# Aprendizado de máquinas

Thiago Rodrigo Ramos

1 de abril de 2025

# Sumário

## Material do curso

Todo o material utilizado neste curso, incluindo códigos e notebooks, pode ser acessado no repositório do GitHub: https://github.com/thiagorr162/curso_aprendizado.

## Referências principais

O conteúdo deste curso é baseado em referências que cobrem tópicos fundamentais de aprendizado de máquina e estatística. Izbicki and dos Santos (2020) introduz o aprendizado de máquina com ênfase em uma abordagem estatística, voltada ao público brasileiro. James et al. (2013) apresentam métodos estatísticos aplicados à aprendizagem supervisionada e não supervisionada, com exemplos em *R* e *Python*. Hastie et al. (2001) abordam técnicas avançadas e a teoria estatística por trás de algoritmos de aprendizado de máquina. Shalev-Shwartz and Ben-David (2014) desenvolvem a teoria do aprendizado e a análise de algoritmos, com foco na compreensão matemática das técnicas. Mohri et al. (2018) tratam de conceitos fundamentais de generalização e estabilidade, além de fornecer uma base teórica para diversos algoritmos modernos.

Todas esses livros podem ser baixadas online de forma legal nos sites dos autores.

# Capítulo 1

# Introdução

*Aprendizado de máquina* é um termo utilizado para descrever sistemas capazes de identificar automaticamente padrões e regularidades em dados (Shalev-Shwartz and Ben-David, 2014). Nos últimos anos, essa área consolidou-se como uma ferramenta indispensável para atividades que envolvem a análise e interpretação de grandes volumes de informação. Hoje em dia, essa tecnologia está presente em nosso cotidiano: motores de busca ajustam seus resultados para atender melhor às nossas consultas (ao mesmo tempo em que exibem anúncios), filtros de *spam* são aperfeiçoados para proteger nossas caixas de e-mail, e sistemas de detecção de fraudes asseguram a integridade de transações financeiras realizadas com cartões de crédito. Além disso, câmeras digitais reconhecem rostos, assistentes virtuais em *smartphones* interpretam comandos de voz e veículos utilizam algoritmos inteligentes para prevenir acidentes. O aprendizado de máquina também desempenha papel crucial em diversas áreas da ciência, como a bioinformática, a medicina e a astronomia.

## 1.1 Um breve histórico do aprendizado estatístico

Como descrito em James et al. (2013), embora o termo *aprendizado estatístico* seja relativamente recente, muitos dos conceitos fundamentais da área foram estabelecidos há bastante tempo. No início do século XIX, surgiu o método dos mínimos quadrados, que representa uma das primeiras formas do que hoje conhecemos como regressão linear. Essa técnica foi aplicada com sucesso, inicialmente, em problemas de astronomia. A regressão linear é amplamente utilizada para prever variáveis quantitativas, como o salário de um indivíduo, por exemplo.

Com o objetivo de prever variáveis qualitativas — como determinar se um paciente sobreviverá ou não, ou se o mercado financeiro terá alta ou queda —, foi proposta em 1936 a análise discriminante linear. Já na década de 1940, autores sugeriram uma abordagem alternativa: a regressão logística. No início dos anos 1970, o conceito de *modelos lineares generalizados* foi introduzido, englobando tanto a regressão linear quanto a logística como casos particulares dentro de uma estrutura mais ampla.

Até o final da década de 1970, diversas técnicas para aprendizado a partir de dados já estavam disponíveis, embora fossem predominantemente lineares, devido às limitações computacionais da época para modelagem de relações não lineares. A partir dos anos 1980, com o avanço da

tecnologia, métodos não lineares passaram a ser mais acessíveis. Nesse período surgiram as árvores de decisão para classificação e regressão, seguidas pelos modelos aditivos generalizados. Ainda nos anos 1980, as redes neurais ganharam destaque, e nos anos 1990, as máquinas de vetor de suporte (*support vector machines*) foram introduzidas.

Desde então, o aprendizado estatístico consolidou-se como um subcampo da estatística dedicado à modelagem e predição em cenários supervisionados e não supervisionados. Nos últimos anos, o progresso na área foi impulsionado pela crescente disponibilidade de softwares poderosos e acessíveis, como a linguagem de programação Python, que é gratuito e de código aberto. Esse avanço vem contribuindo para ampliar o alcance das técnicas de aprendizado estatístico, tornando-as uma ferramenta essencial não apenas para estatísticos e cientistas da computação, mas também para profissionais de diversas outras áreas.

### 1.1.1 Inferência vs Predição

Como descrito em Izbicki and dos Santos (2020), em problemas supervisionados, é importante distinguir entre dois objetivos fundamentais: a inferência e a predição. Essas duas abordagens guiam a forma como modelos são construídos e avaliados.

- **Objetivo inferencial** diz respeito à compreensão da relação entre as covariáveis $\mathbf{x}$ e a variável resposta $Y$. Nesse caso, queremos responder perguntas como: quais covariáveis são mais relevantes para explicar $Y$? Qual a direção e a magnitude do efeito de cada preditor? Esse tipo de análise é útil quando o interesse está em interpretar o modelo, entender a estrutura dos dados ou formular hipóteses científicas.

- **Objetivo preditivo**, por outro lado, está focado em construir uma função $g : \mathbb{R}^d \to \mathbb{R}$ que tenha boa capacidade de prever $Y$ para novas observações não vistas durante o treinamento. O sucesso neste contexto é medido pela capacidade do modelo em generalizar para dados futuros, mesmo que isso ocorra às custas de uma menor interpretabilidade do modelo.

Para ilustrar essas distinções, vejamos dois exemplos práticos:

- No **Exemplo 1.3** (*Isomap face data*), cada observação consiste em uma imagem de um rosto humano, e o objetivo é prever a direção para a qual a pessoa está olhando (variável $y$) com base nos pixels da imagem (variáveis $\mathbf{x}$). Esse é um exemplo puramente preditivo, pois a principal meta é estimar corretamente a direção do olhar em novas imagens. O modelo não busca explicar quais regiões da imagem são mais relevantes ou como cada pixel individual influencia a resposta, mas sim gerar boas predições.

- Já no **Exemplo 1.4** (*Million Song Dataset*), o banco de dados contém informações sobre diversas características de músicas (como timbre, energia, dançabilidade etc.) e o ano de lançamento de cada uma delas. Nesse caso, o problema tem um caráter misto. Por um lado, queremos prever o ano de lançamento a partir das covariáveis disponíveis (objetivo preditivo). Por outro, pode haver interesse em entender como cada característica da música se relaciona com o ano de lançamento, como por exemplo investigar se músicas dos anos 70 são de fato mais "dançantes"do que as atuais (objetivo inferencial).

Portanto, enquanto alguns problemas são essencialmente preditivos ou inferenciais, outros envolvem uma combinação dos dois. Essa distinção é relevante, pois impacta tanto a escolha do modelo quanto a forma de interpretá-lo e validá-lo.

### 1.1.2 As duas culturas de Breiman

Leo Breiman foi um estatístico renomado, conhecido por suas contribuições fundamentais à estatística e ao aprendizado de máquina, incluindo o desenvolvimento de métodos como random forests Breiman (2001a). Em seu influente artigo *"Statistical Modeling: The Two Cultures"* (Breiman, 2001b), Breiman discute duas abordagens distintas para modelagem estatística. Ele inicia seu artigo com o seguinte resumo:

> *"There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."*

O artigo eve grande impacto na comunidade estatística e no campo do aprendizado de máquina. Nele, Breiman argumenta que existem duas culturas distintas na modelagem de dados: a **cultura dos modelos estocásticos**, que assume que os dados são gerados por um modelo probabilístico especificado (como regressão linear, modelos lineares generalizados etc.), e a **cultura algorítmica**, que foca na construção de algoritmos preditivos eficazes sem necessariamente se preocupar com a interpretação ou com a modelagem explícita da distribuição dos dados (como árvores de decisão, random forests, redes neurais, entre outros).

Breiman defende que a estatística tradicional estava excessivamente focada em modelos estocásticos, o que limitava seu impacto em problemas práticos, enquanto métodos algorítmicos — amplamente utilizados fora da estatística, especialmente na ciência da computação — estavam mais bem adaptados para resolver problemas com grandes volumes de dados e alta complexidade.

Hoje em dia, o artigo de Breiman é amplamente citado e considerado um marco que antecipou a ascensão de métodos de aprendizado de máquina dentro da estatística e da ciência de dados. No entanto, sua visão também recebeu críticas. Alguns argumentam que as duas culturas não são mutuamente excludentes e que há um valor significativo na modelagem estatística clássica, especialmente quando a interpretação dos parâmetros e a inferência causal são importantes. Além disso, com o avanço dos métodos de aprendizado estatístico e da estatística bayesiana, muitos pesquisadores propõem abordagens híbridas que combinam modelagem interpretável com o poder preditivo dos algoritmos.

Atualmente, o artigo de Breiman é visto como uma provocação importante que incentivou a comunidade a repensar o papel da estatística em problemas do mundo real, mas também é reconhecido que tanto a modelagem estocástica quanto a preditiva têm seu espaço e relevância dependendo do contexto e dos objetivos da análise.

## 1.2 Algumas tarefas clássicas de aprendizado

A seguir, apresentamos algumas tarefas clássicas de aprendizado de máquina que têm sido amplamente estudadas (Mohri et al., 2018):

- **Classificação**: consiste em atribuir uma categoria a cada item. Por exemplo, na classificação de documentos, o objetivo é rotular cada texto com categorias como política, negócios, esportes ou clima. Já na classificação de imagens, cada imagem pode ser categorizada como carro, trem ou avião. Em geral, o número de categorias é limitado a algumas centenas, mas pode ser consideravelmente maior em tarefas complexas, como reconhecimento óptico de caracteres (OCR), classificação de textos ou reconhecimento de fala.

- **Regressão**: envolve a predição de um valor numérico contínuo para cada item. Exemplos comuns incluem a previsão de preços de ações ou de indicadores econômicos. Diferentemente da classificação, em regressão o erro de uma predição depende da distância entre o valor real e o valor estimado, enquanto na classificação normalmente não há uma medida de proximidade entre as categorias.

- **Ranqueamento**: trata-se de aprender a ordenar itens de acordo com algum critério. Um exemplo típico é o ranqueamento de páginas em um motor de busca, onde o sistema precisa retornar os resultados mais relevantes para uma consulta. Outras aplicações de ranqueamento aparecem em sistemas de extração de informações e em processamento de linguagem natural.

- **Agrupamento (Clustering)**: busca organizar um conjunto de itens em subconjuntos homogêneos. Algoritmos de agrupamento são especialmente úteis na análise de grandes volumes de dados. Na análise de redes sociais, por exemplo, técnicas de clustering são usadas para identificar comunidades ou grupos com características similares dentro de uma rede.

- **Redução de dimensionalidade ou aprendizado de variedades**: refere-se ao processo de transformar uma representação original de dados em uma representação de menor dimensão, preservando certas propriedades estruturais importantes. Um exemplo comum ocorre no pré-processamento de imagens digitais em tarefas de visão computacional.

## 1.3 Exemplos

### 1.3.1 Salários

Nesta análise, utilizamos um conjunto de dados que contém informações sobre salários de trabalhadores da região do Atlântico dos Estados Unidos (Fig. 1.3.1). O foco é explorar como fatores

como idade, nível de escolaridade e o ano em que o salário foi registrado influenciam os valores salariais.

| | year | age | maritl | race | education | region | jobclass | health | health_ins | logwage | wage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2006 | 18 | 1. Never Married | 1. White | 1. < HS Grad | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 2. No | 4.318063 | 75.043154 |
| 1 | 2004 | 24 | 1. Never Married | 1. White | 4. College Grad | 2. Middle Atlantic | 2. Information | 2. >=Very Good | 2. No | 4.255273 | 70.476020 |
| 2 | 2003 | 45 | 2. Married | 1. White | 3. Some College | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 1. Yes | 4.875061 | 130.982177 |
| 3 | 2003 | 43 | 2. Married | 3. Asian | 4. College Grad | 2. Middle Atlantic | 2. Information | 2. >=Very Good | 1. Yes | 5.041393 | 154.685293 |
| 4 | 2005 | 50 | 4. Divorced | 1. White | 2. HS Grad | 2. Middle Atlantic | 2. Information | 1. <=Good | 1. Yes | 4.318063 | 75.043154 |

Figura 1.1: Exemplo de registros do conjunto de dados de salários.

**Exercício 1.** *Utilizando o código nesse link. Faça uma análise do comportamento entre as variáveis de idade e salário. Faça o mesmo para nível de escolaridade e salário.*

### 1.3.2 Mercado de ações

Enquanto o conjunto de dados de salários aborda a previsão de uma variável numérica contínua, neste exemplo o objetivo é prever um resultado qualitativo. Trata-se de um problema clássico de classificação, em que desejamos prever categorias ao invés de valores numéricos. Um exemplo

| | Year | Lag1 | Lag2 | Lag3 | Lag4 | Lag5 | Volume | Today | Direction |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2001 | 0.381 | -0.192 | -2.624 | -1.055 | 5.010 | 1.1913 | 0.959 | Up |
| 1 | 2001 | 0.959 | 0.381 | -0.192 | -2.624 | -1.055 | 1.2965 | 1.032 | Up |
| 2 | 2001 | 1.032 | 0.959 | 0.381 | -0.192 | -2.624 | 1.4112 | -0.623 | Down |
| 3 | 2001 | -0.623 | 1.032 | 0.959 | 0.381 | -0.192 | 1.2760 | 0.614 | Up |
| 4 | 2001 | 0.614 | -0.623 | 1.032 | 0.959 | 0.381 | 1.2057 | 0.213 | Up |

Figura 1.2: Exemplo de registros do conjunto de dados de ações.

interessante envolve dados do mercado financeiro (Fig. 1.2), que incluem as variações diárias do índice S&P 500 ao longo de um período de cinco anos, entre 2001 e 2005. Esse conjunto de dados, que chamaremos de *Smarket*, busca prever a direção do mercado em um determinado dia (se irá subir ou cair), utilizando como variáveis explicativas as mudanças percentuais dos cinco dias anteriores.

