机器学习第二次作业
——决策树

2052902 韩意

一、 问题描述

试用 Python 编程实现基于 C4.5 决策树算法来进行最优划分属性选择的决策树，并为表 4.3 西瓜数据集 3.0 中去掉“密度属性和编号为 9 的西瓜”以后的数据生成一棵决策树。

表 4.3 西瓜数据集 3.0

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

二、 算法实现

本题使用 Python 实现，用到的库包括 Pandas，NumPy 和 Matplotlib，其中 Pandas 用于数据存储，Numpy 用于数据处理，Matplotlib 用于决策树的绘制。

分析题意可知数据包括离散（categorical）和连续（numerical）两种类型，与 sklearn 的输入要求不同，这里为了能够直接对字符串这样的离散特征进行处理，

选择 pandas 来存储这一表格，由于 pandas 有强大的数据处理 API，这样也便于后续决策树的计算生成。

代码中将决策树算法模型实现为了一个类，便于统一的管理与封装。其中，存储决策树的数据结构通过 Python 字典的嵌套结构实现，最终形成的决策树字典如下：

```
tree = {
    '含糖率≤0.126': {
        '是': '坏瓜',
        '否': {
            '根蒂': {
                '蜷缩': '好瓜',
                '稍蜷': {
                    '脐部': {
                        '稍凹': {
                            '含糖率≤0.3035': {
                                '是': '好瓜',
                                '否': '坏瓜'
                            }
                        },
                        '凹陷': '坏瓜'
                    }
                },
                '硬挺': '坏瓜'
            }
        }
    }
}
```

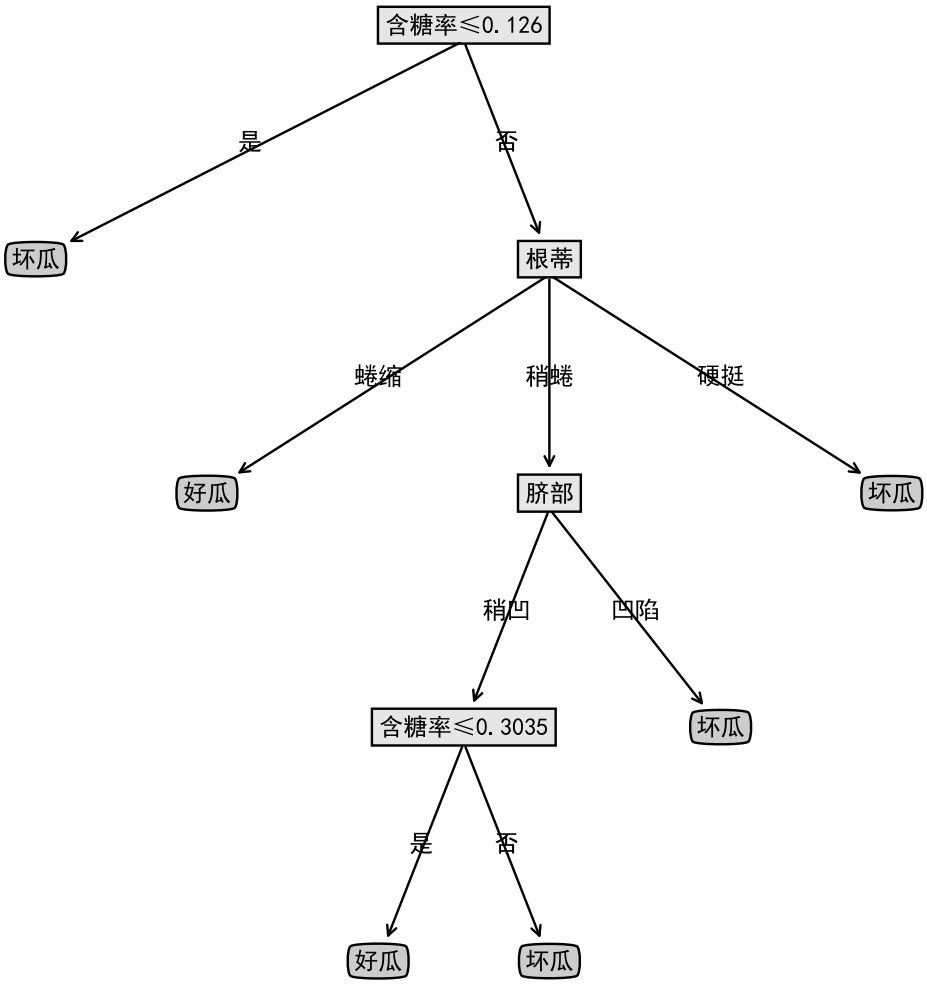
这样可以在可视化前较清晰地查看决策树的层次结构。

注意在使用决策树算法时，连续属性可以重复使用，而离散属性不能再使用，这一点通过 pandas 中 DataFrame 的 drop 方法实现。

C4.5 算法使用信息增益率准则，对可取值数目较少的属性有所偏好，可以看到根结点选择的是连续属性，正是因为连续属性离散化时，可取值数目仅为 2（小于等于和大于两种）。

经测试，该决策树对原始数据集的预测准确率为 100%。

决策树的绘制基于 plt.annotate()函数实现，最终得到的决策树绘制如下，其中叶子结点的颜色稍深，以示区分：


含糖率≤0.126
是
否
坏瓜
根蒂
蜷缩
稍蜷
硬挺
好瓜
脐部
坏瓜
稍凹
凹陷
含糖率≤0.3035
坏瓜
是
否
好瓜
坏瓜