IEE062: Estatística Multivariada II
Exercício Escolar 4

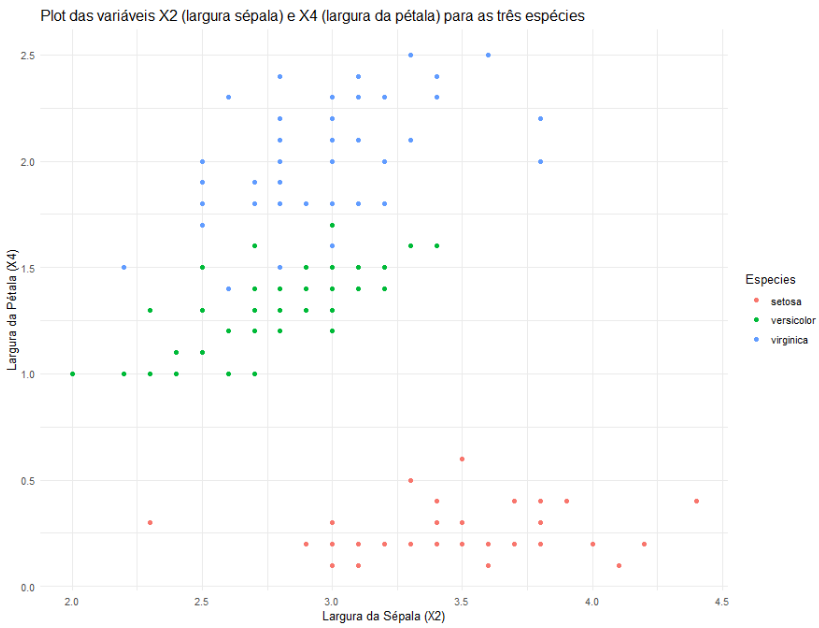
Maria Nilza de Sousa Ramos

Entrega: 19/07/2024

Exercício 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: π_1: *setosa*, π_2: *versicolor* e π_3: *virginica*.

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



Sim, parecem ser normais bivariadas.

(b) Suponha que as amostras sejam de populações normais bivariadas com uma matriz de covariância comum. Teste a hipótese $H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_A: pelo menos um μ_i é diferente dos outros no nível de significância α. A suposição de uma matriz de covariância comum é razoável neste caso? Explique.

Os testes de Pillai, Wilks, Hotelling e Roy tiveram todos os p-valores < 2.2e-16. Rejeitando a hipótese nula aqui, ou seja, pelo menos uma das médias das espécies é diferente das outras.

Ao calcular as covariâncias para cada espécie encontramos valores próximos para Versicolor e Virginica, mas a diferença significativa vista na da Setosa indica que a suposição de uma matriz de covariância comum pode não ser válida ou adequada. Os números podem ser vistos abaixo:

Espécie	**Setosa**		**Versicolor**		**Virginica**	
Variável	X_2	X_4	X_2	X_4	X_2	X_4
X_2	0.143689796	0.009297959	0.09846939	0.04120408	0.10400408	0.04762857
X_4	0.009297959	0.011106122	0.04120408	0.03910612	0.04762857	0.07543265

(c) Classifique a nova observação $x_0^\top = [3.4, 1.75]$ na população π_1, π_2, ou π_3

Escore para Setosa: -105.7821; Escore para Versicolor: -0.631712; Escore para Virginica: -1.445843
$x_0^\top = [3.4, 1.75]$ é classificado como Versicolor por ser o maior discriminante quadrático.

(d) Suponha que as matrizes de covariância Σ_i sejam as mesmas para as três populações normais bivariadas. Construa o escore discriminante linear dado por $\hat{d}_i(x) = \bar{x}_i^\top \Sigma_p^{-1} x - \frac{1}{2}\bar{x}_i^\top \Sigma_p^{-1} \bar{x}_i + \ln(p_i)$, $i = 1, 2, 3$, e use-o juntamente com a regra de classificação abaixo para atribuir $x_0^\top = [3.4, 1.75]$ 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 x em π_i se o escore discriminante linear $\hat{d}_k(x)$ é o maior valor de $\{\hat{d}_1(x), \hat{d}_2(x), \hat{d}_3(x)\}$.