Diferente da tarefa de regressão, aqui o desafio consiste em classificar o movimento do mercado como sendo uma alta (*Up*) ou uma baixa (*Down*). Embora o comportamento passado do índice possa não fornecer uma regra clara para prever o movimento do dia seguinte, pequenas tendências ou padrões podem ser identificados com métodos de aprendizado estatístico.

**Exercício 2.** *Explorar os dados do mercado de ações utilizando esse código.*

# Capítulo 2

# Aprendizado supervisionado

O aprendizado supervisionado é uma das principais áreas do aprendizado de máquina e da estatística, tendo como objetivo construir modelos capazes de prever uma variável de interesse $Y$ a partir de um conjunto de variáveis explicativas $X$. Este paradigma baseia-se em dados rotulados, ou seja, em observações para as quais tanto as covariáveis quanto a variável de resposta são conhecidas.

Ao longo deste capítulo, assumiremos que dispomos de uma amostra de dados $(X_i, Y_i)_{i=1}^n$, em que cada par $(X_i, Y_i)$ é uma realização de um mesmo par de variáveis aleatórias $(X, Y)$. Além disso, adotaremos a hipótese de que estas observações são *i.i.d.* (independentes e identicamente distribuídas). Esta suposição simplifica a análise teórica, permitindo o uso de ferramentas como leis dos grandes números e teoremas de concentração. Na prática, embora a hipótese de *i.i.d.* nem sempre seja completamente satisfeita, ela é uma aproximação útil e bastante comum em aplicações reais.

Nosso objetivo será utilizar esse conjunto de dados (ou uma parte dele) para aprender um modelo preditivo, que denotaremos por $\hat{g}$, de forma que $\hat{g}(X) \approx Y$. O significado da relação de aproximação "$\approx$"será discutido a seguir.

Ao construir um modelo preditivo $\hat{g}(X)$ para estimar uma variável de interesse $Y$, é essencial definir uma métrica que quantifique o erro cometido pelas previsões. Essa métrica é chamada de **função de perda**, e mede a discrepância entre o valor verdadeiro $Y$ e a predição $\hat{g}(X)$. Duas escolhas comuns para problemas com resposta contínua (regressão) são:

$$L(Y, \hat{g}(X)) = \begin{cases} (Y - \hat{g}(X))^2 & \text{(erro quadrático)} \\ |Y - \hat{g}(X)| & \text{(erro absoluto).} \end{cases}$$

A escolha da função de perda impacta diretamente as propriedades do modelo e como ele responde a diferentes tipos de dados ou outliers.

*Erro de teste*, também chamado de *erro de generalização*, é o erro de predição em uma amostra de teste independente:

$$\text{Err}_{\mathcal{D}} = \mathbb{E}\left[L(Y, \hat{g}(X)) \mid \mathcal{D}\right]$$

onde $X$ e $Y$ são sorteados de sua distribuição conjunta (população). Note que $\hat{g}$ **depende de** $\mathcal{D}$**!** Aqui, o conjunto de treinamento $\mathcal{D}$ é fixo, e o erro de teste se refere ao erro para esse conjunto

de treinamento específico. Uma quantidade relacionada é o *erro esperado de predição* (ou erro esperado de teste):

$$\text{Err} = \mathbb{E}\left[L(Y, \hat{g}(X))\right] = \mathbb{E}\left[\text{Err}_{\mathcal{D}}\right].$$

Note que a esperança acima leva em conta toda a aleatoriedade envolvida, incluindo a aleatoriedade do conjunto de treinamento que gerou $\hat{g}$.

Nosso objetivo será a estimação de $\text{Err}_{\mathcal{D}}$, embora veremos que Err é mais acessível do ponto de vista estatístico, e a maioria dos métodos busca estimar efetivamente esse erro esperado. Estimar $\text{Err}_{\mathcal{D}}$ de maneira condicional não é muito viável na prática.

*Erro de treinamento* é a perda média sobre a amostra de treinamento:

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{g}(x_i)).$$

Nosso interesse está em conhecer o erro de teste esperado do modelo $\hat{g}$. À medida que o modelo se torna mais complexo, ele se ajusta melhor aos dados de treinamento e passa a capturar estruturas subjacentes mais complicadas. Com isso, ocorre uma redução no viés, mas um aumento na variância. Existe, portanto, um nível intermediário de complexidade do modelo que minimiza o erro esperado de teste.

**Infelizmente, o erro de treinamento não é uma boa estimativa do erro de teste.** O erro de treinamento tipicamente diminui à medida que a complexidade do modelo aumenta, podendo até atingir zero quando essa complexidade é suficientemente alta. Entretanto, um modelo com erro de treinamento igual a zero está *superajustado* (overfit) aos dados de treinamento e, geralmente, apresentará baixa capacidade de generalização.

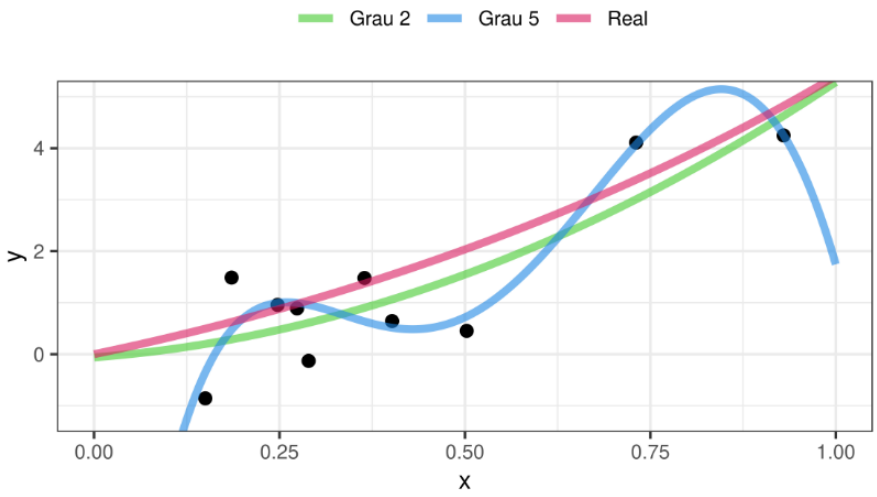



Figura 2.1: Exemplo de sobreajuste. Imagem retirada de (Izbicki and dos Santos, 2020).

Para simplificar a exposição, neste capítulo utilizaremos $Y$ e $\hat{g}(X)$ para representar todos os cenários mencionados, focando no caso de resposta quantitativa com perda quadrática. Em outros contextos, as adaptações necessárias serão diretas.

Ao longo deste capítulo, descreveremos diversos métodos para estimar o erro de teste esperado de um modelo. Tipicamente, o modelo dependerá de um parâmetro ou de um conjunto de parâmetros de ajuste $\alpha$, de modo que podemos escrever a predição como $\hat{g}_\alpha(x)$. O parâmetro $\alpha$ controla a complexidade do modelo, e nosso objetivo é encontrar o valor de $\alpha$ que minimize o erro esperado de teste, ou seja, que produza o menor erro médio. Para simplificar a notação, omitiremos frequentemente a dependência explícita de $\hat{g}(x)$ em $\alpha$.

É importante destacar que temos, na verdade, dois objetivos distintos ao avaliar modelos preditivos:

- **Seleção de modelo:** consiste em comparar diferentes modelos ou configurações a fim de escolher o que apresenta o melhor desempenho.
- **Avaliação de modelo:** uma vez escolhido o modelo final, o objetivo passa a ser estimar o seu erro de predição (ou erro de generalização) em novos dados.

## 2.1 Data Splitting e Validação Cruzada

Em situações onde há abundância de dados, uma abordagem comum é dividir aleatoriamente o conjunto em três partes: um conjunto de treinamento, um conjunto de validação e um conjunto de teste. O conjunto de treinamento é utilizado para ajustar os modelos; o conjunto de validação serve para estimar o erro de predição e realizar a seleção do modelo; e o conjunto de teste é reservado para avaliar o erro de generalização final do modelo escolhido.

> **IMPORTANTE!** Idealmente, o conjunto de teste deve ser mantido isolado — como se estivesse em um "cofre"— e só ser acessado ao final da análise de dados. Caso utilizemos o conjunto de teste de forma repetida durante a seleção do modelo, escolhendo aquele com menor erro no teste, acabaremos subestimando o verdadeiro erro de generalização, às vezes de maneira significativa.

Não há uma regra única para definir a quantidade de observações em cada uma das três divisões, pois isso depende da razão sinal-ruído dos dados e do tamanho da amostra disponível. Uma divisão típica é utilizar cerca de 50% dos dados para treinamento e 25% para validação e teste, respectivamente.


Figura 2.2: Divisão em treino, validação e teste. Tirado de (Hastie et al., 2001).

No `Python`, podemos realizar essa divisão utilizando a biblioteca `scikit-learn`. O procedimento padrão é, primeiramente, dividir os dados em duas partes: treino e teste. Em seguida, subdividir a parte de treino em treino e validação.

Listing 2.1: Divisão em treino, validação e teste

```
from sklearn.model_selection import train_test_split

# Suponha que temos dados X e respostas Y
# X: matriz de covariaveis (n_samples x n_features)
# Y: vetor de respostas (n_samples,)


# Primeira divisao: treino e restante (validacao + teste)
X_train, X_rest, Y_train, Y_rest = train_test_split(
    X, Y, test_size=0.5, random_state=42)

# Segunda divisao: validacao e teste a partir do restante
X_val, X_test, Y_val, Y_test = train_test_split(
    X_rest, Y_rest, test_size=0.5, random_state=42)

print("Treino:", X_train.shape)
print("Validacao:", X_val.shape)
print("Teste:", X_test.shape)
```

No exemplo acima, separamos 25% dos dados para o conjunto de teste e, dos 75% restantes, cerca de 33% foi alocado para validação. Assim, o resultado final aproximado seria: 50% dos dados para treino, 25% para validação e 25% para teste, como discutido anteriormente.

> **ATENÇÃO:** É importante definir o argumento `random_state` para garantir reprodutibilidade da divisão dos dados.

Embora a divisão treino/validação/teste seja bastante comum, em situações onde a quantidade de dados é limitada, desperdiçar uma parte significativa da amostra apenas para validação pode ser custoso. Nesse contexto, uma estratégia amplamente utilizada é a **validação cruzada**, em especial o *k-fold cross-validation*.

A ideia do *k-fold* é dividir o conjunto de dados em $k$ subconjuntos (ou *folds*) de tamanhos aproximadamente iguais. Em cada uma das $k$ iterações, utilizamos $k-1$ desses subconjuntos para treinar o modelo e o subconjunto restante para validá-lo. No final, o erro de validação é calculado como a média dos erros obtidos em cada uma das iterações.

Essa técnica tem como vantagem utilizar o máximo de dados possível para treinamento em cada repetição, reduzindo a variância da estimativa do erro de generalização. Além disso, o *k-fold* ajuda a mitigar a dependência da divisão aleatória dos dados, já que cada observação é utilizada tanto para treino quanto para validação ao longo do processo.

No `Python`, a implementação do *k-fold cross-validation* pode ser feita utilizando a biblioteca `scikit-learn` da seguinte maneira:

Listing 2.2: Validação cruzada k-fold

```
from sklearn.model_selection import KFold

```

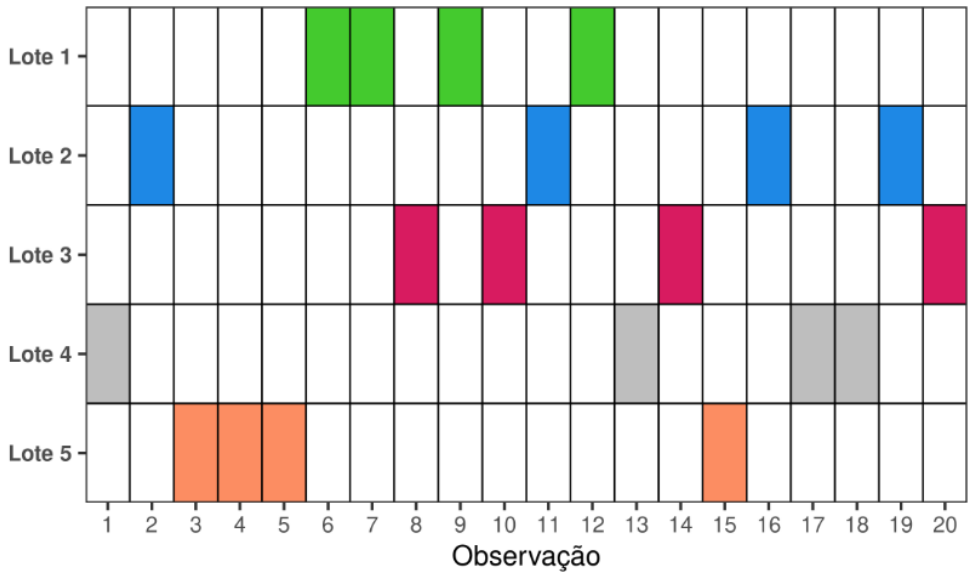



Figura 2.3: Esquema com amostra de tamanho 20 e 5 folds. Retirado de Izbicki and dos Santos (2020).

```
# Suponha que temos dados X e respostas Y
k = 5
kf = KFold(n_splits=k, shuffle=True, random_state=42)

for fold, (train_index, val_index) in enumerate(kf.split(X)):
    X_train, X_val = X[train_index], X[val_index]
    Y_train, Y_val = Y[train_index], Y[val_index]
    print(f"Fold {fold+1}:")
    print("Treino:", X_train.shape, "Validacao:", X_val.shape)
```

No exemplo acima, o conjunto de dados é dividido em $k = 5$ subconjuntos. O argumento `shuffle=True` garante que as observações sejam embaralhadas antes da divisão em folds, e o `random_state` garante a reprodutibilidade dos resultados.

**Exercício 3.** *Acesse o notebook aqui e faça experimentos com o código mudando o tamanho da amostra, random state, etc.*

## 2.2 Problemas de regressão

As origens dos métodos de regressão remontam a mais de 200 anos, com as contribuições de Legendre (1805) e Gauss (1809), que introduziram o método dos mínimos quadrados para modelar o movimento dos planetas ao redor do Sol. Atualmente, a estimação de funções de regressão é um dos pilares fundamentais da estatística.

