编程作业 #2

题目介绍:

编写基础的决策树算法：在提供的代码框架的基础上初步实现 ID3（使用信息增益准则）、C4.5（使用增益率准则）决策树算法，实现预剪枝、后剪枝算法，并对不同决策树算法以及剪枝策略进行简要分析。

数据集介绍:

CelebA 数据集是一个广泛用于人脸识别和人脸属性分析的大型数据集。它包含大约 20 万张名人的人脸图片，每张图片有 40 个属性标签和 5 个标志点位置的注释。本次作业使用 CelebA 数据集的一个子集，其中训练集 250 张图片，测试集 50 张图片，分别属于 10 个不同的人。我们利用每张图片的 40 个属性标签来构建决策树。数据集具体信息可参见 README 文件。

labels.txt 中包含所有属性的名称；train.txt 中是训练集数据，每一行代表一个样本，每列数值依次是各个属性的取值，最后一列是样本的标签；test.txt 中是测试集数据。

作业要求:

（1）在提供的 CodeBase 上实现两种决策树算法，补充脚本文件中的 TODO 部分。

（2）在迷你数据集 melon 上测试代码，使用 ID3 算法，分别给出不进行剪枝操作、只进行预剪枝和只进行后剪枝的可视化结果。

（3）在 CelebA 子集上测试代码，分别给出使用 ID3、C4.5 算法，并进行预剪枝或后剪枝的可视化结果。

（4）分析两种决策树算法以及两种剪枝算法的特点。

（5）代码中还包括 CART 决策树（使用基尼指数准则）算法内容，不做具体要求，不影响最终成绩，有兴趣同学可以自行尝试并进行比较。

提交说明:

提交文件格式及命名要求：

- 学号 + 姓名.zip
  - report.pdf (pdf 版报告)
  - code (代码文件夹)