Fundamentos de Machine Learning para Iniciantes

Ronald A. Mendonça

2025

Contents

0.1 Introdução

Machine Learning (ML), ou aprendizado de máquina, é uma área da inteligência artificial que permite que sistemas computacionais identifiquem padrões e façam previsões a partir de dados, sem a necessidade de programação explícita para cada tarefa. Diferente de softwares tradicionais, onde regras são definidas manualmente, em ML os algoritmos aprendem diretamente com exemplos, ajustando-se para melhorar seu desempenho ao longo do tempo. Essa capacidade transformou campos como medicina, finanças e tecnologia, tornando-se uma das competências mais valiosas do século XXI.

A relevância do Machine Learning está em sua aplicabilidade prática. Empresas utilizam ML para prever vendas, detectar fraudes e personalizar recomendações, enquanto cientistas o aplicam para analisar dados complexos, como sequências genéticas ou padrões climáticos. Para o iniciante, entender os fundamentos de ML abre portas para uma carreira em ciência de dados ou simplesmente para compreender melhor o mundo movido a dados em que vivemos.

Este eBook é voltado para iniciantes que desejam dar os primeiros passos em Machine Learning, sem experiência prévia na área. Para facilidade de entendimento, é recomendável um conhecimento básico de matemática (como médias e equações lineares) e familiaridade mínima com programação, preferencialmente em Python, embora os conceitos sejam explicados de forma acessível. Nosso objetivo é desmistificar o ML, oferecendo uma base teórica para quem quer iniciar a construir modelos simples e entender seus resultados.

Ao longo das próximas páginas, você será introduzido aos conceitos essenciais de Machine Learning, desde a manipulação de dados até a criação de modelos de regressão, classificação e clustering. Cada capítulo combina teoria com exemplos práticos, incluindo trechos de código em Python para ilustrar a aplicação dos métodos discutidos. Ao final, você terá conteúdo suficiente para construir um modelo de machine learning simples e entender os princípios que o sustentam.

Chapter 1

Conceitos Fundamentais: Dados e Aprendizado

1.1 Dados

Em um determinado dia no seu trabalho, você se depara com o desafio de extrair respostas de um conjunto de infromações que, a príncipio, não parecem fazer muito sentido para você. Estruturados ou não, podem ser números, textos, imagens ou qualquer outro tipo de informação. Isso é o que chamamos de dados. Os dados são a base de qualquer modelo de Machine Learning. São coletados, organizados e analisados para extrair informações valiosas.

- **Dados Qualitativos (ou Categóricos):** Representam características ou atributos que não podem ser quantificados numericamente.
  - **Nominais:** Não possuem ordem natural. Ex: cores (vermelho, azul, verde), tipos de frutas.
  - **Ordinais:** Possuem uma ordem ou hierarquia. Ex: níveis de satisfação (ruim, regular, bom, ótimo), ranking de qualidade (A, B, C).
- **Dados Quantitativos (ou Numéricos):** Representam valores numéricos e podem ser medidos.
  - **Discretos:** Assumem valores inteiros e finitos. Ex: número de filhos, número de carros em um estacionamento.
  - **Contínuos:** Podem assumir qualquer valor dentro de um intervalo. Ex: altura, peso, temperatura.

1.2 Importância da Compreensão dos Tipos de Dados

Entender os tipos de dados é essencial para escolher os algoritmos adequados e preparar os dados corretamente. Dados quantitativos podem ser usados diretamente em modelos numéricos, enquanto qualitativos e textuais requerem pré-processamento. Essa distinção será explorada em detalhes nos capítulos seguintes, onde veremos como transformar e utilizar esses dados em tarefas práticas de Machine Learning.

1.3 Tipos de Aprendizado

Os tipos de aprendizado de máquina podem ser classificados em três categorias principais: Supervisionado, Não Supervisionado e Aprendizado por Reforço.

1.3.1 Aprendizado Supervisionado (Supervised Learning)

O aprendizado supervisionado ocorre quando o modelo é treinado em um conjunto de dados rotulado, ou seja, onde a saída correta é conhecida. O objetivo do modelo é aprender um mapeamento entre as entradas e as saídas corretas.

