Aprendizagem Automática 2022/23

Ficha prática 9

Exercício#9.1

Considere o seguinte conjunto de dados, onde cada exemplo é caracterizado por 4 atributos e pertence a uma de 2 classes.

handicap-infants	immigration	crime	duty-free-export	Class
n	n	y	n	republican
n	y	y	n	republican
y	n	n	n	democrat
y	n	n	y	democrat
y	n	y	n	democrat
y	n	y	y	republican
y	y	n	y	democrat
y	y	y	n	republican

Defina a árvore de decisão quando é apresentado o conjunto anterior e a função de impureza é o índice de Gini. apresentando os cálculos. Considere que o índice de Gini para um conjunto D é dado pela fórmula seguinte onde n é o número de classes e pc a probabilidade da classe.

$$gini(D) = 1 - \sum_{c=1}^{n} p_c^2$$

Exercício#8.2 Breast cancer

1. Carregue o conjunto de dados "breast_cancer" usando a função "load_breast_cancer()".

2. Faça uma divisão treino/teste com os valores por omissão (treino=75%, teste=25%).

3. Construa um modelo usando árvores de decisão com os valores por omissão. Verifique o nº de folhas e a profundidade da árvore gerada. Calcule o desempenho sobre ambos os conjuntos.

4. Crie um novo modelo, desta vez aplicando pre-pruning limitando a profundidade da árvore a 4 níveis. Calcule o desempenho sobre ambos os conjuntos para este novo modelo.

5. A árvore pode ser visualizada utilizando a função 'export_graphviz()' do módulo 'sklearn.tree'. Esta função gera um ficheiro no formato '.dot' (formato de texto para guardar grafos). Verifique os parâmetros da função. A chamada seguinte gera um ficheiro 'tree.dot', permite colorir os nós para refletir a classe maioritária e utiliza o nome das classes e atributos para etiquetar a árvore:

```
export_graphviz(tree, out_file="tree.dot",
class_names=["malignant", "benign"], feature_names=cancer.feature_names,
impurity=False, filled=True)
```

Visualize a árvore utilizando o módulo 'graphviz' e analise-a verificando os caminhos que a maioria dos exemplos segue.

6. Outra forma de analisar uma árvore de decisão é calcular a importância dos atributos. Esta medida avalia quão importante é cada atributo na decisão feita pela árvore e varia entre 0 e 1, com 0 a indicar que o atributo não é usado e 1 a indicar que o atributo prevê a classe perfeitamente (as importâncias somam sempre 1). A importância dos atributos é guardada no atributo 'feature_importances_' do modelo gerado. Apresente um gráfico de barras que mostre a importância de cada atributo.