. Seja X um espaço de entrada e $K : X \times X \to \mathbb{R}$ uma função da forma

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_H$$

para algum espaço de Hilbert H e transformação não linear $\phi : X \to \mathbb{R}$, então dizemos que K é um kernel.

Definição 2.3. Uma função $K : X \times X \to \mathbb{R}$ é dita positivo-definida se, para cada escolha de $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$, tem-se:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0. \tag{1}$$

Proposição 2.1. Todo kernel é uma função simétrica e positivo-definida.

Demonstração. Seja $K : X \times X \to \mathbb{R}$ um Kernel, então $K(x, y) = \langle \phi(x), \phi(y) \rangle = \langle \phi(y), \phi(x) \rangle = K(y, x)$ para cada x e y em X por conta da simetria do produto interno. Além disso, dados $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$, temos $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle$. Usando a linearidade do produto interno, podemos ver que $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) = \langle z, z \rangle$, onde $z = \sum_{i=1}^{n} c_i \phi(x_i)$. Note que essa última expressão é o produto interno de um vetor com ele mesmo e, portanto, deve ser maior ou igual a zero. □

O Truque do Kernel consiste em buscar por kernels fáceis de computar cuja transformação não linear implícita seja versátil e consiga separar vários conjunto de dados encontrados na prática. Esses kernels de fato existem e discutiremos alguns deles na próxima seção. Por ora, vejamos um exemplo simples de uma função que é um kernel fácil de computar e que é útil em alguns cenários.

Exemplo 2.1. Seja $X = \mathbb{R}^2$, então $K(x, y) = \langle x, y\rangle^2$ é um Kernel. Com efeito, sejam $x = (x_1, x_2)$ e $y = (y_1, y_2)$ vetores em $\mathbb{R}^2$, então $K(x, y) = \langle x, y\rangle^2 = (x_1y_1+x_2y_2)^2 = x_1^2y_1^2+2x_1x_2y_1y_2 + x_2^2y_2^2$. Note que, se definirmos $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, então $K(x, y) = \langle\phi(x), \phi(y)\rangle_3$, onde $\langle\cdot,\cdot\rangle_3$ é o produto escalar usual de $\mathbb{R}^3$.

Observação. É possível generalizar o exemplo anterior em algumas direções. Na sua forma mais geral, o kernel $K(x, y) = (\langle x, y\rangle_d + c)^n$, em que o espaço de entrada é $\mathbb{R}^d$ e $c \geq 0$, recebe o nome de Kernel Polinomial de grau n.

3. Encontrando Kernels

Veremos na seção seguinte que toda função simétrica e positivo-definida é um kernel. Essa caracterização nos permite desenvolver uma pequena teoria de kernels, como pode ser visto pelos próximos teoremas e corolários.

Teorema 3.1. Sejam K_1 e K_2 kernels definidos sobre o conjunto $X \subseteq \mathbb{R}^d$, $\lambda \in \mathbb{R}_+$, $f : X \to \mathbb{R}$ uma função, $\phi : X \to \mathbb{R}^e$ uma transformação não linear e $K_3 : \mathbb{R}^e \times \mathbb{R}^e \to \mathbb{R}$ um kernel definido sobre $\mathbb{R}^e \times \mathbb{R}^e$, então as seguintes funções também são kernels:

1. $K(x, y) = K_1(x, y) + K_2(x, y)$,
2. $K(x, y) = \lambda K_1(x, y)$,
3. $K(x, y) = K_1(x, y)K_2(x, y)$,
4. $K(x, y) = f(x)f(y)$,
5. $K(x, y) = K_3(\phi(x), \phi(y))$.

