Quinta lista de exercícios

Papagaio-do-mar é o nome comum dado às aves charadriiformes da família dos alcídeos, pertencentes ao gênero *Fratercula*. Existem três espécies de papagaios-do-mar conhecidas: *arctica*, *corniculata* e *cirrhata*. O conjunto `papagaio.txt` apresenta informações sobre o peso (em gramas), o tamanho (em centímetros), a envergadura da asa (em centímetros) e a espécie de 500 papagaios-do-mar. As questões 1 e 2 utilizarão esse conjunto de dados.

Questão 1. (a) Leia o arquivo `papagaio.txt`. Em seguida, comece a analisar os dados a partir das funções `head, tail, str, summary`.

(b) Determine a média e o desvio padrão das variáveis `tamanho_peso` e `envergadura` para cada uma das espécies. Em seguida, exiba numa mesma janela, os bloxplots para o peso de cada uma das espécies. Comente os resultados encontrados.

(c) Conforme pôde ser observado na parte (a), a variável `especie` é do tipo character. Mas, como sabemos, ela deveria ser do tipo factor. Converta esta variável de character para factor.

(d) Divida o conjunto de dados em dois: um para treino e outro para teste. O conjunto de treinamento deve conter 80% dos dados do conjunto inicial.

(e) Calcule o índice de Gini referente à variável `peso` do conjunto de treinamento.

(f) Crie uma árvore de decisão para classificar a espécie de um papagaio-do-mar a partir do seu peso, tamanho e envergadura da asa, isto é, as variáveis de entrada da árvore devem ser `peso`, `tamanho` e `envergadura` e a variável resposta (classificação) deve ser `especie`.

(g) Calcule a sua taxa de acerto para o modelo construído em (f) utilizando nesse cálculo o conjunto de teste. Em seguida, construa a matriz de confusão e, por fim, comente os resultados encontrados.

Questão 2. (a) Divida o conjunto `papagaio.txt` em três data frames: um para cada espécie.

(b) O coeficiente de correlação é uma medida da força e da direção de uma relação linear entre duas variáveis x e y. O símbolo r representa o coeficiente de correlação amostral. Uma fórmula para r é:

$$r = \frac{(\sum_i^n x_i y_i) - n\overline{x}\ \overline{y}}{\sqrt{(\sum_i^n x_i^2) - n\overline{x}^2}\sqrt{(\sum_i^n y_i^2) - n\overline{y}^2}}.$$

em que n é o número de pares de dados. Utilize a equação acima para criar uma função cuja entrada contenha dois vetores `x,y` e cuja saída seja o coeficiente de correlação entre `x` e `y`. Em seguida, utilize sua função para calcular o coeficiente de correlação entre as variáveis `tamanho` e `peso` para cada uma das espécies. Para qual espécie as variáveis estão mais correlacionadas linearmente?

(c) Utilize o conjunto da espécie em que as variáveis `tamanho` e `peso` estão mais correlacionadas para determinar a reta de regressão linear simples entre essas duas variáveis. Considere `tamanho` como a variável independente (x).

(d) Uma variação de 0.5 cm no tamanho da ave provocaria uma variação de quantos gramas no peso da ave? Por que?

(e) A partir da reta determinada em (c), Crie uma função cuja entrada seja um valor x (tamanho de uma ave) e cuja saída seja o valor previsto do peso dessa ave a partir da reta de regressão. Se o valor de x não for adequado para o modelo, a função deve retornar uma mensagem de erro. Avalie sua função nos pontos: $x = 18$ cm e para $x = 41.01$ cm.

Questão 3. O conjunto `olive.txt` apresenta a composição em porcentagem de oito ácidos graxos encontrados na fração lipídica de 572 azeites italianos.

(a) Aplique o modelo de aglomerados hierárquicos com o método ward.D2 para este conjunto e, em seguida, apresente o dendograma resultante do modelo.

(b) Corte o dendograma em uma altura que resulte em 5 diferentes aglomerados. Identifique a proporção de cada região (Sul, Norte, Sardínia) que está dentro de cada um dos cinco aglomerados.

(c) Aplique agora o modelo K-means com $k = 5$. Comente os resultados encontrados.