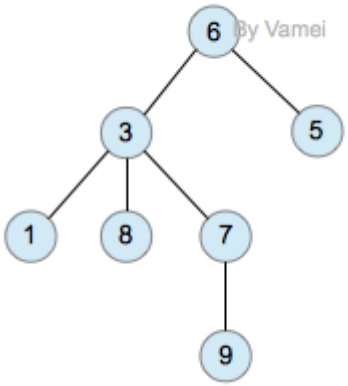
决策树

- 树与二叉树
- 基本流程
- 划分选择
- 剪枝处理
- 连续值与缺失值处理

树与二叉树

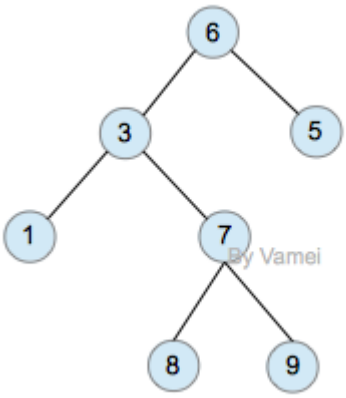
树是有限元素的集合

- 根结点、内部节点、叶结点
- 父结点、子结点
- 层次、深度

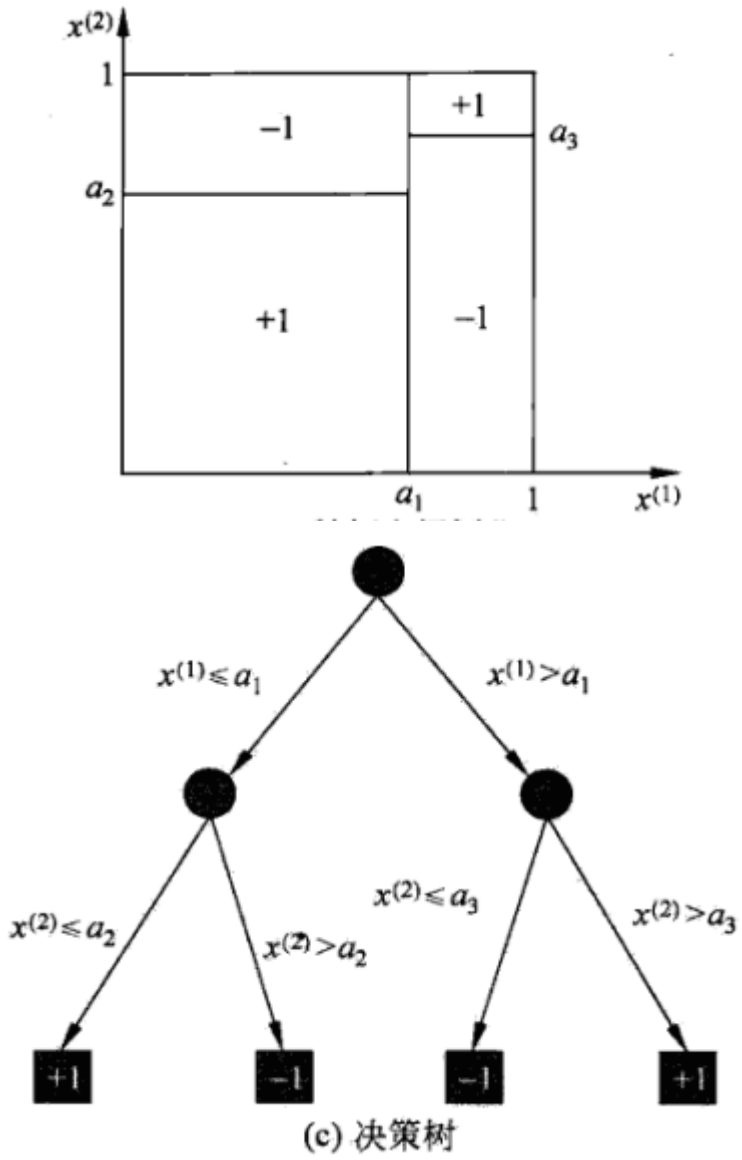


二叉树

- 每个结点最多只能有2个子结点



简单实例


$x^{(2)}$
1
−1
+1
a_3
a_2
+1
−1
a_1
1
$x^{(1)}$
$x^{(1)} \leqslant a_1$
$x^{(1)} > a_1$
$x^{(2)} \leqslant a_2$
$x^{(2)} > a_2$
$x^{(2)} \leqslant a_3$
$x^{(2)} > a_3$
+1
−1
−1
+1
(c) 决策树

基本流程

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
　　属性集 $A = \{a_1, a_2, \ldots, a_d\}$.

