四、决策树

主讲教师：周志华

决策树模型

决策树基于“树”结构进行决策

- 每个“内部结点”对应于某个属性上的“测试”(test)
- 每个分支对应于该测试的一种可能结果（即该属性的某个取值）
- 每个“叶结点”对应于一个“预测结果”

学习过程：通过对训练样本的分析来确定“划分属性”（即内部结点所对应的属性)

预测过程：将测试示例从根结点开始，沿着划分属性所构成的“判定测试序列”下行，直到叶结点

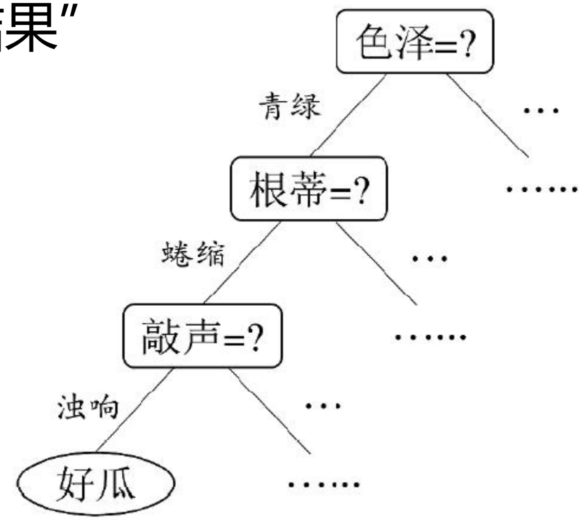


图 4.1 西瓜问题的一棵决策树

基本流程

策略："分而治之"(divide-and-conquer)

自根至叶的递归过程

在每个中间结点寻找一个"划分"(split or test)属性

三种停止条件：

(1) 当前结点包含的样本全属于同一类别，无需划分;

(2) 当前属性集为空, 或是所有样本在所有属性上取值相同，无法划分;

(3) 当前结点包含的样本集合为空，不能划分.

基本算法

输入： 训练集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
属性集 $A = \{a_1, a_2, \ldots, a_d\}$.

过程： 函数 TreeGenerate(D, A)

生成结点 node;

递归返回，情形(1)

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

递归返回，情形(2)

if $A = \varnothing$ **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

利用当前结点的后验分布

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;

递归返回，情形(3)

    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

将父结点的样本分布作为当前结点的先验分布

        以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点
    end if
end for

决策树算法的核心

输出： 以 node 为根结点的一棵决策树

信息增益 (Information Gain)

信息熵 (entropy) 是度量样本集合“纯度”最常用的一种指标

假定当前样本集合 D 中第 k 类样本所占的比例为 p_k，则 D 的信息熵定义为

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

$\mathrm{Ent}(D)$的值越小，则D的纯度越高

计算信息熵时约定：若 $p=0$，则 $p\log_2 p = 0$.

$\mathrm{Ent}(D)$ 的最小值为 0，最大值为 $\log_2 |\mathcal{Y}|$.

信息增益直接以信息熵为基础，计算当前划分对信息熵所造成的变化

信息增益 (Information Gain)

离散属性 a 的取值：$\{a^1, a^2, \ldots, a^V\}$

D^v: D 中在 a 上取值 = a^v 的样本集合

以属性 a 对 数据集 D 进行划分所获得的信息增益为：

$$\mathrm{Gain}(D,a) = \underbrace{\mathrm{Ent}(D)}_{\text{划分前的信息熵}} - \sum_{v=1}^{V} \underbrace{\boxed{\frac{|D^v|}{|D|}}}_{\text{第 } v \text{ 个分支的权重，样本越多越重要}} \underbrace{\mathrm{Ent}(D^v)}_{\text{划分后的信息熵}}$$

ID3算法中使用

一个例子

表 4.1 西瓜数据集 2.0

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

该数据集包含17个训练样例 $|\mathcal{Y}| = 2$，其中正例占 $p_1 = \frac{8}{17}$，反例占 $p_2 = \frac{9}{17}$

