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（内部节点）和（叶子节点）。根节点在类型上属于（内部节点）节点，剪枝节点减去的是（叶子节点）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：（B）

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。（C）

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

答：

利用一系列简单的分类器将样本空间分割为多个局部区域R_t，使得在每个局部区域R_t中训练样本从属于某一个类别ω_j的后验概率$p(\omega_j|\boldsymbol{x}_i), \forall \boldsymbol{x}_i \in R_t$。在测试时，对于落入区域$R_t$的任意样本$\boldsymbol{x} \in R_t$，给出分类结果为$\omega_j$

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。

答：

当该叶子节点的父节点对应的样本集中没有符合该叶子节点特征取值要求的样本，但分裂特征的所有可能取值的集合中又包含了该叶子节点对应的特

征取值时，该叶子节点中没有样本。

此时，该叶子节点返回其父节点中样本占比最多的类别标签。

9. 简述前剪枝与后剪枝的差别及各自的特点

答：

两者的差别：

前剪枝是在决策树的生成过程中同步进行剪枝；后剪枝则是在决策树生成后逐步剪去叶子节点；

特点：

前剪枝训练时间较少，测试时间较少，过拟合风险较低，欠拟合风险较高，泛化能力一般；

后剪枝训练时间较长，测试时间较少，过拟合风险较低，欠拟合风险稳定，泛化能力较好

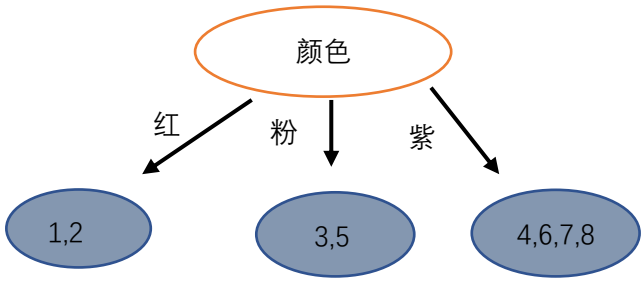
五、计算（画图）题

10. 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

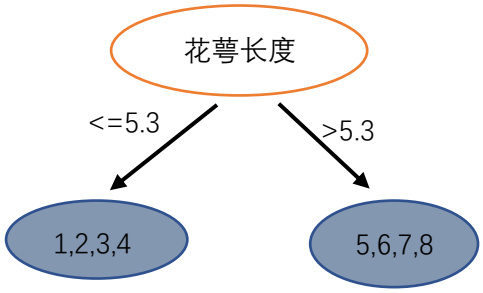
1) 请给出采用颜色特征进行多叉树分裂的结果

答：



2) 请给出采用花瓣长度进行二叉树分裂的任意一个结果

答：



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。

(1) 信息增益：

根节点的经验熵：

$$\hat{H}(Y) = -\sum_{j=1}^{3} P(y=\omega_j)\log P(y=\omega_j) = -\left(\frac{4}{8}\log_2\frac{4}{8}+\frac{2}{8}\log_2\frac{2}{8}+\frac{2}{8}\log_2\frac{2}{8}\right) = 1.5$$

(2) 根据信息增益选择分裂特征

- 香气属性

根据香气特征划分子集集合：

$$S_1 = \{1(se),2(se),3(se),4(se),8(vi)\}, \qquad S_2 = \{5(ve),6(ve),7(vi)\}$$

计算每个子节点的信息熵：

$$\hat{H}(S_1) = -\left(\frac{4}{5}\log_2\frac{4}{5}+\frac{1}{5}\log_2\frac{1}{5}\right) = 0.722,$$

$$\hat{H}(S_2) = -\left(\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}\right) = 0.918$$

加权平均经验熵：

$$\hat{H}_1 = \sum_{k=1}^{2}\left(\frac{N_k}{N}\right)\hat{H}(S_k) = \frac{5}{8}\times 0.722 + \frac{3}{8}\times 0.918 = 0.796$$

属性“香气”信息增益为：

$$G_1 = \hat{H}(Y) - \hat{H}_1 = 1.5 - 0.796 = 0.704$$

- “颜色”属性

根据颜色属性划分子集集合：

$S_1 = \{1(se),2(se)\}$，　$S_2 = \{3(se),5(ve)\}$，　$S_3 = \{4(se),6(ve),7(vi),8(vi)\}$

计算每个子节点的信息熵：

$$\hat{H}(S_1) = 0,$$

$$\hat{H}(S_2) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1.0$$

$$\hat{H}(S_3) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1.5$$

加权平均经验熵：

$$\hat{H}_2 = \sum_{k=1}^{3}\left(\frac{N_k}{N}\right)\hat{H}(S_k) = \frac{2}{8}\times 0 + \frac{2}{8}\times 1.0 + \frac{4}{8}\times 1.5 = 1.0$$