过程： 函数 TreeGenerate(D, A)

```
生成结点 node;
if D 中样本全属于同一类别 C then
    将 node 标记为 C 类叶结点; return
end if
if A = ∅ OR D 中样本在 A 上取值相同 then
    将 node 标记为叶结点, 其类别标记为 D 中样本数最多的类; return
end if
从 A 中选择最优划分属性 a*;
for a* 的每一个值 a*^v do
    为 node 生成一个分支; 令 D_v 表示 D 中在 a* 上取值为 a*^v 的样本子集;
    if D_v 为空 then
        将分支结点标记为叶结点, 其类别标记为 D 中样本最多的类; return
    else
        以 TreeGenerate(D_v, A \ {a*}) 为分支结点
    end if
end for
```

输出： 以 node 为根结点的一棵决策树

决策树的生成是一个递归过程。在决策树基本算法中，有三种情形会导致递归返回

- 当前结点包含的样本全属于同一类别，无需划分
- 当前属性为空，或是所有样本在所有属性上取值相同，无需划分

  把当前结点标记为叶结点，并将其类别设定为该结点所含样本最多的类别
- 当前结点包含的样本集合为空，不能划分

  把当前结点标记为叶结点，将其类别设定为其父结点所含样本最多的类别

划分选择

- 信息增益 - > ID3算法

  属性A对训练数据集D的信息增益Gain(D,A)定义为集合D的经验熵Ent(D)与属性A给定条件下D的经验条件熵Ent(D|A)之差

  表示划分前后不确定性减少的程度，应选择信息增益最大的属性对样本进行划分

$$Gain(D, A) = Ent(D) - Ent(D|A)$$

$$Ent(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} log_2 \frac{|C_k|}{|D|}$$

$$Ent(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} Ent(D_i)$$

- 信息增益比 - > C4.5算法

以信息增益为划分指标，存在偏向选择取值较多的属性的问题

属性A对训练数据集D的信息增益Gain(D,A)定义为其信息增益与训练数据集关于属性A的熵Ent(A)

$$Gain\_ratio(D,A)=\frac{Gain(D,A)}{Ent(A)}$$

$$Ent(A)=-\sum_{i=1}^{V}\frac{|D_i|}{|D|}log_2\frac{|D_i|}{|D|}$$

- 基尼指数 - > CART算法

  描述数据集D的纯度，表示一个随机选中的样本在子集中被分错的可能性，应选择基尼指数最小的进行划分

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

$$Gini(D)=1-\sum_{k=1}^{K}(\frac{|C_k|}{|D|})^2$$

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2)$$

决策树模型——决策树的生成

- ID3算法
- C4.5算法
- CART算法

实例——以二分类为例

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

首先计算属性“色泽”的信息增益：

$$Ent(D) = -(\frac{8}{17}log_2\frac{8}{17} + \frac{9}{17}log_2\frac{9}{17}) = 0.998$$

$$Ent(D_1) = -(\frac{3}{6}log_2(\frac{3}{6}) + \frac{3}{6}log_2(\frac{3}{6})) = 1.000$$

$$Ent(D_2) = -(\frac{4}{6}log_2(\frac{4}{6}) + \frac{2}{6}log_2(\frac{2}{6})) = 0.918$$

$$Ent(D_3) = -(\frac{1}{5}log_2(\frac{1}{5}) + \frac{4}{5}log_2(\frac{4}{5})) = 0.722$$

