# 第五章 决策树

## 一、填空题

1. 大多数决策树的节点类型可以分为(　　)和(　　)。根节点在类型上属于(　　)节点，剪枝节点减去的是(　　)节点？

## 二、判断题

2. 基于基尼指数的分裂结果与基于信息增益比的分裂结果总是相同的。(　　)
3. 信息增益越大的特征信息增益比也越大。(　　)
4. 信息增益比越大的特征不纯度越小。(　　)

## 三、选择题

5. 假设离散特征 A 所有可能的取值为$\{a_1, a_2, a_3\}$, 设当前节点样本集为$S$, 且有$A(\boldsymbol{x}_i) \neq a_2, \forall \boldsymbol{x}_i \in S$，则利用特征 A 进行分裂后产生几个子节点：(　　)

A. 2 个　　B. 3 个　　C. 无法确定　　D. 无法分裂

6. 当前节点内样本集包含 5 个样本，特征数量为 3，其中两个离散特征，1 个连续特征，采用多叉树方案，则当前节点共有多少种可选的分裂方案。

A. 1 种　　B. 2 种　　C. 3 种　　D. 无法确定

## 四、简答题

7. 简述如何利用“分而治之”策略解决复杂非线性分类问题
8. 什么情况下一个叶子节点中会没有样本，此时该叶子节点返回的类别标签如何确定。
9. 简述前剪枝与后剪枝的差别及各自的特点

## 五、计算（画图）题

10. 一下关于 iris 数据库的某个特征增广版本包含 7 个样本，具体情况如下：

| 序号 | 香气 | 颜色 | 花萼长度 | 花萼宽度 | 花瓣长度 | 花瓣宽度 | 类别 |
|---|---|---|---|---|---|---|---|
| 1 | 有 | 红 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 2 | 有 | 红 | 4.9 | 3 | 1.4 | 0.2 | setosa |
| 3 | 有 | 粉 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4 | 有 | 紫 | 5.3 | 3.7 | 1.5 | 0.2 | setosa |
| 5 | 无 | 粉 | 7 | 3.2 | 4.7 | 1.4 | versicolor |
| 6 | 无 | 紫 | 6.4 | 3.2 | 4.5 | 1.5 | versicolor |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7 | 无 | 紫 | 6.3 | 3.3 | 6 | 2.5 | virginica |
| 8 | 有 | 紫 | 5.8 | 2.7 | 5.1 | 1.9 | virginica |

1) 请给出采用颜色特征进行多叉树分裂的结果
2) 请给出采用花瓣长度进行二叉树分裂的任意一个结果
3) 对比香气和颜色两种离散特征，分别依据信息增益、信息增益比和基尼指数给出相应的分裂特征选择结果，并给出计算过程。

11. 根据表 5.1 计算各个特征的信息增益、信息增益比、基尼系数。

| 编号 | 天气 | 温度 | 湿度 | 有无风 | 是否出去玩 |
|---|---|---|---|---|---|
| 1 | 晴 | 热 | 高 | 无 | 否 |
| 2 | 晴 | 热 | 高 | 有 | 否 |
| 3 | 阴 | 热 | 高 | 无 | 是 |
| 4 | 雨 | 温和 | 高 | 无 | 是 |
| 5 | 雨 | 凉爽 | 正常 | 无 | 是 |
| 6 | 雨 | 凉爽 | 正常 | 有 | 否 |
| 7 | 阴 | 凉爽 | 正常 | 有 | 是 |
| 8 | 晴 | 温和 | 高 | 无 | 否 |
| 9 | 晴 | 凉爽 | 正常 | 无 | 是 |
| 10 | 雨 | 温和 | 正常 | 无 | 是 |
| 11 | 晴 | 温和 | 正常 | 有 | 是 |
| 12 | 阴 | 温和 | 高 | 有 | 是 |
| 13 | 阴 | 热 | 正常 | 无 | 是 |
| 14 | 雨 | 温和 | 高 | 有 | 否 |

1) 使用表 5.1 数据集，根据 ID3 决策树算法，手动生成一棵决策树，用于预测是否应该出去玩。请写出根节点的分裂依据计算过程
2) 使用表 5.1 数据集，根据 C4.5 决策树算法，在不考虑剪枝的情况下，手动生成一颗决策树，用于预测是否应该出去玩。请写出根节点的分裂依据计算过程