PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Instituto de Ciências Exatas e Informática
Curso de Ciência da Computação - Coração Eucarístico
Profa.: Camila Laranjeira - **mila.laranjeira@gmail.com**
Disciplina: Inteligência Artificial / 1o Semestre de 2022

Aluna(o):	Gustavo Lopes Rodrigues

Lista 06 - Aprendizado de Máquina

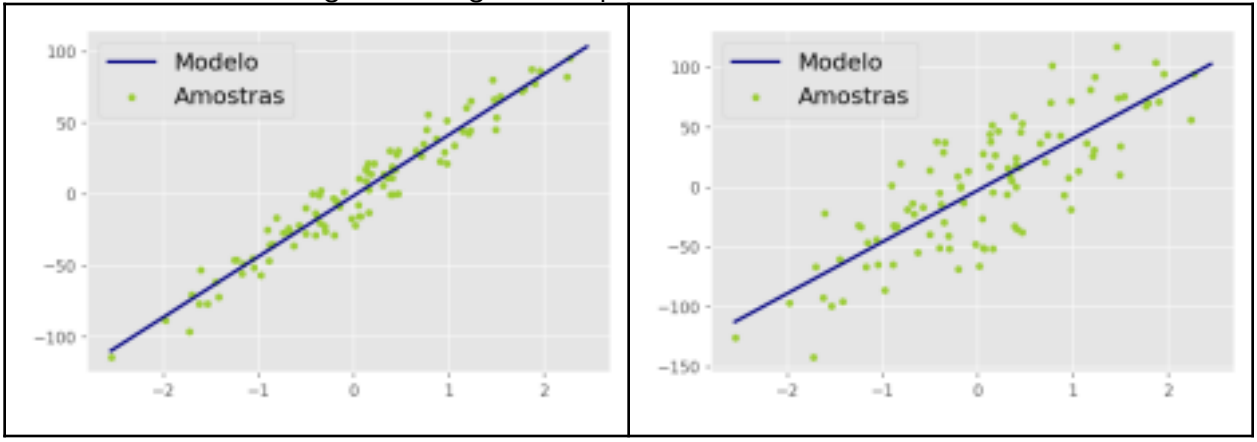
1. Defina em poucas palavras os três principais problemas de aprendizado de máquina: classificação, regressão e clusterização. Forneça exemplos hipotéticos para os três problemas (o problema nem os dados precisam existir).

A “Classificação” é onde temos uma fronteira de divisão para a previsão de classes, por exemplo quando queremos separar pessoas infectadas e não infectadas com covid. “Regressão” é quando transformamos a distribuição de dados em uma linha que faz previsão em uma variável, por exemplo os preços no mercado de valores. “Clusterização” é quando temos que prever múltiplas estruturas de agrupamento, como por exemplo queremos prever a classe birads de um exame de radiografia da mama.

2. Suponha que você queira criar um modelo para filtragem de spam. Proponha uma solução para esse problema em termos de tipo de modelo (classificação, regressão e agrupamento) e tipo de supervisão (não-supervisionado, semi supervisionado, totalmente supervisionado) e justifique as suas escolhas. Por exemplo: quais os atributos a serem preditos? Se supervisionado, de onde viriam os rótulos? Etc.

Para solucionar esse problema, eu usaria um modelo do tipo classificação, pois queremos apenas dividir o problema em duas classes, as mensagens spams e não spams. O tipo de supervisão seria totalmente supervisionado, pois precisaríamos alimentar ao modelo exemplos de email que são spams e não spams e logo este seriam os rótulos: Spams e Não spams. Para fazer a previsão, os atributos seriam todas as palavras contidas no email (título e conteúdo) e então iríamos prever a probabilidade do email em questão ser spam, por conter a palavra X.

3. Considere as duas figuras a seguir e responda.



a) Que tipo de modelo está sendo ajustado?

Está sendo ajustado um modelo de regressão

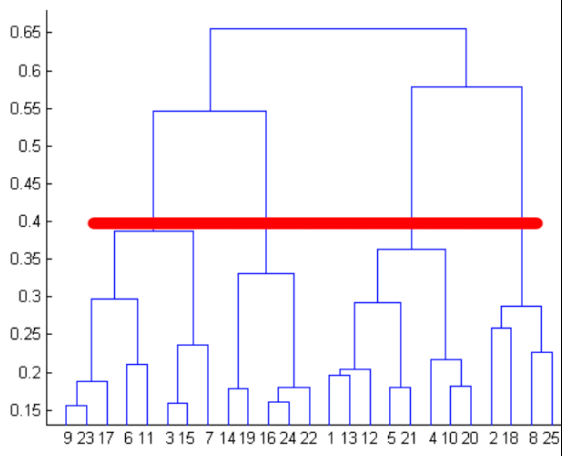
b) Como podemos medir o erro dos modelos apresentados? E qual das distribuições (esq. ou dir.) apresenta o maior erro de acordo com essa métrica? Justifique.

Podemos medir o erro dos modelos a partir da distância do ponto mais distante da reta com a tal reta. Usando essa métrica, a distribuição da direita apresenta o maior erro, pois o ponto mais distante da reta da direita tem uma distância maior do que o ponto mais distante da reta da esquerda.

4. Considere um processo de uso de um conjunto de teste e um conjunto de treino para conduzir as iterações do desenvolvimento do modelo. Em cada iteração, treinamos os dados de treino e avaliamos os dados de teste, usando os resultados da avaliação para orientar escolhas e alterações em vários hiperparâmetros do modelo, como taxa de aprendizado e recursos. Há algo de errado com esta abordagem? Justifique.

O que está errado com esta abordagem é o fato que o treinamento não acontece com todos os dados de treino e teste já em mãos, pois isso pode tornar difícil de encontrar os melhores hiperparâmetros gerais, em vez disso, o que será encontrado será os melhores hiperparâmetros para cada dado de treino separado.

5. Para o dendrograma ao lado, que representa o resultado de um agrupamento aglomerativo, use sua intuição para definir a quantidade de clusters do resultado final. Marque o corte na imagem ao lado e justifique sua resposta abaixo.



O número de clusters no final ficará de 4, o corte na imagem mostrada ficará de tal forma, pois isso irá evitar overfit (muitas classes), e também irá evitar underfit (poucas classes).

6. Execute uma única iteração do K-Means para a distribuição abaixo, que consiste em seis pontos, sendo os pontos 5 e 6 os centróides iniciais. Preencha a tabela abaixo indicando quais pontos pertencem a cada cluster e onde estarão os centróides após uma iteração.

Cluster	Pontos	Centro
1		x=7, y=3
2	2,3,4	x=2, y=3,3