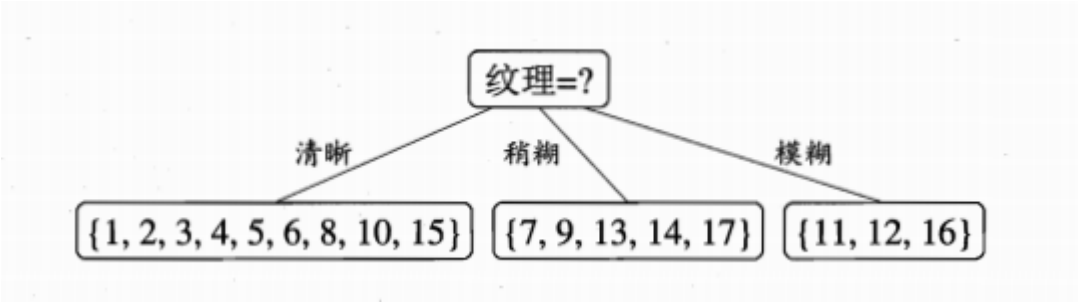
$$Gain(D, \text{色泽}) = Ent(D) - [\frac{6}{17}Ent(D_1) + \frac{6}{17}Ent(D_2) + \frac{5}{17}Ent(D_3)] = 0.109$$

类似的，可计算其他属性的信息增益

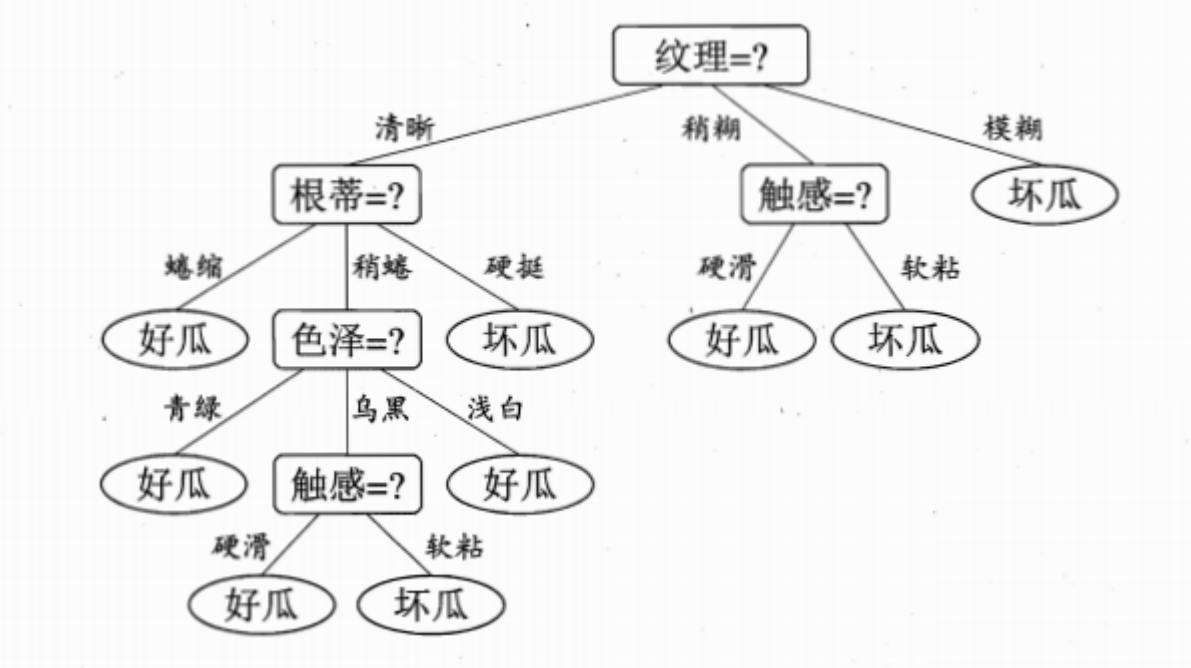
$$Gain(D, \text{根蒂}) = 0.143; Gain(D, \text{敲声}) = 0.141; Gain(D, \text{纹理}) = 0.381;$$

