# 《机器学习与深度学习》课程

# 实 验 报 告



**姓 名： 金家耀**

**专 业：**  人工智能

**学 号： 1193210320**

**江南大学人工智能与计算机学院**

# 决策树分类

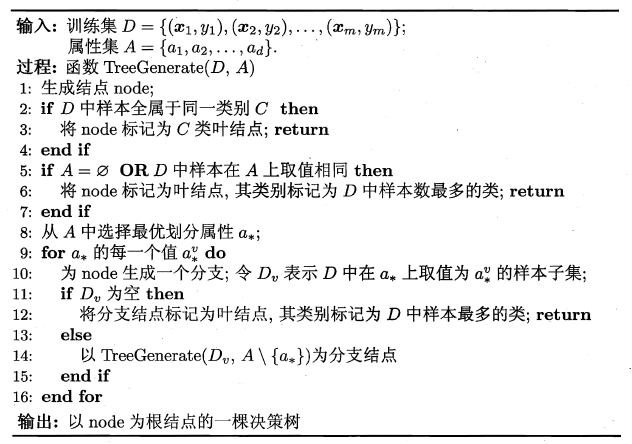
**1实验目的**

决策树是一类常用的机器学习方法，能把特征选择与分步决策相结合来完成分类或预测任务。本实验目的在于加深学生对决策树基本原理及相应算法的理解，

体会其在模式分类、识别中的作用。

**2实验原理**

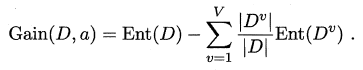
一棵决策树包含一个根结点、若干个内部结点和若干个叶结点; 叶结点对应于决策结果，其他每个结点则对应于一个属性测试;每个结点包含的样本集合根据属性测试的结果被划分到子结点中;根结点包含样本全集. 从根结点到每个叶结点的路径对应了一个判定测试序列. 决策树学习的目的是为了产生一棵泛化能力强，即处理未见示例能力强的决策树，其基本流程如下：

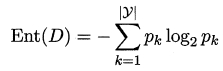


决策树学习的关键是第8 行，即如何选择最优划分属性. 一般而言，随着划分过程不断进行，我们希望决策树的分支结点所包含的样本尽可能属于同一类别，即结点的"纯度" (purity) 越来越高. 度量样本集合纯度常用的指标包括：

1）信息增益（ID3决策树学习算法采用）

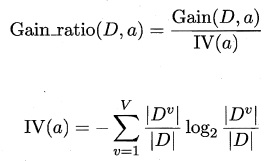
属性α对样本集D 进行划分所获得的"信息增益"为：





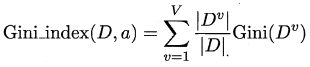
2）增益率（C4.5决策树学习算法）

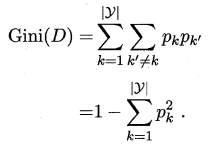
属性α对样本集D 进行划分所获得的"增益率"为：



3）基尼指数（CART决策树学习算法采用）

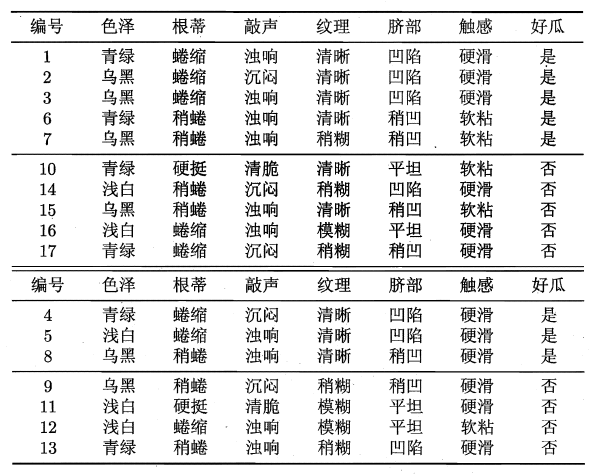
属性α的基尼指数定义为：





**3实验内容**

利用如下所示的西瓜数据集进行实验，训练生成决策树并进行测试。表中双线上部为训练集，双线下部为测试集。



**4实验要求**

1. 根据实验原理中给出的基本流程设计实现决策树学习算法，要求度量样本集合纯度指标采用“信息增益”，画出决策树并给出相应的测试结果。
2. 将度量样本集合纯度指标分别替换为“增益率”和“基尼指数”，给出相应的决策树，并比较分类结果和分类精度，体会不同集合纯度指标对算法的影响；

**5实验代码和结果**