Embora as primeiras soluções para este problema sejam antigas, apenas nas últimas décadas, com o avanço das tecnologias de computação e armazenamento, novas abordagens puderam ser

exploradas. Em especial, o crescimento exponencial da quantidade de dados disponíveis tem impulsionado o desenvolvimento de métodos que fazem menos suposições sobre o comportamento real dos fenômenos estudados.

Esse cenário trouxe novos desafios: por exemplo, métodos clássicos muitas vezes não conseguem lidar adequadamente com bancos de dados em que o número de variáveis excede o número de observações, uma situação comum nos contextos atuais. Além disso, aplicações envolvendo dados complexos — como imagens ou textos — têm se tornado frequentes e demandam técnicas mais sofisticadas.

De modo geral, o objetivo de um modelo de regressão é capturar a relação entre uma variável aleatória de interesse $Y \in \mathbb{R}$ e um vetor de covariáveis $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$. O foco está em estimar a chamada função de regressão, definida por

$$r(\mathbf{x}) := \mathbb{E}[Y \mid X = \mathbf{x}].$$

A motivação para estudar essa função está relacionada ao problema de minimizar o erro quadrático. Para ilustrar, considere uma variável aleatória $Z$ e a função objetivo

$$\phi(t) = \mathbb{E}[(Z - t)^2],$$

em que buscamos o valor $t \in \mathbb{R}$ que minimiza $\phi(t)$. Derivando em relação a $t$ e igualando a zero, obtemos:

$$\phi'(t) = \mathbb{E}[2(Z - t)] = 0 \quad \Leftrightarrow \quad t = \mathbb{E}[Z].$$

Portanto, o valor ótimo de $t$ que minimiza o erro quadrático é justamente a esperança de $Z$.

Esse raciocínio se estende naturalmente ao contexto de regressão. Nosso objetivo passa a ser encontrar uma função $g(\mathbf{x})$ que minimize

$$\phi(g) = \mathbb{E}[(Y - g(X))^2] = \mathbb{E}\left[\mathbb{E}\left[(Y - g(\mathbf{x}))^2 \mid X = \mathbf{x}\right]\right].$$

Fixando $X = \mathbf{x}$, $g(\mathbf{x})$ se comporta como um número, e, pelo argumento anterior, a minimização local de $\mathbb{E}[(Y - g(\mathbf{x}))^2 \mid X = \mathbf{x}]$ ocorre quando $g(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}]$.

Assim, a função de regressão $r(\mathbf{x})$ é, sob a métrica de erro quadrático, a melhor escolha para aproximar $Y$ em função de $\mathbf{x}$.

**Exercício 4.** *Seja*

$$\varphi(t) = \mathbb{E}[|Z - t|].$$

*Encontre o $t$ que minimiza a expressão acima.*

### 2.2.1 Decomposição de viés-variância

Suponha que o modelo verdadeiro seja dado por $Y = g^*(X) + \varepsilon$, onde $\varepsilon$ é um ruído com $\mathbb{E}[\varepsilon] = 0$, independente das covariáveis $X$. Nosso objetivo é estudar o erro quadrático esperado de um modelo preditivo $\hat{g}(x)$ que tenta estimar $y$ a partir de $x$.

O erro de predição é medido pela perda quadrática $(y - \hat{g}(x))^2$, e buscamos decompor o valor esperado desse erro em três componentes: variância, viés ao quadrado e ruído irredutível, considerando $x$ fixado.

A função ótima $g^*$ é definida como:

$$g^*(x) = \arg\min_g \mathbb{E}\left[(y - g(x))^2\right].$$

Para a decomposição, começamos adicionando e subtraindo $g^*(x)$ no termo de erro esperado:

$$\mathbb{E}[(y - \hat{g}(x))^2] = \mathbb{E}[(y - g^*(x) + g^*(x) - \hat{g}(x))^2].$$

Expandindo o quadrado, obtemos:

$$\mathbb{E}[(y - g^*(x))^2] + \mathbb{E}[(g^*(x) - \hat{g}(x))^2] + 2\mathbb{E}[(y - g^*(x))(g^*(x) - \hat{g}(x))].$$

O primeiro termo, $\mathbb{E}[(y - g^*(x))^2]$, representa a variância do ruído $\varepsilon$ e, portanto, é irredutível. O segundo termo, $\mathbb{E}[(g^*(x) - \hat{g}(x))^2]$, captura o erro introduzido pelo modelo $\hat{g}(x)$. O terceiro termo é nulo, pois $y - g^*(x) = \varepsilon$ tem média zero e é independente de $\hat{g}(x)$.

Portanto, temos a seguinte decomposição:

$$\mathbb{E}[(y - \hat{g}(x))^2] = \underbrace{\mathbb{E}[(y - g^*(x))^2]}_{\text{ruído irreduzível}} + \mathbb{E}[(g^*(x) - \hat{g}(x))^2].$$

Agora, o termo $\mathbb{E}[(g^*(x) - \hat{g}(x))^2]$ pode ser decomposto em viés e variância:

$$\begin{aligned}\mathbb{E}[(g^*(x) - \hat{g}(x))^2] &= \mathbb{E}[(g^*(x) - \mathbb{E}[\hat{g}(x)] + \mathbb{E}[\hat{g}(x)] - \hat{g}(x))^2] \\ &= (g^*(x) - \mathbb{E}[\hat{g}(x)])^2 + \mathbb{V}ar(\hat{g}(x)),\end{aligned}$$

onde o termo cruzado é novamente nulo por independência.

Finalmente, temos a seguinte decomposição clássica:

$$\mathbb{E}[(y - \hat{g}(x))^2] = \underbrace{\mathbb{E}[(y - g^*(x))^2]}_{\text{ruído irredutível}} + \underbrace{(g^*(x) - \mathbb{E}[\hat{g}(x)])^2}_{\text{viés}^2} + \underbrace{\mathbb{V}ar(\hat{g}(x))}_{\text{variância}}.$$

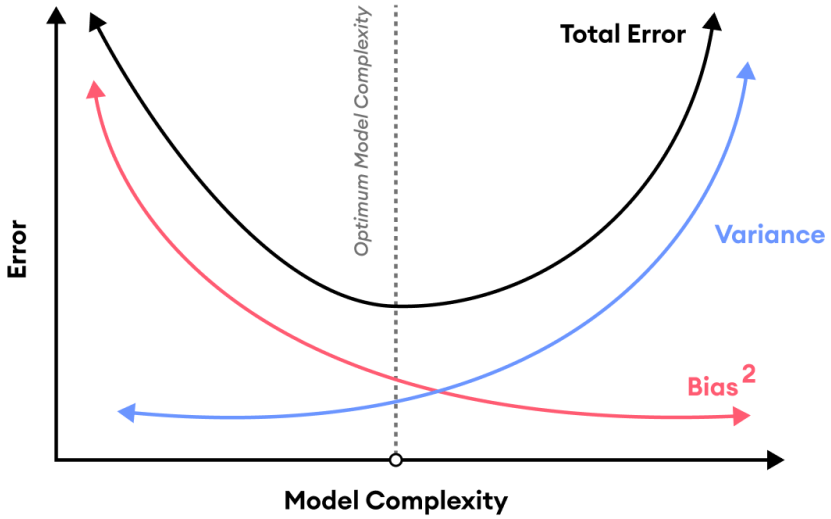
Essa decomposição nos permite entender um dos conceitos fundamentais em aprendizado de máquina e estatística: o **trade-off viés-variância**. Em termos simples, existe uma relação inversa entre o viés e a variância de um modelo. Modelos mais simples tendem a ter baixo variância, pois suas previsões mudam pouco ao variar a amostra de treinamento, mas frequentemente apresentam alto viés por não capturarem toda a complexidade da função $g^*(x)$. Por outro lado, modelos mais complexos conseguem ajustar melhor os dados e reduzir o viés, mas tendem a ter alta variância, pois são mais sensíveis a flutuações nas amostras de treino. O desafio central em modelagem preditiva é encontrar um equilíbrio entre essas duas quantidades, de modo a minimizar o erro total. Esse equilíbrio é essencial para garantir que o modelo generalize bem para novos dados, sem ser excessivamente simples (subajuste) ou excessivamente complexo (sobreajuste).

**Resumindo:** no caso da perda quadrática, o erro total pode ser decomposto como:

$$\text{Erro Total} = \textbf{Erro Irredutível} + \textbf{Viés}^2 + \textbf{Variância}.$$

**Exercício 5.** *Acesse o notebook aqui e faça experimentos com o código mudando o tamanho da amostra, random state, grau dos polinômios, etc. Além disso, aplique o método de validação cruzada k-fold.*

## 2.3 Problemas de classificação


Total Error
Optimum Model Complexity
Variance
Bias 2
Error
Model Complexity

# Capítulo 3

# Regressão

## 3.1 Mínimos quadrados e regressão linear

O *método dos mínimos quadrados* tem suas origens no início do século XIX e está intimamente ligado à história da astronomia e da estatística. Ele foi introduzido formalmente por Carl Friedrich Gauss, que o utilizava desde 1795 em seus trabalhos com órbitas planetárias, embora o primeiro a publicar sobre o método tenha sido Adrien-Marie Legendre, em 1805.

Legendre apresentou o método em seu trabalho sobre o cálculo de órbitas cometárias, propondo uma técnica para ajustar curvas a dados experimentais minimizando a soma dos quadrados dos erros. Poucos anos depois, em 1809, Gauss publicou seu famoso livro *Theoria Motus Corporum Coelestium*, onde apresentou uma justificativa probabilística do método com base na distribuição normal dos erros.

Desde então, os mínimos quadrados tornaram-se uma das ferramentas fundamentais em estatística, ciência de dados e análise numérica, sendo aplicados em regressão linear, ajuste de modelos, filtragem de sinais e muitos outros contextos científicos e tecnológicos.

Motivados por esses contextos, consideremos agora o seguinte problema: queremos encontrar um vetor $\beta$ tal que

$$y = X\beta.$$

Essa expressão possui solução se, e somente se, $y \in \operatorname{Ran} X$. Mas o que podemos fazer quando queremos resolver uma equação que não tem solução? Por exemplo, se $y = X\beta + \varepsilon$ onde $\varepsilon$ é um erro aleatório de medição?

À primeira vista, essa pode parecer uma pergunta boba, pois se não há solução, então não há solução. No entanto, situações em que desejamos resolver uma equação sem solução podem surgir naturalmente — por exemplo, quando a equação vem de dados experimentais. Se não houver nenhum erro, o vetor do lado direito $y$ pertence à imagem de $X$, e a equação é consistente.

Porém, na prática, é impossível eliminar completamente os erros de medição. Assim, pode acontecer de uma equação que teoricamente deveria ser consistente não possuir solução.

O que fazer, então, nessa situação?

A ideia mais simples é escrever o erro na forma

$$\|X\beta - y\|$$

e tentar encontrar o vetor $\beta$ que minimiza essa quantidade. Se conseguirmos encontrar um $\beta$ tal que o erro seja zero, então o sistema é consistente e temos uma solução exata. Caso contrário, obtemos a chamada *solução de mínimos quadrados*.

O nome *mínimos quadrados* vem do fato de que minimizar $\|X\beta - y\|$ é equivalente a minimizar

$$\|X\beta - y\|^2 = \sum_{k=1}^{m} |(X\beta)_k - y_k|^2 = \sum_{k=1}^{m} \left| \sum_{j=1}^{n} X_{k,j}\beta_j - y_k \right|^2,$$

isto é, estamos minimizando a soma dos quadrados de funções lineares.

Existem diversas maneiras de encontrar a solução de mínimos quadrados. Se estivermos em $\mathbb{R}^n$, e tudo for real, podemos ignorar os valores absolutos. Nesse caso, basta calcular as derivadas parciais em relação a cada $\beta_j$ e encontrar onde todas elas se anulam — o que nos dará o ponto de mínimo.

**Exercício 6.** *Encontre a solução do problema de mínimos quadrados derivando e igualando à zero.*

Existe uma forma mais simples de encontrar o mínimo. De fato, ao considerarmos todos os vetores $\beta$, o vetor $X\beta$ percorre todo o espaço imagem de $X$, ou seja, $\operatorname{Ran} X$. Portanto, minimizar $\|X\beta - y\|$ equivale a calcular a menor distância de $y$ até $\operatorname{Ran} X$.

Assim, $\|X\beta - y\|^2$ é mínima se, e somente se,

$$X\beta = P_{\operatorname{Ran} X} y,$$

onde $P_{\operatorname{Ran} X}$ denota a projeção ortogonal de $y$ sobre o subespaço imagem de $X$.

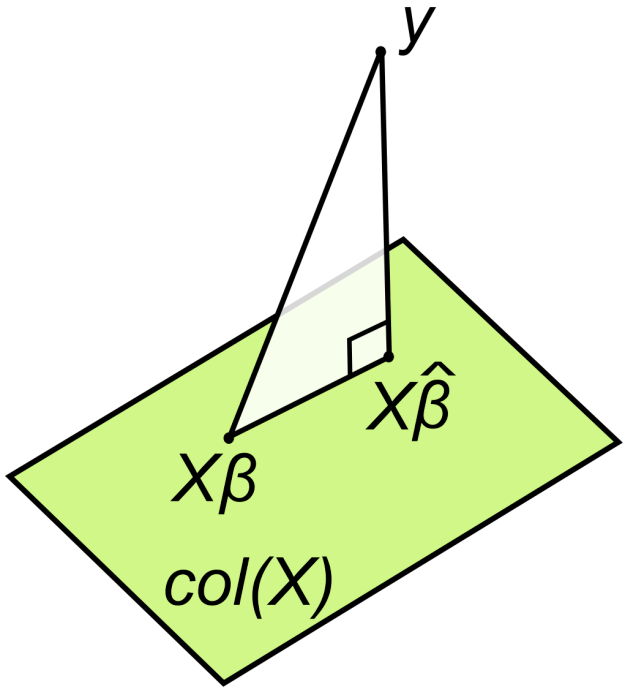



Figura 3.1: Projeção ortogonal de $y$ no espaço coluna de $X$.

