实验报告

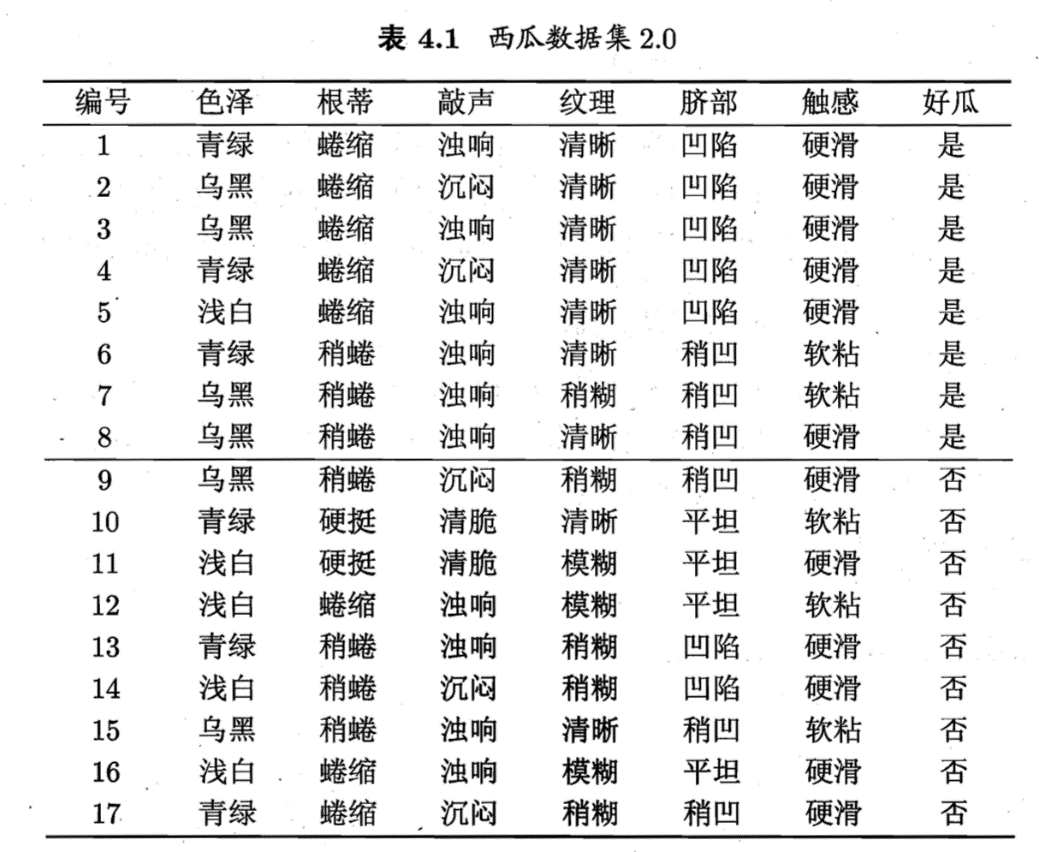
实验项目：决策树生成

实验日期：2020/4/18

1. 数据处理

本次实验采用《机器学习》表4.1 西瓜数据集2.0作为实验数据（如图1）。考虑到存储方便等问题，我们将每一样本作为一个向量进行存储，向量的不同维度是该样本在不同属性上的取值（如图2），为了使最终可视化结果直观，我们将样本类别直接作为向量的一个维度，并未进行编码。同时，我们也存储了一个列表用以保存所有的属性名称，最终的数据集结果如图3。

**图 1西瓜数据集2.0**





**图 2不同属性的值的编码**



**图 3数据存储格式**

1. 程序说明

·data\_set()

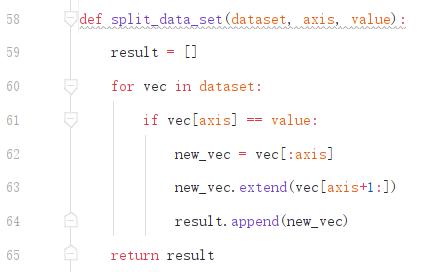
生成并返回图3所示的数据集和属性标签

·shannon\_ent(dataset)

