**1.实验目的**

构建一棵基于 ID3 算法的决策树,并用该决策树对小数据集进行预测,以此熟悉掌握决策树的 ID3 算法,然后训练学习一些大规模数据并且完成决策,并期望利用该决策树对于一些样例做出正确的预测。

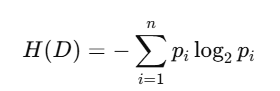
**2.算法核心**

决策树ID3算法的核心是基于信息增益选择最优属性进行分类。它主要包括以下几个核心步骤：

计算信息熵（Entropy）：

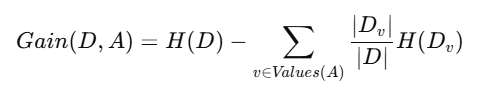
信息熵是衡量数据混乱程度的指标。数据越纯（同一类别占比越大），熵越低；数据越混乱（类别分布越均匀），熵越高。

熵公式：



计算信息增益（Information Gain）：

信息增益是通过某个属性划分数据后，熵的减少量，表示属性对分类的贡献。信息增益公式：



选择最优属性：

对所有候选属性，计算其信息增益，选择信息增益最大的属性作为当前节点的划分属性。

递归构建决策树：

根据最优属性划分数据，生成子节点；

对每个子节点重复上述步骤，直到满足停止条件（如数据纯度达到一定水平或没有更多属性可用）。

停止条件：数据集中所有样本属于同一类别；没有剩余的候选属性；达到预设的树深度限制。

核心思想：

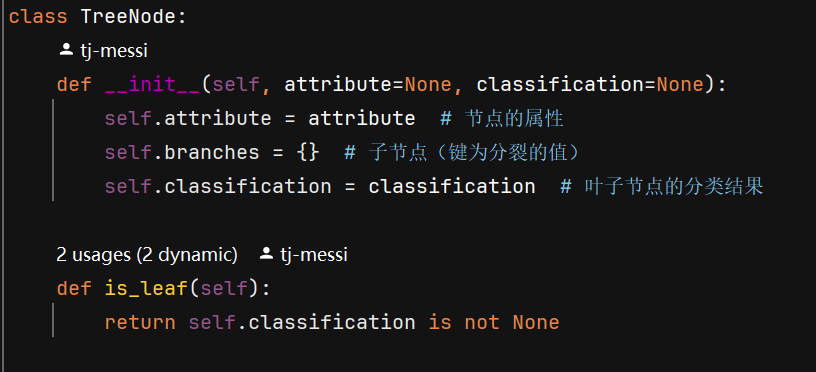
通过逐步减少数据的不确定性，递归地将数据划分为更纯的子集，最终形成一个以最优属性为分裂点的树形结构，用于分类或决策。

**3.实验过程**

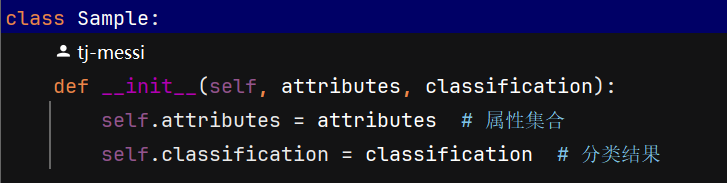
**3.1数据结构定义**

TreeNode表示决策树的节点，包含属性 attribute（当前节点的划分属性）、branches（分支字典）和 classification（叶子节点的分类结果）。