Se conhecemos uma base ortogonal $\mathbf{v}_1, \ldots, \mathbf{v}_n$ de $\operatorname{Ran} X$, podemos calcular $P_{\operatorname{Ran} X} y$ pela fórmula:

$$P_{\operatorname{Ran} X} y = \sum_{k=1}^{n} \frac{\langle y, \mathbf{v}_k \rangle}{\|\mathbf{v}_k\|^2} \mathbf{v}_k.$$

Caso só tenhamos uma base qualquer de $\operatorname{Ran} X$, é necessário utilizar o processo de Gram–Schmidt para ortogonalizá-la antes de aplicar a fórmula.

Existe, no entanto, uma alternativa mais direta. A condição de que $X\beta = P_{\operatorname{Ran} X} y$ é equivalente a exigir que o vetor $y - X\beta$ seja ortogonal a $\operatorname{Ran} X$, ou seja,

$$y - X\beta \perp \text{coluna de } X.$$

Seja $X = [a_1, a_2, \ldots, a_n]$, onde cada $a_k$ é uma coluna. A condição acima é equivalente a:

$$\langle y - X\beta, a_k \rangle = 0, \quad \text{para } k = 1, 2, \ldots, n.$$

Ou, de forma matricial:

$$X^*(y - X\beta) = 0,$$

o que é equivalente à chamada *equação normal*:

$$X^* X\beta = X^* y.$$

A solução dessa equação nos fornece a solução de mínimos quadrados da equação $X\beta = y$. Note que a solução é única se, e somente se, $X^*X$ for inversível.

Agora, se $\beta$ é uma solução da *equação normal* $X^*X\beta = X^*y$ (ou seja, uma solução de mínimos quadrados da equação $X\beta = y$), então $X\beta = P_{\operatorname{Ran} X} y$. Assim, para encontrar a projeção ortogonal de $y$ sobre o espaço coluna $\operatorname{Ran} X$, basta resolver a equação normal $X^*X\beta = X^*y$ e multiplicar a solução por $X$.

Se o operador $X^*X$ for inversível, então a solução da equação normal é dada por:

$$\beta = (X^*X)^{-1} X^* y,$$

e, portanto, a projeção ortogonal $P_{\operatorname{Ran} X} y$ pode ser escrita como:

$$P_{\operatorname{Ran} X} y = X(X^*X)^{-1} X^* y.$$

Como isso vale para todo $y$, obtemos a expressão matricial da projeção ortogonal sobre o espaço coluna de $X$:

$$P_{\operatorname{Ran} X} = X(X^*X)^{-1} X^*.$$

**Observação 1.** *Note que a expressão*

$$P_{\operatorname{Ran} X} = X(X^*X)^{-1} X^*.$$

*é uma generalização matricial para a projeção ortogonal em vetores da forma*

$$\frac{vv^T}{v^T v} = v(v^T v)^{-1} v^T,$$

*já que para um vetor $x$ qualquer,*

$$\frac{\langle v, x \rangle}{\|v\|^2} v = \frac{v^T x}{\|v\|^2} v = \frac{vv^T x}{\|v\|^2} = \frac{vv^T}{\|v\|^2} x.$$

**Exemplo 1.** *Suponha que sabemos que a relação entre $x$ e $y$ é dada por uma parábola da forma*

$$y = a + bx + cx^2,$$

*e queremos ajustar essa parábola aos dados observados. Os coeficientes desconhecidos $a, b, c$ devem satisfazer o sistema:*

$$a + bx_k + cx_k^2 = y_k, \quad k = 1, 2, \ldots, n.$$

*Em forma matricial, esse sistema pode ser escrito como:*

$$\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

*Por exemplo, para os dados do exemplo anterior, devemos resolver a equação de mínimos quadrados:*

$$\begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

*Então calculamos:*

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 3 \\ 4 & 1 & 0 & 1 & 9 \end{pmatrix} \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 2 & 18 \\ 2 & 18 & 26 \\ 18 & 26 & 114 \end{pmatrix}.$$

*E também:*

$$X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 3 \\ 4 & 1 & 0 & 1 & 9 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 9 \\ -5 \\ 31 \end{pmatrix}.$$

*Portanto, a equação normal $X^T X\beta = X^T y$ é:*

$$\begin{pmatrix} 5 & 2 & 18 \\ 2 & 18 & 26 \\ 18 & 26 & 114 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 9 \\ -5 \\ 31 \end{pmatrix},$$

*cuja solução única é:*

$$a = \frac{86}{77}, \quad b = -\frac{62}{77}, \quad c = \frac{43}{154}.$$

*Portanto, a parábola que melhor se ajusta aos dados é:*

$$y = \frac{86}{77} - \frac{62}{77}x + \frac{43}{154}x^2.$$

### 3.1.1 Resolução Numérica

A forma fechada da solução dos mínimos quadrados, dada por

$$(X^\top X)\hat{\beta} = X^\top y \Rightarrow \hat{\beta} = (X^\top X)^{-1} X^\top y,$$

embora útil em termos analíticos, pode ser instável numericamente e custosa para grandes dimensões. Assim, é comum empregar abordagens numéricas mais robustas.

***Fatoração QR.*** Um método bastante utilizado é a decomposição $QR$, na qual a matriz $X \in \mathbb{R}^{n\times p}$ é escrita como $X = QR$, onde $Q$ possui colunas ortonormais (ou seja, $Q^\top Q = I$) e $R \in \mathbb{R}^{p\times p}$ é triangular superior. Essa fatoração evita a inversão direta da matriz $X^\top X$ e proporciona maior estabilidade numérica.

Como $X^\top X = R^\top Q^\top QR = R^\top R$, temos:

$$(X^\top X)\hat{\beta} = X^\top y \iff R^\top R\hat{\beta} = R^\top Q^\top y \iff R\hat{\beta} = Q^\top y.$$

Portanto, basta resolver um sistema linear triangular com matriz $R$, o que é computacionalmente eficiente. O custo total permanece na ordem de $\mathcal{O}(d^3)$. Em situações onde $d$ é grande, métodos iterativos como o gradiente conjugado também podem ser considerados (ver *Golub e Loan, 1996*).

### 3.1.2 Um pouco de inferência

Perceba que a teoria acima não depende de informações sobre os dados — a única coisa que fizemos foi encontrar a melhor aproximação de $y$ no espaço gerado pelas colunas de $X$. Ou seja, tratamos $X$ e $y$ como vetores fixos, e a projeção ortogonal $P_{\mathrm{Ran}\, X} y$ é puramente uma construção geométrica.

Suponha agora que temos um modelo probabilístico associado aos dados:

$$Y = \langle x, \beta\rangle + \varepsilon, \quad \varepsilon \sim N(0,1),$$

onde $x \in \mathbb{R}^p$ é um vetor fixo (não aleatório), $\beta \in \mathbb{R}^p$ é o vetor de parâmetros desconhecido, e $\varepsilon$ é um erro aleatório com distribuição normal padrão.

Se coletamos $n$ observações $(x_i, Y_i)$, $i = 1, \ldots, n$, podemos escrever o modelo vetorialmente como:

$$y = X\beta + \varepsilon,$$

onde:

- $y \in \mathbb{R}^n$ é o vetor de respostas;
- $X \in \mathbb{R}^{n\times p}$ é a matriz cujas linhas são os vetores $x_i^T$;
- $\varepsilon \sim N(0, I_n)$ é o vetor de erros independentes com variância 1.

Neste caso, a estimativa de mínimos quadrados:

$$\hat{\beta} = (X^*X)^{-1}X^*y$$

é também o *estimador de máxima verossimilhança* de $\beta$ sob o modelo gaussiano. Além disso, por ser combinação linear de variáveis gaussianas, $\hat{\beta}$ é também uma variável aleatória com distribuição normal.

**Teorema 1.** *Se $y = X\beta + \varepsilon$, com $\varepsilon \sim N(0, I_n)$, então:*

$$\hat{\beta} = (X^*X)^{-1}X^*y \sim N\left(\beta, (X^*X)^{-1}\right).$$

*Demonstração.* Note que:

$$\hat{\beta} = (X^*X)^{-1}X^*y = (X^*X)^{-1}X^*(X\beta + \varepsilon) = \beta + (X^*X)^{-1}X^*\varepsilon.$$

Como $\varepsilon \sim N(0, I_n)$, e $(X^*X)^{-1}X^*\varepsilon$ é uma combinação linear de gaussianas, então:

$$\hat{\beta} \sim N\left(\beta, (X^*X)^{-1}X^*X(X^*X)^{-1}\right) = N\left(\beta, (X^*X)^{-1}\right).$$

□

Essa propriedade permite que façamos *inferência estatística* sobre os coeficientes $\beta$. Por exemplo, para testar a hipótese nula:

$$H_0 : \beta_j = 0,$$

podemos usar o fato de que:

$$\hat{\beta}_j \sim N(\beta_j, \sigma_j^2), \quad \text{com } \sigma_j^2 = [(X^*X)^{-1}]_{jj}.$$

Ou seja, o valor padronizado

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\sigma_j^2}} \sim N(0,1) \quad \text{sob } H_0.$$

Isso nos permite construir intervalos de confiança e calcular valores-$p$.

**Exemplo 2.** *Suponha que queremos um intervalo de confiança para $\beta_j$ com nível de confiança $1-\alpha$. Como $\hat{\beta}_j \sim N(\beta_j, \sigma_j^2)$, temos:*

$$\mathbb{P}\left(\hat{\beta}_j - z_{\alpha/2}\sqrt{\sigma_j^2} \leq \beta_j \leq \hat{\beta}_j + z_{\alpha/2}\sqrt{\sigma_j^2}\right) = 1 - \alpha,$$

*onde $z_{\alpha/2}$ é o quantil superior de ordem $1-\alpha/2$ da normal padrão.*

Para mais detalhes sobre inferência de parâmetros, ver (James et al., 2013, Capítulo 3).

### 3.1.3 Viés e variância do estimador

Como vimos no primeiro capítulo, um conceito central na análise de métodos de aprendizado de máquina é compreender o comportamento de viés e variância do modelo.

Assumimos novamente que os dados seguem o modelo

$$y = \langle x, \beta \rangle + \varepsilon,$$

com $\mathbb{E}[\varepsilon] = 0$ e $\mathrm{Var}(\varepsilon) = \sigma^2$. Seja $\hat{\beta} = (X^\top X)^{-1} X^\top y$ a solução obtida por mínimos quadrados. Então:

$$\mathbb{E}[\hat{\beta}] = (X^\top X)^{-1} X^\top \mathbb{E}[y] = (X^\top X)^{-1} X^\top X\beta = \beta,$$

isto é, o estimador é não-viesado. Para a variância, temos:

$$\begin{aligned}
\hat{\beta} - \beta &= (X^\top X)^{-1} X^\top y - \beta \\
&= (X^\top X)^{-1} X^\top (X\beta + \varepsilon) - \beta \\
&= (X^\top X)^{-1} X^\top \varepsilon,
\end{aligned}$$

portanto, a variância de $\hat{\beta}$ é:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathbb{E}\left[(X^\top X)^{-1} X^\top \varepsilon\varepsilon^\top X (X^\top X)^{-1}\right] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}[\varepsilon\varepsilon^\top] X (X^\top X)^{-1} \\
&= \sigma^2 (X^\top X)^{-1}.
\end{aligned}$$

Podemos agora calcular o erro quadrático esperado associado ao uso de um vetor qualquer $\beta_0$. Temos:

$$\mathbb{E}_y\left[\|y - X\beta_0\|^2\right] = \sigma^2 + (\beta_0 - \beta)^\top X^\top X (\beta_0 - \beta).$$

Em particular, tomando $\beta_0 = \hat{\beta}$, obtemos:

$$\begin{aligned}
\mathbb{E}_y\left[\|y - X\hat{\beta}\|^2\right] &= \sigma^2 + \mathbb{E}\left[\left((X^\top X)^{-1} X^\top \varepsilon\right)^\top X^\top X \left((X^\top X)^{-1} X^\top \varepsilon\right)\right] \\
&= \sigma^2 + \mathbb{E}\left[\varepsilon^\top X (X^\top X)^{-1} X^\top X (X^\top X)^{-1} X^\top \varepsilon\right] \\
&= \sigma^2 + \mathbb{E}\left[\varepsilon^\top X (X^\top X)^{-1} X^\top \varepsilon\right] \\
&= \sigma^2 + \mathbb{E}\left[\mathrm{tr}\left(\varepsilon\varepsilon^\top X (X^\top X)^{-1} X^\top\right)\right] \\
&= \sigma^2 + \mathrm{tr}\left(\mathbb{E}[\varepsilon\varepsilon^\top] X (X^\top X)^{-1} X^\top\right) \\
&= \sigma^2 + \sigma^2\, \mathrm{tr}\left(X (X^\top X)^{-1} X^\top\right) \\
&= \sigma^2 + \sigma^2\, \mathrm{tr}(H),
\end{aligned}$$

onde $H = X(X^\top X)^{-1}X^\top$ é a matriz de projeção (ou "hat matrix"). Como $H$ é idempotente e simétrica, temos $\text{tr}(H) = p$, o número de parâmetros (ou colunas de $X$). Logo:

$$\mathbb{E}_y \left[\|y - X\hat{\beta}\|^2\right] = \sigma^2 + \sigma^2 p.$$

Isso nos mostra que o erro aumenta com o número de covariáveis, ou seja, com a complexidade do modelo.

## Bibliografia

Bach (2024)

# Capítulo 4

# Tópicos avançados

## 4.1 Descida dupla

# Apêndice A

# Revisão

Nesta seção, faremos uma breve revisão de alguns conceitos matemáticos importantes.

## A.1 Álgebra linear

Ao longo deste material, adotaremos a seguinte notação:

- $n$: número de observações.
- $p$: número de variáveis preditoras.
- $x_{ij}$: valor da $j$-ésima variável na $i$-ésima observação, com $i = 1, \ldots, n$ e $j = 1, \ldots, p$.

