

# Introdução ao Aprendizado de Máquina

## 1 Introdução ao Aprendizado de Máquina

O aprendizado de máquina é uma subárea da inteligência artificial que se concentra no desenvolvimento de algoritmos que permitem que sistemas aprendam a partir de dados e melhorem seu desempenho ao longo do tempo. Existem três tipos principais de aprendizado de máquina: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

## 2 Tipos de Aprendizado de Máquina

### 2.1 Aprendizado Supervisionado

**Definição:** O aprendizado supervisionado envolve o uso de um conjunto de exemplos rotulados, onde cada exemplo é descrito por um conjunto de **características de entrada**  $\{X_1, X_2, \dots, X_m\}$  e um conjunto de **características-alvo**  $\{Y_1, Y_2, \dots, Y_k\}$ . O objetivo é aprender uma função que mapeia as características de entrada para as características-alvo.

**Aplicações:**

- **Predição de Atividades:** Um smartwatch pode prever a atividade do usuário (ex: dormir, caminhar) com base na frequência cardíaca e no movimento.
- **Previsão de Enchentes:** Prever a probabilidade de uma área inundar com base na topografia, clima e uso do solo.
- **Reconhecimento de Escrita:** Prever palavras que alguém escreveu à mão em um tablet com base na forma das letras.

### 2.2 Aprendizado Não Supervisionado

**Definição:** No aprendizado não supervisionado, o modelo é treinado em dados que não possuem rótulos. O objetivo é identificar padrões ou estruturas ocultas nos dados.

**Aplicações:**

- **Agrupamento de Clientes (Customer Segmentation):** Técnicas como K-means são usadas para segmentar clientes em grupos com características semelhantes.
- **Redução de Dimensionalidade:** Algoritmos como Análise de Componentes Principais (PCA) reduzem a quantidade de variáveis em um conjunto de dados, preservando as características mais importantes.
- **Detecção de Anomalias:** Identificar padrões incomuns em dados, que podem ser sinais de fraudes ou defeitos em processos industriais.

### 2.3 Aprendizado por Reforço

**Definição:** No aprendizado por reforço, um agente aprende a tomar decisões por meio de interações com o ambiente. O agente recebe recompensas ou punições com base em suas ações, aprendendo a maximizar as recompensas cumulativas ao longo do tempo.

**Aplicações:**

- **Robótica:** Treinar robôs para realizar tarefas complexas, como caminhar ou pegar objetos, por tentativa e erro.

- **Jogos:** Algoritmos como AlphaGo, que superam jogadores humanos em jogos de tabuleiro ou vídeo games.
- **Sistemas de Recomendação:** Adaptar continuamente as sugestões com base nas interações e preferências dos usuários ao longo do tempo.

### 3 Principais Tarefas do Aprendizado Supervisionado

O aprendizado supervisionado pode ser aplicado em várias tarefas principais, cada uma com seus próprios objetivos e métodos.

#### 3.1 Classificação

Na classificação, o objetivo é categorizar as entradas em uma ou mais classes discretas. Exemplos incluem:

- **Diagnóstico Médico:** Classificar pacientes como tendo ou não uma determinada doença com base em sintomas e exames.
- **Detecção de Spam:** Classificar emails como spam ou não-spam.

#### 3.2 Regressão

A tarefa de regressão envolve prever um valor contínuo para as entradas. Exemplos incluem:

- **Previsão de Preços:** Prever o preço de uma casa com base em características como tamanho, localização, etc.
- **Previsão de Vendas:** Estimar o número de vendas de um produto com base em campanhas de marketing e outras variáveis.

#### 3.3 Detecção de Anomalias

Detecção de anomalias é a tarefa de identificar entradas que se desviam significativamente do padrão esperado. Exemplos incluem:

- **Fraude em Cartões de Crédito:** Identificar transações que são possivelmente fraudulentas.
- **Monitoramento de Equipamentos:** Detectar falhas potenciais em equipamentos industriais antes que ocorram.

#### 3.4 Classificação Multiclasse e Multirrotulo

Além da classificação binária, onde as entradas são atribuídas a uma de duas classes, existem problemas mais complexos, como:

- **Classificação Multiclasse:** Onde cada entrada pode ser classificada em uma de várias classes (por exemplo, classificação de espécies animais).
- **Classificação Multirrotulo:** Onde cada entrada pode pertencer a várias classes simultaneamente (por exemplo, marcar tópicos relevantes em um artigo).

### 4 Avaliação de Predições

#### 4.1 Funções de Perda

Dado um conjunto de exemplos  $E$  e uma característica-alvo  $Y$ , seja  $\hat{Y}(e)$  a predição para o exemplo  $e$ . A **perda** é uma medida de quão próxima a predição está do valor real. Algumas funções de perda comuns são:

- **Perda 0-1 (L0):**

$$\text{loss}(p, a) = \begin{cases} 1 & \text{se } p \neq a \\ 0 & \text{se } p = a \end{cases}$$

- **Perda Absoluta (L1):**

$$\text{loss}(p, a) = |p - a|$$

- **Perda Quadrática (L2):**

$$\text{loss}(p, a) = (p - a)^2$$

- **Perda de Pior Caso ( $L_\infty$ ):**

$$\text{loss}_\infty = \max_{e \in E} |\hat{Y}(e) - Y(e)|$$

- **Verossimilhança:**

$$L(\theta) = \prod_{i=1}^n P(Y_i|X_i; \theta)$$

A verossimilhança é usada para estimar os parâmetros  $\theta$  do modelo que maximizam a probabilidade dos dados observados  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ . Em problemas de classificação, o objetivo é maximizar essa função de verossimilhança para encontrar os parâmetros do modelo.