提供方法 is\_leaf() 判断节点是否为叶子节点。

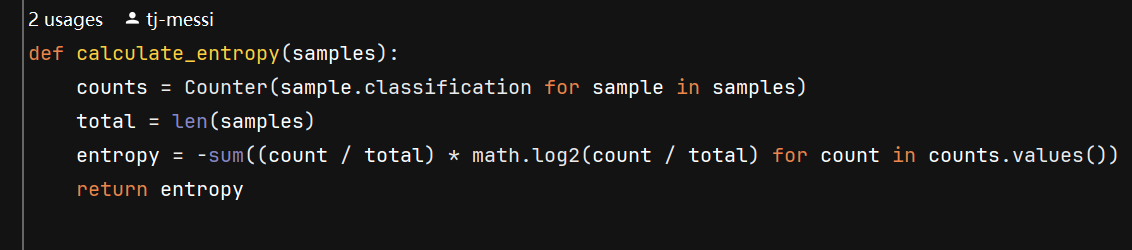


Sample表示一个样本数据，包括 attributes（属性值列表）和 classification（分类结果）。

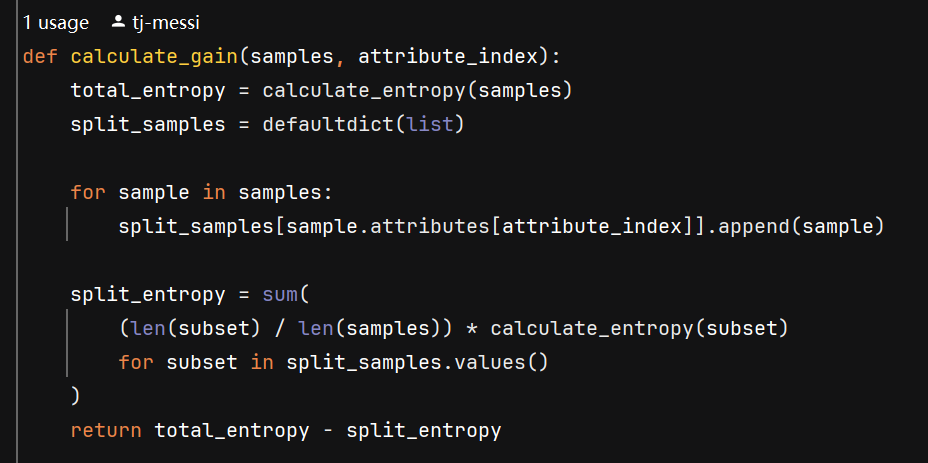


**3.2功能函数定义**

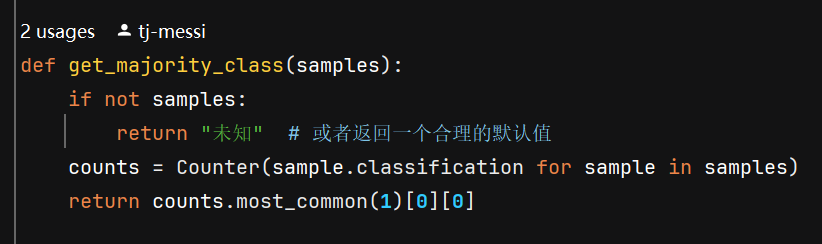
calculate\_entropy(samples)：计算给定样本集的熵，用于衡量数据的不确定性。



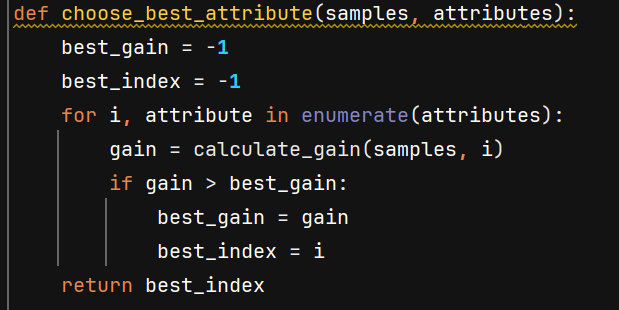
calculate\_gain(samples, attribute\_index)：计算某个属性对样本集的信息增益。



get\_majority\_class(samples)：返回样本集中数量最多的类别，作为当前节点的分类结果。此时要注意samples中没有分类好的先别输出。



choose\_best\_attribute(samples, attributes)：遍历所有属性，计算每个属性的信息增益。 返回信息增益最大的属性的索引，作为当前节点的划分属性。