Representamos os dados como uma matriz $X \in \mathbb{R}^{n \times p}$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Cada linha de $X$ é um vetor $x_i \in \mathbb{R}^p$, representando as variáveis da $i$-ésima observação:

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}.$$

Também podemos considerar as colunas de $X$, escritas como $x_j \in \mathbb{R}^n$:

$$x_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

Assim, a matriz $X$ pode ser expressa de duas formas:

$$X = (x_1\, x_2\, \cdots\, x_p) \quad \text{ou} \quad X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

O símbolo $T$ representa a transposta de vetores ou matrizes, por exemplo:

$$X^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}.$$

Denotamos a variável resposta (ou target) por $y_i$, para a $i$-ésima observação. O vetor completo de respostas é:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

O conjunto de dados observados é formado por pares $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

**Exercício 7.** *Considere o conjunto de dados de salários, exemplificado abaixo:*

| | year | age | maritl | race | education | region | jobclass | health | health_ins | logwage | wage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2006 | 18 | 1. Never Married | 1. White | 1. < HS Grad | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 2. No | 4.318063 | 75.043154 |
| **1** | 2004 | 24 | 1. Never Married | 1. White | 4. College Grad | 2. Middle Atlantic | 2. Information | 2. >=Very Good | 2. No | 4.255273 | 70.476020 |
| **2** | 2003 | 45 | 2. Married | 1. White | 3. Some College | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 1. Yes | 4.875061 | 130.982177 |
| **3** | 2003 | 43 | 2. Married | 3. Asian | 4. College Grad | 2. Middle Atlantic | 2. Information | 2. >=Very Good | 1. Yes | 5.041393 | 154.685293 |
| **4** | 2005 | 50 | 4. Divorced | 1. White | 2. HS Grad | 2. Middle Atlantic | 2. Information | 1. <=Good | 1. Yes | 4.318063 | 75.043154 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2995** | 2008 | 44 | 2. Married | 1. White | 3. Some College | 2. Middle Atlantic | 1. Industrial | 2. >=Very Good | 1. Yes | 5.041393 | 154.685293 |
| **2996** | 2007 | 30 | 2. Married | 1. White | 2. HS Grad | 2. Middle Atlantic | 1. Industrial | 2. >=Very Good | 2. No | 4.602060 | 99.689464 |
| **2997** | 2005 | 27 | 2. Married | 2. Black | 1. < HS Grad | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 2. No | 4.193125 | 66.229408 |
| **2998** | 2005 | 27 | 1. Never Married | 1. White | 3. Some College | 2. Middle Atlantic | 1. Industrial | 2. >=Very Good | 1. Yes | 4.477121 | 87.981033 |
| **2999** | 2009 | 55 | 5. Separated | 1. White | 2. HS Grad | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 1. Yes | 4.505150 | 90.481913 |

3000 rows × 11 columns

*Descreva quem é a matriz de dados $X$, quem é $n$, quem é $p$, quem é o vetor resposta* $\mathbf{y}$. ***Dica:*** *tem uma pegadinha.*

### A.1.1 Multiplicações

Nessa seção, vamos estudar fatos importantes sobre multiplicações envolvendo matrizes. Para mais detalhes, o leitor pode ver o excelente livro Trefethen and Bau (1997).

### Matriz-vetor

Seja $x_j$ a $j$-ésima coluna de $X$, um $n$-vetor. Então, a equação $y = Xb$ pode ser reescrito como:

$$y = Xb = \sum_{j=1}^{n} x_j b_j. \tag{A.1}$$

Essa equação pode ser representada esquematicamente da seguinte forma:

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = b_1 \begin{bmatrix} \\ x_1 \\ \\ \end{bmatrix} + b_2 \begin{bmatrix} \\ x_2 \\ \\ \end{bmatrix} + \cdots + b_p \begin{bmatrix} \\ x_p \\ \\ \end{bmatrix}.$$

Na equação acima, $y$ é expresso como uma combinação linear das colunas de $X$. Desa forma, podemos resumir essas diferentes descrições do produto matriz-vetor da seguinte forma. Como matemáticos, estamos acostumados a interpretar a fórmula $Xb = y$ como uma afirmação de que $X$ age sobre $b$ para produzir $y$. A forma acima, por outro lado, sugere a interpretação de que $b$ age sobre $X$ para produzir $y$.

### Matriz-Matriz

Para o produto matriz-matriz $B = AC$, *cada coluna de $B$ é uma combinação linear das colunas de $A$*. Para demonstrar esse fato, começamos com a fórmula usual para produtos de matrizes. Se $A$ é uma matriz de dimensão $\ell \times n$ e $C$ é de dimensão $n \times p$, então $B$ será de dimensão $\ell \times p$, com entradas definidas por

$$B_{ij} = \sum_{k=1}^{n} A_{ik} C_{kj}. \tag{A.2}$$

Aqui, $B_{ij}$, $A_{ik}$ e $C_{kj}$ são elementos de $B$, $A$ e $C$, respectivamente. Escrito em termos de colunas, o produto é

$$\begin{bmatrix} B_1 & B_2 & \cdots & B_n \end{bmatrix} = A \begin{bmatrix} C_1 & C_2 & \cdots & C_n \end{bmatrix},$$

que implica em:

$$B_j = AC_j = \sum_{k=1}^{m} C_{kj} A_k. \tag{A.3}$$

Note que isso é só uma generalização da multiplicação anterior, já que $B_j = AC_j$ e podemo utilizar a formulação Matriz-Vetor da seção anterior.

Um exemplo simples de um produto matriz-matriz é o *produto externo*. Este é o produto de um vetor coluna $u$ de dimensão $n$ com um vetor linha $v$ de dimensão $p$; o resultado é uma matriz $n \times p$ de posto 1. O produto externo pode ser escrito como:

$$\begin{bmatrix} \\ u \\ \\ \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} v_1 u & v_2 u & \cdots & v_n u \end{bmatrix} = \begin{bmatrix} v_1 u_1 & \cdots & v_n u_1 \\ \vdots & \ddots & \vdots \\ v_1 u_m & \cdots & v_n u_m \end{bmatrix}.$$

As colunas são todas múltiplos do mesmo vetor $u$ e, da mesma forma, as linhas são todas múltiplos do mesmo vetor $v$.

### A.1.2 Mudança de base

Ao escrever o produto $b = X^{-1}y$, é importante não deixar que a notação de matriz inversa obscureça o que realmente está acontecendo! Em vez de pensar em $b$ como o resultado da aplicação de $X^{-1}$ a $y$, devemos entendê-lo como o vetor único que satisfaz a equação $Xb = y$.

Uma coisa importante de se notar é que como $XX^{-1}y = y$, então se $z = X^{-1}y$, temos que:

$$y = \sum z_i x_i,$$

isto é, as coordenadas do vetor $z = X^{-1}y$ indicam os coeficientes necessários para escrever $y$ na base dada pelas colunas de $X$.

#### Aplicações

Com as ideias desenvolvidas nessa seção, somos capazes de desenvolver várias transformações de forma rápida. Por exemplo, suponha que queremos uma matriz $C$ cuja primeira coluna é a primeira coluna de $A$ duplicada, e as outras colunas são iguais as de $A$. Pela Seção de multiplicação Matriz-Matriz, queremos então que

$$\begin{aligned} C_1 &= 2A_1 + 0A_2 + \dots 0A_n = A[2, 0, \dots, 0]^T \\ &\vdots \\ C_i &= A_i = A[0, 0, \dots, 1, \dots, 0]^T, \end{aligned}$$

logo, $C = AB$ onde $B = \text{diag}(2, 1, \dots, 1)$.

Suponha agora que $D$ é igual a M, porém com a linha 3 somada com a linha 1. Note que a gente só sabe trabalhar com operações nas colunas, então a primeira coisa é transformas linhas em colunas, fazendo $A^T$, logo

$$\begin{aligned} D_1 &= A_1^T + A_3^T = A^T[1, 0, 1, \dots, 0]^T \\ &\vdots \\ D_i &= A_i^T = A^T[0, 0, \dots, 1, \dots, 0]^T. \end{aligned}$$

Logo,

$$D = A^T \begin{pmatrix} 1, 0, \dots, 0 \\ 0, 1, \dots, 0 \\ 1, 0, \dots, 0 \\ \vdots \\ 0, 0, \dots, 1. \end{pmatrix} = A^T M$$

Como queremos uma expressão em termos de $A$, podemos fazer $D^T = M^T A$.

Ou seja, operações nas colunas de uma matriz são feitas à direita e operações com linhas são feitas à esquerda transposta.

**Exercício 8.** *Considere:* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. *Verifique que as multiplicações definidas acima de fato tem o comportamento esperado descrito no texto.*

**Exercício 9.** *Considere:* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. *Calcule as multiplicações necessárias para dobrar a coluna 1 somada com menos a coluna 2 e fazer linha 2 mais o dobro da linha 1.*

*Faça os cálculos explícitos par amostrar que suas multiplicações estão corretas.*

### A.1.3 Produtos internos

Nos espaços de dimensão 2 ou 3, definimos o comprimento de um vetor $\mathbf{x}$ (ou seja, a distância de sua extremidade até a origem) usando o teorema de Pitágoras. Por exemplo, no espaço $\mathbb{R}^3$, temos:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + x_3^2}.$$

Essa fórmula pode ser naturalmente estendida para qualquer dimensão $n$, definindo a *norma* de um vetor $\mathbf{x} \in \mathbb{R}^n$ como:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

O termo *norma* é frequentemente usado como uma forma mais técnica ou refinada de se referir ao comprimento de um vetor.

O produto interno que definimos para $\mathbb{R}^n$ e $\mathbb{C}^n$ satisfaz as seguintes propriedades fundamentais:

1. **Simetria (conjugada)**: $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$; no caso real, isso equivale à simetria usual: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$;
2. **Linearidade**: $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$, para todos os vetores $\mathbf{x}, \mathbf{y}, \mathbf{z}$ e escalares $\alpha, \beta$;
3. **Não-negatividade**: $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ para todo $\mathbf{x}$;
4. **Não-degenerescência**: $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ se, e somente se, $\mathbf{x} = \mathbf{0}$.

Seja $V$ um espaço vetorial (real ou complexo). Um *produto interno* em $V$ é uma função que associa a cada par de vetores $\mathbf{x}, \mathbf{y} \in V$ um escalar $\langle \mathbf{x}, \mathbf{y} \rangle$ que satisfaz as propriedades 1 a 4 acima.

No caso real, assumimos que $\langle \mathbf{x}, \mathbf{y} \rangle$ é sempre real. Já em espaços complexos, o produto interno pode assumir valores complexos.

Chamamos de *espaço com produto interno* o par $(V, \langle \cdot, \cdot \rangle)$ formado por um espaço vetorial $V$ e um produto interno definido sobre ele. Dado um produto interno, podemos definir a norma de um vetor por:

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

**Exemplo 3.** *Seja $V = \mathbb{R}^n$ ou $\mathbb{C}^n$. Já vimos que o produto interno pode ser definido como*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^* \mathbf{x} = \sum_{k=1}^{n} x_k \overline{y_k}.$$

*Esse produto interno é conhecido como o produto interno padrão em $\mathbb{R}^n$ ou $\mathbb{C}^n$.*

Ao longo do texto, usaremos a letra $\mathbb{F}$ para representar tanto $\mathbb{R}$ quanto $\mathbb{C}$. Assim, qualquer afirmação sobre o espaço $\mathbb{F}^n$ é válida para ambos os casos: $\mathbb{R}^n$ e $\mathbb{C}^n$.

**Exemplo 4.** *Recordemos que, para uma matriz quadrada $A$, o traço é definido como a soma dos elementos da diagonal principal:*

$$\operatorname{tr}(A) := \sum_{k=1}^{n} a_{k,k}.$$

*Seja $M_{m\times n}$ o espaço das matrizes $m \times n$. Definimos o produto interno de Frobenius por:*

$$\langle A, B \rangle = \operatorname{tr}(B^* A).$$

*É possível verificar que esse produto interno satisfaz as propriedades 1 a 4, ou seja, é de fato um produto interno.*

*Observe que:*

$$\operatorname{tr}(B^* A) = \sum_{j,k} A_{j,k} \overline{B_{j,k}},$$

*o que mostra que esse produto interno coincide com o produto interno padrão em $\mathbb{C}^{mn}$.*

**Exemplo 5.** *Seja $V = L^2([a,b])$, o espaço das funções complexas mensuráveis ao quadrado integrável no intervalo $[a,b]$, ou seja:*

$$L^2([a,b]) = \left\{ f : [a,b] \to \mathbb{C} \;\middle|\; \int_a^b |f(t)|^2 \, dt < \infty \right\}.$$

*Definimos o produto interno entre duas funções $f, g \in L^2([a,b])$ por:*

$$\langle f, g \rangle = \int_a^b f(t) \, \overline{g(t)} \, dt.$$

*Esse produto interno satisfaz as propriedades fundamentais (conjugada simetria, linearidade, não-negatividade e não-degenerescência), tornando $L^2([a,b])$ um espaço com produto interno.*

*A norma induzida por esse produto interno é:*

$$\|f\| = \left( \int_a^b |f(t)|^2 \, dt \right)^{1/2}.$$

**Exercício 10.** *Mostre que nos exemplos acima, os produtos acima de fato satisfazem as propriedades de produto-interno.*

A seguir apresentamos alguns resultados importantes sobre produtos internos.

**Lema 1.** *Seja* $\mathbf{x}$ *um vetor em um espaço com produto interno* $V$*. Então* $\mathbf{x} = 0$ *se, e somente se,*

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0 \quad \textit{para todo } \mathbf{y} \in V. \tag{1.1}$$

**Corolário 1.** *Sejam* $\mathbf{x}, \mathbf{y}$ *vetores em um espaço com produto interno* $V$*. A igualdade* $\mathbf{x} = \mathbf{y}$ *vale se, e somente se,*

$$\langle \mathbf{x}, \mathbf{z} \rangle = \langle \mathbf{y}, \mathbf{z} \rangle \quad \textit{para todo } \mathbf{z} \in V.$$

**Corolário 2.** *Sejam* $A, B : X \to Y$ *dois operadores lineares. Suponha que*

$$\langle A\mathbf{x}, \mathbf{y} \rangle = \langle B\mathbf{x}, \mathbf{y} \rangle \quad \textit{para todo } \mathbf{x} \in X \textit{ e } \mathbf{y} \in Y.$$

*Então,* $A = B$*.*

Um dos resultados mais importantes com relação ao produto interno é a desigualdade de Cauchy-Schwarz:

**Teorema 2** (Desigualdade de Cauchy–Schwarz)**.** *Sejam* $\mathbf{x}, \mathbf{y}$ *vetores em um espaço com produto interno. Então:*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|.$$

*Demonstração.* Vamos apresentar uma demonstração que, embora não seja a mais curta, revela bem a origem das ideias principais.