信息增益：

$$G_2 = \hat{H}(Y) - \hat{H}_2 = 1.5 - 1.0 = 0.5$$

根据信息增益$G_1 > G_2$，所以选择“香气”属性分裂特征。

(3) 根据信息增益比选择分裂特征

“香气”属性经验熵：

$$\hat{H}\left("香气"\right) = -\sum_{k=1}^{2}\frac{N_k}{N}\log\frac{N_k}{N} = -\left(\frac{5}{8}\log\frac{5}{8} + \frac{3}{8}\log\frac{3}{8}\right) = 0.954$$

信息增益比为：

$$R_1 = \frac{0.704}{0.954} = 0.738$$

“颜色”属性经验熵：

$$\hat{H}\left("颜色"\right) = -\sum_{k=1}^{3}\frac{N_k}{N}\log\frac{N_k}{N} = -\left(\frac{2}{8}\log\frac{2}{8} + \frac{2}{8}\log\frac{2}{8} + \frac{4}{8}\log\frac{4}{8}\right) = 1.5$$

信息增益比为：

$$R_2 = \frac{0.5}{1.5} = 0.33$$

信息增益比$R_1 > R_2$，所以选择“香气”属性进行特征分裂

(4) 根据基尼指数选择分裂特征

计算“香气”属性的信息增益：

$$Gini\left("香气 = 有"\right) = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.320$$

$$Gini\left("香气 = 无"\right) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$$

$$Gini\_score("香气") = \frac{5}{8}\times 0.320 + \frac{3}{8}\times 0.444 = 0.37$$

计算“颜色”属性的信息增益：

$$Gini\left("颜色 = 红"\right) = 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0$$

$$Gini\left("颜色 = 粉"\right) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$Gini\left("颜色 = 紫"\right) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.625$$

$$Gini\_score("颜色") = \frac{2}{8} \times 0 + \frac{2}{8} \times 0.5 + \frac{4}{8} \times 0.625 = \mathbf{0.437}$$

由于$Gini_score("香气") < Gini_score("颜色")$，所以按照“香气”特征进行分裂。

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

第一次分裂：

样本集的熵：

$$\hat{H}(Y) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right) = 0.940$$

计算天气属性的信息增益，计算每个子节点的信息熵：

$$\hat{H}\left("天气 = 晴"\right) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971,$$

$$\hat{H}\left("天气 = 阴"\right) = -\left(\frac{4}{4}\log_2\frac{4}{4} + 0\right) = 0$$

$$\hat{H}\left("天气=雨"\right)=-\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5}\right)=0.971$$

加权平均经验熵：

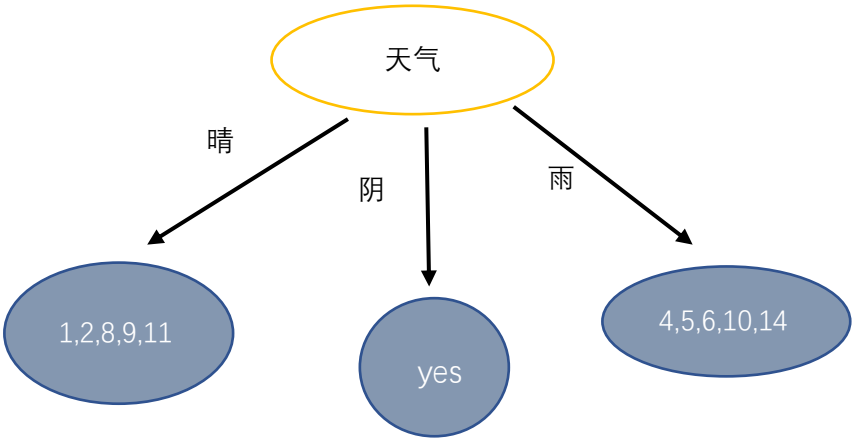
$$\hat{H}_1=\sum_{k=1}^{3}\left(\frac{N_k}{N}\right)\hat{H}(S_k)=\frac{5}{14}\times 0.971+\frac{4}{14}\times 0+\frac{5}{14}\times 0.971=0.694$$

$$G_1=\hat{H}(Y)-\hat{H}_1=0.940-0.694=\mathbf{0.246}$$

同理可以算出温度、湿度、风三个属性的信息增益：

$$G_2=\mathbf{0.151},\ G_3=\mathbf{0.048},\ G_4=\mathbf{0.029}$$

所以第一次分裂按照天气属性进行分裂。



第二次分裂：

对上图中左侧第一个结点进行分裂：

样本集的熵：

$$\hat{H}(Y)=-\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log_2\frac{2}{5}\right)=0.971$$