- Previsão de preços de imóveis
- Diagnóstico médico
- Classificação de e-mails como spam ou não spam

1.3.2 Aprendizado não Supervisionado (Unsupervised Learning)

No tipo de aprendizado não supervisionado, os dados não são rotulados e o modelo deve identificar padrões e estruturas nos dados de forma autônoma.

- Segmentação de clientes
- Compressão de dados
- Detecção de anomalias

1.3.3 Aprendizado por Reforço (Reinforcement Learning)

O aprendizado por reforço se baseia em um agente que aprende interagindo com um ambiente e recebendo recompensas ou penalidades com base em suas ações.

- Jogos e Inteligência Artificial (xAI, OpenAI, DeepSeek)
- Controle de robôs

Aprendizado	**Descrição**	**Tipo de Dados**
Supervisionado	Dados rotulados	Qualitativos e Quantitativos
Não Supervisionado	Usa dados não rotulados para encontrar padrões	Quantitativos
Por Reforço	Aprendizado por tentativa e erro, com recompensas	Qualitativos e Quantitativos

Table 1.1: Comparação entre os tipos de aprendizado de máquina e os tipos de dados utilizados.

Chapter 2

Estatística Essencial para Machine Learning

2.1 Estatística Descritiva

A estatística descritiva é uma parte fundamental da análise de dados, pois fornece uma visão geral das características principais de um conjunto de dados. Ela envolve o resumo e a descrição dos dados por meio de medidas numéricas, gráficos e tabelas. As principais medidas incluem:

- **Média:** A média aritmética é a soma de todos os valores dividida pelo número total de valores.
- **Mediana:** O valor que separa os dados em duas metades, onde 50% dos dados estão abaixo e 50% estão acima.
- **Moda:** O valor que aparece com mais frequência em um conjunto de dados.
- **Variância:** Uma medida da dispersão dos dados em relação à média. A variância é calculada como a média dos quadrados das diferenças entre cada valor e a média.
- **Desvio Padrão:** Uma medida da dispersão dos dados em relação à média. Quanto maior o desvio padrão, mais espalhados estão os dados.
- **Correlação:** Uma medida que indica a força e a direção da relação entre duas variáveis. A correlação varia de -1 a 1, onde -1 indica uma correlação negativa perfeita, 0 indica nenhuma correlação e 1 indica uma correlação positiva perfeita.

2.1.1 Exemplo: Calculando a Média e o Desvio Padrão de Temperaturas Diárias

Suponha que registramos as temperaturas máximas diárias (°C) de uma semana em uma cidade. Os dados coletados são: 20, 22, 19, 23, 21, 20, 24. Vamos calcular a média e o desvio padrão passo a passo.

Cálculo da Média

A média ($\bar{x}$) é obtida somando todos os valores e dividindo pelo número de observações (n). Matematicamente:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

Para os dados fornecidos:

- Soma dos valores: $20 + 22 + 19 + 23 + 21 + 20 + 24 = 149$,
- Número de observações: $n = 7$,
- Média: $\bar{x} = \frac{149}{7} \approx 21.29$.

Portanto, a temperatura média diária é aproximadamente 21,29°C.

Cálculo do Desvio Padrão

O desvio padrão (σ) mede a variabilidade dos dados em relação à média. Para uma população, é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{2.2}$$

Passo a passo:

1. Calcule as diferenças entre cada valor e a média ($x_i - \bar{x}$):
   - $20 - 21.29 = -1.29$,
   - $22 - 21.29 = 0.71$,
   - $19 - 21.29 = -2.29$,
   - $23 - 21.29 = 1.71$,
   - $21 - 21.29 = -0.29$,
   - $20 - 21.29 = -1.29$,
   - $24 - 21.29 = 2.71$.
2. Eleve cada diferença ao quadrado:
   - $(-1.29)^2 = 1.6641$,
   - $0.71^2 = 0.5041$,
   - $(-2.29)^2 = 5.2441$,
   - $1.71^2 = 2.9241$,
   - $(-0.29)^2 = 0.0841$,
   - $(-1.29)^2 = 1.6641$,
   - $2.71^2 = 7.3441$.
3. Some os quadrados: $1.6641 + 0.5041 + 5.2441 + 2.9241 + 0.0841 + 1.6641 + 7.3441 = 19.4288$,
4. Divida pelo número de observações: $\frac{19.4288}{7} \approx 2.7755$,
5. Tire a raiz quadrada: $\sigma = \sqrt{2.7755} \approx 1.66$.