Comecemos com o caso real. Se $\mathbf{y} = 0$, a desigualdade é trivial. Assim, podemos supor que $\mathbf{y} \neq 0$. Pelas propriedades do produto interno, para qualquer escalar $t$ temos:

$$0 \leq \|\mathbf{x} - t\mathbf{y}\|^2 = \langle \mathbf{x} - t\mathbf{y}, \mathbf{x} - t\mathbf{y} \rangle = \|\mathbf{x}\|^2 - 2t\langle \mathbf{x}, \mathbf{y} \rangle + t^2\|\mathbf{y}\|^2.$$

Essa desigualdade vale para todo $t \in \mathbb{R}$, em particular para

$$t = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2},$$

o que nos leva a:

$$0 \leq \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2},$$

ou seja,

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \|\mathbf{x}\|^2 \cdot \|\mathbf{y}\|^2.$$

Portanto, obtemos a desigualdade desejada.

Para o caso complexo, uma das estratégias é considerar o mesmo argumento acima com $t$ complexo (escolhendo, por exemplo, $t = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2}$), ou então proceder de forma análoga usando:

$$\|\mathbf{x} - t\mathbf{y}\|^2 = \|\mathbf{x}\|^2 - t\langle \mathbf{y}, \mathbf{x} \rangle - \bar{t}\langle \mathbf{x}, \mathbf{y} \rangle + |t|^2\|\mathbf{y}\|^2.$$

A escolha de $t = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2}$ minimiza a expressão acima, o que nos leva novamente a:

$$0 \leq \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2},$$

ou seja,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|.$$

Esse raciocínio mostra completamente a validade da desigualdade. A justificativa anterior serviu apenas para motivar a escolha do valor específico de $t$. □

**Lema 2** (Desigualdade triangular)**.** *Para quaisquer vetores* $\mathbf{x}, \mathbf{y}$ *em um espaço com produto interno, vale:*

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

*Demonstração.*

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2|\langle \mathbf{x}, \mathbf{y} \rangle| \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\| \cdot \|\mathbf{y}\| = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \end{aligned}$$

Tomando a raiz quadrada dos dois lados, obtemos a desigualdade desejada. □

**Lema 3** (Identidades de polarização)**.** *Sejam* $\mathbf{x}, \mathbf{y} \in V$*. É possível recuperar o produto interno a partir da norma usando as seguintes fórmulas:*

- *Se $V$ é um espaço com produto interno real, então:*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4} \left( \|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 \right).$$

- *Se $V$ é um espaço com produto interno complexo, então:*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4} \sum_{\alpha \in \{1, -1, i, -i\}} \alpha \|\mathbf{x} + \alpha \mathbf{y}\|^2.$$

**Exercício 11.** *Prove* ***todos*** *os resultados anteriores que não possuem provas.*

### A.1.4 Ortogonalidade

**Definição 1.** *Dois vetores* $\mathbf{u}$ *e* $\mathbf{v}$ *são chamados ortogonais (ou também perpendiculares) se*

$$\langle \mathbf{u}, \mathbf{v} \rangle = 0.$$

*Usamos a notação* $\mathbf{u} \perp \mathbf{v}$ *para indicar que os vetores são ortogonais.*

Note que, se os vetores $\mathbf{u}$ e $\mathbf{v}$ forem ortogonais, então vale a seguinte identidade, conhecida como identidade pitagórica:

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 \quad \text{se } \mathbf{u} \perp \mathbf{v}.$$

**Definição 2.** *Dizemos que um vetor* $\mathbf{v}$ *é ortogonal a um subespaço $E$ se* $\mathbf{v}$ *for ortogonal a todos os vetores* $\mathbf{w} \in E$*.*

*Analogamente, dizemos que dois subespaços $E$ e $F$ são ortogonais se todos os vetores de $E$ são ortogonais a todos os vetores de $F$, ou seja,* $\langle \mathbf{e}, \mathbf{f} \rangle = 0$ *para todo* $\mathbf{e} \in E$ *e* $\mathbf{f} \in F$*.*

O próximo lema mostra como verificar se um vetor é ortogonal a um subespaço gerado por um conjunto finito de vetores.

**Lema 4.** *Seja $E$ o subespaço gerado pelos vetores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$. Então, um vetor $\mathbf{v}$ é ortogonal a $E$ se, e somente se,*

$$\mathbf{v} \perp \mathbf{v}_k, \quad \textit{para todo } k = 1, 2, \ldots, r.$$

**Definição 3.** *Um sistema de vetores $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ é dito ortogonal se quaisquer dois vetores distintos do sistema forem ortogonais entre si, ou seja,*

$$\langle \mathbf{v}_j, \mathbf{v}_k \rangle = 0 \quad \textit{para } j \neq k.$$

*Se, adicionalmente, $\|\mathbf{v}_k\| = 1$ para todo $k$, então o sistema é chamado de ortonormal.*

**Lema 5** (Identidade de Pitágoras generalizada)**.** *Sejam $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ um sistema ortogonal. Então:*

$$\left\| \sum_{k=1}^{n} \alpha_k \mathbf{v}_k \right\|^2 = \sum_{k=1}^{n} |\alpha_k|^2 \cdot \|\mathbf{v}_k\|^2.$$

Essa fórmula assume uma forma particularmente simples no caso de sistemas ortonormais, pois nesse caso $\|\mathbf{v}_k\| = 1$ para todo $k$.

**Definição 4.** *Um sistema ortogonal (ou ortonormal) $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, que também forma uma base, é chamado de base ortogonal (ou base ortonormal).*

É claro que, se $\dim V = n$, então qualquer sistema ortogonal de $n$ vetores não nulos em $V$ é automaticamente uma base ortogonal.

Como vimos anteriormente, para encontrar as coordenadas de um vetor em uma base arbitrária, normalmente é necessário resolver um sistema linear. No entanto, no caso de uma base ortogonal, isso pode ser feito de forma muito mais simples.

Suponha que $\mathbf{v}_1, \ldots, \mathbf{v}_n$ seja uma base ortogonal, e que

$$\mathbf{x} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_n \mathbf{v}_n = \sum_{j=1}^{n} \alpha_j \mathbf{v}_j.$$

Tomando o produto interno de ambos os lados com $\mathbf{v}_1$, obtemos:

$$\langle \mathbf{x}, \mathbf{v}_1 \rangle = \sum_{j=1}^{n} \alpha_j \langle \mathbf{v}_j, \mathbf{v}_1 \rangle = \alpha_1 \langle \mathbf{v}_1, \mathbf{v}_1 \rangle = \alpha_1 \|\mathbf{v}_1\|^2.$$

(já que $\langle \mathbf{v}_j, \mathbf{v}_1 \rangle = 0$ para $j \neq 1$)

Portanto,

$$\alpha_1 = \frac{\langle \mathbf{x}, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2}.$$

De forma semelhante, multiplicando ambos os lados por $\mathbf{v}_k$, obtemos:

$$\langle \mathbf{x}, \mathbf{v}_k \rangle = \sum_{j=1}^{n} \alpha_j \langle \mathbf{v}_j, \mathbf{v}_k \rangle = \alpha_k \|\mathbf{v}_k\|^2,$$

então

$$\alpha_k = \frac{\langle \mathbf{x}, \mathbf{v}_k \rangle}{\|\mathbf{v}_k\|^2}. \tag{A.4}$$

Para encontrar as coordenadas de um vetor em uma base ortogonal, não é necessário resolver um sistema linear — as coordenadas são dadas diretamente pela fórmula (A.4).

Retomando a definição de projeção ortogonal da geometria clássica no plano (bidimensional), podemos introduzir a seguinte definição. Seja $E$ um subespaço de um espaço com produto interno $V$.

**Definição 5.** *Seja* $\mathbf{v}$ *um vetor. A sua projeção ortogonal sobre o subespaço* $E$*, denotada por* $P_E\mathbf{v}$*, é o vetor* $\mathbf{w}$ *tal que:*

1. $\mathbf{w} \in E$;
2. $\mathbf{v} - \mathbf{w} \perp E$.

Usaremos a notação $\mathbf{w} = P_E\mathbf{v}$ para representar a projeção ortogonal de $\mathbf{v}$ sobre $E$.

**Teorema 3.** *Seja* $\mathbf{w} = P_E\mathbf{v}$ *a projeção ortogonal de* $\mathbf{v}$ *sobre o subespaço* $E$*. Então,* $\mathbf{w}$ *é o ponto de* $E$ *mais próximo de* $\mathbf{v}$*, ou seja, para todo* $\mathbf{x} \in E$*:*

$$\|\mathbf{v} - \mathbf{w}\| \leq \|\mathbf{v} - \mathbf{x}\|.$$

*Além disso, se existir* $\mathbf{x} \in E$ *tal que*

$$\|\mathbf{v} - \mathbf{w}\| = \|\mathbf{v} - \mathbf{x}\|,$$

*então* $\mathbf{x} = \mathbf{w}$.

*Demonstração.* Seja $\mathbf{y} = \mathbf{w} - \mathbf{x}$. Então:

$$\mathbf{v} - \mathbf{x} = \mathbf{v} - \mathbf{w} + \mathbf{w} - \mathbf{x} = \mathbf{v} - \mathbf{w} + \mathbf{y}.$$

Como $\mathbf{v} - \mathbf{w} \perp E$ e $\mathbf{y} \in E$, segue que $\mathbf{v} - \mathbf{w} \perp \mathbf{y}$. Assim, pelo teorema de Pitágoras:

$$\|\mathbf{v} - \mathbf{x}\|^2 = \|\mathbf{v} - \mathbf{w}\|^2 + \|\mathbf{y}\|^2 \geq \|\mathbf{v} - \mathbf{w}\|^2.$$

A igualdade ocorre se, e somente se, $\|\mathbf{y}\| = 0$, ou seja, $\mathbf{y} = 0$, o que implica $\mathbf{x} = \mathbf{w}$. □

**Proposição 1.** *Sejam* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ *uma base ortogonal do subespaço* $E$*. Então, a projeção ortogonal de um vetor* $\mathbf{v}$ *sobre* $E$ *é dada por:*

$$P_E\mathbf{v} = \sum_{k=1}^{r} \alpha_k \mathbf{v}_k, \quad \textit{onde} \quad \alpha_k = \frac{\langle \mathbf{v}, \mathbf{v}_k \rangle}{\|\mathbf{v}_k\|^2}.$$

*Em outras palavras:*

$$P_E\mathbf{v} = \sum_{k=1}^{r} \frac{\langle \mathbf{v}, \mathbf{v}_k \rangle}{\|\mathbf{v}_k\|^2} \mathbf{v}_k. \tag{A.5}$$

Note que a fórmula dos coeficientes $\alpha_k$ coincide com a da equação (A.4), isto é, essa fórmula continua válida mesmo se o sistema $\{\mathbf{v}_k\}$ for apenas ortogonal (e não base), pois ela projeta $\mathbf{v}$ sobre o subespaço gerado pelos vetores.

*Demonstração.* Definimos:

$$\mathbf{w} := \sum_{k=1}^{r} \alpha_k \mathbf{v}_k, \quad \text{com} \quad \alpha_k = \frac{\langle \mathbf{v}, \mathbf{v}_k \rangle}{\|\mathbf{v}_k\|^2}.$$

Queremos mostrar que $\mathbf{v} - \mathbf{w} \perp E$. É é suficiente mostrar que

$$\mathbf{v} - \mathbf{w} \perp \mathbf{v}_k, \quad \text{para todo } k = 1, 2, \ldots, r.$$

Para isso, calculamos:

$$\begin{aligned} \langle \mathbf{v} - \mathbf{w}, \mathbf{v}_k \rangle &= \langle \mathbf{v}, \mathbf{v}_k \rangle - \langle \mathbf{w}, \mathbf{v}_k \rangle \\ &= \langle \mathbf{v}, \mathbf{v}_k \rangle - \left\langle \sum_{j=1}^{r} \alpha_j \mathbf{v}_j, \mathbf{v}_k \right\rangle \\ &= \langle \mathbf{v}, \mathbf{v}_k \rangle - \sum_{j=1}^{r} \alpha_j \langle \mathbf{v}_j, \mathbf{v}_k \rangle. \end{aligned}$$

Como o sistema $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ é ortogonal, temos $\langle \mathbf{v}_j, \mathbf{v}_k \rangle = 0$ para $j \neq k$, e $\langle \mathbf{v}_k, \mathbf{v}_k \rangle = \|\mathbf{v}_k\|^2$. Logo:

$$\langle \mathbf{v} - \mathbf{w}, \mathbf{v}_k \rangle = \langle \mathbf{v}, \mathbf{v}_k \rangle - \alpha_k \|\mathbf{v}_k\|^2 = \langle \mathbf{v}, \mathbf{v}_k \rangle - \langle \mathbf{v}, \mathbf{v}_k \rangle = 0.$$

Portanto, $\mathbf{v} - \mathbf{w} \perp \mathbf{v}_k$ para todo $k$, e segue que $\mathbf{v} - \mathbf{w} \perp E$. Assim, $\mathbf{w} = P_E \mathbf{v}$. □

**Observação 2.** *Retomando a definição de produto interno em $\mathbb{C}^n$ e $\mathbb{R}^n$, podemos deduzir da fórmula (A.5) que a matriz da projeção ortogonal $P_E$ sobre um subespaço $E \subseteq \mathbb{C}^n$ (ou $\mathbb{R}^n$) é dada por:*

$$P_E = \sum_{k=1}^{r} \frac{1}{\|\mathbf{v}_k\|^2} \mathbf{v}_k \mathbf{v}_k^*, \tag{A.6}$$

*onde os vetores coluna $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ formam uma base ortogonal de $E$.*

### Ortogonalização de Gram–Schmidt

Suponha que temos um conjunto linearmente independente de vetores $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. O método de Gram–Schmidt constrói a partir desse conjunto um sistema ortogonal $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ tal que:

$$\operatorname{span}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n).$$

Além disso, para todo $r \leq n$, temos:

$$\operatorname{span}(\mathbf{x}_1, \ldots, \mathbf{x}_r) = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_r).$$

O algoritmo segue os seguintes passos:

1. Defina $\mathbf{v}_1 := \mathbf{x}_1$. Seja $E_1 := \operatorname{span}(\mathbf{v}_1)$.

2. Defina
$$\mathbf{v}_2 := \mathbf{x}_2 - P_{E_1}\mathbf{x}_2 = \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2}\mathbf{v}_1.$$
Seja $E_2 := \text{span}(\mathbf{v}_1, \mathbf{v}_2)$. Como $\mathbf{x}_2 \notin E_1$, temos $\mathbf{v}_2 \neq 0$.