Demonstração. Fixe $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$. As funções dos itens 1,2,3 e 5 são todas simétricas por conta da simetria de K_1, K_2 e K_3, enquanto a função do item 4 é simétrica por conta da comutatividade da multiplicação. Agora, vamos mostrar que cada uma dessas funções é positivo-definida. Para 1, observe que:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j(K_1(x_i, x_j)+K_2(x_i, x_j)) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_jK_1(x_i, x_j)+\sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_jK_2(x_i, x_j) \geq 0,$$

uma vez que K_1 e K_2 são ambas positivo-definidas. Assim, K é um kernel. O segundo item é análogo. Para o terceiro item, é necessário usar alguns fatos relacionados ao produto de Schur de duas matrizes[2]. O quarto item é imediato se fizermos a fatoração certa:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_jf(x_i)f(x_j) = (\sum_{i=1}^{n} c_if(x_i))^2 \geq 0.$$

Por fim, o item 5 já foi explorado na Proposição 2.1. □

[2]Há uma prova do presente teorema na seção 3.3.2 da referência [Cristianini 2000].

Exemplo 3.1. Usando o teorema anterior, conseguimos mostrar que $K(x,y) = (\langle x,y\rangle_d + c)^n$, em que $\langle\cdot,\cdot\rangle_d$ é o produto interno usual de $\mathbb{R}^d$ e $c \geq 0$, é um Kernel. Com efeito, vamos mostrar que $K_1(x,y) = \langle x,y\rangle_d + c$ é um kernel e então usaremos o fato que $K(x,y) = K_1(x,y)^n$ para concluir que este é um kernel. Seja $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ a transformação não linear dada por $\phi(x) = (x_1, x_2, ..., x_n, \sqrt{c})$, então $K_1(x,y) = \langle\phi(x), \phi(y)\rangle_{d+1}$, o que nos permite concluir que K_1 e K são kernels.

Corolário 3.1. Seja $K_1 : X \times X \to R$ um kernel sobre $X \times X$ e p um polinômio com coeficientes não negativos, então $K(x,y) = p(K_1)(x,y)$ é um kernel.

Demonstração. A demonstração é análoga ao exemplo anterior. □

Teorema 3.2. Seja $(K_n)_{n\in\mathbb{N}}$ uma sequência de kernels sobre $X \times X$ tal que, para todo x e y em X, $\lim_{n\to\infty} K_n(x,y)$ exista, então podemos definir o limite pontual dos K_n:

$$K(x,y) = \lim_{n\to\infty} K_n(x,y).$$

Assim definida, K é um kernel.

Demonstração. Sejam $x_1, x_2, ..., x_n \in X$ e $c_1, c_2, ..., c_n \in \mathbb{R}$, então $K(x,y) = \lim_{n\to\infty} K_n(x,y) = \lim_{n\to\infty} K_n(y,x) = K(y,x)$, uma vez que toda função K_n é simétrica. Adicionalmente, temos que:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \lim_{n\to\infty} K_n(x_i,x_j).$$

Podemos tirar o limite para fora da última expressão, o que nos dá:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) = \lim_{n\to\infty}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_n(x_i,x_j).$$

Como cada K_n é positivo-definida, temos que todo elemento da sequência a_n dada por $a_n = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_n(x_i,x_j)$ é maior ou igual a zero. Portanto,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i,x_j) = \lim_{n\to\infty} a_n \geq 0.$$

Podemos concluir que K é um kernel. □

Exemplo 3.2. O Kernel Gaussiano é dado por $K(x,y) = K(x,y) = e^{-\gamma\|x-y\|^2}$, onde $\gamma = \frac{1}{2\sigma^2}$. Vamos verificar que é mesmo um kernel. De fato, note que podemos reescrevê-lo como $K(x,y) = e^{-\gamma\|x\|^2} e^{-\gamma\|y\|^2} e^{2\gamma\langle x,y\rangle}$. Dessa forma, se definirmos $f(x) = e^{-\gamma\|x\|^2}$, então $K_1(x,y) = f(x)f(y)$, o produto dos dois primeiros termos da reformulação de K, é um kernel pelo item 4 do Teorema 3.1. Pelo item 3 do mesmo teorema, sabemos que basta verificar que $K_2(x,y) = e^{2\gamma\langle x,y\rangle}$ também é um kernel para que possamos concluir que K é um kernel. Usando a expansão de e^x em série de potência, vejamos que $K_2(x,y) = \sum_{n=0}^{\infty} \frac{K_3(x,y)}{n!}$, onde $K_3(x,y) = 2\gamma\langle x,y\rangle$ é um kernel. Pelos Corolário 3.1 e Teorema 3.2, somos capazes de concluir que K_2 e, portanto, K são kernels.

