Aprendizado de máquinas

Thiago Rodrigo Ramos

18 de março 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. O livro 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.

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.

1.2 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.

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.*

	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: Exemplo de registros do conjunto de dados de salários.

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 interessante envolve dados do mercado financeiro (Fig. 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.

	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 2: Exemplo de registros do conjunto de dados de ações.

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

2 Revisão matemática

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

2.1 Álgebra linear

Ao longo deste material, adotaremos a seguinte notação:

- n: número de observações (ou amostras).
- 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 3. *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.*

2.1.1 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:

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

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

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

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:

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

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.

2.1.2 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.

2.1.3 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 4. *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 5. *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.

2.1.4 Decomposição em valores singulares

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 σ_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 $A = U\Sigma V^T$.

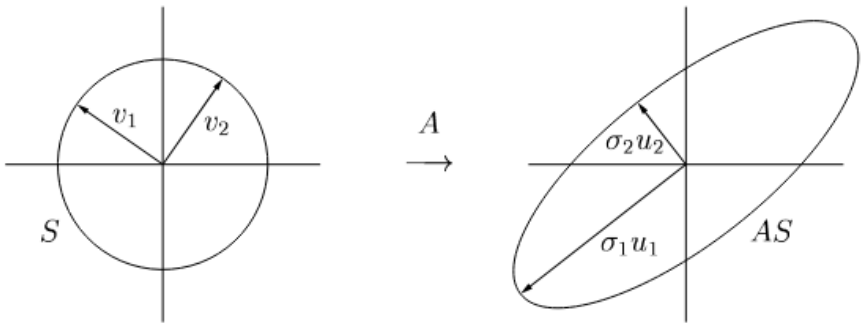


Figure 4.1. *SVD of a* 2×2 *matrix.*

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:

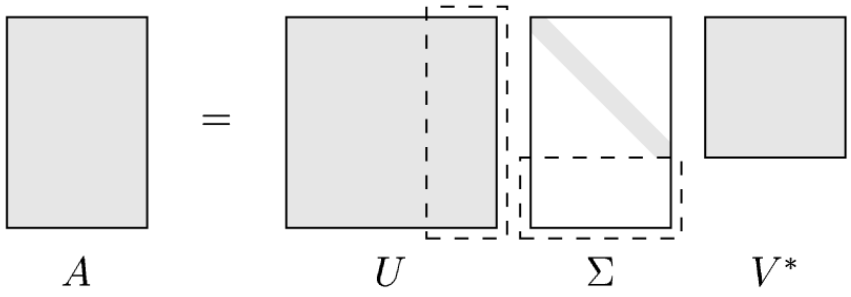
$$A = U\Sigma V^*,$$

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 σ_j da matriz Σ 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 1 (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 σ_j forem distintos, então os vetores singulares à esquerda $\{u_j\}$ e à direita $\{v_j\}$ também são unicamente determinados, a menos de sinais.*



2.2 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** Ω: Ω 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 σ-álgebra, ou seja, um conjunto de subconjuntos de Ω que contém Ω 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 $\mathbb{P}[\Omega] = 1$, $\mathbb{P}[\varnothing] = 0$ 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".

2.2.1 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 1 (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 2 (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 3 (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 4 (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:*

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

2.2.2 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.

2.2.3 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 6. *Prove os resultados acima.*

2.2.4 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 2 (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 7. *Prove as desigualdades de Markov.*

2.2.5 Variância e a desigualdade de chebyshev

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

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

O desvio padrão de X é denotado por σ_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 8. *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 3 (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 9. *Prove a desigualdade de Chebyshev.*

2.2.6 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 10. *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 11. *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 1. *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 $\text{Cov}(X, g(X)) = 0$?*

Sabemos que

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

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

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

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 12. *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

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

2.2.7 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 4 (Lei Fraca dos Grandes Números)**.** *Seja* $(X_n)_{n\in\mathbb{N}}$ *uma sequência de variáveis aleatórias independentes, todas com a mesma esperança* μ *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 13. *Prove a Lei Fraca dos Grandes números utilizando a desigualdade de Chebyshev.*

Teorema 5 (Teorema Central do Limite)**.** *Seja $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias i.i.d. com esperança μ e desvio padrão σ. 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 2. *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.*

2.2.8 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 14. *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 15. *Seja X uma variável aleatória com distribuição normal padrão, ou seja, $X \sim N(0,1)$. Mostr que A função geradora de momentos de X é dada por por:*

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

A Ferramentas computacionais

A.1 Python

A.2 Poetry

A.3 Git

Referências

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. 2

Izbicki, R. and dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. 2

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. 2, 3

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. The MIT Press, 2nd edition. 2, 4

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning : from theory to algorithms*. 2, 3

Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra*. SIAM. 7