Aprendizado de máquinas

Thiago Rodrigo Ramos

16 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.

1 Introdução

Aprendizado de máquina é um termo utilizado para descrever sistemas capazes de identificar automaticamente padrões e regularidades em dados [SSBD14]. 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 [JWHT13], 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 [MRT18]:

- **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 $i = 1, \dots, n$ e $j = 1, \dots, p$.

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 [TB97].

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

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

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

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

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

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

A Ferramentas computacionais

A.1 Python

A.2 Poetry

A.3 Git

Referências

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. 3

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018. 4

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning : from theory to algorithms*. 2014. 3

[TB97] Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997. 7