机器学习实验 - 2023 春

实验二：决策树和随机森林

实验内容：决策树，随机森林
作业提交截止时间：2023/04/01 23:59:59

环境要求：
Python，numpy 支持多维度的数组和矩阵运算，pandas 数据处理和分析工具，Matplotlib 图形化工具，sklearn 机器学习库。

任务一：决策树

本任务中你将使用决策树来预测泰坦尼克号的乘客是否能够生存。通过乘客的相关信息来构建合适的决策树，以此判断乘客的生还情况，并尝试不同的决策树参数，观察其对决策树的影响（如节点分裂标准，树的最大深度等），尝试使用网格搜索的方法找到较优的参数，最后将决策树可视化出来。
文件 ex2data.csv 包含我们的线性回归问题的数据集。数据参数如下：PassengerId（乘客 ID），Name（姓名），Ticket（船票信息）；Survived（获救情况）其值为 1 或 0，代表着获救或未获救；Pclass（乘客等级），Sex（性别）值为 male 或者 female，Embarked（登船港口）其值为 S，Q，C；Age（年龄），SibSp（堂兄弟妹个数），Parch（父母与小孩的个数）；Fare（票价）是数值型数据；Cabin（船舱）则为文本型数据。决策树相关参数和网格搜索参数代码介绍在 demo.py 可见。

(1) 请将 70%的数据用作训练集，30%的数据用作测试集，使用留出法对以上模型进行验证。
(2) 请对生成的决策树进行剪枝，并可视化剪枝前/剪枝后的决策树，比较两者区别。

任务二：随机森林

本任务中你将使用随机森林来预测泰坦尼克号的乘客是否能够生存。通过乘客的相关信息来构建合适的随机森林，以此判断乘客的生还情况，并尝试不同的随机森林参数，观察其对随机森林的影响（如森林中的决策树数量，节点分裂标准，树的最大深度等），尝试使用网格搜索的方法找到较优的参数。
数据集同上。随机森林相关参数代码介绍在 demo.py 可见。

请使用 10 折交叉验证法对以上模型进行验证。

作业提交内容：

1. 作业代码
2. 实验结果及分析 （PDF 格式）
3. 请将以上文件打包并以 zip 格式压缩，命名为 学号+姓名+第几次实验.zip，邮件发送到 facanhe@163.com