

人工智能导论课程实验报告

|  |  |
| --- | --- |
| 实验名称： | 实验五  决策树算法ID3 |
| 实验日期： | 2024-5-2 |
| 实验地点： | 西部片区4号楼301 |
|  | |
| 学号： | 33920212204567 |
| 姓名： | 任宇 |
| 专业年级： | 软工2021级 |
| 学年学期： | 2023-2024学年第二学期 |

1. **实验目的**

编程实现决策树算法ID3；理解算法原理。

1. **实验内容**

利用taindata.txt的数据（75\*5，第5列为标签）进行训练，构造决策树；利用构造好的决策树对testdata.txt的数据进行分类，并输出分类准确率。

1. **实验过程**
2. ID3算法的基本原理：

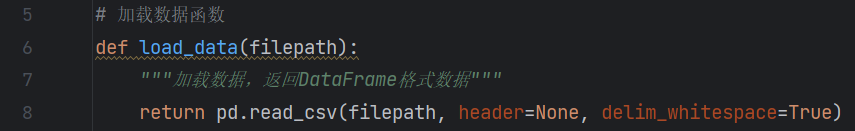
ID3算法的核心思想就是以信息增益度量属性选择，选择分裂后信息增益最大的属性进行分裂。

对于特征属性为连续值，可以如此使用ID3算法：先将D中元素按照特征属性排序，则每两个相邻元素的中间点可以看作潜在分裂点，从第一个潜在分裂点开始，分裂D并计算两个集合的期望信息，具有最小期望信息的点称为这个属性的最佳分裂点，其信息期望作为此属性的信息期望。

