决策树算法

**一、实验目的**

1、掌握决策树算法的相关概念；

2、掌握决策树求解问题的流程；

3、能够编写出决策树算法预测问题的代码；

4、能够分析实验结果，对算法进行评估。

**二、实验原理**

决策树（decisiontree）是一种常用的有监督算法，之所以名字带有“树”，因为其算法的实现过程类似于一棵树的生长过程，它从根部（根结点）开始，逐步到进行分支（内部结点），最终到一片片末端的树叶（叶结点）。在决策树里，所有的训练样本构成了根结点，每个内部节点对应一个特征（属性）测试，用来区分不同标签的样本。经过特征测试后，这个节点包含的所有样本根据特征测试的结果。决策树有一种“天然”高方差特征，会随着训练集的变化出现较大的变动，导致各模型对样本观测X0的预测差异较大，表现出预测的高方差。

**三、实验环境**

Anaconda（支持Python3.7），使用Jupyternotebook基于网页的交互性文本编辑器，安装Numpy库、Pandas库、Matplotlib库、Scikit-learn库和可视化包GraphViz。

**四、实验内容及步骤**

**1、实验内容**

**（1）用决策树模型预测在某种天气情况下，是否会出去运动？**（训练数据：play.txt）构建决策树模型，并对决策树进行可视化，用graphviz 将决策树呈现出来。分类决策树中有很多参数（criterion，splitter，max\_depth，min\_samples\_split，max\_features）分析这些参数对模型的影响？

**（2）用决策树模型预测泰坦尼克生还判断？**（训练数据：train.csv、测试数据：test.csv），一个旅客是否能够逃生与他的Pclass、Name、Sex、Age、SibSp、Parch、Ticket、Fare、Cabin、Embarked等均有一定联系，能否找到这些特征和他们能否逃生之间的规律，选择对分类结果有关键作用的特征。构建决策树模型，并对决策树进行可视化，用graphviz将决策树呈现出来。

**2、实验步骤**

(1)导入Numpy库、Matplotlib库、Pandas库、Scikit-learn库和graphviz包。

(2)导入Scikit-learn提供的决策树模块DecisionTreeClassifier()

(3)使用Pandas库读取数据文件（数据清洗：数据类型转换、删除冗余数据等）

(4)构造决策树模型

(5)模型的评估和预测

(6)模型参数的选择、调整（选择最优参数）

(7)决策树的可视化、用graphviz将决策树呈现出来。

**五、实验代码**