3. Defina
$$\mathbf{v}_3 := \mathbf{x}_3 - P_{E_2}\mathbf{x}_3 = \mathbf{x}_3 - \frac{\langle \mathbf{x}_3, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2}\mathbf{v}_1 - \frac{\langle \mathbf{x}_3, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2}\mathbf{v}_2.$$
Seja $E_3 := \text{span}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$.

4. Suponha que já tenhamos construído vetores ortogonais $\mathbf{v}_1, \ldots, \mathbf{v}_r$, tais que
$$E_r := \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_r) = \text{span}(\mathbf{x}_1, \ldots, \mathbf{x}_r).$$
Defina:
$$\mathbf{v}_{r+1} := \mathbf{x}_{r+1} - \sum_{k=1}^{r} \frac{\langle \mathbf{x}_{r+1}, \mathbf{v}_k \rangle}{\|\mathbf{v}_k\|^2}\mathbf{v}_k.$$
Note que $\mathbf{x}_{r+1} \notin E_r$, então $\mathbf{v}_{r+1} \neq 0$.

Continuando esse processo até $r = n$, obtemos um sistema ortogonal de vetores $\mathbf{v}_1, \ldots, \mathbf{v}_n$ tal que:
$$\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \text{span}(\mathbf{x}_1, \ldots, \mathbf{x}_n).$$

**Exercício 12.** *Aplique o método de Gram-Schmidit para os vetores* $\{(1,1,1), (0,1,2), (1,0,2)\}$.

### 3.3. Complemento ortogonal. Decomposição $V = E \oplus E^\perp$

**Definição 6.** *Seja $E$ um subespaço de um espaço com produto interno $V$. O complemento ortogonal de $E$, denotado por $E^\perp$, é o conjunto de todos os vetores ortogonais a $E$:*
$$E^\perp := \{\mathbf{x} \in V : \mathbf{x} \perp E\}.$$

Se $\mathbf{x}, \mathbf{y} \perp E$, então qualquer combinação linear $\alpha\mathbf{x} + \beta\mathbf{y}$ também está em $E^\perp$ (consegue ver por quê?). Logo, $E^\perp$ é um subespaço.

Pela definição de projeção ortogonal, qualquer vetor $\mathbf{v} \in V$ admite uma decomposição única da forma:
$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2, \quad \text{com } \mathbf{v}_1 \in E \text{ e } \mathbf{v}_2 \perp E \text{ (ou seja, } \mathbf{v}_2 \in E^\perp).$$
Neste caso, temos $\mathbf{v}_1 = P_E\mathbf{v}$.

Essa afirmação pode ser escrita simbolicamente como:
$$V = E \oplus E^\perp,$$
o que expressa precisamente que todo vetor admite a decomposição única acima.

A proposição a seguir mostra uma propriedade fundamental do complemento ortogonal:

**Proposição 2.** *Seja $E$ um subespaço. Então:*
$$(E^\perp)^\perp = E.$$

**Exercício 13.** *Prove* ***todos*** *os resultados anteriores que não possuem provas.*

### A.1.5 Decomposição em valores singulares

Para mais detalhes na parte básica de álgebra linear, ver Treil (2015).

**Definição 7.** *Um operador autoadjunto $A : X \to X$ é chamado de definido positivo se*

$$\langle Ax, x\rangle > 0 \quad \forall x \neq 0,$$

*e é chamado de semidefinido positivo se*

$$\langle Ax, x\rangle \geq 0 \quad \forall x \in X.$$

*Usaremos a notação $A \succ 0$ para operadores definidos positivos e $A \succeq 0$ para operadores semidefinidos positivos.*

O teorema a seguir descreve operadores definidos positivos e semidefinidos positivos.

**Teorema 4.** *Seja $A = A^*$. Então:*

1. *$A \succ 0$ se, e somente se, todos os autovalores de $A$ são positivos.*
2. *$A \succeq 0$ se, e somente se, todos os autovalores de $A$ são não-negativos.*

**Corolário 3.** *Seja $A = A^* \succeq 0$ um operador semidefinido positivo. Então, existe um único operador semidefinido positivo $B$ tal que $B^2 = A$. Tal $B$ é chamado de raiz quadrada (positiva) de $A$ e é denotado por $\sqrt{A}$ ou $A^{1/2}$.*

Considere um operador $A : X \to Y$. O seu *quadrado hermitiano* $A^*A$ é um operador semidefinido positivo atuando em $X$. De fato,

$$(A^*A)^* = A^*A^{**} = A^*A$$

e

$$\langle A^*Ax, x\rangle = \langle Ax, Ax\rangle = \|Ax\|^2 \geq 0 \quad \forall x \in X.$$

Portanto, existe uma (única) raiz quadrada semidefinida positiva $R = \sqrt{A^*A}$. Esse operador $R$ é chamado de *módulo* do operador $A$, e frequentemente é denotado por $|A|$.

**Definição 8.** *Os autovalores de $|A|$ são chamados de valores singulares de $A$. Em outras palavras, se $\lambda_1, \lambda_2, \ldots, \lambda_n$ são autovalores de $A^*A$, então $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}$ são os valores singulares de $A$.*

Considere um operador $A : X \to Y$ e denote por $\sigma_1, \sigma_2, \ldots, \sigma_n$ os valores singulares de $A$, contando multiplicidades. Suponha ainda que $\sigma_1, \sigma_2, \ldots, \sigma_r$ sejam os valores singulares não nulos de $A$. Isso significa, em particular, que $\sigma_k = 0$ para $k > r$.

Pela definição de valores singulares, os números $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ são autovalores de $A^*A$. Seja $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ uma base ortonormal de autovetores de $A^*A$, com $A^*A\mathbf{v}_k = \sigma_k^2\mathbf{v}_k$.

**Proposição 3.** *O sistema*

$$\mathbf{w}_k := \frac{1}{\sigma_k} A\mathbf{v}_k, \quad k = 1, 2, \ldots, r$$

*é um sistema ortonormal.*

**Exercício 14.** *Prove a proposição acima.*

**Intuição geométrica**

A decomposição em valores singulares (SVD) é uma fatoração matricial importante tanto para o desenvolvimento de algoritmos quanto para a interpretação conceitual em álgebra linear. Um dos principais insights geométricos da SVD é que *a imagem da esfera unitária sob qualquer matriz $n \times p$ é uma hiperelipse.*

Consideramos a matriz $A$ real e o espaço $\mathbb{R}^n$. A hiperelipse pode ser entendida como a generalização de uma elipse para dimensões superiores. Formalmente, em $\mathbb{R}^n$, é a superfície obtida ao esticar a esfera unitária em $n$ direções ortogonais $\{u_1, \ldots, u_n\}$ por fatores $\sigma_1, \ldots, \sigma_n$. Esses fatores são chamados de *semieixos principais* e são as quantidades $\{\sigma_i u_i\}$. Quando $A$ tem posto $r$, exatamente $r$ dos $\sigma_i$ serão não nulos. Em particular, se $n \geq p$, no máximo $p$ deles serão positivos.

A esfera unitária $S$ em $\mathbb{R}^p$ é mapeada por $A$ em uma hiperelipse no espaço $\mathbb{R}^n$. Os valores singulares $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ são as medidas dos semieixos principais. Associados a esses valores, temos:

- Os **vetores singulares à esquerda** de $A$, que são os vetores $\{u_1, \ldots, u_p\}$ na imagem $AS$.
- Os **vetores singulares à direita** de $A$, que são os vetores $\{v_1, \ldots, v_p\} \subset S$, isto é, os vetores da esfera que correspondem às pré-imagens dos semieixos principais.

Temos então que $Av_i = \sigma_i u_i$, que pode ser escrito como $AV = U\Sigma$, como $V$ é unitária, temos então que $A = U\Sigma V^T$.

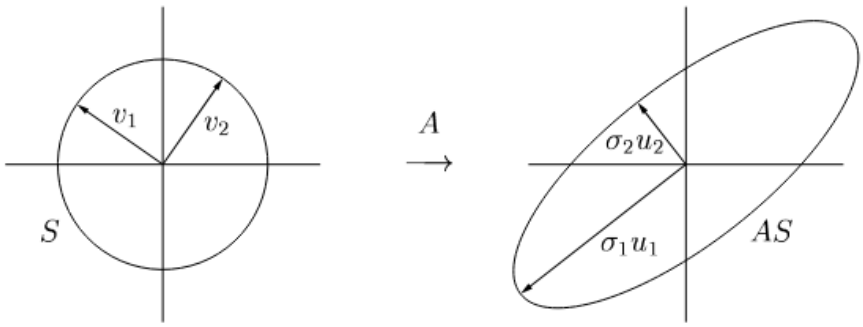



Figure 4.1. *SVD of a $2 \times 2$ matrix.*

Figura A.1: Retirado de (Trefethen and Bau, 1997).

Formalmente, sejam $n$ e $p$ números inteiros quaisquer, sem a necessidade de $n \geq p$. Dada uma matriz $A \in \mathbb{R}^{n \times p}$ (não necessariamente de posto completo), a *decomposição em valores singulares* (SVD) de $A$ é uma fatoração da forma:

$$A = U\Sigma V^*,$$

onde:

- $U \in \mathbb{R}^{n \times n}$ é uma matriz unitária;
- $V \in \mathbb{R}^{p \times p}$ é uma matriz unitária;
- $\Sigma \in \mathbb{R}^{n \times p}$ é uma matriz diagonal (ou quase-diagonal).

Os elementos diagonais $\sigma_j$ da matriz $\Sigma$ são não-negativos e ordenados em ordem não crescente, ou seja, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, com $p = \min(n, p)$.

**Teorema 5** (Teorema de Existência e Unicidade)**.** *Toda matriz $A \in \mathbb{R}^{n \times p}$ admite uma decomposição em valores singulares da forma $A = U\Sigma V^*$. Além disso, os valores singulares $\{\sigma_j\}$ são unicamente determinados. Se $A$ for quadrada e os $\sigma_j$ forem distintos, então os vetores singulares à esquerda $\{u_j\}$ e à direita $\{v_j\}$ também são unicamente determinados, a menos de sinais.*

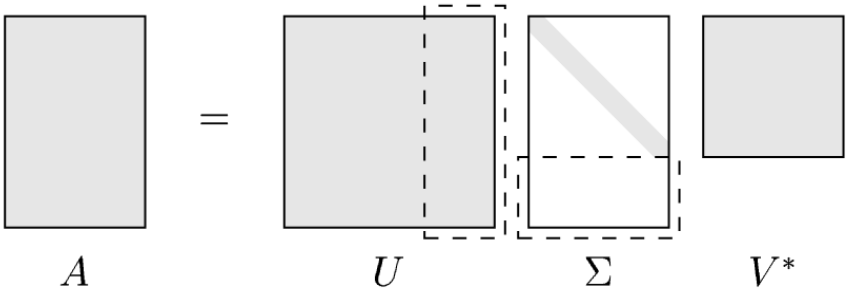



Figura A.2: Ilustração do SVD. Retirado de (Trefethen and Bau, 1997).

**Bibliografia**

## A.2 Probabilidade

Um *espaço de probabilidade* é uma tupla composta por três elementos: o *espaço amostral*, o *conjunto de eventos* e uma *distribuição de probabilidade*:

- **Espaço amostral** $\Omega$: $\Omega$ é o conjunto de todos os eventos elementares ou resultados possíveis de um experimento. Por exemplo, ao lançar um dado, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- **Conjunto de eventos** $\mathcal{F}$: $\mathcal{F}$ é uma $\sigma$-álgebra, ou seja, um conjunto de subconjuntos de $\Omega$ que contém $\Omega$ e é fechado sob complementação e união enumerável (e, consequentemente, também sob interseção enumerável). Um exemplo de evento é: "o dado mostra um número ímpar".
- **Distribuição de probabilidade** $\mathbb{P}$: $\mathbb{P}$ é uma função que associa a cada evento de $\mathcal{F}$ um número em $[0, 1]$, tal que $\mathbb{P}[\Omega] = 1$, $\mathbb{P}[\emptyset] = 0$ e, para eventos mutuamente exclusivos $A_1, \ldots, A_n$, temos:

$$\mathbb{P}\left[A_1 \cup \cdots \cup A_n\right] = \sum_{i=1}^{n} \mathbb{P}[A_i].$$

A distribuição de probabilidade discreta associada ao lançamento de um dado justo pode ser definida como $\mathbb{P}[A_i] = 1/6$ para $i \in \{1, \ldots, 6\}$, onde $A_i$ é o evento "o dado mostra o valor $i$".

### A.2.1 Variáveis aleatórias

Uma variável aleatória $X$ é uma função $X : \Omega \to \mathbb{R}$ mensurável, ou seja, tal que para qualquer intervalo $I$, o subconjunto $\{\omega \in \Omega : X(\omega) \in I\}$ pertence ao conjunto de eventos.

A *função de massa de probabilidade* de uma variável aleatória discreta $X$ é a função $x \mapsto \mathbb{P}[X = x]$.

