TRABALHO FINAL DE ESTATÍSTICA COMPUTACIONAL

Exercício 1. O conjunto de dados `SBI.csv` contém informações de mais de 2 mil crianças que compareceram aos serviços de emergência de um hospital com febre e que foram submetidas a um teste para detecção de infecção bacteriana grave. As seguintes variáveis estão incluídas no conjunto:

- id - número do paciente;
- fever_hours - duração da febre em horas;
- age - idade da criança;
- sex - sexo da criança (M, F);
- wcc - contagem de células brancas;
- prevAB - antibióticos anteriores (yes, no);
- sbi - infecção bacteriana grave (Not Applicable, UTI, Pneum e Bact);
- pct - procalcitonina;
- crp - proteína c-reativa.

A variável sbi possui 4 categorias: Not Applicable, UTI, Pneum e Bact. Not Applicable significa que o teste deu negativo; já as outras categorias indicam a existência de infeccção bacteriana grave.

(a) Acresente ao conjunto de dados uma nova coluna chamada `infection`. Essa variável será `yes` se a criança foi diagnosticada com infecção grave e `no` caso contrário. Lembre-se que essa variável deve ser do tipo categórica (factor).

(b) Retire do conjunto de dados as variáveis `X`, `id` e `sbi`.

(c) Separe o conjunto de dados em dois novos conjuntos, um para treino e um para teste. O conjunto para treino deverá ter 80% dos dados iniciais.

(d) Crie um modelo de árvore de decisão para classificar a variável `infection` a partir das outras variáveis do conjunto de treinamento. Plote a representação gráfica da árvore resultante. A partir do conjunto de teste e da função `predict`, verifique a acurácia do modelo. Por fim, crie uma matriz de confusão da previsão versus respostas verdadeiras.

(e) Repita o item acima (exceto a parte referente à representação gráfica) para um modelo de floresta aleatória.

Exercício 2. Na Estação Antártica Palmer, pesquisadores fizeram medições em três espécies diferentes de pinguins: Adélie, Chinstrap e Gentoo. Os dados obtidos estão no arquivo **pinguim.csv**. O conjunto possui as seguintes variáveis: sexo, espécie, ilha onde o pinguim habita, peso em gramas (`body_mass_g`), tamanho da asa em milímetros (`flipper_length_mm`), tamanho da crista dorsal do bico em milímetros (`culmen_length_mm`) e profundidade da crista dorsal do bico em milímetros (`culmen_depth_mm`).

(a) Plote o gráfico de `flipper_length_mm` versus `body_mass_g`.

(b) Utilize a função `cor()` para calcular a correlação entre as variáveis `flipper_length_mm` e `body_mass_g`.

(c) A partir das duas respostas anteriores, responda: há alguma relação entre `flipper_length_mm` e `body_mass_g`? Quão forte é essa relação? A relação é positiva ou negativa?

(d) Utilize a função `lm()` para determinar a reta do modelo de regressão linear simples. Considere `flipper_length_mm` como a variável explanatória x e `body_mass_g` como a variável resposta y.

(e) Explique o coeficiente angular da reta encontrada em (d).

(f) A partir do modelo linear encontrado em (d), qual seria o peso médio de um pinguim que possui uma asa de 204 mm? Você poderia utilizar esse modelo para estimar o peso médio de um pinguim que tivesse uma asa de 168 mm? Justifique sua resposta.

(g) De acordo com a variável `island`, os pinguins habitam 3 ilhas diferentes. Divida o conjunto em três outros conjuntos de tal forma que cada novo conjunto contenha apenas pinguins de uma única ilha.

(h) A partir do conjunto que contém apenas os pinguins da ilha Biscoe, crie um conjunto que contenha apenas as fêmeas; em seguida, utilize esse conjunto das fêmeas da ilha Biscoe e o modelo de aglomerados hierárquicos com o método ward.D2 para agrupar os pinguins. Em seguida, plote o dendograma referente ao modelo criado.

(i) A partir da análise do gráfico plotado em (h), em qual altura você cortaria o dendograma? Justifique sua resposta. Quantos aglomerados resultaram do corte? Identifique a proporção de cada espécie dentro de cada aglomerado. Comente os resultados obtidos.

(j) Utilize o conjunto do item (h) para construir um modelo k-means com $k = 2$ e outro modelo k-means com $k = 3$. Comente os resultados obtidos.

(k) Considere o modelo obtido em (d) com $k = 3$. Plote um gráfico em que o eixo x é dado pela variável flipper lenght e o eixo y é dado pela variável body mass e cada aglomerado tenha uma cor diferente. Por fim, acrescente a este gráfico, os centróides de cada aglomerado.

Exercício 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.

Exercício 4. Considere três urnas com as seguintes configurações: a urna I contém 6 bolas pretas, 3 brancas e 4 vermelhas; a urna II contém 3 bolas pretas, 5 brancas e 2 vermelhas; a urna III contém 4 bolas pretas, 2 brancas e 2 vermelhas. Lança-se um dado equilibrado. Se sair 5, uma bola da urna I é retirada; se sair 1, 4 ou 6, então uma bola da urna II é retirada; se sair 2 ou 3, então uma bola da urna III é retirada. Estime a probabilidade da bola retirada ser vermelha.

Exercício 5. Uma urna contém bilhetes numerados de 1 até 30, todos do mesmo tamanho. Considere o seguinte experimento: retirar um bilhete da urna, registrar o resultado e devolver o bilhete para a urna. Você continuará a sortear bilhetes até que todos os números sejam retirados. Seja N o número de sorteios que você precisou realizar até que todos os números fossem registrados. Utilize o Método de Monte Carlo para estimar $E[N]$, isto é, para estimar a esperança de N.

Exercício 6. Um dado será lançado até que o número 4 seja obtido pela terceira vez. Seja X a variável aleatória que conta o número de lançamentos que foram necessários para obter o número 4 pela terceira vez.

(a) Estime, via Monte Carlo, a esperança de X.

(b) Estime, via Monte carlo, a probabilidade de X ser menor do que 10.