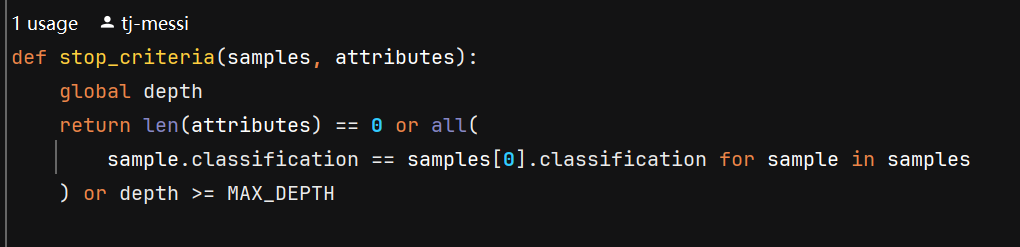


**构建决策树：**

stop\_criteria(samples, attributes)：

判断是否满足停止条件：

1. 属性集合为空。
2. 所有样本的分类相同。
3. 达到最大深度限制。

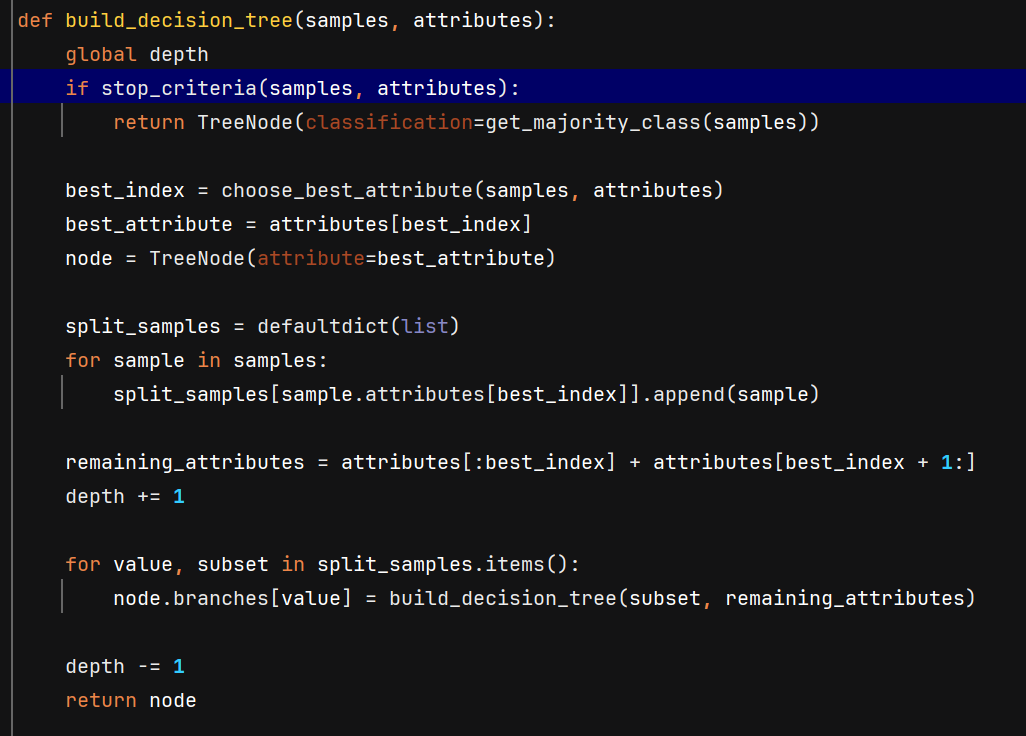


build\_decision\_tree(samples, attributes)：递归构建决策树的主函数：

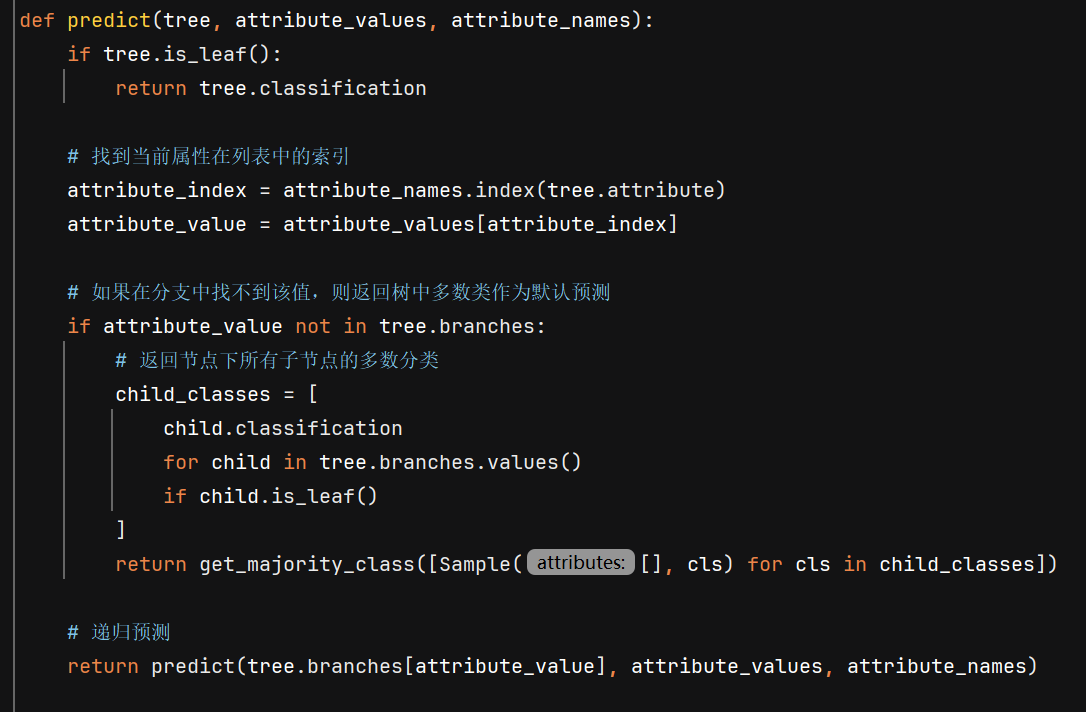
1.判断是否满足停止条件，若满足则创建叶子节点。

2.选择最优划分属性，基于该属性将样本划分为子集。

