# Aprendizado de máquinas

Thiago Rodrigo Ramos

16 de março de 2025

## Sumário

# 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.

# A Python

## Referências


[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. 2

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018. 3

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning : from theory to algorithms*. 2014. 2