Avanti Bootcamp - Atividade 1

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

Aprendizado de máquina ou Machine Learning (ML) é uma subárea da Inteligência Artificial que se caracteriza como uma conjunto de técnicas e algoritmos que em sua construção visam tornar o computador uma entidade inteligente, capaz de tomar decisões próprias e aprender com estas. A principal diferença das técnicas desta área para técnicas convencionais de programação é a presença de dados para realizar tal aprendizado, enquanto em outras utilizam uma programação pragmática com instruções claras sobre o que a máquina deve executar, aqui, são usados conjuntos de dados e fórmulas matemáticas para realizar o treinamento de modelos, que ao fim generalizarão tais dados, e, assim, consiguirão aprender com estes e tomar as próprias decisões sem a necessidade de uma intervenção direta de um programador. Por exemplo: Em um sistema de cadastro de produtos um desenvolvedor especifica cada passo e decisão que este deve tomar, enquanto em um sistema de classificação de bananas, o computador utiliza um modelo que já foi previamente treinado para identificar e classificar a banana em uma imagem.

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

- Conjunto de treinamento:

  É um conjunto de dados, que pode ser composto por imagens ou dados de texto, retirado de um dataset que será utilizado para a construção de um modelo. A finalidade deste é o treinamento do modelo, por isso carrega este nome, ou seja, o modelo utilizará os dados presentes nele para aprender informações importantes para a realização da sua tarefa. O dataset original é dividido de forma que parte dele fique destinado para este conjunto de treinamento. Esta divisão pode ocorrer de várias formas, contudo as técnicas mais utilizadas são a *train_test_split* e a *kfold*. A primeira técnica divide o dataset em subconjuntos conforme porcentagens passadas. Já a segunda técnica é mais usada quando o dataset utilizado não possui muitos dados disponíveis, nela, o dataset é dividido em *folds* e a cada iteração um fold é escolhido para teste enquanto o restante fica disponível para treinamento do modelo.

- Conjunto de validação:

É um conjunto de dados, que tal qual o anterior mencionado, é construído a partir do dataset usado no estudo. Além disso, sua composição segue conforme o conjunto de treinamento, podendo ser composto por dados de imagens ou textos. Tem como finalidade o ajuste de hiperparâmetros do modelo. É geralmente construído a partir de dataset que contém um grande volume de dados.

- Conjunto de teste:

  É a parcela do dataset utilizada para validar o aprendizado do modelo. É preciso avaliar o desempenho do modelo e assim validar que seu aprendizado foi, no mínimo, satisfatório. Dessa forma, é retirado do dataset uma parcela variada de dados, contendo todos os tipos de dados, para que, ao fim do treinamento, submeta-se o modelo a testes. Os testes são conduzidos passando dados que o modelo nunca tenha visto, ou seja, que não foram usados em seu treinamento ou em sua validação, para que este realize suas predições. Estas, por fim, são comparadas com os dados presentes no conjunto de teste e vistos sua validade.

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

Acredito que a primeira coisa seria avaliar a situação da base e dos dados ausentes, observando a sensibilidade e a importância deste. Após isso avaliaria qual técnica seria a melhor para resolver o problema, se seria a remoção da linha contendo o dado ausente, a inserção da média, mediana ou do valor mais frequente no lugar ou até mesmo outras técnicas.

04. 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 tabela que expõe a quantidade de instâncias classificadas pelo modelo em cada classe. Nas colunas e linhas ficam todas as classes avaliadas. As classes das colunas representam aquelas preditas pelo modelo e as das linhas aquelas que sabemos que as instâncias pertencem, ou seja, as verdadeiras. Para ser preenchida compara-se a classe predita com a classe verdadeira de uma instância. Para compreendermos melhor, suponha um modelo que classifica imagens de carros e motos, onde ‘0’ representa a classe carros e ‘1’ representa a classe motos. Dessa forma, a matriz de confusão teria 2 linhas e duas colunas.

Ao realizar a comparação se o modelo classifica uma instância como ‘0’ e realmente a sua classe é ‘0’, então a célula (0,0) da matriz seria incrementada em uma unidade. Entretanto, se o modelo classifica a instância como ‘0’, mas na realidade sua classe é a ‘1’, então a célula (1, 0) seria incrementada em uma unidade. Já se o modelo classifica como ‘1’

mas na verdade a classe desta instância é ‘0’ a célula incrementada é a (0, 1). Por fim, se o modelo classifica como ‘1’ e realmente a instância é ‘1’, então a célula incrementada é a (1,1).

Assim, temos que a diagonal principal da matriz representa os acertos do modelo e as outras células representam os erros. Cada célula da matriz possui nomenclatura própria: Verdadeiro positivo, célula (0,0); Falso Positivo, célula (1,0); Falso Negativo, célula (0, 1) e Verdadeiro Negativo, célula (1, 1,). Da matriz de confusão, então, é possível calcular métricas de avaliação do modelo como a acurácia, o recall, o f1-score e a precisão.

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

Acredito que toda área que tenha a infraestrutura para desenvolver esse tipo de sistema, ou seja, dados, é interessante. Mas, particularmente, vejo um grande potencial na agricultura, pois as possibilidades de aplicação dessas técnicas são enormes e ainda estão começando a serem exploradas. Além disso, acho que a área espacial é interessante, não só porque ela já detém enormes quantidades de dados, mas justamente porque sistemas desses são necessários, visto a ausência de humanos fora da terra.