MAE0330 - Análise Multivariada de Dados
2º Semestre/2019

4ª Lista de Exercícios

Informações Importantes

- Data de entrega: 13/11 (quarta-feira)
- Forma de entrega: exclusivamente pelo sistema e-Disciplinas (a lista deverá ser entregue no sistema até 23:50 do dia 11/10).
- Podem ser feitas em grupos de **no máximo** 2 alunos.

1. Sejam as densidades

$$f_1 = (1 - |x|), |x| \leq 1,$$

e

$$f_2(x) = (1 - |x - 0.5), -0,5 \leq x \leq 1,5.$$

   (a) Faça o gráfico das densidades.

   (b) Obtenha as regiões de classificação quando $p_1 = p_2$ e e $c(1|2) = c(2|1)$.

   (c) Obtenha as regiões de classificação quando $p_1 = 0,2$ e e $c(1|2) = c(2|1)$.

2. Considere três populações bivariadas com a mesma matriz de covariância, e médias dadas por: $\boldsymbol{\mu}_1^T = (0; 0)$, $\boldsymbol{\mu}_2^T = (0; -1)$ e $\boldsymbol{\mu}_3^T = (1; 0)$. A matriz de covariância comum é:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix}.$$

   (a) Obtenha as regiões de classificação das observações pelo método de Fisher.

   (b) Suponha que as probabilidades a priori das três populações são 1/2, 1/3 e 1/6, respectivamente. Obtenha as funções discriminantes pelo método geral. Faça suposições necessárias.

3. Os dados no arquivo T11-5.DAT são bastante conhecidos e referem-se a medidas da pétala e sépala de amostras de três diferentes espécies de flores do gênero *iris* (*iris setosa*, *iris versicolor* e *iris virginica*). O arquivo contém 5 colunas: comprimento da sépala (X_1), largura da sépala (X_2), comprimento da pétala (X_3), largura da pétala (X_4) e, por fim, a espécie da flor (1 - *setosa*; 2 - *versicolor*; 3 - *virginica*). Considere as variáveis X_2: largura da sépala e X_4: largura da pétala.

   (a) Faça o diagrama de dispersão de X_2 e X_4, diferenciando as observações das três espécies diferentes.

   (b) Assumindo que as distribuição de X_2 e $X4$ seja normal bivariada, obtenha os escores discrimantes quadráticos com $p_1 = p_2 = p_3$. Classifique uma nova observação $\boldsymbol{x}_0^T = (3,5; 1,75)$ em uma das três populações.

   (c) Assumindo que as distribuição de X_2 e $X4$ seja normal bivariada com a mesma matriz de covariância, calcule os escores discriminantes lineares com $p_1 = p_2 = p_3$. Classifique uma nova observação $\boldsymbol{x}_0^T = (3,5; 1,75)$ em uma das três populações. Compare os resultados com o item anterior. Qual dos métodos de classificação você escolheria? Justifique.

4. Gere dois grupos de 100 observações de uma distribuição normal bivariada com mesma matriz de covariância $\boldsymbol{\Sigma}$ e vetores de médias $\boldsymbol{\mu}_j$, $j = 1, 2$, diferentes para os grupos (indique a *semente* adotada na simulação).

   (a) Divida os dados em uma subamostra de "treinamento/estimação" (cerca de 70% dos dados) e uma de validação/predição (o restante das observações).

   (b) Obtenha a função discriminante linear de Fisher e o critério de classificação. Faça as suposições necessárias. Obtenha a matriz de classificação usando os dados de validação. Comente.

   (c) Obtenha a função discriminante supondo $c(2|1) = 50$ e $c(1|2) = 100$, em que $c(i|j)$ é o custo de classificar uma observação em π_i quando ela é na verdade de π_j . Ainda, suponha que 20% da população total pertence a π_1. Compare os critérios de classificação com e sem essas informações.

   (d) Gere 20 novas observações da seguinte maneira:

      - Associe essa nova observação a π_1 com probabilidade 0,2 e a π_2 com probabilidade 0,8;
      - Se a nova observação pertence a π_1, gere uma observação de uma normal bivariada com média $\boldsymbol{\mu}_1$ e matriz de covariância $\boldsymbol{\Sigma}$;
      - Se a nova observação pertence a π_2, gere valores de uma normal bivariada com média $\boldsymbol{\mu}_2$ e matriz de covariância $\boldsymbol{\Sigma}$.