Assim, o desvio padrão das temperaturas é aproximadamente 1,66°C.

Interpretação

A média de 21,29°C indica a temperatura típica da semana, enquanto o desvio padrão de 1,66°C mostra que as temperaturas variam, em média, 1,66°C para mais ou menos em relação à média. Esses valores ajudam a entender a consistência do clima e podem ser usados em modelos de Machine Learning para previsões futuras.

Cálculo com Python

Os mesmos cálculos podem ser realizados de forma eficiente usando Python com a biblioteca NumPy. O código a seguir mostra como:

```
# Importando a biblioteca NumPy
import numpy as np

# Dados das temperaturas diárias
temperaturas = [20, 22, 19, 23, 21, 20, 24]

# Calcula a média
media = np.mean(temperaturas)
print(f"Média das temperaturas: {media:.2f}°C")

# Calcula o desvio padrão
desvio_padrao = np.std(temperaturas)
print(f"Desvio padrão das temperaturas: {desvio_padrao:.2f}°C")
```

Executando o código, obtemos uma média de 21,29°C e um desvio padrão de 1,67°C, valores consistentes com o cálculo manual (pequenas diferenças ocorrem devido a arredondamentos).

Visualização com Curva Normal

A distribuição das temperaturas pode ser visualizada em uma curva normal, que mostra como os dados se distribuem em torno da média. A Figura 2.1 apresenta o histograma das temperaturas e a curva normal correspondente, gerada com Python e as bibliotecas Matplotlib e SciPy.

Interpretação

A média de 21,29°C e o desvio padrão de 1,67°C indicam que as temperaturas variam moderadamente em torno de um valor central. A curva normal na Figura 2.1 ilustra essa distribuição, sendo útil em Machine Learning para entender a variabilidade dos dados antes de aplicar modelos preditivos.

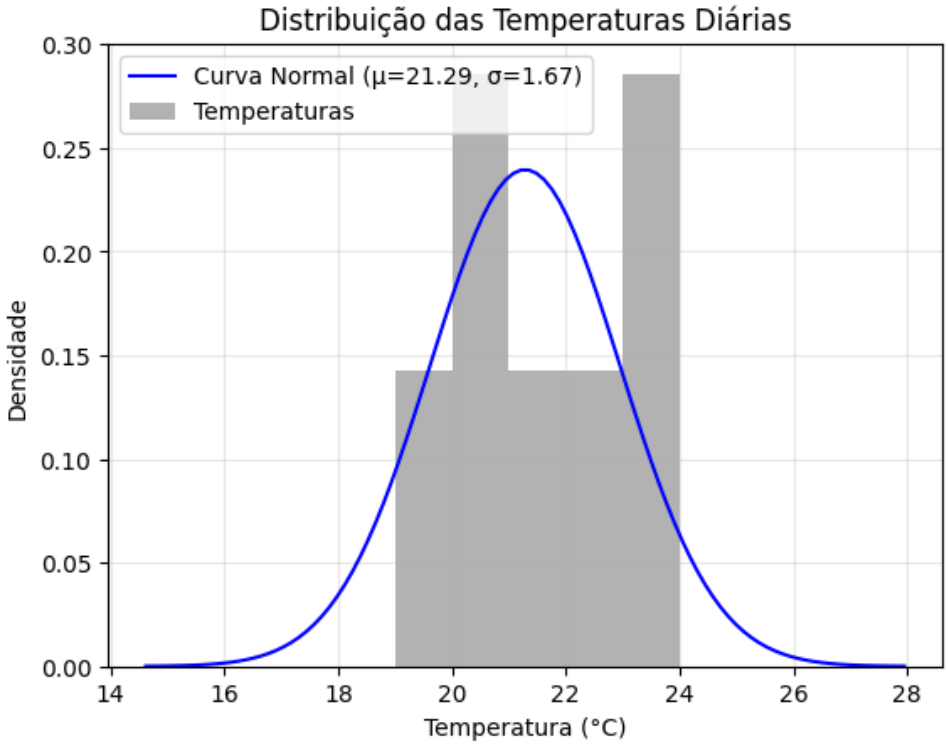

Figure 2.1: Distribuição das temperaturas diárias com curva normal (média = 21,29°C, desvio padrão = 1,67°C).