3.递归构建每个子节点，并将子节点挂载到当前节点上。

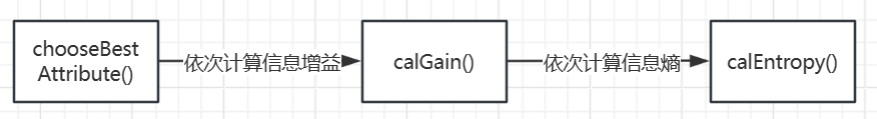


predict(tree, attribute\_values, attribute\_names)： 输入测试样本的属性值，通过递归从根节点到叶子节点进行预测。如果遇到测试样本的属性值不在训练集中，返回当前节点所有子节点的多数类。

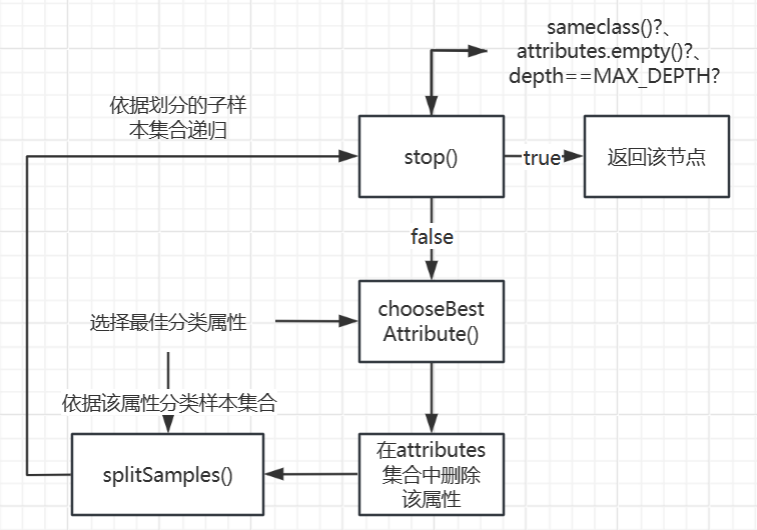


calculate\_accuracy(tree, samples, attribute\_names)：遍历样本集，利用决策树预测每个样本的分类结果，与实际分类对比，计算分类准确率。

