Segunda lista de exercícios

Fundamentos de R, análise e manipulação de dados, modelos de classificação

Exercício 1. Considere o seguinte jogo: Steven e Garnit escolherão, cada um, uma sequência de tamanho 3 em que cada entrada da sequência é cara ou coroa; logo em seguida, uma moeda será lançada três vezes; se aparecer a sequência de um dos jogadores, este jogador vence e o jogo acaba; caso não apareça a sequência de nenhum deles, a moeda é lançada pela quarta vez e os três últimos lançamentos são analisados; se nestes três últimos lançamentos aparecer a sequência de um dos jogadores, este jogador vence e o jogo acaba. Se isto não acontecer, a moeda é lançada pela quinta vez e os três últimos resultados são analisados; se aparecer a sequência de um dos jogadores, este jogador vence e o jogo acaba. Este processo é realizado até que apareça a sequência que um dos dois escolheu; se aparecer primeiro a sequência de Steven, ele ganha; se aparecer primeiro a sequência de Garnit, ela vence. Convencione que cara seja 1 e que coroa seja zero. Supondo que Steven escolheu a sequência (0, 1, 0) e que Garnit escolheu a sequência (0, 0, 1), simule uma partida deste jogo. A simulação deve retornar `steven` caso Steven tenha vencido ou deve retornar `garnit` caso contrário. Replique o experimento 10 mil vezes e calcule a média de vitórias de Garnit. Comente o resultado obtido. (6 pontos)

Observação: Suponha que os três primeiros lançamentos foram (1,0,0). Logo, ninguém ganhou e a moeda é lançada pela quarta vez. Suponha que o quarto lançamento foi 0; logo os três últimos lançamentos foram (0,0,0) e ninguém ganhou. Na quinta vez saiu 1 e, portanto, os três últimos lançamentos foram (0,0,1) e o jogo acaba com vitória de Garnit. As sequências (0, 1, 0), (1, 0, 1, 0) e (1, 1, 0, 1, 0) fazem Steven vitorioso; as sequências (0, 0, 1), (0, 0, 0, 1) e (1, 0, 0, 0, 1) fazem Garnit vitoriosa.

Exercício 3. Harold Frederick Shipman (Nottingham, 14 de janeiro de 1946 — Wakefield, 13 de janeiro de 2004), conhecido como "Doutor Morte", foi um médico e assassino em série britânico condenado pela morte de muitos pacientes entre as décadas de 1970 e 1990. Dr. Shipman é, talvez, o assassino em série mais prolífico da História Moderna. O arquivo `dados.txt` contém informações sobre o sexo, a idade, o local da morte (casa do paciente; hospital; casa de repouso) e o ano da morte das vítimas de Shipman. Antes de responder as questões abaixo, abra o arquivo `dados.txt` e compreenda sua estrutura. Importe o arquivo para o R e utilize-o para responder os seguintes itens.

(a) Escolha um gráfico apropriado para representar as frequências das categorias da variável sexo. Comente os resultados encontrados.

(b) Apresente o histograma da variável idade em 8 (argumento `bins` na geometria do histograma) intervalos. Comente os resultados obtidos. Analise este gráfico para cada gênero.

(c) Apresente o boxplot da variável idade. Comente os resultados obtidos.

(d) Apresente um gráfico para representar o local da morte. Comente os resultados obtidos.

(e) Analise graficamente o ano da morte das vítimas de Harold Shipman.

(f) Com base nas informações obtidas nos itens anteriores, escreva um parágrafo sobre o padrão e o perfil das vítimas de Harold Shipman.

Exercício 3. Os arquivos `treino_baleias.txt` e `teste_baleias.txt` contém informações sobre as características de algumas espécies de baleias. Os conjuntos de dados possuem, ao todo, 248 observações (198 para treino, 50 para teste). As variáveis incluídas nestes conjuntos de dados são:

- **especie**: indica a espécie da baleia e é uma variável categórica;
- **comprimento**: indica o comprimento da baleia em metros e é uma variável numérica contínua;
- **peso**: indica o peso da baleia em quilos e é uma variável numérica contínua;
- **profundidade_maxima**: indica a profundidade máxima mergulhada pela baleia em metros e é uma variável numérica contínua;
- **volume_cranio**: indica o volume do crânio da baleia em centímetros cúbicos e é uma variável numérica contínua.

Os itens de (a) até (e) devem ser respondidos usando apenas os dados de `treino_baleias.txt`. Em (f), os dois conjuntos devem ser utilizados.

(a) Crie um conjunto para cada espécie de baleia; cada data frame criado deverá conter apenas baleias de uma espécie.

(b) Calcule a média, a variância, o desvio padrão e o coeficiente de variação para a variável peso para cada espécie de baleia. Comente os resultados obtidos.

(c) Apresente o histograma da variável peso para a espécie de baleia azul. Comente os resultados obtidos.

(d) Apresente numa mesma janela os boxplots para cada espécie para a variável comprimento. Comente os resultados obtidos.

(e) Apresente um gráfico de dispersão de `comprimento` versus `profundidade_maxima`. Cada espécie deve ser registrada por uma cor diferente.

(f) Com base em todas as informações anteriores, construa um modelo de árvore de decisão a partir de estruturas condicionais e de repetição para prever a espécie de uma baleia com base nas variáveis numéricas do estudo. Justifique as escolhas das variáveis e dos pontos de corte escolhidos. Por fim, utilize o conjunto do arquivo `teste_baleias.txt` para calcular a taxa de acerto. Comente o resultado obtido.

(g) Utilize gráficos de dispersão para registrar por linhas horizontais e verticais os pontos de cortes escolhidos em sua árvore de decisão. As espécies de baleias devem ser registradas por diferentes cores.

(h) Crie um modelo de classificação KNN para classificar as baleias. Utilize $K = 1$ e depois $K = 3$. Compare os resultados dos dois modelos KNN.

Exercício 4. O conjunto `cogumelos.csv` contém informações sobre 23 espécies de cogumelos dos gêneros Agaricus e Lepiota, retiradas do Guia de Campo da Sociedade Audubon para Cogumelos da América do Norte (1981). Cada espécie é classificada (`class`) como comestível (edible = e) ou venenosa (poisonous = p). Detalhes sobre cada uma das variáveis do conjunto estão no Kaggle ou em UC Irvine Machine Learning Repository.

Embaralhe o conjunto e, em seguida, separe-o em treinamento (80%) e teste (20%). Estude o conjunto de treinamento a partir de uma análise gráfica (nesta parte faça algumas perguntas interessantes e encontre um gráfico que ajudará na sua resposta; exemplos: quantas espécies venenosas há no treinamento? e comestíveis?; a forma, a cor ou o odor pode influenciar na classificação? etc). A partir das conclusões e observações obtidas, crie um modelo de árvore de decisão para classificar um cogumelo como comestível ou venenoso. Avalie a taxa de acerto e comente o resultado obtido.