实验六：决策树

一、 说明

- 实验采用 jupyter notebook, 请填写完代码后提交完整的 ipynb 文件
- 文件命名规则：班级_姓名_ML2018_HW6.ipynb，如计科 1701_张三_ML2018_HW6.ipynb
- 提交方式：采用在线提交至：

  http://pan.csu.edu.cn:80/invitation/52e27eff-ba4a-42ed-a0e2-cb538b298bdb
- 实验提交截止日期：2018.12.16 24：00

二、 实验内容

决策树（decision tree）是一种基本的分类与回归方法。决策树模型呈树形结构，在分类问题中，表示基于特征对实例进行分类的过程。它可以认为是 if-then 规则的集合，也可以认为是定义在特征空间与类空间上的条件分布。其主要优点是模型具有可读性，分类熟读快。学习时，利用训练数据，根据损失函数最小化的原则建立决策模型。预测时，对新的数据，利用决策树模型进行分类。决策树学习通常包括 3 个步骤：特征选择、决策树的生成和决策树的修剪。

本实验指导用户实现 ID3、C4.5 和 CART 三种经典的决策树算法。

三、 实验目标

- 熟悉决策树算法的 3 个步骤，特征选择、决策树的生成和决策树的修剪。
- 熟悉 ID3 使用的特征选择策略——信息增益，并能够实现 ID3 算法。
- 熟悉 C4.5 使用的特征选择策略——信息增益率，并能够实现 C4.5 算法。
- 熟悉 CART 使用的特征选择策略——基尼系数和平方误差，并能够实现 CART 算法。
- 了解决策树的剪枝算法。

四、 实验操作步骤

本实验需要用到的 python 环境包括

名称	版本
Python	3.6.5
Numpy	1.14.3
matplotlib	2.2.2
jupyter	1.0.0

1. 启动 jupyter notebook 使用

参照实验一的任务指导书，使用 jupyter notebook 打开本实验的 DecisionTree.ipynb 文件。

2. 完成实验任务

任务 1 实现 get_max_num_class 函数。

任务 2 实现 split_data_set 函数。

任务 3 实现 count_values 函数。

任务 4 实现 generate_tree 函数。

任务 5 实现 compute_entropy 函数。

任务 6 实现 get_best_feature_id3 函数。

任务 7 实现 inference 函数。

任务 8 实现 get_best_feature_c45 函数。

任务 9 实现 compute_tree_cc 函数。

任务 10 实现 pruning 函数。

任务 11 实现函数。

任务 12 实现 compute_gini_index 函数。

任务 13 实现 get_best_feature_cart_decision 函数。

任务 14 实现 generate_cart_classifier 函数。

任务 15 实现 cart_classifier_inference 函数。

任务 16 实现 split_X 函数。

任务 17 实现 get_splitting_variable_and_point 函数。

任务 18 实现 generate_cart_regressor 函数。

任务 19 实现 cart_regressor_inference 函数。