计算香农熵

·split\_data\_set(dataset, axis, value)

对于给定的dataset，按照axis下标的属性的value值对数据集进行分割，返回的数据集是所有axis下标的属性值为value的样本除去该属性后的向量集合



**图 4分割数据集函数**

result列表保存分割后的向量，首先遍历数据集，找到下标axis的属性值为value的所有样本（60-61行），接着该向量进行切片分割，新向量保存在result中。最后返回result。

·best\_split(dataset)

对于给定的数据集，依次计算按照该数据集不同属性进行划分所带来的信息增益，选择能够使得信息增益最大的划分属性的下标进行返回。



**图 5选取最佳划分方式函数**

72行依次对每一个属性进行遍历，74-79行计算按照该属性划分的信息增益，80-82行判断是否为目前的最佳划分。遍历结束后，返回最佳划分属性的下标。

·vote(classlist)

对应于递归构造函数中的一个原子操作。若所有特征遍历完成，则采用投票方式，最终类别与多数类别保持一致。

·creat\_tree(dataset,label)

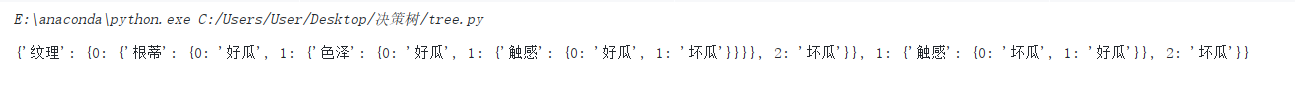
递归构造决策树。



**图 6递归构造决策树函数**

首先，kinds列表保存当前dataset中的样本的列表标签。102-103行是递归操作的原子操作之一，若当前数据集所有样本类别一致，则将其样本类别作为叶子节点返回。104-105行是递归操作的另一个原子操作，若当前dataset已经遍历完所有特征，无法继续划分，则按照投票方法选择多数特征作为叶子节点返回。

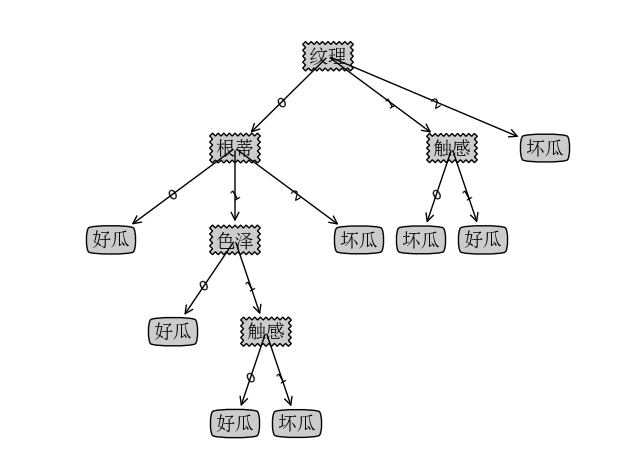
参考网上的指点，我们使用字典的格式保存最终的决策树结构。递归过程中，我们每次选择一个最佳划分特征，对数据集按照该特征的不同值进行划分，对于划分后的数据集，我们再递归的调用该函数继续向下构建决策树，直到满足原子操作的条件，则返回叶子节点。

1. 实验结果

**图 7实验结果**

图7为实验结果。

由于字典形式的结果不够直观，我们借鉴了《机器学习实战》中的代码，对其进行了可视化的呈现。可视化的代码为graph.py，具体使用方法在程序中做出了说明。



**图 8可视化实验结果**

1. 总结

在这一次实验中，我们使用代码实现了决策树的生成过程，对于理论性的知识进行了实战的练习，加深了对于决策树方法的理解。