# 机器学习决策树实验内容

## 一、 实验内容

详细内容见提供的 jupyter notebook

### 1.1 决策树处理分类任务

1. 使用 sklearn.tree.DecisionTreeClassifier 完成 dota2 比赛结果预测问题

2. 计算最大深度为 10 时，十折交叉验证的精度(accuracy)，查准率(precision)，查全率(recall)，F1 值

3. 绘制最大深度从 1 到 10 的决策树十折交叉验证精度的变化图

4. 选做：通过调整参数，得到一个泛化能力最好的模型

### 1.2 决策树处理回归任务

1. 使用 sklearn.tree.DecisionTreeRegressor 完成 kaggle 房价预测问题

2. 计算最大深度为 10 的决策树， 训练集上十折交叉验证的 MAE 和 RMSE

3. 绘制最大深度从 1 到 30，决策树在训练集和测试集上 MAE 的变化曲线

4. 选择一个合理的树的最大深度，并给出理由

### 1.3 实现决策树

使用 LendingClub Safe Loans 数据集：

1. 实现信息增益、信息增益率、基尼指数三种划分标准

2. 使用给定的训练集完成三种决策树的训练过程

3. 计算三种决策树在最大深度为 6 时在训练集和测试集上的精度，查准率，查全率， F1值

### 1.4 实现预剪枝

使用 Lending Club Safe Loans 数据集：

1. 实现使用信息增益率划分的预剪枝

2. 计算出带有预剪枝和不带预剪枝的决策树的精度，查准率，查全率和 F1 值(最大深度为6，使用信息增益率)，保留 4 位小数，四舍五入

3. 选做：使用Echarts绘制决策树

4. 选做：绘制其他的决策树

### 1.5 选做： 实现后剪枝

1. 实现带有后剪枝的决策树

2. 数据集随意

3. 最后对比剪枝和不剪枝的差别

## 二、数据介绍

### 2.1 LendingClub Safe Loads 数据集

文件名：

1. 原始数据： lending-club-data.csv

2. 字段说明： 无

数据来源： <https://www.coursera.org/learn/ml-classification>

读取方式：

import pandas as pd

loans = pd.read\_csv('data/lendingclub/lending-club-data.csv', low\_memory=False)

### 2.2 白葡萄酒质量数据集

文件名：

1. 原始数据： winequality-white.csv

2. 字段说明： 葡萄酒字段说明.txt

数据来源： <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

数据无需预处理，读取即可使用

import pandas as pd

data = pd.read\_csv('data/wine\_quality/winequality-white.csv', delimiter=";")

### 2.3 kaggle 房价预测

文件名：

1. 原始数据： kaggle\_hourse\_price\_train.csv

2. 字段说明： kaggle 房价预测字段说明.txt

3. 预处理脚本： preprocess.py

4. 处理后数据： kaggle 房价预测处理后数据.csv

数据来源： <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

数据预处理步骤：

该数据集是真实数据集含有大量的缺失值和噪声数据，提供的预处理脚本只对含缺失值的特征进行丢弃处理，学生可根据自己的需要进行其他方式的处理。

读取方式：

import pandas as pd

data = pd.read\_csv('data/kaggle\_house\_price\_prediction/kaggle\_hourse\_price\_train.csv')

### 2.4 Dota2 Games Results Data Set Dota2 游戏结果数据集

文件名：

1. 原始数据： dota2Train.csv

2. 字段说明： dota2 比赛结果字段说明.txt

数据来源： <http://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results>

数据无需预处理，读取即可使用

import numpy as np

data = np.loadtxt('data/dota2Dataset/dota2Train.csv', delimiter=',')

### 2.5 spambase 垃圾邮件数据集

文件名：

1. 原始数据： spambase.data

2. 字段说明： spambase 数据说明.txt

数据来源： <http://archive.ics.uci.edu/ml/datasets/spambase>

数据无需预处理，读取即可使用

import numpy as np

data = np.loadtxt('data/spambase/spambase.data', delimiter = ",")

三、 准备工作  
**安装 pyecharts，我们需要安装这个库绘制决策树。**

**Pyecharts使用可见：**<https://pyecharts.org/#/zh-cn/intro>

**pip 安装**：$ pip(3) install pyecharts

（anaconda可在anaconda prompt中进行安装）

**源码安装：**

$ git clone https://github.com/pyecharts/pyecharts.git

$ cd pyecharts

$ pip install -r requirements.txt

$ python setup.py install

（或者执行 python install.py）