4. O Teorema de Moore-Aronszajn e os espaços de Hilbert de Reprodução

Nesta seção, provamos o teorema de Moore-Aronszajn e estabelecemos a caracterização alternativa de kernels como funções simétricas e positivo-definidas.

Definição 4.1. Dizemos que um espaço de Hilbert H é um espaço de Hilbert de Reprodução se for um espaço de funções $f : X \to \mathbb{R}$ e todo funcional avaliação $T_x : H \to \mathbb{R}$ dado por $T_X(f) = f(x)$ for contínuo.

Observação. $P = C[a, b]$ com a norma $p = 2$ é um espaço com produto interno. Sabemos que todo subespaço de dimensão finita de P é um espaço de Hilbert de funções. Isso motiva vários exemplos, como pode ser visto a seguir.

Exemplo 4.1. Seja Y um subespaço de dimensão finita de $C[a, b]$, então $H = (Y, \langle \cdot, \cdot \rangle_{Y \times Y})$ é um espaço de Hilbert. Por construção, H é um espaço de Hilbert de funções, o que significa que podemos definir os funcionais avaliação sobre H. Adicionalmente, como toda transformação linear entre espaços normados de dimensão finita é contínua, temos que todos os funcionais avaliação sobre H são contínuos. Isso nos permite concluir que todo subespaço Y de fimensão finita de $C[a, b]$ é um EHR.

Apenas através de sua definição, não é possível ver qual a relevância de espaços de Hilbert de Reprodução para o estudo de kernels. Para isto, precisamos da seguinte proposição.

Proposição 4.1. Seja H um EHR com funções definidas sobre um conjunto X, então é possível construir um kernel $K : X \times X \to \mathbb{R}$ tal que:

1. $\forall x : K_x := K(x, \_) \in H$.
2. $\forall x : \forall f : X \to \mathbb{R} \in H : f(x) = \langle f, K_x \rangle$.

Demonstração. Seja $T_x : H \to \mathbb{R}$ o funcional avaliação para o ponto x, então, pelo Teorema de Representação de Riesz, há uma única função $K_x \in H$ tal que $T_x(f) = \langle f, K_x \rangle$. Isso nos permite definir uma transformação não linear $\phi(x) = K_x$, que, por sua vez, nos permite definir um kernel $K(x, y) = \langle \phi(x), \phi(y) \rangle = \langle K_x, K_y \rangle$. Como K_y é uma função de H, temos $K_x(y) = T_y(K_x) = \langle K_x, K_y \rangle = K(x, \_)$. Assim, as funções K_x obtidas pelo Teorema de Riesz satisfazem ambas as condições do enunciado. □

Observação. Com a proposição anterior, fica claro a razão por trás do nome espaço de Hilbert de Reprodução. De fato, todo EHR H vem equipado com um kernel que reproduz as funções de H. É possível introzudir o conceito de EHR diretamente através dessa condição, entretanto, assim não seria possível dar exemplos de maneira tão rápida como fizemos.

Um fato muito interessante é que a recíproca da Proposição 4.1 também vale. Isto é, dado um kernel $K : X \times X \to \mathbb{R}$ definido sobre um conjunto X, é possível construir um único espaço de Hilbert de Reprodução H_K cujo kernel associado é exatamente K. Assim, podemos concluir que todo kernel está unicamente associado a um EHR H_K e vice-versa. Esse fato elucida a natureza dos kernels. De fato, com esse resultado, sabemos que todo método de kernel opera sobre um EHR cujos elementos são combinações lineares[3] das funções K_x induzidas pelo kernel. Vejamos o seguinte teorema.