根结点的信息熵为

$$\mathrm{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{8}{17}\log_2\frac{8}{17} + \frac{9}{17}\log_2\frac{9}{17}\right) = 0.998$$

一个例子 (续)

表 4.1 西瓜数据集 2.0

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

以属性“色泽”为例，其对应的3个子集分别为：

D^1(色泽=青绿)

D^2(色泽=乌黑)

D^3(色泽=浅白)

对D^1(色泽=青绿)，正例3/6，反例3/6

于是：$\mathrm{Ent}(D^1) = -(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}) = 1.000$

一个例子 (续)

D²(色泽=乌黑)，
正例4/6，反例2/6

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

D³(色泽=浅白)，
正例1/5，反例4/5

$$\mathrm{Ent}(D^3) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722$$

表 4.1 西瓜数据集 2.0

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

于是，属性“色泽”的信息增益为

$$\begin{aligned}\mathrm{Gain}(D, \text{色泽}) &= \mathrm{Ent}(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \\ &= 0.998 - \left(\frac{6}{17}\times 1.000 + \frac{6}{17}\times 0.918 + \frac{5}{17}\times 0.722\right) = 0.109\end{aligned}$$

一个例子 (续)

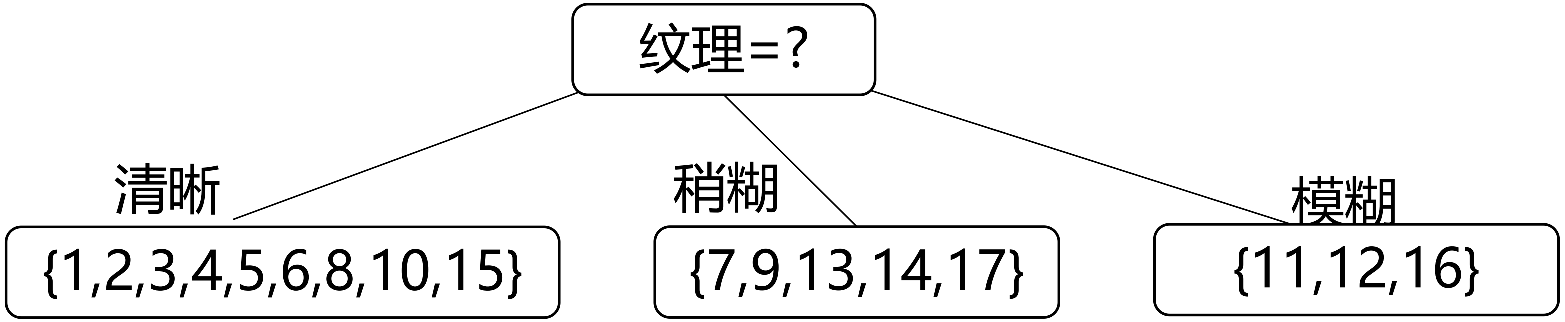
类似的，其他属性的信息增益为

$$\mathrm{Gain}(D,\text{根蒂}) = 0.143 \qquad \mathrm{Gain}(D,\text{敲声}) = 0.141$$

$$\mathrm{Gain}(D,\text{纹理}) = 0.381 \qquad \mathrm{Gain}(D,\text{脐部}) = 0.289$$

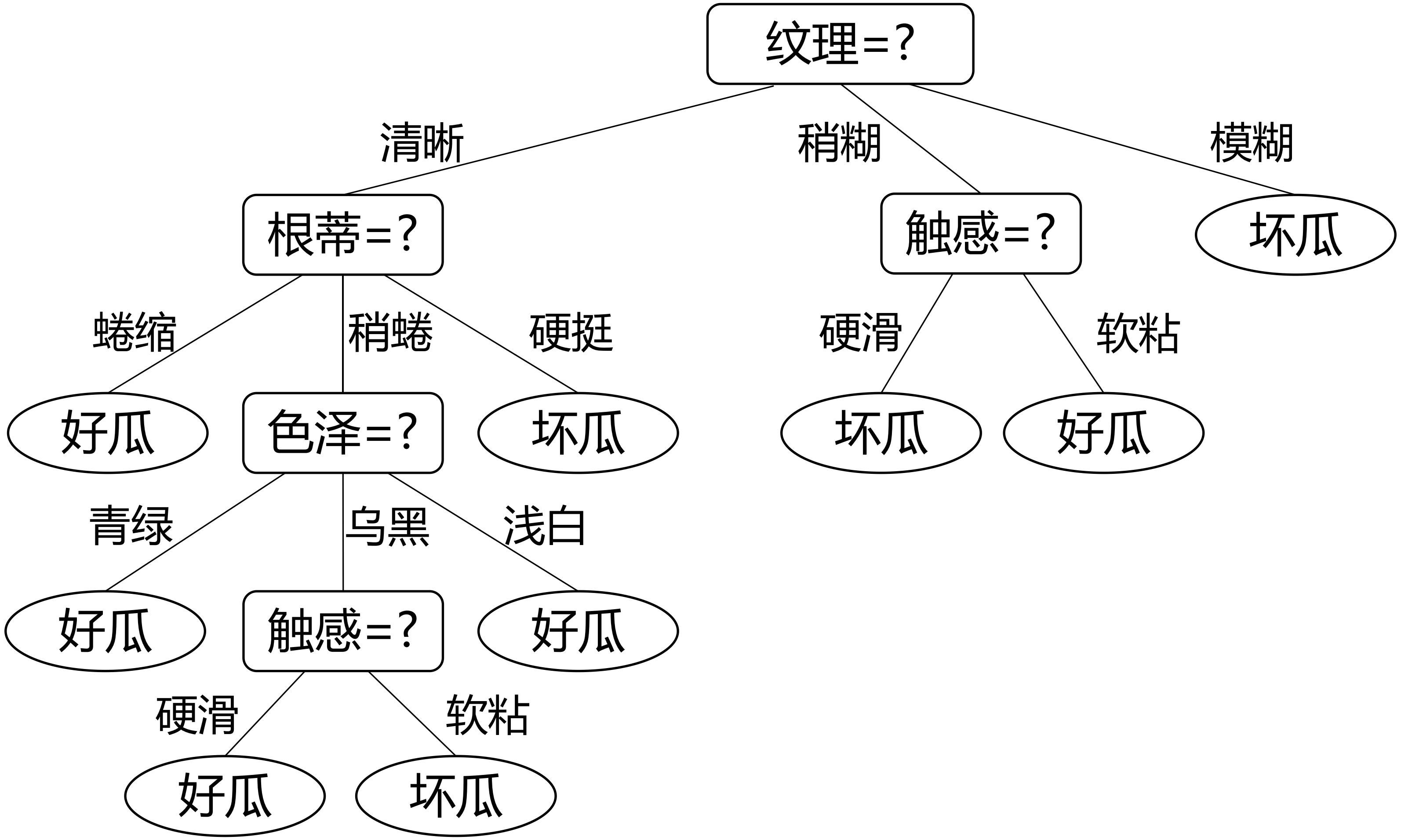
$$\mathrm{Gain}(D,\text{触感}) = 0.006$$

属性“纹理”的信息增益最大，被选为划分属性



一个例子 (续)

对每个分支结点做进一步划分，最终得到决策树



信息增益 (Information Gain)

离散属性 a 的取值：$\{a^1, a^2, \ldots, a^V\}$

D^v: D 中在 a 上取值 = a^v 的样本集合

以属性 a 对 数据集 D 进行划分所获得的信息增益为：

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

划分前的信息熵

划分后的信息熵

第 v 个分支的权重，
样本越多越重要

ID3算法中使用

增益率 (Gain Ratio)

