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):	Lucas Santiago de Oliveira

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).

Modelos de classificação são usados para prever em qual grupo um novo elemento se encontra. É necessário ter todos os rótulos de um grupo grande de dados previamente para conseguir identificar a qual grupo um novo elemento pertence. Por exemplo, prever se uma imagem é de um pássaro ou de um avião com um conjunto de dados de treino com várias fotos de ambos.

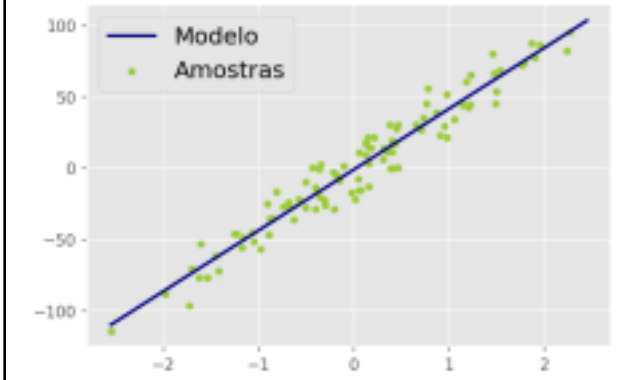
Modelos de regressão são baseados em encontrar uma função que possa prever com certa acurácia onde estará o próximo dado. Por exemplo, prever qual será a quantidade de emissão de carbono gerada por um novo motor baseando-se na quantidade de cilindros e na potência do motor.

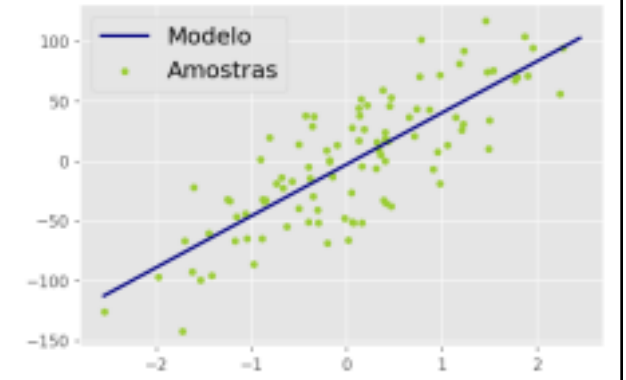
Clusterização é um modelo baseado em agrupar valores próximos em grandes grupos (clusters). Por exemplo, agrupar fotos parecidas com base em suas características semelhantes sem conhecimento prévio de nenhum rótulo.

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.

Modelos baseados em classificação seriam ótimos. Poderia permitir que os usuários da plataforma rotulassem (totalmente supervisionado) e-mails que eles consideram como spam e aqueles emails não rotulados seriam considerados como não spam. Com isso, seria possível treinar um modelo de tempos em tempos que procurasse por elementos semelhantes entre todos os emails marcados. Os atributos usados no treinamento seriam palavras ou conjunto de palavras (não necessariamente frases inteiras) para descobrir quais são as expressões mais usadas em spams.

3. Considere as duas figuras a seguir e responda.





a) Que tipo de modelo está sendo ajustado?

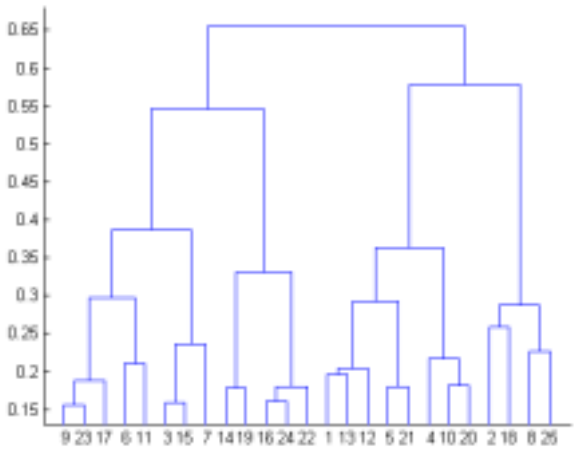
Regressão Linear

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.

Uma forma de calcular o erro desse modelo é usando um cálculo de distância entre o ponto real da linha que foi predita. Com isso, o segundo modelo possui um erro maior que o primeiro, pois quando um novo ponto for adicionado no modelo a chance dele estar próximo da linha predita será menor do que o modelo da esquerda. Dessa forma, para não se ter erros seria necessário que a linha passasse por cima de todos os pontos - em um cenário real isso seria quase impossível, os dados que temos que trabalhar nunca são tão simples assim :P -. Considerando que não é comum encontrar um cenário com poucos erros, encontrar poucos erros pode significar um overfitting do modelo.

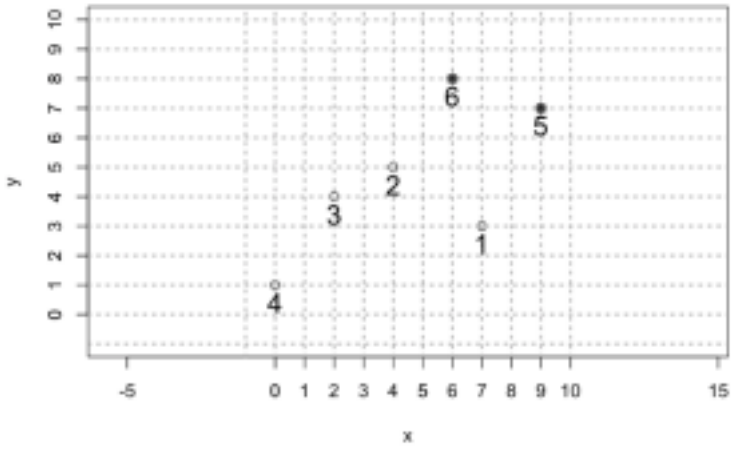
4. Considere um processo de uso de um conjunto de teste e um conjunto de treino para conduzir as iterações do desenvolvimento do modelo. Em cada iteração, treinamos os dados de treino e avaliamos os dados de teste, usando os resultados da avaliação para orientar escolhas e alterações em vários hiperparâmetros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

Essa abordagem está correta. A ideia da separação dos dados de treinamento e teste é feita para termos uma base de teste sólida e realista, se treinamos uma IA com uma parte do banco de dados e na hora de predizer alguns valores verificamos esses dados preditos com os valores do teste, teremos, então, um dado real que pode mostrar qual o nível de acerto do modelo que construímos. Só há um ponto importante para ser destacado, é sim interessante alterar os hiperparâmetros para encontrar o melhor modelo, mas um ajuste excessivo desses hiper parâmetros pode fazer um overfitting do modelo. Esse problema pode acarretar em um modelo treinado para acertar excessivamente os dados de teste e errar bastante os dados reais num cenário fora dos dados que já possuímos no nosso banco de teste e treino.

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.	

(Não consegui desenhar no dendograma então vou escrever :-))
A princípio eu cortaria o dendograma no ponto 0.45, pois a partir desse ponto dá para notar que todos os outros pontos que estavam nos dados originais, não estão próximos. Uma maior distância entre ligações do dendograma representa que os dados originais estão distantes entre si. Com isso, a melhor posição de corte será onde tiver poucos clusters (tiver poucas linhas sobrando, pois todas as outras já foram unidas) e linhas muito grandes no eixo Y (isso representa a distância entre os clusters).

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	1, 5	(8, 5)	
2	4, 3, 2, 6	(3, 4.5)	