**3.3程序总流程**



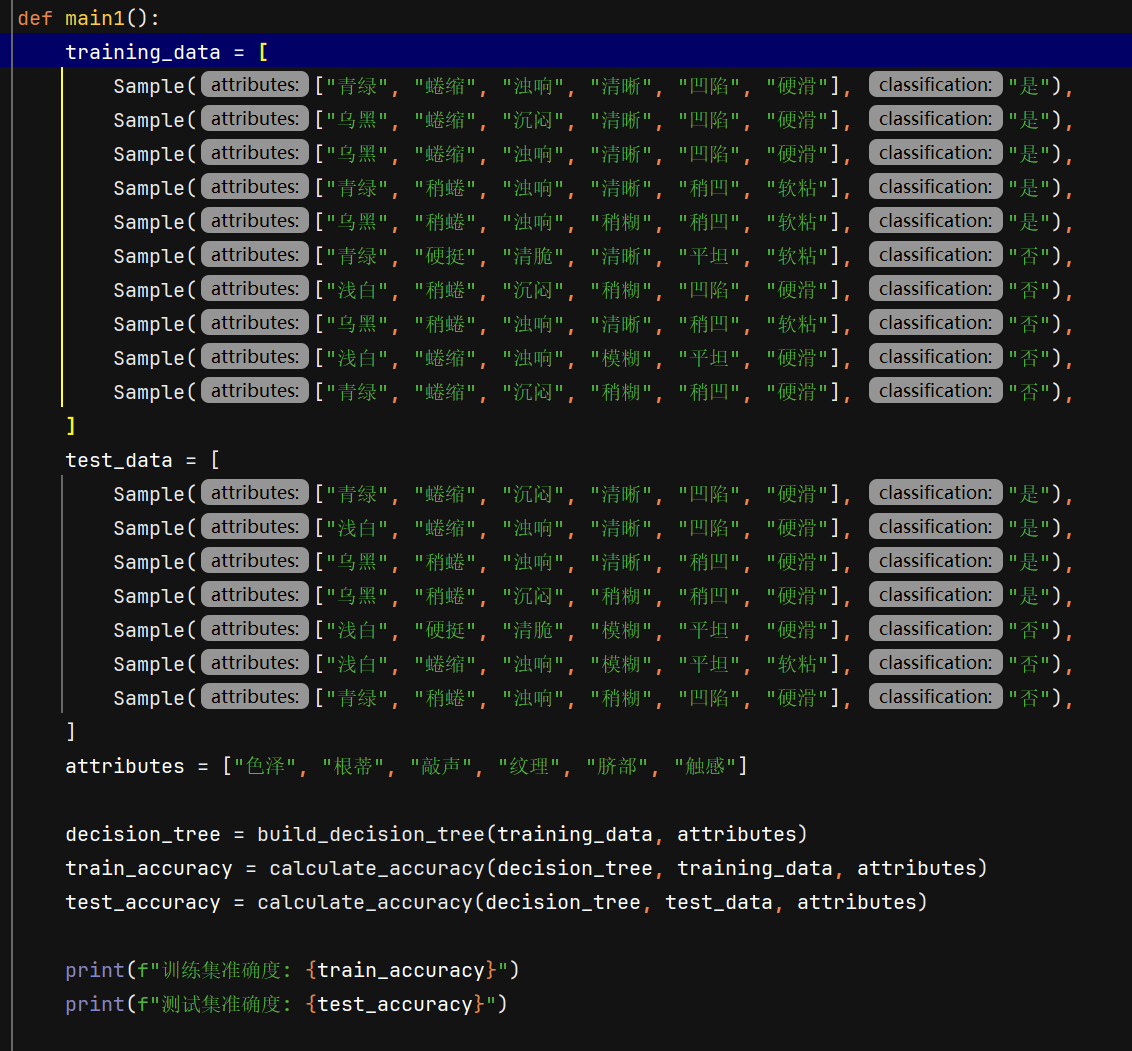
先计算信息熵关系挑选出最佳分类属性



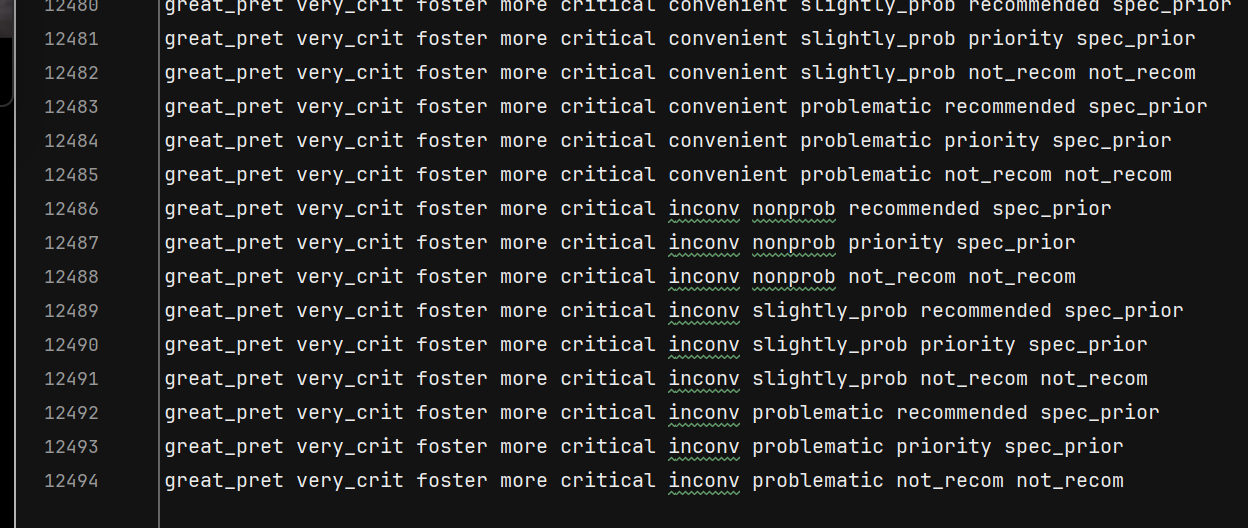
选择了分类属性之后用这个最佳属性来分类样本。

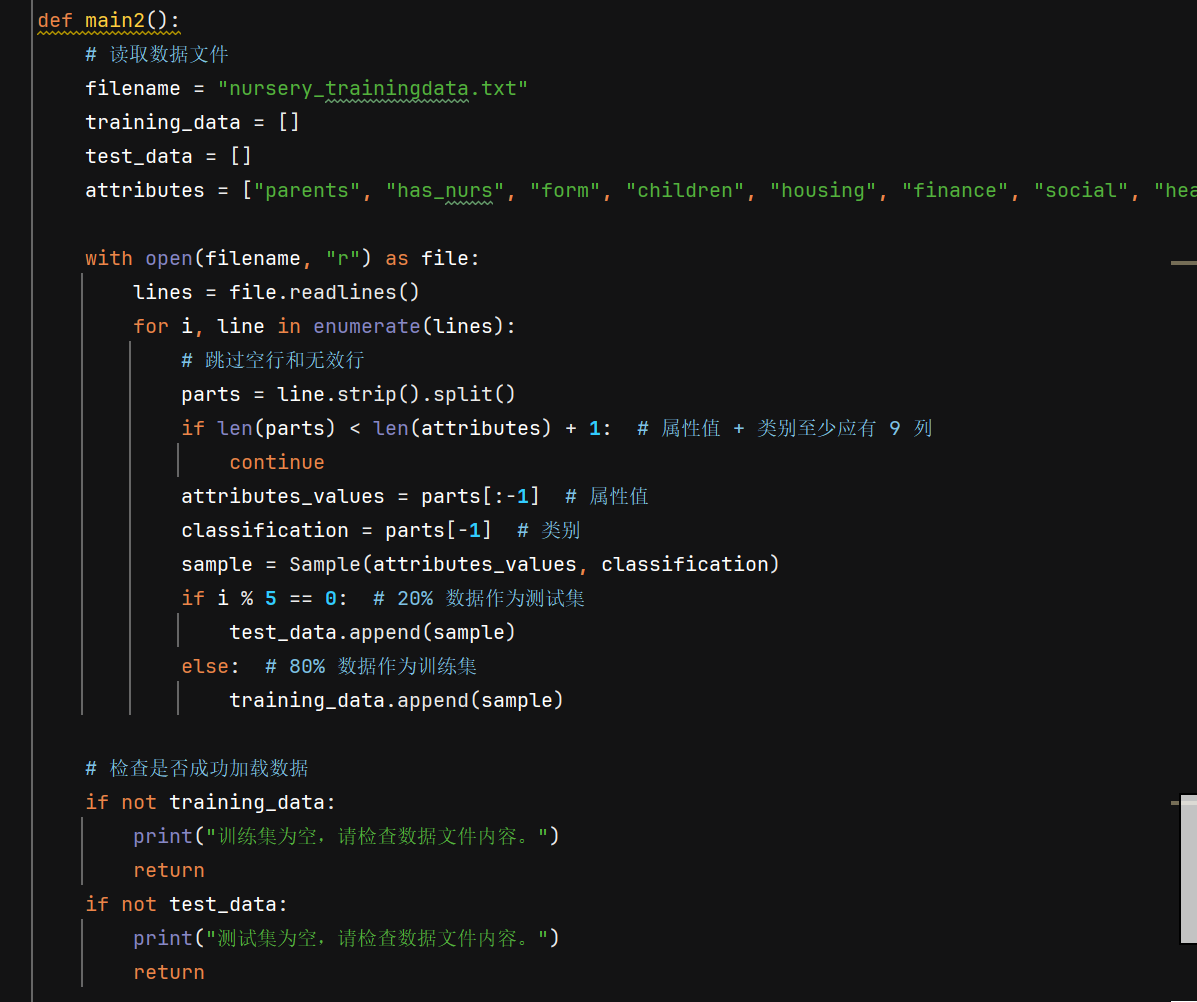
**4.选择数据**

选择了一个小样本数据和一个大数据。



之后选择了一个大数据nursery trainingdata



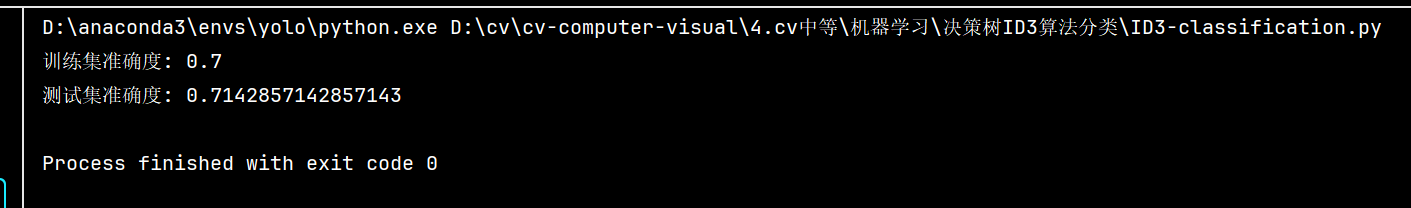