$$Gain(D, \text{脐部}) = 0.381; Gain(D, \text{触感}) = 0.006;$$

显然，属性“纹理”的信息增益最大，于是被选为划分属性。划分结果如下：



接下来进行进一步划分，最终结果如图



剪枝处理

- 预剪枝 训练时间开销小，但有欠拟合的风险

  在决策树生成过程中，对每个结点在划分之前先进行估计，若当前结点的划分不能带来泛化性能提升，则停止划分并将当前结点标记为叶结点

- 后剪枝 泛化性能更好，但训练时间开销大

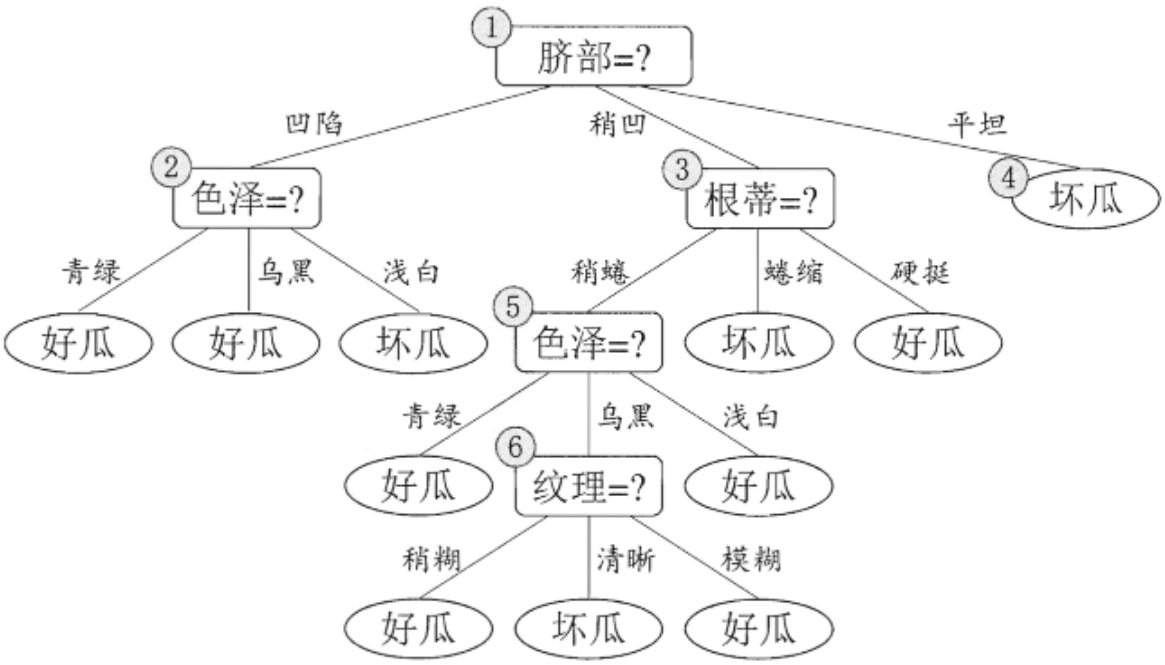
  先生成一颗完整的决策树，然后自底向上的对结节点进行考察，若将该结点对应的子树替换为叶结点带来繁华性能的提升，则将该结点替换为叶结点

举例来说，首先将西瓜数据集分成训练集和验证集

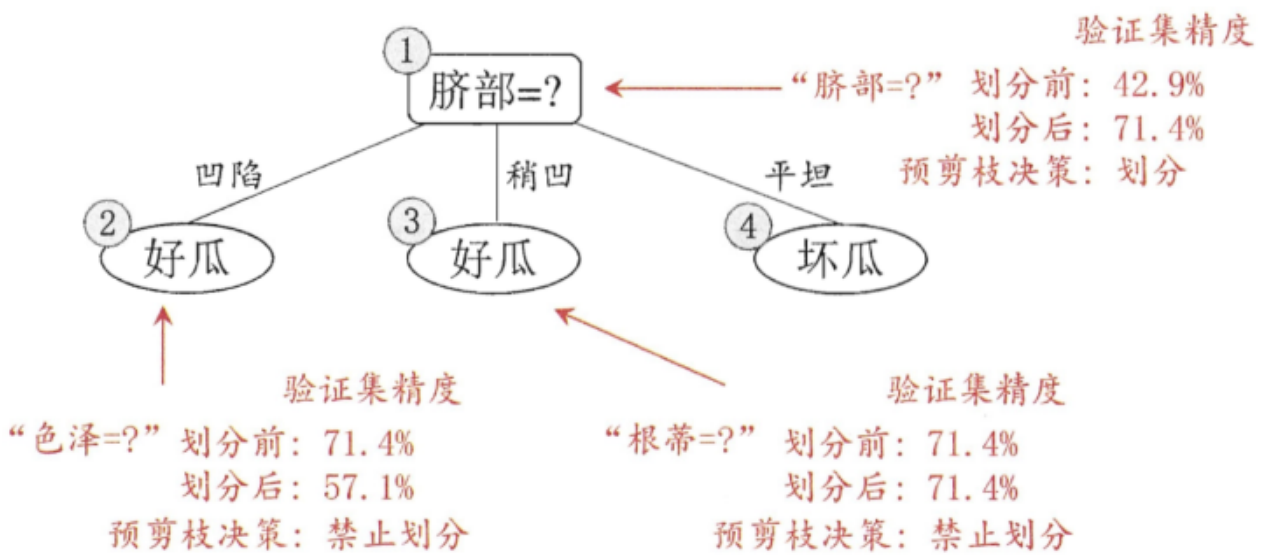
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

