Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática

Ficha Teórico-Prática 6

Classificação: Árvores de Decisão

Objetivos:

- Modelos de árvores de regressão, usando Python;
- Avaliação dos modelos.

1. O conjunto de dados "`BreastCancer.csv`" 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.
   a) Comece por carregar o dataset "`BreastCancer.csv`". Verifique a sua dimensão e obtenha um sumário dos dados.
   b) Usando os gráficos apropriados, analise os vários atributos do conjunto de dados.
   c) Separe o conjunto de dados inicial em dois subconjuntos treino e teste, segundo o método *holdout*, (70% treino/30% teste), aplique a função "*DecisionTreeClassifier*" da biblioteca "*scikit-learn*" sobre os dados de treino para gerar um modelo de classificação e visualize a árvore de decisão.
   d) Apresente a matriz de confusão e a taxa de acerto do modelo gerado.
   e) Repita o processo anterior de aprendizagem/teste 10 vezes (com amostras diferentes em cada repetição) colecionando, em cada iteração, a percentagem de acerto obtida pela respetiva árvore. Apresente o valor médio da percentagem de acerto nas 10 repetições e o respetivo desvio padrão.
   f) Elabore uma função para apresentar a matriz de confusão e as medidas de avaliação: taxa de acerto (accuracy), recall, precision e F1 de um modelo.
   g) Repita novamente o processo de aprendizagem usando agora o método *k-fold cross validation* e a função anterior para obter as medidas de avaliação de cada modelo.
   h) Obtenha o valor médio e o respetivo desvio padrão das medidas obtidas anteriormente.

2. Considere o problema do exemplo típico de jogar golfe (`golf.csv`), que tem como objetivo prever se há condições para jogar (ou não) golfe. Realize a análise descrita no exercício 1.

Exercícios complementares

1. Um conceito importante nas árvores de decisão é a Entropia. Apresente uma definição de entropia e explique a forma como se relaciona com as árvores de decisão.

2. Outro conceito muito importante nas árvores de decisão é o Ganho de Informação. Dê uma definição de ganho de informação e explique a forma como se relaciona com as árvores de decisão, nomeadamente com o conceito de "dividir para reinar".

3. As árvores de decisão podem ser utilizadas para apoio ao diagnóstico médico. Na Tabela seguinte foi recolhida a informação recolhida pelo médico. Neste exemplo, o médico fez 8 perguntas sobre os sintomas dos pacientes (resposta: 'S'/'N'), por exemplo, 1 = Dor de Cabeça? 2= Febre?, 3 = Problemas digestivos? Neste desafio pretende-se a criação de uma árvore binária de decisão baseada nesta tabela de conhecimento os médicos para suportar o diagnóstico.

1	2	3	4	5	6	7	8	Diagnóstico
S	S	N	S	N	S	S	S	Gripe
S	N	S	S	S	N	N	S	Saudável
S	N	S	N	S	N	S	N	Morte Certa
S	N	N	S	S	N	S	N	Morte Certa