Aplique as regras de classificação de (b) e (c) nessas novas observações. Compare as proporções de classificações erradas e corretas obtidas com as classificações de (b) e (c). Discuta os resultados.

5. Considere os dados no arquivo primate.scapulae.txt utilizados na lista 2. Relembre que esses dados são referentes a medidas feitas na escápula de cinco diferentes gêneros de primatas Hominoidea (*Hylobates*, *Pong*, *Pan*, *Gorilla* e *Homo*). As medidas estão nas variáveis AD.BD, AD.CD, EA.CD, Dx.CD, SH.ACR, EAD, β e γ. As cinco primeiras medidas são índices e as três últimas são ângulos. O ângulo γ não está disponível para os primatas *Homo* e, portanto, não deve ser usado na análise (relembre que as medidas faltantes não estão representadas por NA nos dados). Com auxílio computacional, considerando apenas as 7 medidas das escápulas disponíveis, obtenha a melhor regra de classificação dentre as discutidas em sala. Utilize as taxas de classificação incorreta e correta para comparação entre os métodos. Discuta os resultados.

6. Considere o arquivo de dados Carseats disponível no pacote ISLR no R. A descrição dos dados pode ser obtida digitando-se ?Carseats após o carregamento do pacote. Assuma que o interesse está em predizer vendas (Sales - variável contínua) usando árvore de regressão.

   (a) Divida os dados em dados de treinamento e teste, deixando 70% das observações no banco de dados de treinamento.

   (b) Ajuste uma árvore de regressão nos dados de treinamento. Faça um gráfico da árvore e interprete.

   (c) Obtenha as somas de quadrado dos erros de predição dos dados de treinamento e depois nos dados de teste.

   (d) Ajuste um modelo de regressão no banco de treinamento (o melhor que você encontrar para predição). Faça a predição dos valores para os dados de teste e compare com os resultados da árvore.

   Observação: Códigos em R para obtenção de árvores de regressão podem ser encontrados no texto disponível no item "Material de Apoio" no e-disciplinas.

7. Considere 51 objetos $O_1, O_2, \ldots, O_{51}$ organizados em uma linha reta, sendo que o j-ésimo objeto está localizado em um ponto com coordenada igual a j. Defina a medida de similaridade s_{ij} entre os objetos O_i e O_j por:

$$s_{ij} = \begin{cases} 9, & \text{se } i = j \\ 8, & \text{se } 1 \leq |i-j| \leq 3 \\ 7, & \text{se } 4 \leq |i-j| \leq 6 \\ \vdots & \\ 1, & \text{se } 22 \leq |i-j| \leq 24 \\ 0, & \text{se } 1|i-j| \geq 25 \end{cases}.$$

Converta as similaridades em dissimilaridades δ_{ij} pela transformação

$$\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}.$$

Utilize o método de escalonamento multidimensional clássico nesta matriz de dissimilaridade obtida. Faça o gráfico da solução obtida em duas dimensões e interprete o resultado.

8. A tabela a seguir apresenta as distâncias entre sítios arqueológicos de diferentes períodos. As distâncias foram calculadas com base em frequências de diferentes tipos de cerâmicas encontradas nos sítios.

   (a) Dadas as distâncias, utilizando escalonamento multidimensional não-métrico, obtenha o *stress* para $q = 3, 4$ e 5 dimensões. Faça um gráfico do *stress* mínimo versus q. Discuta o número de dimensões que é necessário para uma boa representação dos dados.

   (b) Obtenha as coordenadas dos pontos em duas dimensões e faça o gráfico.

   (c) Obtenha as coordenadas dos pontos em duas dimensões utilizando o escalonamento multidimensional clássico e faça o gráfico.

	Sítio Arqueológico								
Sítio	P1980918 (1)	P1931131 (2)	P1550960 (3)	P1530987 (4)	P1361024 (5)	P1351005 (6)	P1340945 (7)	P1311137 (8)	P1301062 (9)
(1)	0								
(2)	2,202	0							
(3)	1,004	2,025	0						
(4)	1,108	1,943	0,233	0					
(5)	1,122	1,870	0,719	0,541	0				
(6)	0,914	2,070	0,719	0,679	0,539	0			
(7)	0,914	2,186	0,452	0,681	1,102	0,916	0		
(8)	2,056	2,055	1,986	1,990	1,963	2,056	2,027	0	
(9)	1,608	1,722	1,358	1,168	0,681	1,005	1,719	1,991	0