[3]Na verdade, limites de sequências de combinações lineares.

Teorema 4.1. (Moore-Aronszajn) Seja $K : X \times X \to \mathbb{R}$ uma função simétrica e positivo-definida sobre um conjunto X, então existe um único espaço de Hilbert de Reprodução H_K cujo kernel associado é K.

Demonstração. Começamos definindo $K_x = K(x, \_)$ para cada $x \in X$ e construindo o espaço $H_0 = \mathrm{Span}\{K_x\}$. Vamos definir um produto interno sobre H_0. Para isto, sejam f, g e h funções em H_0 e λ um escalar. Podemos escrever $f = \sum_{i=1}^{n} a_i K_{x_i}, g = \sum_{j=1}^{m} b_j K_{y_j}$ e $h = \sum_{i=1}^{n} c_i K_{x_i}$, então

$$\langle f, g\rangle_0 = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j K(x_i, y_j),$$

define um produto interno. De fato, primeiro note que $\langle\cdot,\cdot\rangle_0$ está bem definida, pois

$$\langle f, g\rangle_0 = \sum_{i=1}^{n} a_i g(x_i) = \sum_{j=1}^{m} b_j f(y_j).$$

Além disso, conseguimos ver que $\langle\cdot,\cdot\rangle_0$ é bilinear:

$$\langle f + \lambda\cdot h, g\rangle_0 = \sum_{i=1}^{n}\sum_{j=1}^{m}(a_i + \lambda c_i) b_j K(x_i, y_j) = \langle f, g\rangle_0 + \lambda\langle h, g\rangle_0.$$

Essa função também claramente é simétrica e $\langle f, f\rangle_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$ por conta de K ser positivo-definida. Nos resta mostrar que $\langle f, f\rangle_0 = 0 \Rightarrow f = 0$. Para isto, precisamos do fato que a desigualdade de Cauchy-Schwarz também vale para formas bilineares simétricas e positivo-definidas (a função $\langle\cdot,\cdot\rangle_0$ satisfaz essas condições). Assim, temos:

$$|f(x)| = |\langle f, K_x\rangle_0| \leq \|f\|_0 \|K_x\|_0 = 0.$$

Portanto, $\langle f, f\rangle_0 = 0 \Rightarrow f = 0$ e $\langle\cdot,\cdot\rangle_0$ é um produto interno. Podemos concluir que $P = (H_0, \langle\cdot,\cdot\rangle_0)$ é um pré-espaço de Hilbert. Veja que temos quase todas as condições do teorema satisfeitas: P é um pré-espaço de Hilbert que possui um kernel de reprodução K. O que falta é a completude de P. Para isto, precisaremos trabalhar um pouco mais.

Começamos notando que toda sequência de Cauchy $(f_n)_{n\in N}$ em H_0 admite limite pontual. De fato, veja que:

$$|f_n(x) - f_m(x)| = |\langle f_n - f_m, K_x\rangle_0| \leq \|f_n - f_m\|_0 \|K_x\|.$$

Assim, se fixarmos x, $(f_n(x))_{n\in\mathbb{N}}$ é uma sequência de Cauchy em $\mathbb{R}$. Pela completude de $\mathbb{R}$, sabemos que $f_n(x) \to f(x)$ para algum número $f(x) \in \mathbb{R}$. Isso nos permite definir o limite pontual de toda sequência de Cauchy:

$$f(x) = \lim_{n\to\infty} f_n(x).$$

Iremos denotar por H o conjunto de todos os limites de sequências de Cauchy em H_0. Claramente, $H_0 \subseteq H$, uma vez que, para toda função $f \in H$, podemos montar a sequência de Cauchy $(f)_{n\in\mathbb{N}}$ cujo limite pontual é a própria f. Assim definido, H também é um espaço vetorial. De fato, $0 \in H$ e é possível ver que, se $(f_n)_{n\in\mathbb{N}}$ e $(g_n)_{n\in\mathbb{N}}$ são sequências de Cauchy com limites pontuais f e g e λ é um escalar, então $(f_n + \lambda \cdot g_n)_{n\in\mathbb{N}}$ é uma sequência de Cauchy cujo limite pontual é $f + \lambda \cdot g$. Vamos agora definir um produto interno sobre H de forma que H se torne um espaço de Hilbert de Reprodução. Com efeito, sejam f e g funções em H, então $f = \lim_{n\to\infty} f_n$ e $g = \lim_{n\to\infty} g_n$, em que $(f_n)_{n\in\mathbb{N}}$ e $(g_n)_{n\in\mathbb{N}}$ são sequências de Cauchy em H_0. Podemos definir um produto interno $\langle\cdot,\cdot\rangle$ sobre H da seguinte maneira:

$$\langle f, g\rangle = \lim_{n\to\infty} \langle f_n, g_n\rangle_0.$$

Precisamos mostrar que este limite existe e que não depende da escolha das sequências $(f_n)_{n\in\mathbb{N}}$ e $(g_n)_{n\in\mathbb{N}}$. Começamos mostrando que o limite existe. Faremos isso mostrando que $(\langle f_n, g_n\rangle_0)_{n\in\mathbb{N}}$ é uma sequência de Cauchy em $\mathbb{R}$. Com efeito,

$$|\langle f_n, g_n\rangle_0 - \langle f_m, g_m\rangle_0| = |\langle f_n, g_n\rangle_0 - \langle f_n, g_m\rangle_0 + \langle f_n, g_m\rangle_0 - \langle f_m, g_m\rangle_0|,$$

que podemos agupar para obter:

$$|\langle f_n, g_n\rangle_0 - \langle f_m, g_m\rangle_0| = |\langle f_n, g_n - g_m\rangle_0 - \langle f_n - f_m, g_m\rangle_0|.$$

Aplicando a desigualdade triangular e depois Cauchy-Schwarz, obtemos:

$$|\langle f_n, g_n\rangle_0 - \langle f_m, g_m\rangle_0| \leq \|f_n\|_0 \|g_n - g_m\|_0 + \|f_n - f_m\|_0 \|g_m\|_0 .$$

Como as sequências $(f_n)_{n\in\mathbb{N}}$ e $(g_n)_{n\in N}$ são de Cauchy, sabemos que são limitadas. Sejam M_1 e M_2 números reais tais que $\|f_n\|_0 \leq M_1$ e $\|g_n\|_0 \leq M_2$ para todo n. Com isso,

$$|\langle f_n, g_n\rangle_0 - \langle f_m, g_m\rangle_0| \leq M_1 \|g_n - g_m\|_0 + \|f_n - f_m\|_0 M_2$$

e a sequência em questão é Cauchy. Portanto, $\lim_{n\to\infty}\langle f_n, g_n\rangle_0$ existe. Vamos agora mostrar que este limite independe da escolha das sequências $(f_n)_{n\in\mathbb{N}}$ e $(g_n)_{n\in\mathbb{N}}$. Para isto, sejam $(f'_n)_{n\in\mathbb{N}}$ e $(g'_n)_{n\in\mathbb{N}}$ outras sequências de Cauchy em H_0 tais que $f_n \to f$ e $g_n \to g$ pontualmente, então $|\langle f_n, g_n\rangle_0 - \langle f'_n, g'_n\rangle_0| \to 0$ por um argumento análogo ao que demos para mostrar que $\lim_{n\to\infty}\langle f_n, g_n\rangle_0$ existe. Portanto, $\langle\cdot,\cdot\rangle$ está bem definido. Não é difícil verificar que tal função é simétrica e bilinear usando as propriedades de $\langle\cdot,\cdot\rangle_0$. Note que, para toda f e $g \in H_0$, temos:

