Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática

Ficha Teórico-Prática 5 Objetivos:

•Modelos de K-Vizinhos-mais -próximos, usando Phyton;

•Análise e discussão dos resultados.

Modelos de regressão lineares. Árvores de regressão.

Objetivos:

- Modelos de regressão linear simples e múltipla, usando o Python;
- Modelos de árvores de regressão, usando o Python;
- Avaliação dos modelos.

1. Pretende-se avaliar o impacto que o orçamento em publicidade em três canais (youtube, facebook e newsletter) têm sobre as vendas de uma empresa. Os dados disponíveis são o orçamento em publicidade em milhares de dólares e o montante das vendas. A publicidade em cada um dos canais foi repetida 200 vezes com diferentes orçamentos e as vendas observadas foram recolhidas. O objetivo é prever as vendas futuras da empresa usando modelos de regressão lineares e árvores de regressão.
   - **a.** Comece por carregar o *dataset* "`Advertising.csv`".
   - **b.** Analise os dados.
   - **c.** Separe o conjunto de dados inicial em dois subconjuntos treino e teste, segundo o critério *holdout* (70% treino/30% teste).
   - **d.** Obtenha um modelo de regressão linear simples usando apenas um dos canais de publicidade.
     - **I.** Apresente a função linear resultante.
     - **II.** Visualize a reta correspondente ao modelo de regressão linear simples e o respetivo diagrama de dispersão.
     - **III.** Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) do modelo sobre os 30% casos de teste.
   - **e.** Repita as alíneas anteriores, com um modelo de regressão linear múltipla usando os três canais
   - **f.** Simplifique o modelo.
   - **g.** Obtenha a árvore de regressão usando a função *DecisionTreeRegressor()* da *Scikit-Learn* para prever as vendas futuras da empresa em função dos orçamentos em publicidade nos três canais.

h. Visualize a árvore de decisão.

i. Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) da árvore de regressão sobre o conjunto de teste.

2. Considere o *dataset* "`swiss.csv`" referente a um estudo demográfico realizado na Suíça em 1888. Para várias regiões do país registaram-se uma série de variáveis, como, o índice de fertilidade, a percentagem de população afeta à agricultura, percentagem de militares que receberam a nota máxima num teste do exército, percentagem de pessoas com mais do que a escola primária, a percentagem da população católica, e a percentagem de nascimentos que sobrevivem menos do que um ano (índice de mortalidade infantil). Este *dataset* contém informações, para 47 províncias francófonas na Suíça, relativamente aos seguintes atributos:
    - Fertility: Ig, Índice de fertilidade
    - Agriculture: % de homens envolvidos na agricultura como ocupação
    - Examination: % de pessoas que obtiveram a nota máxima no exame do exército
    - Education: % de pessoas com habilitações literárias para além do ensino primário
    - Catholic: % da população que é católica (em oposição à protestante)
    - Infant.Mortality: % de nados-vivos que vivem menos de 1 ano.

    a. Comece por carregar o *dataset* "`swiss.csv`"

    b. Analise os dados.

    c. Analise graficamente a distribuição dos valores da variável mortalidade infantil

    d. Obtenha a árvore de regressão que relacione a variável Infant.Mortality com os outros atributos

    e. Visualize a árvore de regressão.

    f. Separe o conjunto de dados inicial em dois subconjuntos treino e teste, segundo o critério *holdout* (70% treino/30% teste).

    g. Obtenha a árvore de regressão para prever a mortalidade infantil.

    h. Visualize a árvore de regressão.

    i. Calcule o erro médio absoluto (MAE) para o modelo sobre o conjunto de teste.

Exercício Complementar

1. Considere o dataset "*Boston Housing*". Pretende-se prever o preço típico das casas de uma área de Boston usando modelos de regressão lineares e árvores de regressão. Este *dataset* consiste numa amostra de 506 observações, cada uma descrita por 14 atributos, sendo MEDV, a variável objetivo.
    a) Comece por carregar o dataset "`housing.csv`".

b) Analise os dados.

c) Separe o conjunto de dados inicial em dois subconjuntos treino e teste, segundo o critério holdout (70% treino/30% teste).

   i. Analise graficamente a distribuição dos valores da variável Número médio de quartos por habitação (RM)

   ii. Obtenha um modelo de regressão linear simples considerando o Número médio de quartos por habitação (RM). Apresente a função linear resultante.

   iii. Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) do modelo sobre os 30% casos de teste.

   iv. Obtenha a árvore de regressão para prever o valor médio das casas.

   v. Visualize a árvore de regressão.

   vi. Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) da árvore de regressão sobre o conjunto de teste.