Uma distribuição é dita *absolutamente contínua* quando possui uma *função densidade de probabilidade* $f$ associada, tal que, para todo $a, b \in \mathbb{R}$:

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f(x)dx.$$

**Exemplo 6** (Binomial)**.** *Uma variável aleatória $X$ segue uma distribuição binomial $B(n, p)$ com $n \in \mathbb{N}$ e $p \in [0, 1]$ se, para $k \in \{0, 1, \ldots, n\}$,*

$$\mathbb{P}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}.$$

**Exemplo 7** (Uniforme)**.** *Uma variável aleatória $X$ segue uma distribuição uniforme $U(a, b)$ no intervalo $(a, b)$ se,*

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq x \leq b \\ 0 & \text{caso contrário.} \end{cases}$$

**Exemplo 8** (Normal)**.** *Uma variável aleatória $X$ segue uma distribuição normal $N(\mu, \sigma^2)$ com $\mu \in \mathbb{R}$ e $\sigma > 0$ se sua densidade for dada por:*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

*A distribuição normal padrão é $N(0, 1)$, com média zero e variância unitária.*

**Exemplo 9** (Laplace)**.** *Uma variável aleatória $X$ segue uma distribuição de Laplace com parâmetro de localização $\mu \in \mathbb{R}$ e parâmetro de escala $b > 0$ se sua densidade for:*

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right).$$

### A.2.2 Probabilidade condicional e independência

A probabilidade condicional do evento $A$ dado o evento $B$ é definida como a razão entre a probabilidade da interseção $A \cap B$ e a probabilidade de $B$, desde que $\mathbb{P}[B] \neq 0$:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Dois eventos $A$ e $B$ são ditos independentes quando a probabilidade conjunta $\mathbb{P}[A \cap B]$ pode ser fatorada como o produto $\mathbb{P}[A]\mathbb{P}[B]$:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B].$$

De forma equivalente, a independência entre $A$ e $B$ pode ser expressa afirmando que $\mathbb{P}[A \mid B] = \mathbb{P}[A]$, sempre que $\mathbb{P}[B] \neq 0$.

Além disso, uma sequência de variáveis aleatórias é dita *i.i.d.* (independentes e identicamente distribuídas) quando todas as variáveis da sequência são mutuamente independentes e seguem a mesma distribuição de probabilidade.

### A.2.3 Algumas fórmulas importantes

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \qquad \text{(regra da soma)}$$

$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \mathbb{P}[A_i] \qquad \textbf{(desigualdade da união)}$$

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\mathbb{P}[A]}{\mathbb{P}[B]} \qquad \text{(fórmula de Bayes)}$$

$$\mathbb{P}\left[\bigcap_{i=1}^{n} A_i\right] = \mathbb{P}[A_1]\mathbb{P}[A_2 \mid A_1] \cdots \mathbb{P}\left[A_n \mid \bigcap_{i=1}^{n-1} A_i\right] \qquad \text{(regra da cadeia)}$$

**Exercício 15.** *Prove os resultados acima.*

### A.2.4 Esperança e desigualdade de Markov

A esperança ou valor esperado de uma variável aleatória $X$ é denotada por $\mathbb{E}[X]$ e, no caso discreto, é definida como

$$\mathbb{E}[X] = \sum_{x} x\, \mathbb{P}[X = x]. \tag{C.9}$$

No caso contínuo, quando $X$ possui uma função densidade de probabilidade $f(x)$, a esperança é dada por

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Além disso, dado uma função qualquer $g$, temos que:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx.$$

Uma propriedade fundamental da esperança é sua linearidade. Isto é, para quaisquer variáveis aleatórias $X$ e $Y$ e constantes $a, b \in \mathbb{R}$, temos:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \tag{C.10}$$

A seguir, apresentamos um limite superior simples para uma variável aleatória não-negativa em função de sua esperança, conhecido como a *Desigualdade de Markov*.

**Teorema 6** (Desigualdade de Markov)**.** *Seja $X$ uma variável aleatória não-negativa ($X \geq 0$ quase certamente) com valor esperado $\mathbb{E}[X] < \infty$. Então, para todo $t > 0$, temos:*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

**Exercício 16.** *Prove as desigualdades de Markov.*

### A.2.5 Variância e a desigualdade de Chebyshev

A variância de uma variável aleatória $X$ é denotada por $\text{Var}[X]$ e definida como

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

O desvio padrão de $X$ é denotado por $\sigma_X$ e definido como

$$\sigma_X = \sqrt{\text{Var}[X]}.$$

Para qualquer variável aleatória $X$ e qualquer constante $a \in \mathbb{R}$, as seguintes propriedades básicas são válidas:

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

$$\text{Var}[aX] = a^2 \,\text{Var}[X].$$

Além disso, se $X$ e $Y$ forem independentes, então

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y].$$

**Exercício 17.** *Prove as identidade acima.*

A seguinte desigualdade, conhecida como *Desigualdade de Chebyshev*, fornece um limite para a probabilidade de uma variável aleatória se desviar de sua esperança em função do seu desvio padrão.

**Teorema 7** (Desigualdade de Chebyshev)**.** *Seja $X$ uma variável aleatória com valor esperado $\mu = \mathbb{E}[X]$ e variância finita $\text{Var}(X) = \sigma^2$. Então, para todo $\varepsilon > 0$, vale:*

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

**Exercício 18.** *Prove a desigualdade de Chebyshev.*

### A.2.6 Covariância

A covariância entre duas variáveis aleatórias $X$ e $Y$ é denotada por $\text{Cov}(X, Y)$ e definida por

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

**Exercício 19.** *Prove que*

$$\text{Cov}(X, Y) = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right].$$

Dizemos que $X$ e $Y$ são *não correlacionadas* quando $\text{Cov}(X, Y) = 0$. Se $X$ e $Y$ forem independentes, então certamente são não correlacionadas, mas a recíproca nem sempre é verdadeira.

**Exercício 20.** *Seja $X$ uniforme no intervalo $[-1, 1]$ e seja $Y = X^2$. Mostre que $\text{Cov}\,(X, Y) = 0$ mas $X, Y$ não são independentes.*

**Observação 3.** *Considere uma variável aleatória contínua $X$ centrada em zero, ou seja, $\mathbb{E}[X] = 0$, com densidade de probabilidade par e definida em um intervalo do tipo $(-a, a)$, com $a > 0$. Seja $Y = g(X)$ para uma função $g$. A questão é: para quais funções $g(X)$ temos $\mathrm{Cov}(X, g(X)) = 0$?*

*Sabemos que*

$$\mathrm{Cov}(X, g(X)) = \mathbb{E}[Xg(X)] - \mathbb{E}[X]\mathbb{E}[g(X)].$$

*Como $\mathbb{E}[X] = 0$, segue que $\mathrm{Cov}(X, g(X)) = \mathbb{E}[Xg(X)]$. Denotando a densidade de $X$ por $f(x)$, temos*

$$\mathrm{Cov}(X, g(X)) = \int_{-a}^{a} xg(x)f(x)dx.$$

*Uma maneira de garantir que $\mathrm{Cov}(X, g(X)) = 0$ é exigir que $g(x)$ seja uma função par. Assim, $xg(x)f(x)$ será uma função ímpar e a integral em $(-a, a)$ se anulará, ou seja,*

$$\int_{-a}^{a} xg(x)f(x)dx = 0.$$

*Portanto, $\mathrm{Cov}(X, f(X)) = 0$ e como $Y = g(X)$, teremos que ambas são dependentes.*

*Dessa forma, podemos concluir que a distribuição precisa de $X$ não afeta a condição, desde que $p(x)$ seja simétrica em torno da origem. Qualquer função par $f(\cdot)$ satisfará $\mathrm{Cov}(X, f(X)) = 0$.*

A covariância é uma forma bilinear simétrica e semi-definida positiva, com as seguintes propriedades:

- **Simetria**: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ para quaisquer variáveis $X$ e $Y$.
- **Bilinearidade**: $\mathrm{Cov}(X + X', Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X', Y)$ e $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y)$ para qualquer $a \in \mathbb{R}$.
- **Semi-definida positiva**: $\mathrm{Cov}(X, X) = \mathrm{Var}[X] \geq 0$ para qualquer variável $X$.

Além disso, vale a desigualdade de Cauchy-Schwarz, que afirma que para variáveis $X$ e $Y$ com variância finita,

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}.$$

**Exercício 21.** *Prove os resultados acima.*

A matriz de covariância de um vetor de variáveis aleatórias $\mathbf{X} = (X_1, \ldots, X_p)$ é a matriz em $\mathbb{R}^{n\times n}$ denotada por $\mathbf{C}(\mathbf{X})$ e definida por

$$\mathbf{C}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\right].$$

Portanto, $\mathbf{C}(\mathbf{X})$ é a matriz cujos elementos são $\mathrm{Cov}(X_i, X_j)$. Além disso, é imediato mostrar que

$$\mathbf{C}(\mathbf{X}) = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}]^\top.$$

### A.2.7 Teoremas assintóticos

Em muitas aplicações de probabilidade e estatística, estamos interessados no comportamento de sequências de variáveis aleatórias quando o número de observações tende ao infinito. Os *teoremas assintóticos* fornecem resultados fundamentais que descrevem como certos estimadores ou somas de variáveis aleatórias se comportam no limite, ou seja, quando o tamanho da amostra $n$ cresce indefinidamente.

**Teorema 8** (Lei Fraca dos Grandes Números)**.** *Seja $(X_n)_{n\in\mathbb{N}}$ uma sequência de variáveis aleatórias independentes, todas com a mesma esperança $\mu$ e variância $\sigma^2 < \infty$. Definindo a média amostral por*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

*então, para qualquer $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) = 0.$$

**Exercício 22.** *Prove a Lei Fraca dos Grandes números utilizando a desigualdade de Chebyshev.*

**Teorema 9** (Teorema Central do Limite)**.** *Seja $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias i.i.d. com esperança $\mu$ e desvio padrão $\sigma$. Definimos a média amostral como*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

*e a variância da média como $\sigma_n^2 = \sigma^2/n$. Então, a variável padronizada $(\overline{X}_n - \mu)/\sigma_n$ converge em distribuição para uma normal padrão $N(0,1)$. Mais precisamente, para todo $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma_n} \leq t\right) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

**Observação 4.** *Apesar dos teoremas assintóticos, como a Lei Fraca dos Grandes Números e o Teorema Central do Limite, serem fundamentais para entender o comportamento de sequências de variáveis aleatórias quando $n \to \infty$, na prática, em aprendizado de máquina, o número de amostras n nem sempre é grande o suficiente para que esses resultados sejam aplicáveis com segurança. Por outro lado, desigualdades como as de Markov e Chebyshev fornecem limites válidos para qualquer valor finito de n. Essas desigualdades são exemplos de desigualdades de concentração, que nos permitem controlar a probabilidade de desvios em torno da média de uma variável aleatória. A teoria de concentração será crucial em tópicos futuros, pois fornece ferramentas importantes para analisar o desempenho de algoritmos em cenários onde o regime assintótico não pode ser garantido.*

### A.2.8 Função geradora de momentos

A esperança $\mathbb{E}[X^p]$ é chamada de $p$-ésimo momento da variável aleatória $X$. A *função geradora de momentos* de uma variável aleatória $X$ é uma ferramenta importante, pois permite obter seus diferentes momentos por meio de diferenciação em zero. Essa função é crucial tanto para descrever a distribuição de $X$ quanto para analisar suas propriedades.

A função geradora de momentos de uma variável aleatória $X$ é a função $M_X : t \mapsto \mathbb{E}[e^{tX}]$, definida para os valores de $t \in \mathbb{R}$ tais que a expectativa exista (seja finita).

**Exercício 23.** *Mostre que se $M_X$ for diferenciável em zero, então o p-ésimo momento de X é dado por* $\mathbb{E}[X^p] = M_X^{(p)}(0)$.

**Exercício 24.** *Seja X uma variável aleatória com distribuição normal padrão, ou seja,* $X \sim N(0,1)$. *Mostre que a função geradora de momentos de X é dada por por:*

$$M_X(t) = e^{\frac{t^2}{2}}.$$

## Bibliografia

# Apêndice B

# Guia de desigualdades

# Apêndice C

# Ferramentas computacionais

## C.1 Git

O **Git** é um sistema de controle de versão distribuído amplamente utilizado no desenvolvimento de software. Ele permite que diversos desenvolvedores trabalhem simultaneamente em um projeto, acompanhando as alterações feitas no código, revertendo mudanças, criando ramificações (*branches*) e colaborando de forma eficiente.

Com o Git, o histórico de alterações de um projeto é armazenado localmente, o que possibilita o trabalho off-line e oferece grande flexibilidade na manipulação de versões.

### Principais comandos do Git

A seguir, apresentamos os comandos mais básicos e úteis do Git:

- `git init`
  Inicializa um novo repositório Git em um diretório.
- `git clone <URL>`
  Clona um repositório remoto (por exemplo, do GitHub) para a máquina local.
- `git status`
  Exibe o estado atual do repositório: arquivos modificados, não rastreados etc.
- `git add <arquivo>`
  Adiciona um ou mais arquivos ao *staging area*, preparando-os para o commit.
- `git commit -m "mensagem"`
  Registra as mudanças preparadas com uma mensagem descritiva.
- `git log`
  Mostra o histórico de commits do repositório.
- `git diff`
  Exibe as diferenças entre arquivos modificados e o último commit.

- `git branch`
  Lista todas as branches do projeto.
- `git checkout <branch>`
  Alterna para outra branch existente.
- `git merge <branch>`
  Mescla o conteúdo de uma branch à branch atual.
- `git pull`
  Atualiza o repositório local com as alterações do repositório remoto.
- `git push`
  Envia os commits locais para o repositório remoto.

### Exemplo de fluxo básico

```
git init
git add arquivo.txt
git commit -m "Adiciona arquivo inicial"
git remote add origin https://github.com/usuario/repositorio.git
git push -u origin main
```

## C.2 Python

## C.3 Poetry

# Referências Bibliográficas


Bach, F. (2024). *Learning Theory from First Principles*. Adaptive Computation and Machine Learning series. MIT Press. 28

Breiman, L. (2001a). Random forests. *Machine Learning*, 45:5–32. 9

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231. 9

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA. 5, 15

Izbicki, R. and dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. 5, 8, 14, 17

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. 5, 7, 26

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. The MIT Press, 2nd edition. 5, 10

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning : from theory to algorithms*. 5, 7

Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra*. SIAM. 32, 44, 45

Treil, S. (2015). *Linear Algebra Done Wrong*. 43