Bootcamp Avanti – Machine Learning

Atividade 1

Aluno: José Lucas da Silva Pinheiro

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

É uma subárea da Inteligência Artificial que tem o objetivo de simular o aprendizado humano através do uso de dados e algoritmos. A partir desse aprendizado o sistema deve ser capaz de tomar decisões sem a necessidade de programá-las para cada tarefa específica.

O algoritmo é responsável por identificar padrões nos dados analisados e posteriormente realizar tarefas que podem ser de classificação ou regressão, por exemplo.

Por exemplo, um sistema para identificar a presença de cachorros em uma imagem. Para esse modelo deve ser utilizado um conjunto de dados com imagens que contenham cachorros e outro conjunto com imagens que não contenham cachorros.

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

Conjunto de treinamento: é o conjunto de dados utilizado para treinar o modelo. A partir dele que o algoritmo irá aprender os padrões mais relevantes. Usando o exemplo anterior, seria possível separar cerca de 70% dos dados para treinamento, onde seria metade das imagens contendo cachorros e a outra metade não, que essas seriam as duas classes.

Conjunto de validação: é um conjunto separado dos dados que é utilizado para ajustar determinados parâmetros do modelo e monitorar o desempenho dele durante o treinamento. Por exemplo, dos 30% restantes da base de dados, poderia ser utilizado metade, 15%, para a validação.

Conjunto de teste: é o conjunto utilizado no final para testar e avaliar o desempenho do modelo. Lembrando que esse conjunto não deve conter a mesma porção de dados utilizados para o treinamento e validação. Utilizando o mesmo exemplo, o conjunto de teste seria composto pelos 15% restantes da base de dados.

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

Dependendo da quantidade, remover as linhas ou colunas com dados ausentes de forma que não interfira significativamente o tamanho do conjunto de dados. Outra opção seria substituir os valores ausentes pela mediana dos dados na determinada coluna.

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

A matriz de confusão é uma ferramenta que avalia o desempenho do modelo. Ela apresenta de forma visual a distribuição das previsões feitas pelo modelo para cada caso real testado. A partir desse resultado apresentado é possível calcular várias métricas, como acurácia, precisão e recall.

A matriz de confusão possui um formato NxN, onde N é o número de classes.

		Valor predito	
		Sim	**Não**
Real	**Sim**	**Verdadeiro Positivo (TP)**	**Falso Negativo (FN)**
	Não	**Falso Positivo (FP)**	**Verdadeiro Negativo (TN)**

TP: Modelo previu a classe positiva corretamente.
FN: Modelo previu a classe negativa, mas a observação era da classe positiva.
FP: Modelo previu a classe positiva, mas a observação era da classe negativa.
TN: Modelo previu a classe negativa corretamente.

Podemos exemplifcar:

		Valor predito	
		Cachorro	**Não cachorro**
Real	**Cachorro**	**40 (TP)**	**10 (FN)**
	Não cachorro	**8 (FP)**	**42 (TN)**

TP: O modelo identificou corretamente 40 imagens como cachorros.
FN: 10 imagens foram classificadas incorretamente como não cachorros.
FP: 8 imagens foram classificadas incorretamente como cachorros.
TN: O modelo indentificou corretamente 42 como não cachorros.

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 aplicada nas mais diversas áreas. Mas dentre as citadas, acho bastante interessante a aplicação na agricultura e na saúde.

Na agricultura, como:

- Detecção de doenças: utilizando visão computacional e a análise de imagens tiradas por drones, por exemplo, seria possível detectar sinais de doenças nas plantas.

- Classificação de frutas ou vegetais: através de algoritmos de visão computacional é possível classificar frutas e vegetais de acordo com seu tamanho, cor e presença de defeitos.

Na saúde:

- Detecção de câncer: analisando imagens médicas, como radiografias e tomografias, para identificar sinais precoces de câncer.