Universidade Federal do Amazonas
Instituto de Ciências Exatas
Curso de Bacharelado em Estatística
IEE062: Estatística Multivariada II
Exercício Escolar 4
Entrega: 19/07/2024

1. Considere o banco de dados *iris* do programa R. Trabalhe apenas com as variáveis X_2 = largura sépica (sepal width) e X_4 = largura da pétala (petal width) para as três espécies

$$\pi_1 : setosa$$
$$\pi_2 : versicolor$$
$$\pi_3 : virginica$$

   (a) Plote os dados sobre o espaço (x_2, x_4). As observações para os três grupos parecem ser normais bivariadas?

   (b) Suponha que as amostras sejam de populações normais bivariadas com uma matriz de covariância comum. Teste a hipótese $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ versus pelo menos um é diferente dos outros no nível de significância. A suposição de uma matriz de covariância comum é razoável neste caso? Explique.

   (c) Supondo que as populações sejam normais bivariadas, construa os escores discriminantes quadráticos $\hat{d}_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^\top \mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \ln(p_i),\ i = 1,2,3$, com $p_1 = p_2 = p_3 = 1/3$. Usando a regra abaixo classifique a nova observação $\mathbf{x}_0^\top = [3.4\ 1.75]$ na população π_1, π_2, ou $\pi_i 3$.
   Classificar uma observação $\mathbf{x}$ em π_i se o escore quadrático

$$\hat{d}_k^Q(\mathbf{x}) = \text{maior valor de } \{\hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), \hat{d}_3^Q(\mathbf{x})\}$$

   (d) Suponha que as matrizes de covariância $\boldsymbol{\Sigma}_i$ sejam as mesmas para as três populações normais bivariadas. Construa o escore discriminante linear dado por $\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i^\top \mathbf{S}_p^{-1}\mathbf{x} - \frac{1}{2}\bar{\mathbf{x}}_i^\top \mathbf{S}_p^{-1}\bar{\mathbf{x}}_i + \ln(p_i),\ i = 1,2,3$ e use-o juntamente com a regra de classificação abaixo para atribuir $\mathbf{x}_0^\top = [3.4\ 1.75]$ a uma das populações. Tome $p_1 = p_2 = p_3 = 1/3$. Compare os resultados das letras (c) e (d). Qual abordagem você prefere? Explique.
   Classificar uma observação $\mathbf{x}$ em π_i se o escore discriminante linear

$$\hat{d}_k(\mathbf{x}) = \text{maior valor de } \{\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \hat{d}_3(\mathbf{x})\}$$

   (e) Assumindo matrizes iguais de covariância e populações normais bivariadas, e supondo que $p_1 = p_2 = p_3 = 1/3$ alocar $\mathbf{x}_0^\top = [3.4\ 1.75]$ a π_1, π_2 ou π_3 usando a regra
   Classificar uma observação $\mathbf{x}$ em π_i se

$$\hat{d}_{ki}(\mathbf{x}) = (\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)^\top \mathbf{S}_p^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_i)^\top \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_k + \bar{\mathbf{x}}_i) \geq \ln\left(\frac{p_i}{p_k}\right) \text{ para todo } i \neq k.$$

Compare o resultado com aquele da letra (d). Delinear as regiões de classificação no seu gráfico da letra (a) determinado pelas funções lineares de $\hat{d}_{ki}(\mathbf{x})$.

(f) Usando os escores discriminantes lineares da letra (d), classifique as observações da amostra. Calcule a **taxa de erro aparente** (TEA) pelo método de ressubstituição e pelo método de validação cruzada (Pseudo-jackknife - Método de validação cruzada). Ao usar o método de validação cruzada a TEA é denominada de **taxa de erro atual esperada estimada** (TEAE) que é uma medida obtida de futuras amostras. A TEAE indica como a função de classificação da amostra será executada em amostras futuras.