计算温度属性的信息增益，计算每个子节点的信息熵：

$$\hat{H}\left("温度=热"\right)=-\left(\frac{2}{2}\log_2\frac{2}{2}\right)=0,$$

$$\hat{H}\left("温度=温和"\right)=-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)=1,$$

$$\hat{H}\left("温度=凉爽"\right)=-\left(\frac{1}{1}\log_2\frac{1}{1}\right)=0$$

加权平均经验熵：

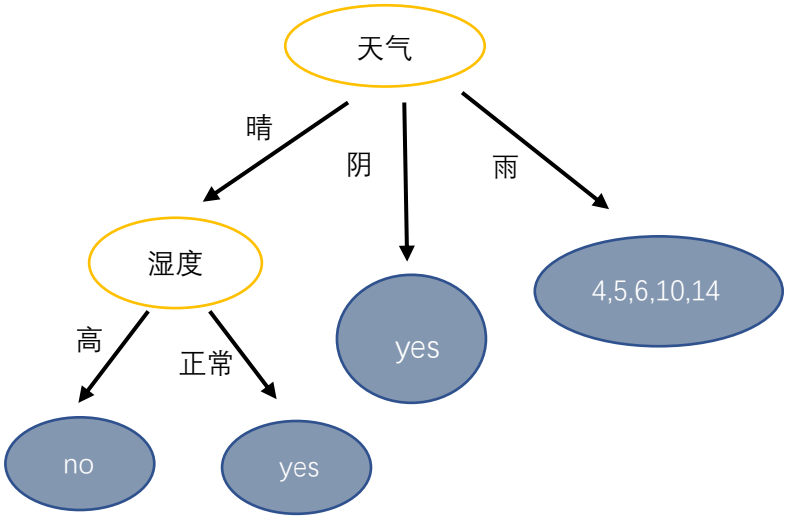
$$\hat{H}_1=\sum_{k=1}^{3}\left(\frac{N_k}{N}\right)\hat{H}(S_k)=\frac{2}{5}\times 0+\frac{2}{5}\times 1+\frac{1}{5}\times 0=0.4$$

$$G_1=\hat{H}(Y)-\hat{H}_1=0.971-0.4=\mathbf{0.371}$$

同理可以算出湿度、风两个属性的信息增益：

$$G_2=\mathbf{0.971},\ G_3=\mathbf{0.02}$$

所以这一结点按照“湿度”属性进行分裂。



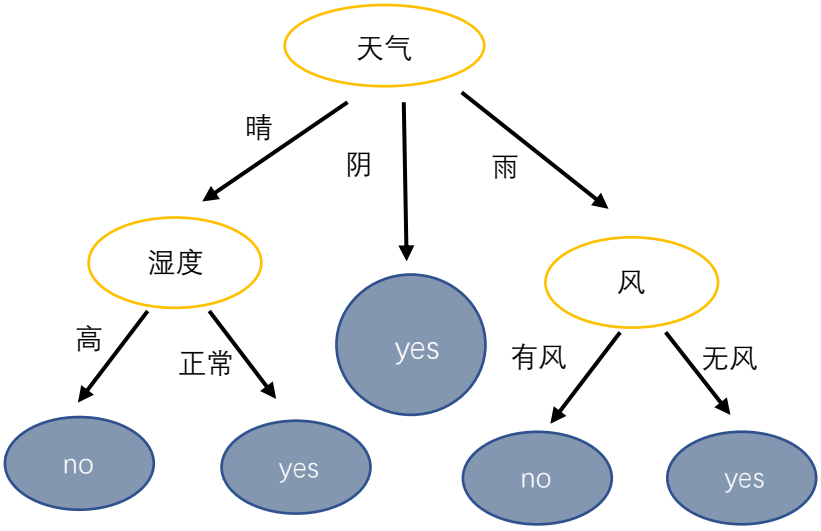
根节点下左侧第二个子节点是叶子节点，不需要再分裂。

对天气是雨的分裂出的结点进行分裂：

对温度、湿度和风计算信息增益：

$$G_1=\mathbf{0.02}\ \ G_2=\mathbf{0.02},\ G_3=\mathbf{0.951}$$

所以按照“风”属性进行分裂



至此，所有新分裂出的节点均为叶子节点，无需再分，决策树生成过程完成。

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