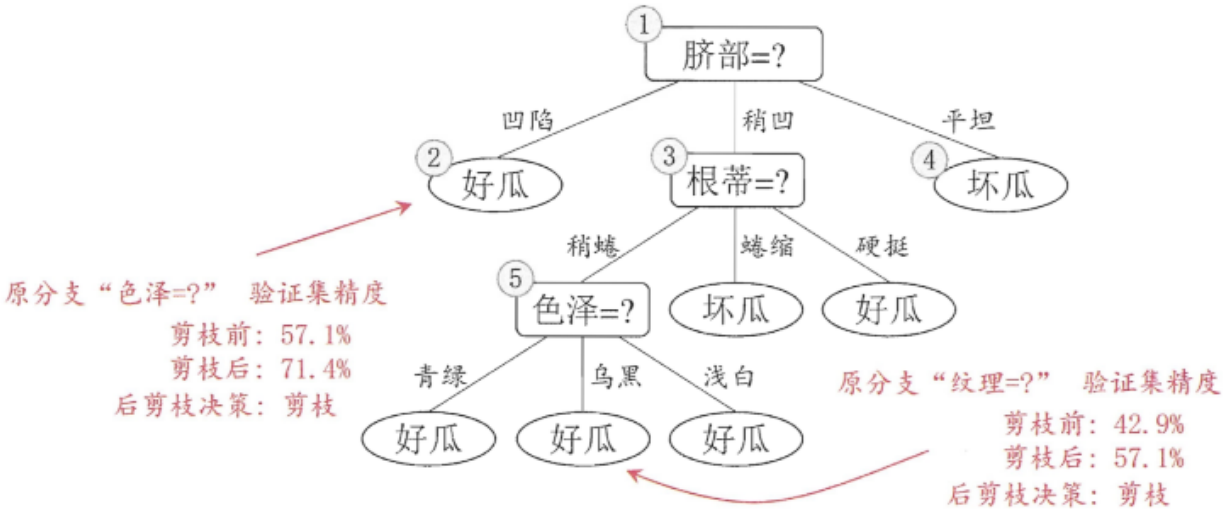
生成的未剪枝的决策树如下



预剪枝



后剪枝



连续值与缺失值处理

- 连续值

  取每个区间的中点，找到其中信息增益最大的点，作为划分点。

  对连续属性A，可考察包含n-1个元素的候选划分点集合：

$$T_a = \{\frac{a^i + a^{i+1}}{2} | 1 \leq i \leq n-1\}$$

  然后选取最优的划分点进行样本集合划分：

$$Gain(D, A) = \max_{t \in T_a} Gain(D, A, t) = \max_{t \in T_a} Ent(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda)$$

- **For continuous attributes**
  - Algorithm
    - Find the best threshold according to the splitting criterion

Weights:
5.2kg(B),6.6kg(B),7kg(G),8.6kg(B),11.3kg(G)
Thresholds:
5.9kg, 6.8kg, 7.8kg, 9.9gkg

$$\text{Gain}(D,a) = \max_{t\in T_a} \text{Gain}(D,a,t) \qquad \text{Ent}(D) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

$$= \max_{t\in T_a} \text{Ent}(D) - \sum_{\lambda\in\{-,+\}} \frac{|D_t^\lambda|}{|D|}\text{Ent}(D_t^\lambda)$$

$$\text{Gain}(D,a,t_1 = 5.9) = 0.971 - \frac{1}{5}\left[-\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right)\right] - \frac{4}{5}\left[-\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)\right] = \mathbf{0.1710}$$