1. 使用ID3算法编程解决testdata.txt数据集分类问题：

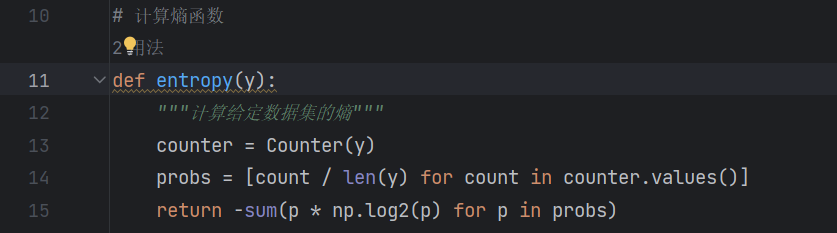
* 加载数据（load\_data函数）

使用pandas库从文件中读取数据，并返回一个DataFrame对象。



* 计算熵（entropy函数）

使用信息熵来度量数据集的不确定性，这是选择最优分割点的关键依据。Counter 用于计算每个类别的频数，进而计算概率和熵值。



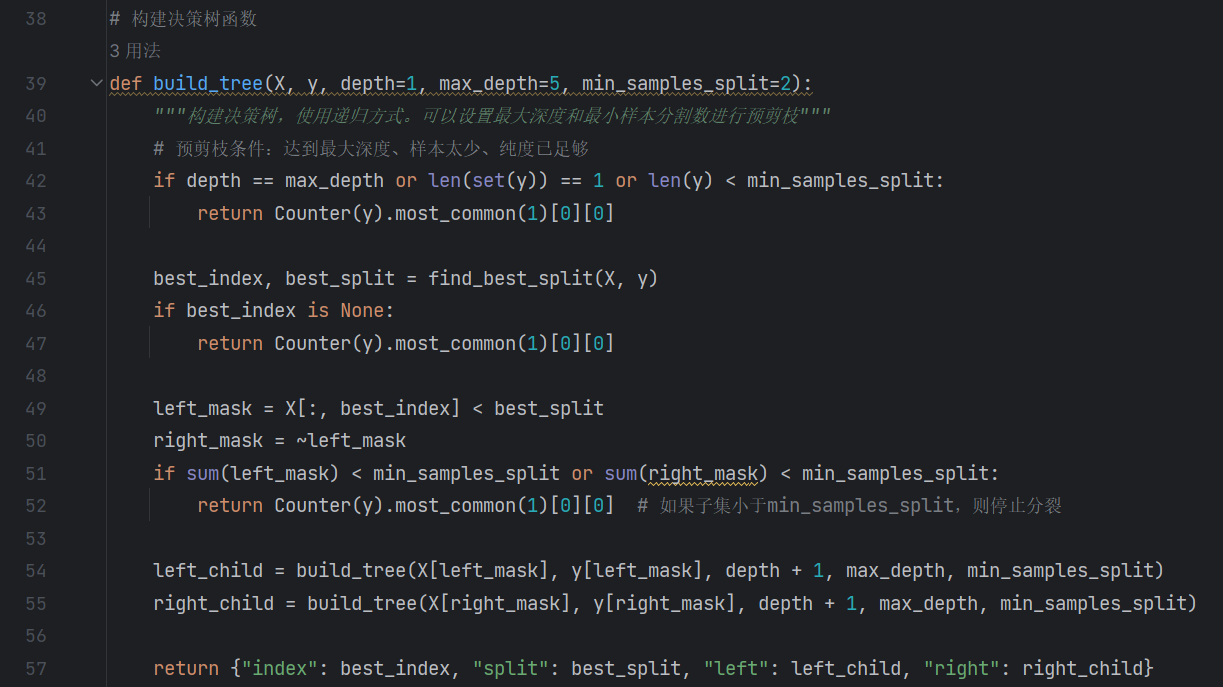
* 寻找最佳分割点（find\_best\_split函数）

遍历所有特征及其可能的分割值（连续变量的中点），计算每个可能分割点的信息熵。选择能最大化信息增益（即最小化分割后的熵）的分割点作为最佳分割点。



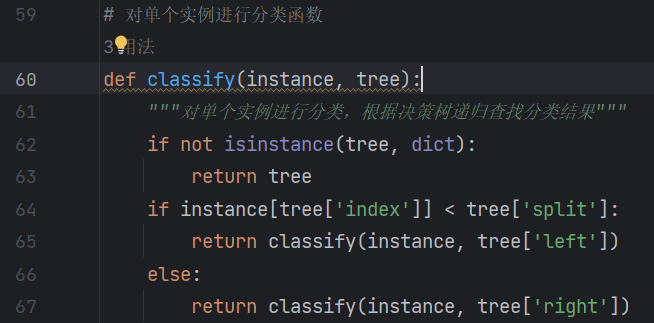
* 构建决策树（build\_tree函数）

递归构建决策树，直到满足停止条件：达到最大深度、样本太少或样本纯度已足够（所有样本属于同一类别）。这个函数使用了预剪枝技巧，如限定树的最大深度（max\_depth）和最小样本分割（min\_samples\_split），这有助于防止过拟合并提高模型的泛化能力。



