

Atividade 01 - Conceitos Básicos de Machine Learning

Alexia Rocha Sales

24/03/2025

1. Explique, com suas palavras, o que é machine learning?

Machine learning é uma área da inteligência artificial que permite que computadores aprendam padrões a partir de dados sem que precisem ser programados explicitamente para cada tarefa. Os algoritmos de machine learning analisam grandes quantidades de informação, identificam padrões e fazem previsões ou decisões com base nesses padrões. Isso é amplamente utilizado em reconhecimento de imagens, análise de textos, recomendações de conteúdo e muitas outras aplicações.

2. Explique o conceito de conjunto de treinamento, conjunto de validação e conjunto de teste em machine learning.

Conjunto de treinamento: é a parte dos dados utilizada para treinar o modelo, ou seja, para ajustar os parâmetros do algoritmo e permitir que ele aprenda padrões.

Conjunto de validação: é usado para avaliar o desempenho do modelo durante o treinamento e ajudar a ajustar hiperparâmetros, evitando problemas como overfitting (quando o modelo aprende padrões específicos demais e não generaliza bem para novos dados).

Conjunto de teste: é um conjunto separado que avalia o desempenho final do modelo depois que ele foi treinado e validado. Esse conjunto simula como o modelo se comportará com dados nunca antes vistos.

3. Explique como você lidaria com dados ausentes em um conjunto de dados de treinamento.

Lidar com dados ausentes é essencial para garantir que o modelo não tenha problemas na fase de treinamento. Algumas abordagens comuns incluem:

Remoção de registros incompletos: quando a quantidade de dados ausentes é pequena e sua exclusão não compromete o conjunto de dados.

Preenchimento com valores médios, medianos ou modais: quando os dados ausentes são numéricos, pode-se substituí-los por média ou mediana. No caso de variáveis categóricas, pode-se usar a moda (valor mais frequente). Modelos de imputação: algoritmos como k-Nearest Neighbors (k-NN) ou regressão podem prever valores ausentes com base nos demais dados.

Uso de indicadores binários: criar uma nova variável que indique se um valor estava ausente ou não, ajudando o modelo a capturar padrões.

4. O que é uma matriz de confusão e como ela é usada para avaliar o desempenho de um modelo preditivo?

Uma matriz de confusão é uma tabela que resume o desempenho de um modelo de classificação, comparando previsões feitas pelo modelo com os valores reais. Ela é organizada em quatro categorias:

- Verdadeiros Positivos (VP): quando o modelo previu corretamente a classe positiva.

- Falsos Positivos (FP): quando o modelo previu a classe positiva incorretamente.
- Verdadeiros Negativos (VN): quando o modelo previu corretamente a classe negativa.
- Falsos Negativos (FN): quando o modelo previu incorretamente a classe negativa.

A matriz de confusão é usada para calcular métricas como precisão, revocação e F1-score, ajudando a entender onde o modelo está errando e como ele pode ser melhorado.

5. Em quais áreas (tais como construção civil, agricultura, saúde, manufatura, entre outras) você acha mais interessante aplicar algoritmos de machine learning?

Machine learning pode ser aplicado em diversas áreas, mas algumas que considero mais interessantes incluem:

- Saúde: diagnóstico de doenças com base em imagens médicas, previsão de surtos de doenças e personalização de tratamentos.
- Agricultura: monitoramento de plantações por imagens de satélite, detecção de pragas e previsão de produtividade.
- Astronomia: identificação de exoplanetas, previsão de fenômenos espaciais e classificação de galáxias.
- Ciência de materiais: descoberta de novos materiais com propriedades específicas, como novos supercondutores e materiais para energia renovável.
- Física: análise de grandes volumes de dados experimentais, como aqueles gerados por aceleradores de partículas.