Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática

Ficha Teórico-Prática 8

Classificação: K-Vizinhos-mais -próximos

Objetivos:

- Modelos de K-Vizinhos-mais -próximos, usando Phyton;
- Análise e discussão dos resultados.

1. O **Abalone** é um molusco com uma concha peculiar em forma de orelha forrada a madrepérola. O seu valor económico está positivamente correlacionado com a sua idade, sendo por isso importante determinar a idade com precisão. No entanto, os produtores estimam a idade deste molusco cortando a concha e através de um microscópio contam o número de anéis na concha. Este processo além de demorado é pouco preciso e aumenta o custo do molusco. O objetivo é prever a idade do **abalone** através de modelos usando as medições físicas do molusco.
    a. Comece por carregar o ficheiro (”`abalone.data`”), verifique a sua dimensão e obtenha um sumário dos dados.
    b. Usando os gráficos apropriados explore os vários atributos do conjunto de dados e realize as seguintes transformações aos dados:
        i. Conversão do atributo categórico **Sex** para numérico
        ii. Normalização dos dados
    c. Separe o conjunto de dados em dois subconjuntos treino e teste, segundo o critério *Holdout*, (70% treino/30% teste).
    d. Aplique o algoritmo K-vizinhos-mais-próximos para prever o atributo **Rings** usando os valores ímpares de K no intervalo [1, 50]. Recolha para cada valor de K o RMSE da previsão. Verifique qual o valor de k que minimiza o RMSE.
    e. Sabendo que o valor de **Rings + 1.5** corresponde à idade em anos, derive um novo atributo **Age** a partir do atributo **Rings** e discretize este novo atributo em duas classes: **Young** e **Adult**.

f. Aplique o algoritmo K-vizinhos-mais-próximos para prever o atributo **Age**, para **K** no intervalo [1, 50]. Recolha para cada valor de **K** a taxa de acerto (accuracy) da previsão. Verifique qual o valor de **K** que maximiza a taxa de acerto.

g. Usando o método de treino *k-fold cross validation* obtenha modelos de previsão do atributo Age com:

   i. O algoritmo K-vizinhos-mais-próximos e o valor de k obtido na alínea anterior

   ii. Um modelo árvore de regressão e obtenha a média e o desvio padrão taxa de acerto dos modelos.

h. Verifique se a diferença de desempenho entre os modelos obtidos anteriormente é estatisticamente significativa.

2. Considere o *dataset* "`BreastCancer.csv`" da Ficha 6. O conjunto de dados "***BreastCancer"*** a analisar contém atributos que foram obtidos a partir de imagens digitalizadas de pequenas amostras de massa mamária de pacientes e descrevem as características dos núcleos celulares presentes nessas imagens. O objetivo é determinar a qual das duas classes (benigna ou maligna) o tumor pertence. Realize a análise descrita do exercício 1.