MAE 5905: Introdução à Ciência de Dados

Prova 1. Primeiro Semestre de 2023. Entregar 19/05/2023.

1. (a) Considere o caso de duas populações exponenciais, uma com média 1 e outra com média 0,5. Supondo $\pi_1 = \pi_2$, encontre o classificador de Bayes. Quais são as probabilidades de classificação incorreta? Construa um gráfico, mostrando a fronteira de decisão e as regiões de classificação em cada população. Generalize para o caso das médias serem $\alpha > 0$ e $\beta > 0$, respectivamente.

(b) Simule 200 observações de cada distribuição exponencial da parte (a). Usando os dados para estimar os parâmetros, supostos agora desconhecidos, obtenha o classificador de Bayes, a fronteira de decisão e as probabilidades de classificação incorreta com a regra obtida no exercício anterior. Compare os resultados com aqueles obtidos no item (a).

2. Considere os dados do arquivo **disco** e a variável resposta $y = 1$ se o disco estiver deslocado e $y = 0$, caso contrário. Use a função discriminante linear de Fisher para obter um classificador. Tome o conjunto de treinamento aquele contendo as primeiras 80 observações e o conjunto de teste contendo as demais 24 observações. Obtenha um classificador tendo como variável preditora a distância aberta e outro tendo como preditores as duas distâncias. Use a função **lda()** do pacote **MASS**. Interprete os resultados e escolha o melhor classificador usando a acurácia como base. Obtenha a sensibilidade e especificidade de cada classificador.

3. Use o mesmo conjunto de dados do problema anterior e distância aberta como variável preditora. Use LOOCV e o classificador KNN, com vizinhos mais próximos de 1 a 5.

(a) Qual o melhor classificador baseado na acurácia?

(b) obtenha a matriz de confusão e realize o teste de McNemar.

(c) Obtenha a sensibilidade e a especificidade e explique seus significados nesse caso

4. O conjunto de dados **Auto** do pacote **ISLR** contém as seguintes variáveis:

mpg: miles per gallon
cylinders: Number of cylinders between 4 and 8
displacement: Engine displacement (cubic inches)
horsepower: Engine horsepower
weight: Vehicle weight (lbs.)
acceleration: Time to accelerate from 0 to 60 mph (sec.)
year: Model year (modulo 100)
origin: Origin of car (1. American, 2. European, 3. Japanese)
name: Vehicle name

(a) Divida os dados em conjunto de treinamento(S) e conjunto de teste (T).

(b) Ajuste um modelo aos dados de S tendo **horsepower** como preditor e **mpg** como resposta. Obtenha os EMQ de treinamento e faça o diagnóstico do modelo. O que você nota no gráfico dos resíduos contra valores ajustados? Obtenha o EQM de teste.

(c) Agora inclua (**horsepower**)^2 no modelo e proceda como no item (b). Qual modelo você escolheria? Justifique.

(d) Ajuste um modelo de regressão **ridge** aos dados de S, tendo **mpg** com resposta e **displacement**, **horsepower**, **weight** e **acceleration** como preditores, com λ escolhido por VC. Obtenha o EQM de teste.

(e) Ajuste um modelo de regressão **lasso** e proceda como em (d). Quais coeficientes foram zerados?

(f) Comente sobre os resultados obtidos em (d) e (e), baseados no R^2 e EQMI.