来训练。

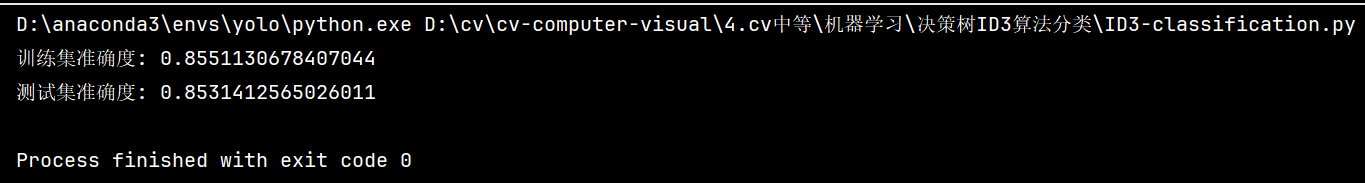
**5.结果分析**

。训练结果如下：

小数据：



大数据：



综合以上两个案例:case 1:西瓜的小样例(共 10 个训练样例);case 2:nursery 的大样例(共 12000 个训练样例),发现: .当样例较小时,生成的决策树对于训练集的准确度很高,但此时的决策树的泛化能力显然较弱;当样例较大时,生成的决策树对于训练集的准确度并不是很高,但此时该决策树的泛化能力较强;

上述训练样例符合决策树的特点：容易过拟合。

**6.总结**

本实验根据 ID3 算法构建了一棵基本的决策树;构建决策树包括构建了相关的数据类型(结构体),构建了生成决策树的 buildDecisionTree()等函数,同时构建了一些利用生成的决策树生成数据的函数,如 predictData()(计算准确率);

待改进: 1.采用了 ID3 算法,故比较难以处理连续的特征,如:工资、长度等。同时

ID3 算法决策的核心是信息增益(贪心),故其会倾向于选择取值较多的属性,故可

能并不是整体上的最优解;2.本程序并未考虑到缺失值的问题,留待后续学习了 ID4.5 后自行改进;