UNIVERSIDADE FEDERAL DE UBERLÂNDIA

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

LISTA FINAL DE ESTATÍSTICA COMPUTACIONAL

Questão 1. O conjunto de dados `diabetes.txt` contém diversas variáveis que podem ser úteis para análises e modelagem estatística relacionadas ao diabetes. Algumas das variáveis incluídas são idade, número de gravidezes, concentração de glicose no plasma, concentração de insulina sérica no plasma sanguíneo, pressão sanguínea, espessura da dobra cutânea do tríceps, índice de massa corporal, histórico familiar de diabetes e resultados de testes médicos (tem ou não diabetes). Muitas das variáveis do conjunto podem ser utilizadas para prever o risco de diabetes em pacientes, identificar fatores de risco associados à doença e realizar análises epidemiológicas. Importe o conjunto para o R e, em seguida, crie um conjunto de treinamento e de teste. Para estudar o problema e construir o modelo de classificação, utilize sempre o treinamento.

(a) Realize um estudo descritivo a partir de análises gráficas para compreender as diferenças entre pacientes que possuem diabetes e aqueles que não possuem. Apresente os gráficos e escreva um pequeno texto indicando os resultados obtidos.

(b) Crie um modelo de árvore de decisao e, em seguida, plote a árvore do modelo (esta representação visual da árvore deverá ser entregue junto com o script). A partir representação visual da árvore de decisão, crie uma função cuja entrada seja as informações de um paciente e a saída seja o diagnóstico do paciente (sem ou com diabetes). Esta função deverá ser construída utilizando uma estrutura condicional como aprendido na primeira parte do curso. Utilize esta função para classificar todos pacientes do teste. Calcule a acurácia do modelo.

(c) Crie agora um modelo de floresta aleatória e encontre a acurácia do modelo.

(d) Para cada um dos modelos, dado que um paciente tenha diabetes, qual a probabilidade de cada um dos modelos classificar corretamente o diagnóstico deste paciente?

(e) Faça um pequeno texto para comparar os dois modelos construídos (por exemplo, com relação as medidas calculadas, qual se saiu melhor?).

Questão 2. O cerebelo é uma estrutura cerebral altamente convoluta localizada abaixo dos hemisférios cerebrais. Acredita-se que esta estrutura intrigante facilite a aquisição e uso de dados sensoriais pelo resto do cérebro, particularmente nas áreas motoras. Estudos sugerem que o cerebelo pode aumentar proporcionalmente com o tamanho do corpo dos animais, enquanto outras partes do cérebro são claramente representadas de forma diferente entre as espécies. O arquivo `cerebelo.csv` apresenta os dados de 15 espécies de mamíferos mostrando os pesos em gramas de seus corpos e cerebelos, junto com os valores transformados em logaritmo na base 10.

(a) Faça um gráfico de dispersão do peso do cerebelo y em relação ao peso do corpo x e outro gráfico de dispersão utilizando os valores transformados, isto é, outro gráfico em que o logaritmo do peso do cerebelo estará em y e o logaritmo do peso do corpo estará em x. Descreva e compare ambos os gráficos de dispersão. O que você aprendeu sobre a relação entre o peso do cerebelo e o peso do corpo a partir desses gráficos?

(b) Calcule o coeficiente de correlação entre as variáveis peso do cerebelo e peso do corpo.

(c) Calcule o coeficiente de correlação entre as variáveis logaritmo do peso do cerebelo e logaritmo do peso do corpo.

(d) Compare os dois resultados obtidos anteriormente.

(e) Encontre a equação da reta de regressão usando os valores transformados em logaritmo. Qual a equação da reta? Analise a saída do modelo `lm` a partir da função `summary()`; comente os resultados dos testes de hipóteses que foram feitos em relação ao modelo. Por fim, acrescente a reta de regressão no gráfico obtido em (a).

(f) Realize um teste de hipótese para verificar se os resíduos seguem uma distribuição Normal. Comente os resultados obtidos.

(g) Preveja o peso do cerebelo em gramas de uma espécie que pesa 100.000 g. Certifique-se de transformar os gramas em unidades logarítmicas primeiro e depois de volta para gramas.

Questão 3. O conjunto `olive.txt` apresenta a composição em porcentagem de oito ácidos graxos encontrados na fração lipídica de 572 azeites italianos além de indicar a região italiana onde cada azeite foi produzido. O objetivo deste exercício é clusterizar os 572 azeites em grupos a partir do modelo K-means.

(a) Como o modelo K-means é construído a partir do cálculo de distâncias entre observações, é preciso padronizar as variáveis numéricas dos dados antes da modelagem. Utilizando a função `scale()`, padronize os dados e guarde a padronização num objeto chamado `dados_padronizados`.

(b) Construa para este conjunto um modelo K-means com $k = 3$. Extraia os aglomerados resultados (cluster) e insira este vetor no conjunto como uma variável categórica chamada `cluster_k3`. Por fim, crie um gráfico de barras em que $x = cluster\_k3$ e o preenchimento dessas barras será feito pela variável `region`. Comente os resultados obtidos no gráfico.

(c) Repita o item anterior para $k = 4$ e $k = 5$.