$$\langle f, g\rangle = \langle f, g\rangle_0,$$

uma vez que podemos escolher as sequências $(f)_{n\in\mathbb{N}}$ e $(g)_{n\in\mathbb{N}}$ para computar o produto interno em H. Isso nos permite concluir que $\langle 0, 0\rangle = 0$. Por fim, veja que K ainda é um kernel de reprodução para H, uma vez que $K_x \in H_0 \subseteq H$ para todo x e $\langle f, K_x\rangle = \lim_{n\to\infty}\langle f_n, K_x\rangle_0 = f_n(x) = f(x)$. Assim, se $f \in H$ é tal que $\langle f, f\rangle = 0$, podemos inferir que:

$$|f(x)| = |\langle f, K_x\rangle| \leq \|f\| \, \|K_x\| = 0,$$

e $f = 0$. Portanto, H é um pré-espaço de Hilbert com um kernel de reprodução. Veremos agora que H é completo e, portanto, um EHR com kernel associado K. Precisaremos do fato que H_0 é denso em H. De fato, seja $f \in H$, então há uma sequência de funções $(f_n)_{n\in\mathbb{N}}$ em H_0 tal que $f_n \to f$ pontualmente. É possível mostrar[4] que isso implica em $\|f - f_n\| \to 0$, o que nos permite concluir que H_0 é denso em f.

Por fim, seja $(f_n)_{n\in\mathbb{N}}$ uma sequência de Cauchy em H. Pela densidade de H_0 em H, podemos montar outra sequência $(f'_n)_{n\in\mathbb{N}}$, desta vez de elementos em H_0, tal que $\|f_n - f'_n\| \to 0$. Isso nos permite mostrar que $(f'_n)_{n\in\mathbb{N}}$ é de Cauchy:

$$\|f'_n - f'_m\| \leq \|f'_n - f_n\| + \|f_n - f_m\| + \|f_m - f'_m\| \, .$$

Dado $\varepsilon > 0$, claramente conseguimos escolher N grande o suficiente tal que $\|f'_n - f_n\| < \varepsilon/3, \|f_n - f_m\| < \varepsilon/3$ e $\|f_m - f'_m\| < \varepsilon/3$ para todo $m, n \geq N$. Assim, $(f'_n)_{n\in\mathbb{N}}$ é de Cauchy, o que significa que possui limite pontual em H. Seja $f = \lim_{n\to\infty} f_n$, então

$$\|f - f_n\| \leq \|f - f'_n\| + \|f'_n - f_n\| \, .$$

O lado direito claramente vai para zero quando $n \to \infty$, o que nos permite concluir que H é completo. Agora, nos voltamos a questão da unicidade. Seja G outro EHR com kernel de reprodução K, então $H_0 \subset G$. Não é difícil verificar que G também deve conter o completamento H de H_0. Basta então verificar que toda função de G também está em H. Para isto, seja $f \in G$, como H é um subespaço fechado de G, podemos usar a decomposição ortogonal que vimos no Capítulo 3 para escrever $G = H \oplus H^{\perp}$. Assim, há uma único decomposição de f como $f = g + h$, em que $g \in H$ e $h \in H^{\perp}$. Como K é o kernel de reprodução tanto de H quanto de G, temos:

$$f(x) = \langle f, K_x\rangle_G = \langle g, K_x\rangle_G + \langle h, K_x\rangle_G.$$

h é ortogonal à K_x, o que nos permite concluir que $f(x) = \langle g, K_x\rangle_G = \langle g, K_x\rangle_H = g(x)$. Assim, $f \in H$ e $G = H$, como queríamos demonstrar. □

[4]Uma prova similar a este fato pode ser encontrada na referência [Vert 2020].

4.1. Consequências da abordagem por EHR

A abordagem de Métodos de Kernel por espaços de Hilbert de Reprodução abre espaço para alguns resultados interessantes. Entre eles, se destaca o Teorema do Representante, cujo enunciado pode ser visto a seguir.

