Atividade 1 - Machine Learning - Vinicius Moreira Maia

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

Machine Learning é uma área da Inteligência Artificial que se preocupa em utilizar recursos computacionais para a criação de ferramentas que aprendem padrões diversos quando submetidas ao treinamento a partir de uma massa de dados produzida tanto pelo homem, como por outras máquinas e até mesmo fenômenos naturais. A partir do reconhecimento dos padrões, estas ferramentas, já “educadas”, servem para realizar uma série de computações, tais como previsões e classificações, que podem nos auxiliar na tomada de decisões de negócios, governamentais, de saúde e assim por diante.

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

- Conjunto de Treinamento refere-se ao conjunto de dados históricos utilizados como entrada dos algoritmos de Machine Learning. Dados históricos referem-se aos dados provenientes de ambientes de bases de dados analíticas (também chamados, muitas vezes, de ambientes OLAP) que, diferente do que ocorre em bases de dados transacionais, não há a exclusão de dados. Portanto, dados de treino são a entrada do algoritmo que produz o modelo de ML, e geralmente representam a maior porção dos dados disponíveis.

- O Conjunto de Validação é como se fosse uma camada intermediária entre o treinamento e a avaliação, no sentido de que o modelo produzido com o conjunto de treinamento opera bem com este conjunto de dados em específico, e para ajustar de maneira mais precisa os parâmetros, é preciso apresentar ao modelo dados que não foram utilizados no treinamento.

- O Conjunto de Testes é uma massa de dados utilizada como entrada dos modelos já prontos, simulando situações preditivas reais.

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

Primeiro eu definiria alguma métrica para avaliar a proporção de dados ausentes em uma coluna, pensando aqui em dados estruturados. Em seguida, eu definiria uma regra para ou substituir os dados ausentes com um valor indicativo de ausência, ou até mesmo excluir toda a coluna da base. Tudo dependerá do controle definido nas métricas e nas regras.

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

Matriz de Confusão é como se fosse um sumário de execução de um modelo, utilizada para averiguar e quantificar sua acurácia. Por exemplo, em um modelo de classificação eu preciso reconhecer, em uma massa de imagens, todas aquelas que tem a representação de um gato. A Matriz entra justamente para avaliar a quantidade de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

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

Eu sou bolsista de análise de dados atualmente em um banco. Não sou exatamente da área de ciência de dados ou de machine learning (ainda), mas a utilização que eles fazem de modelos de ML no ambiente de riscos muito me intriga e desperta minha curiosidade. Em tudo relacionado à concessão de empréstimos, análises de crédito e coisas do tipo, eles utilizam modelos de ML para a tomada de decisões baseada em dados. Portanto acredito que modelos bem treinados e avaliados são ferramentas poderosas nas tomadas de decisões quando se trata da utilização do dinheiro público.