1.2 Introdução Teórica - Ciência de Dados

1. Qual a diferença entre aprendizagem supervisionada e não supervisionada? Explique um pouco de cada.

   Resposta: Na aprendizagem supervisionada, nós temos a saída de cada instância analisada, ou seja, temos a "classe" (label ou rótulo) à qual aquela instância pertence, no caso de problemas de classificação, ou o valor o qual desejamos prever, no caso de problemas de regressão, e usamos os dados para tentar prevê-las. Na aprendizagem não supervisionada, nós não temos a saída de cada instância, então precisamos utilizar de técnicas diferentes para as análises e previsões.

2. O que você entende por validação cruzada?

   Resposta: A validação cruzada(cross validation) é uma técnica que ajuda a melhorar e avaliar a generalização de um modelo. Por exemplo, o k-Fold é um método de cross validation onde dividimos o nosso conjunto de dados em k subconjuntos e depois alternamos, utilizando um subconjunto para teste e k - 1 subconjuntos para treino, sendo esse processo repetido k vezes, alternando os subconjuntos a cada iteração de maneira circular até que todos os subconjuntos tenham sido utilizados para teste.

3. O que é uma regressão linear?

   Resposta: Com relação à machine learning, a regressão linear é uma técnica baseada no aprendizado supervisionado que tenta usar pontos dos dados para encontrar a melhor linha de ajuste possível para modelar os mesmos e pode ser usada para descobrir relações entre variáveis e realizar previsões. Seu objetivo é encontrar a equação de uma linha de regressão que minimize o erro à partir de uma função de custo(normalmente é utilizada a média dos erros quadrados) da diferença entre o valor verdadeiro e o valor que foi previsto.

4. O que são outliers? Apresente, ao menos, uma forma de lidar.

   Resposta: Os outliers são "pontos fora curva", ou seja, valores que extremos que se desviam daquilo que se observa sobre os dados e pode indicar um erro na coleta dos mesmos. Para lidar com outliers, podemos simplesmente removê-los do nosso conjunto de dados ou usar técnicas de clusterização para tentar achar uma aproximação que possa "corrigir" o valor.