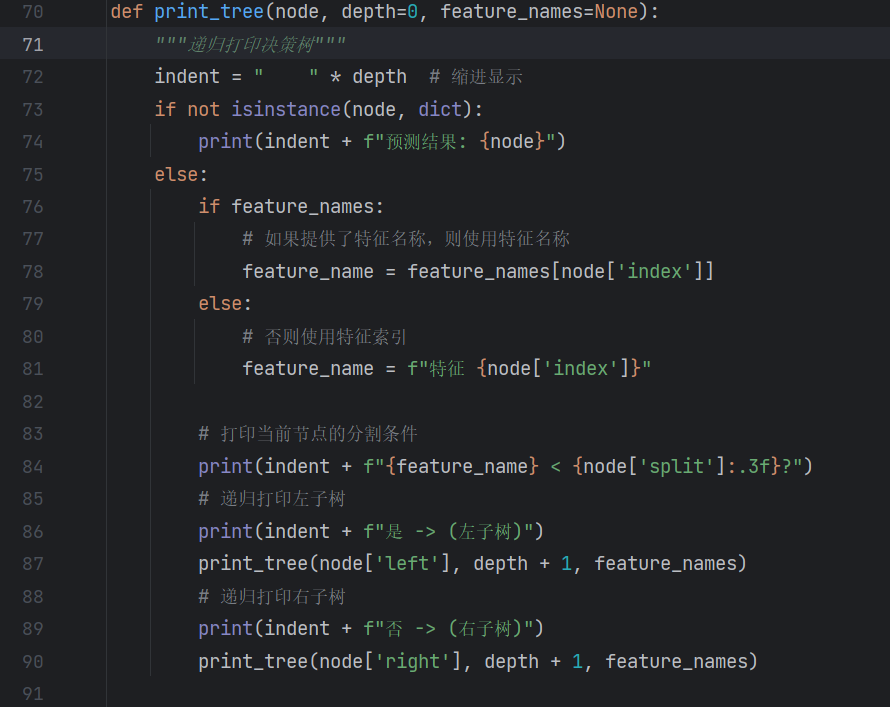
* 分类函数（classify函数）

对给定实例进行分类，通过递归遍历决策树来确定实例的类别。



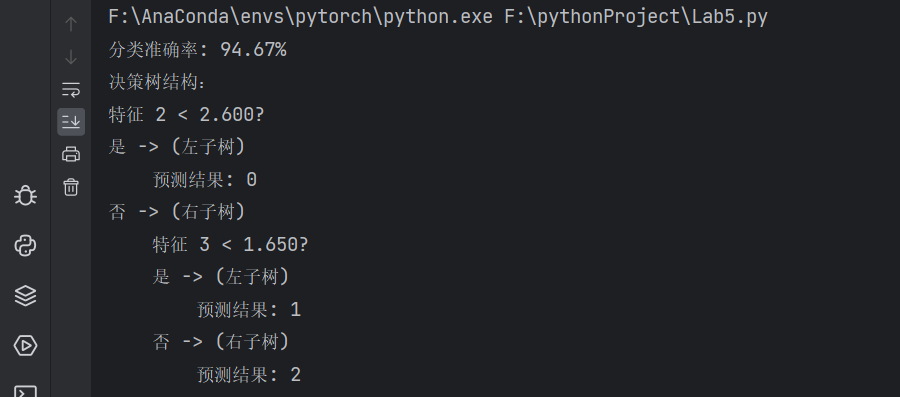
* 打印决策树（print\_tree函数）

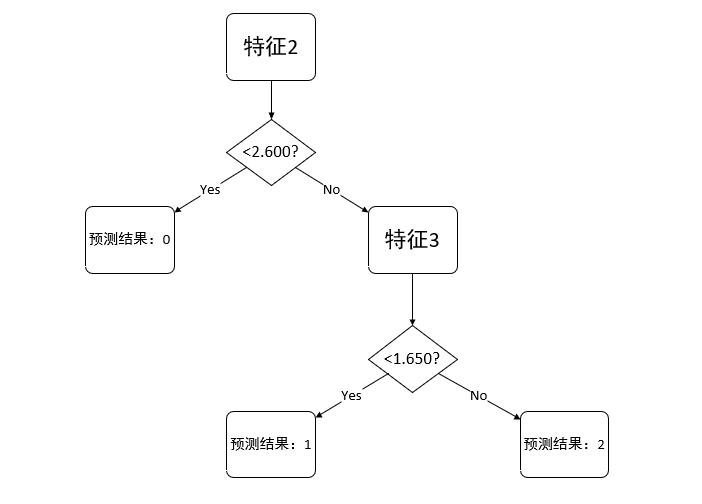
递归打印决策树的结构，清晰显示每个决策节点的条件，便于理解模型的决策路径。



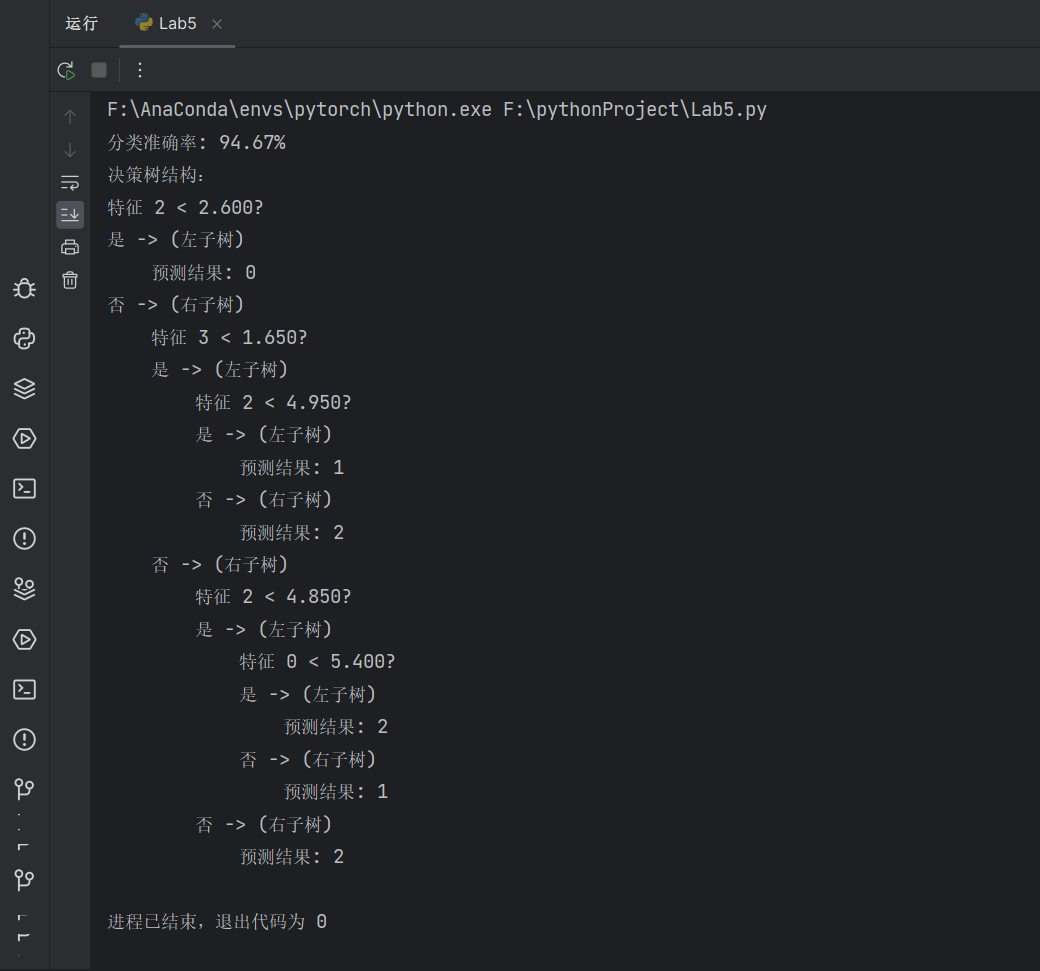
1. 运行结果：

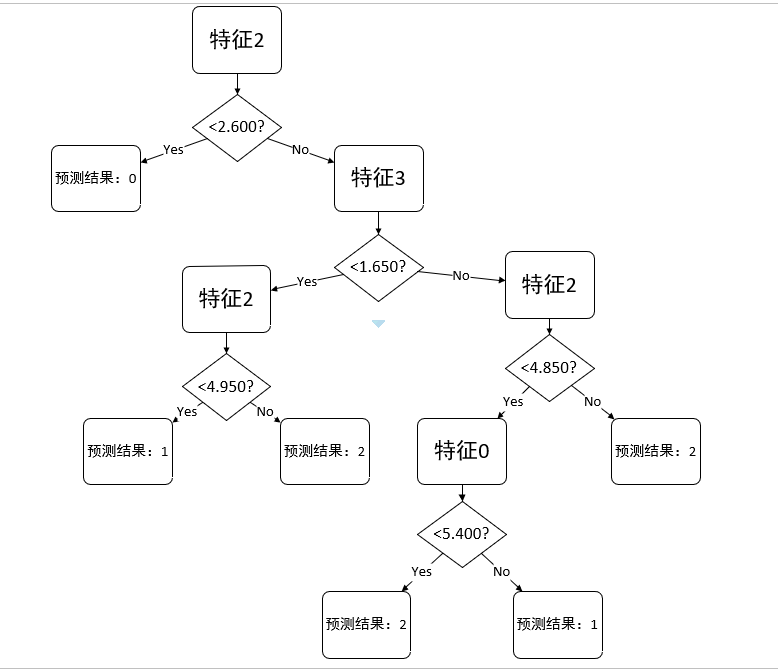
构建决策树，设置最小分裂样本数为2，最大深度为3：





构建决策树，设置最小分裂样本数为1，最大深度为5：





尽管因为最小分裂样本数和最大深度不同而构造了不同的决策树，但二者准确度一样，这也许是因为样本数据量太小。

1. **实验思考及心得**

完成这次实验让我深入理解了决策树ID3算法的核心原理及其在实际问题中的应用。通过亲手编程实现ID3算法，我不仅学会了如何处理连续属性、如何计算信息增益，还学到了如何通过预剪枝技巧防止决策树过拟合。这次实验加深了我对机器学习模型泛化能力和模型复杂度之间关系的认识。尤其是在相同准确率下构建不同复杂度的决策树，让我明白了在实际应用中如何权衡模型的简洁性与性能，以及模型参数对学习结果的影响。这不仅提升了我的实际操作能力，也激发了我进一步探索更复杂数据科学问题的兴趣。