$$\text{Gain}(D,a,t_2 = 6.8) = 0.971 - \frac{2}{5}\left[-\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right)\right] - \frac{3}{5}\left[-\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right] = \mathbf{0.4200}$$

$$\text{Gain}(D,a,t_3 = 7.8) = 0.971 - \frac{3}{5}\left[-\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)\right] - \frac{2}{5}\left[-\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)\right] = \mathbf{0.0200}$$

$$\text{Gain}(D,a,t_4 = 9.9) = 0.971 - \frac{4}{5}\left[-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right)\right] - \frac{1}{5}\left[-\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right)\right] = \mathbf{0.3220}$$

- 缺失值

  如果仅对无缺失值的样本进行学习，显然是对数据信息极大的浪费。

  两个问题：

  - 如何在属性值缺失的情况下进行划分属性选择？
  - 给定划分属性，若样本在该属性上的值缺失，如何对样本进行划分？

  为每个样本x赋予一个权重w_x,并定义

$$\rho = \frac{\sum_{x\in\tilde{D}} w_x}{\sum_{x\in D} w_x}$$

$$\tilde{r_v} = \frac{\sum_{x\in\tilde{D}^v} w_x}{\sum_{x\in\tilde{D}} w_x}$$

  这样就可将信息增益的公式推广为：

$$Gain(D,A) = \rho \times Gain(\tilde{D},A) = \rho \times (Ent(\tilde{D}) - \sum_{k=1}^{K} \tilde{r}^v Ent(\tilde{D}))$$

  其中

$$Ent(\tilde{D}) = -\sum_{k=1}^{K} \tilde{p_k} log_2 \tilde{p_k}$$

- **For missing attributes**

Initialize the weights of the samples $\boldsymbol{x}$ as $w_{\boldsymbol{x}} = 1$, and score the attributes based on the complete samples

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	–	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	–	是
3	乌黑	蜷缩	–	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	–	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	–	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	–	稍凹	硬滑	是
9	乌黑	–	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	–	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	–	否
12	浅白	蜷缩	–	模糊	平坦	软粘	否
13	–	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	–	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	–	沉闷	稍糊	稍凹	硬滑	否

$$\mathrm{Ent}(D, a_1) = -\frac{6}{14}\log\left(\frac{6}{14}\right) - \frac{8}{14}\log\left(\frac{8}{14}\right) = 0.985$$

$$\mathrm{Ent}(\tilde{D}^1) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1.000 \quad \text{青绿}$$

$$\mathrm{Ent}(\tilde{D}^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918 \quad \text{乌黑}$$

$$\mathrm{Ent}(\tilde{D}^3) = -\left(\frac{0}{4}\log_2\frac{0}{4} + \frac{4}{4}\log_2\frac{4}{4}\right) = 0.000 \quad \text{浅白}$$

$$\mathrm{Gain}(\tilde{D}, \text{色泽}) = \mathrm{Ent}(\tilde{D}) - \sum_{v=1}^{3}\tilde{r}_v\,\mathrm{Ent}(\tilde{D}^v)$$
$$= 0.985 - \left(\frac{4}{14}\times 1.000 + \frac{6}{14}\times 0.918 + \frac{4}{14}\times 0.000\right)$$
$$= 0.306\,.$$

$$\tilde{r}_v = \frac{\sum_{\boldsymbol{x}\in\tilde{D}^v} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x}\in\tilde{D}} w_{\boldsymbol{x}}}$$

$$\mathrm{Gain}(D, \text{色泽}) = \rho \times \mathrm{Gain}(\tilde{D}, \text{色泽}) = \frac{14}{17}\times 0.306 = 0.252$$

$$\rho = \frac{\sum_{\boldsymbol{x}\in\tilde{D}} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x}\in D} w_{\boldsymbol{x}}}$$

$\mathrm{Gain}(D, \text{色泽}) = 0.252$; $\mathrm{Gain}(D, \text{根蒂}) = 0.171$;
$\mathrm{Gain}(D, \text{敲声}) = 0.145$; $\mathrm{Gain}(D, \text{纹理}) = 0.424$;
$\mathrm{Gain}(D, \text{脐部}) = 0.289$; $\mathrm{Gain}(D, \text{触感}) = 0.006$.

NEU-ISE