**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。

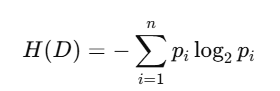
**2.算法核心**

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）：

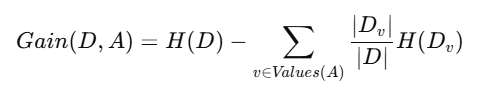
信息熵是衡量数据混乱程度的指标。数据越纯（同一类别占比越大），熵越低；数据越混乱（类别分布越均匀），熵越高。

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：

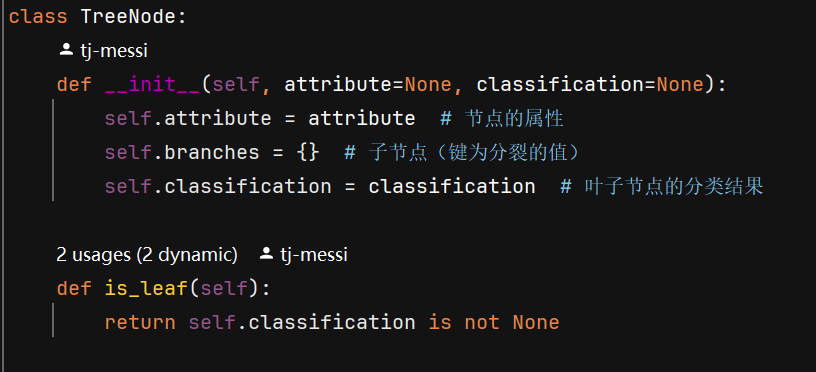
通过逐步减少数据的不确定性，递归地将数据划分为更纯的子集，最终形成一个以最优属性为分裂点的树形结构，用于分类或决策。

**3.实验过程**

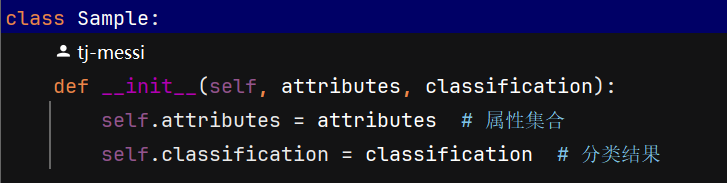
**3.1数据结构定义**

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）。



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