Teorema 4.2. (Teorema do Representante) Sejam $K : X \times X \to \mathbb{R}$ um kernel com EHR associado H_K, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \in X \times \mathbb{R}$ um conjunto de dados, $g : [0, \infty) \to \mathbb{R}$ uma função estritamente crescente e $E : (X \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$ uma função de erro, então podemos definir o seguinte funcional de erro regularizado sobre H_K:

$$f \mapsto E((x_1, y_1, f(x_1)), (x_2, y_2, f(x_2)), ..., (x_n, y_n, f(x_n))) + g(\|f\|).$$

Dessa forma, qualquer função f^* que atinja o valor mínimo deste funcional admite uma representação da forma:

$$f^*(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i).$$

Demonstração. Uma demonstração para este teorema pode ser encontrada na referência [Schölkopf et al. 2001]. □

Observação. O processo de treinamento de alguns métodos de kernel (como o kSVM) pode ser formulado em termos de problemas de otimização; isto é, em termos de encontrar um valor mínimo para um funcional de erro:

$$f^* = \underset{f \in H_K}{\operatorname{argmin}} \{E((x_1, y_1, f(x_1)), (x_2, y_2, f(x_2)), ..., (x_n, y_n, f(x_n))) + g(\|f\|)\}.$$

Nestes casos, o Teorema do Representante nos diz que só é necessário buscar a solução f^* em um subespaço de dimensão finita: $\operatorname{Span}\{K_{x_1}, K_{x_2}, ..., K_{x_n}\}$. Isso reduz um problema de otimização de dimensão potencialmente infinita (como no caso do Kernel Gaussiano) para um problema de dimensão finita, possibilitando implementações numéricas como programas de computador.

5. Conclusões

Neste trabalho, ilustramos como é possível demonstrar rigorosamente que um método de kernel funciona em conexão com qualquer kernel através da prova de corretude do kPerceptron. De fato, para justificar o funcionamento de outros métodos de kernel, bastaria apresentar suas corretudes em termos de espaços com produto interno (ou EHR), sem assumir que o espaço de entrada em que operam é $\mathbb{R}^d$ para algum d. Além disso, apresentamos a razão pela qual o Kernel Gaussiano e os kernels polinomiais são kernels. Por fim, mostramos que todo kernel e, portanto, todo Método de Kernel, está ligado a um espaço de Hilbert de Reprodução de maneira canônica.

Há várias extensões naturais desse trabalho. De fato, seria possível abordar outros métodos de kernel, como o kPCA e kSVM. Em outra direção, seria possível discutir mais a fundo sobre alguns dos kernels, como o Gaussiano e o Laplaciano. Seria interessante construir os EHR associados a esses kernels e então utilizar métodos da Análise Funcional para revelar sua estrutura. Valiosas também seriam as adições com respeito às propriedades estatísticas do Kernel Gaussiano. Essas questões serão abordadas em trabalhos futuros.

Referências

Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, pages 337–404.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA. Association for Computing Machinery.

Cristianini, N. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge New York.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, pages 441–458.

Minsky, M. (1969). *Perceptrons; an introduction to computational geometry*. MIT Press, Cambridge, Mass.

Navot, A. (2006). *On the role of feature selection in machine learning*. PhD thesis, Citeseer.

SCHLKOPF, B. (2018). *LEARNING WITH KERNELS : support vector machines, regularization, optimization, and beyond*. MIT Press, Place of publication not identified.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Vert, J.-P. (2020). Aronszajn's theorem.

Vocaturo, E., Perna, D., and Zumpano, E. (2019). Machine learning techniques for automated melanoma detection. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2310–2317.

Zaremba, S. (1907). L'équation biharmonique et une class remarquable de functions fondamentales harmoniques. *Bulletin International de l'Acadmie des Sciences de Cracovie*, pages 147–196.