MAE 5905: Introdução à Ciência de Dados

Prova 2. Primeiro Semestre de 2023. Entregar em 11/07/2023.

1. (3,0 pontos). Considere o conjunto de dados **Boston** do pacote MASS, contendo $n = 506$ amostras e $p = 14$ variáveis. Considere o conjunto de treinamento contendo as primeiras 253 amostras e o conjunto teste contendo as amostras restantes. Ajuste uma árvore de regressão, considerando a variável **medv** como resposta.

(a) Ajuste um modelo de árvore aos dados de treinamento. Verifique se é necesário podar a árvore.
(b) Use a árvore não podada para fazer previsões para o conjunto teste. Calcule o EQM.
(c) Use bagging, florestas e boosting e comente sobre o melhor ajuste.

2. (4,0 pontos) Considere o conjunto de dados **OJ** do pacote **ISLR**.

(a) Criar um conjunto de treinamento contendo uma amostra de 800 observações e um conjunto teste contendo as observações restantes.
(b) Ajuste um classificador SVM ao conjunto de treinamento usando **cost=0.01**, tendo **Purchase** como resposta e as outras variáveis como preditoras. Use a função **summary()** e descreva os resultados obtidos.
(c) Quais são as taxas de erros de treinamento e de teste?
(d) Use a função **tune()** para selecionar um **cost** ótimo. Considere valores no intervalo 0.01 a 10.
(e) Calcule as taxas de erro para este novo valor de **cost**.
(f) Repita (b)-(e) usando SVM com kernel radial. Use o valor default para **gamma**.
(g) Repita (b)-(e) com um kernel polinomial com **degree=2**.
(h) Qual procedimento parece dar os melhores resultados para esses dados?

3. (3,0 pontos) Considere o conjunto de dados **food-texture**, que pode ser encontrado em openmv.net/info/food-texture. Os dados estão no formato csv. Leia com atenção o significado de cada variável.

(a) Faça uma análise de componentes principais (ACP). Escreva cada CP como função das variáveis originais. Tente interpretar cada componente que você vai reter. Faça os gráficos apropriados.

(b) Faça uma análise fatorial (AF) com dois fatores. Para isso, considere AF em três situações: sem rotação, com rotação varimax e com rotação promax.

(c) Faça os gráficos apropriados e comente sobre qual rotação é mais apropriada para melhor interpretar os fatores.

(d) Faça uma análise de componentes independentes (ACI). Escreva cada CI como função das variáveis originais. Tente interpretar cada componente que você vai reter.