- **Log-Verossimilhança:**

$$\log L(\theta) = \sum_{i=1}^n \log P(Y_i|X_i; \theta)$$

A log-verossimilhança é a versão logarítmica da função de verossimilhança. Ela é comumente utilizada porque transforma o produto de probabilidades em uma soma, o que facilita a derivação e a otimização durante o treinamento do modelo. Maximizar a log-verossimilhança é equivalente a maximizar a verossimilhança.

## 4.2 Erro Médio

O erro médio em um conjunto de dados  $E$  para uma predição  $\hat{Y}$  é dado por:

$$\frac{1}{|E|} \sum_{e \in E} \text{loss}(\hat{Y}(e), Y(e))$$

onde  $|E|$  é o número de exemplos em  $E$ .

## 5 Métricas de Avaliação de Classificadores

### 5.1 Taxa de Verdadeiros Positivos (Recall)

A taxa de verdadeiros positivos é definida como a proporção de positivos reais que foram corretamente classificados:

$$\text{Recall} = \frac{tp}{tp + fn}$$

### 5.2 Taxa de Falsos Positivos

A taxa de falsos positivos é definida como a proporção de negativos reais que foram incorretamente classificados como positivos:

$$\text{Taxa de Falsos Positivos} = \frac{fp}{fp + tn}$$

## 6 Áreas de Aplicação e Desafios

Os métodos de aprendizado supervisionado, não supervisionado e por reforço são amplamente aplicados em várias áreas como saúde, finanças, e segurança, mas enfrentam desafios como overfitting, necessidade de grandes volumes de dados e questões éticas relacionadas à privacidade e viés.

## 7 Avaliação de Modelos de Aprendizado Supervisionado

A avaliação de modelos de aprendizado supervisionado é crucial para garantir que o modelo generalize bem para novos dados, ou seja, dados que não foram usados durante o treinamento. Abaixo estão alguns dos principais conceitos e práticas para a avaliação de modelos:

### 7.1 Divisão de Conjuntos de Dados

Uma prática comum para avaliar modelos supervisionados é dividir o conjunto de dados em dois ou três subconjuntos: treinamento, validação, e teste.

- **Conjunto de Treinamento:** Usado para treinar o modelo, ou seja, ajustar os parâmetros do modelo de acordo com os dados.
- **Conjunto de Validação:** Usado para ajustar hiperparâmetros e evitar overfitting. A validação permite comparar diferentes configurações de modelos para selecionar a melhor.
- **Conjunto de Teste:** Usado para avaliar o desempenho do modelo em dados não vistos, garantindo que o modelo generalize bem.

Uma divisão típica pode ser 60% dos dados para treinamento, 20% para validação, e 20% para teste, mas isso pode variar dependendo do tamanho e da natureza do conjunto de dados.

### 7.2 Validação Cruzada (Cross-Validation)

Para garantir uma avaliação mais robusta, especialmente em casos onde o conjunto de dados é pequeno, é comum usar validação cruzada. Na validação cruzada, o conjunto de dados é dividido em várias partes (ou *folds*). O modelo é treinado em  $k - 1$  partes e testado na parte restante. Esse processo é repetido  $k$  vezes, e o desempenho final é a média dos resultados em todas as partes.

### 7.3 Métricas de Avaliação

Além da função de perda, existem várias métricas que podem ser usadas para avaliar a performance de modelos de aprendizado supervisionado:

- **Acurácia:** Proporção de previsões corretas sobre o total de exemplos.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

onde  $TP$  são os verdadeiros positivos,  $TN$  são os verdadeiros negativos,  $FP$  são os falsos positivos, e  $FN$  são os falsos negativos.

- **Precisão (Precision):** Proporção de previsões positivas corretas entre todas as previsões positivas feitas pelo modelo.

$$\text{Precisão} = \frac{TP}{TP + FP}$$

- **Revocação (Recall):** Proporção de positivos reais que foram corretamente identificados pelo modelo.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** A média harmônica entre a precisão e a revocação, útil para casos de classes desbalanceadas.

$$\text{F1-Score} = \frac{2 \cdot \text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$$

- **Erro Quadrático Médio (MSE):** Média dos quadrados das diferenças entre os valores previstos e os valores reais, usada principalmente em regressão.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

onde  $n$  é o número de exemplos,  $\hat{Y}_i$  é a predição para o  $i$ -ésimo exemplo, e  $Y_i$  é o valor real do  $i$ -ésimo exemplo.

- **Coeficiente de Determinação ( $R^2$ ):** Mede a proporção da variância nos dados dependentes que é explicada pelo modelo, também usada em regressão.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

onde  $\bar{Y}$  é a média dos valores reais.

## 7.4 Curvas ROC e AUC

Para problemas de classificação, especialmente quando há um desbalanceamento entre as classes, as curvas ROC (Receiver Operating Characteristic) e a métrica AUC (Area Under the Curve) são amplamente utilizadas. A curva ROC plota a taxa de verdadeiros positivos (Recall) contra a taxa de falsos positivos para diferentes limiares de decisão. O AUC mede a área sob essa curva, indicando a capacidade do modelo em distinguir entre as classes.

## 7.5 Overfitting e Underfitting

- **Overfitting:** Ocorre quando o modelo se ajusta muito bem aos dados de treinamento, capturando ruídos e padrões não generalizáveis, resultando em baixa performance no conjunto de teste.
- **Underfitting:** Ocorre quando o modelo é muito simples para capturar os padrões nos dados de treinamento, resultando em baixa performance tanto no treinamento quanto no teste.

Para evitar overfitting, técnicas como regularização (ex: L1, L2), poda de árvores (no caso de árvores de decisão), e métodos de ensemble (ex: Random Forest, Gradient Boosting) podem ser aplicadas.