MAE 5905: Introdução à Ciência de Dados

Prova 1. Primeiro Semestre de 2024. Entregar 02/05/2024.

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. Simule um conjunto de dados com $n = 500$ e $p = 2$, tal que as observações pertençam a duas classes com uma fronteira de decisão não linear. Por exemplo, você pode usar:

$>$ x1=runif(500)-0.5
$>$ x2=runif(500)-0.5
$> y = 1 * (x1 \wedge 2 - x2 \wedge 2 > 0)$.

(a) Faça um gráfico das observações, com símbolos (ou cores) de acordo com cada classe.

(b) Separe os dados em conjunto de treinamento e de teste. Obtenha o classificador de margem máxima, tendo X_1 e X_2 com preditores. Obtenha as previsões para o conjunto de teste e a acurácia do classificador.

(c) Obtenha o classificador de margem flexível, tendo X_1 e X_2 com preditores. Obtenha as previsões para o conjunto de teste e a taxa de erros de classificação.

(d) Obtenha o classificador de margem não linear, usando um kernel apropriado. Calcule a taxa de erros de classificação.

(e) Compare os dois últimos classificadores.