Tipo de Escore	Abordagem Linear	Abordagem Quadrática
Setosa	23.41569	-105.7821
Versicolor	55.82686	-0.631712
Virginica	55.27001	-1.445843

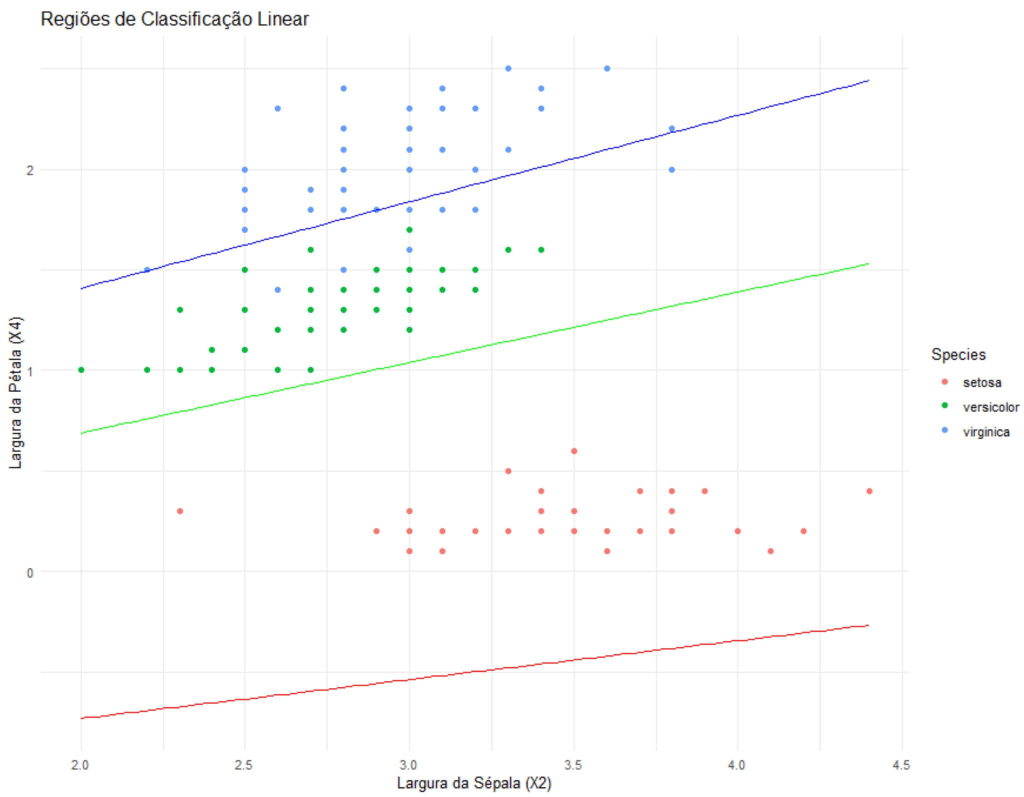
Para o ponto $x_0 = [3.4, 1.75]$ ambas classificaram como sendo da espécie Versicolor. Já foi mostrado anteriormente que há evidências de que as matrizes de covariância são diferentes entre os grupos, neste caso a abordagem quadrática seria a de preferência. Apesar de considerar a linear mais simples, podemos ver que nesta o valor que a diferenciava da Virginica foi extremamente perto. Senti mais segurança na quadrática, se a amostra fosse pequena provavelmente seria o contrário.

(e) Assumindo matrizes iguais de covariância e populações normais bivariadas, e supondo que $p_1 = p_2 = p_3 = 1/3$, alocar $x_0^\top = [3.4, 1.75]$ a π_1, π_2 ou π_3 usando a regra:

Classificar uma observação x em π_i se $\hat{d}_{ki}(x) = (\bar{x}_k - \bar{x}_i)^\top \Sigma_p^{-1} x - \frac{1}{2}(\bar{x}_k - \bar{x}_i)^\top \Sigma_p^{-1} (\bar{x}_k + \bar{x}_i) \geq \ln\left(\frac{p_i}{p_k}\right)$ 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 $\hat{d}_{ki}(x)$.

A classificação final foi "versicolor", assim também como na letra d).



Acima da linha azul, temos as virgínicas.
Acima da linha verde, temos as versicolor.
Acima da linha vermelha, temos as setosas.

(f) Usando os escores discriminantes lineares da letra (d), classifique as observações da amostra. Calcule a taxa de erro aparente (TEA) pelo método de re-substituição 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 TEA (taxa de erro aparente) é de 3.33%, mostrando erros na classificação da própria amostra. A TEAE (taxa de erro esperada) é de 4% usando validação cruzada, indicando que o modelo deve errar um pouco mais em novas amostras.