信息增益：对可取值数目较多的属性有所偏好

有明显弱点，例如：考虑将“编号”作为一个属性

增益率：$\mathrm{Gain\_ratio}(D,a) = \dfrac{\mathrm{Gain}(D,a)}{\mathrm{IV}(a)}$

其中 $\mathrm{IV}(a) = -\sum_{v=1}^{V} \dfrac{|D^v|}{|D|} \log_2 \dfrac{|D^v|}{|D|}$

属性 a 的可能取值数目越多 (即 V 越大)，则 $\mathrm{IV}(a)$ 的值通常就越大

启发式：先从候选划分属性中找出信息增益高于平均水平的，再从中选取增益率最高的

C4.5算法中使用

基尼指数 (Gini Index)

$$\mathrm{Gini}(D) = \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'}$$

反映了从 D 中随机抽取两个样例，其类别标记不一致的概率

$$= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2 \,.$$

Gini(D) 越小，数据集 D 的纯度越高

属性 a 的基尼指数：$\mathrm{Gini\_index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Gini}(D^v)$

在候选属性集合中，选取那个使划分后基尼指数最小的属性

CART算法中使用

划分选择 vs. 剪枝

研究表明: 划分选择的各种准则虽然对决策树的尺寸有较大影响，但对泛化性能的影响很有限

例如信息增益与基尼指数产生的结果，仅在约 2% 的情况下不同

剪枝方法和程度对决策树泛化性能的影响更为显著

在数据带噪时甚至可能将泛化性能提升 25%

Why?

剪枝 (pruning) 是决策树对付“过拟合”的主要手段!

剪枝

为了尽可能正确分类训练样本，有可能造成分支过多 → 过拟合

可通过主动去掉一些分支来降低过拟合的风险

基本策略：

- 预剪枝 (pre-pruning): 提前终止某些分支的生长
- 后剪枝 (post-pruning): 生成一棵完全树，再“回头”剪枝

剪枝过程中需评估剪枝前后决策树的优劣 ⟶ 第 2 章

现在我们假定使用“留出法”

缺失值

现实应用中，经常会遇到属性值“缺失”(missing)现象

仅使用无缺失的样例？ → 对数据的极大浪费

使用带缺失值的样例，需解决：

Q1：如何进行划分属性选择?

Q2：给定划分属性，若样本在该属性上的值缺失，如何进行划分?

基本思路：样本赋权，权重划分

一个例子

仅通过无缺失值的样例来判断划分属性的优劣

学习开始时，根结点包含样例集 D 中全部17个样例，权重均为 1

表 4.4 西瓜数据集 2.0α

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

以属性“色泽”为例，该属性上无缺失值的样例子集 $\tilde{D}$ 包含 14 个样例，信息熵为

$$\mathrm{Ent}(\tilde{D}) = -\sum_{k=1}^{2} \tilde{p}_k \log_2 \tilde{p}_k = -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985$$

一个例子

令 $\tilde{D}^1$，$\tilde{D}^2$，$\tilde{D}^3$ 分别表示在属性“色泽”上取值为“青绿”“乌黑”以及“浅白”的样本子集，有

$$\mathrm{Ent}(\tilde{D}^1) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1.000 \quad \mathrm{Ent}(\tilde{D}^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918$$

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

因此，样本子集 $\tilde{D}$ 上属性“色泽”的信息增益为

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

无缺失值样例中属性 a 取值为 v 的占比

于是，样本集 D 上属性“色泽”的信息增益为

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

无缺失值样例占比

一个例子

类似地可计算出所有属性在数据集上的信息增益

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

进入“纹理=清晰”分支

进入“纹理=稍糊”分支

进入“纹理=模糊”分支

样本权重在各子结点仍为1

在“纹理”上出现缺失值，样本 8, 10 同时进入三个分支，三分支上的权重分别为 7/15，5/15，3/15

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

权重划分