明理 精工

笃学 致远

第7章 决策树

Decision Tree

电子工程学院、人工智能学院

college of Electronic Engineering、college of Artificial Intelligence

7.1 基本原理

7.2 划分属性选择

7.3 剪枝处理

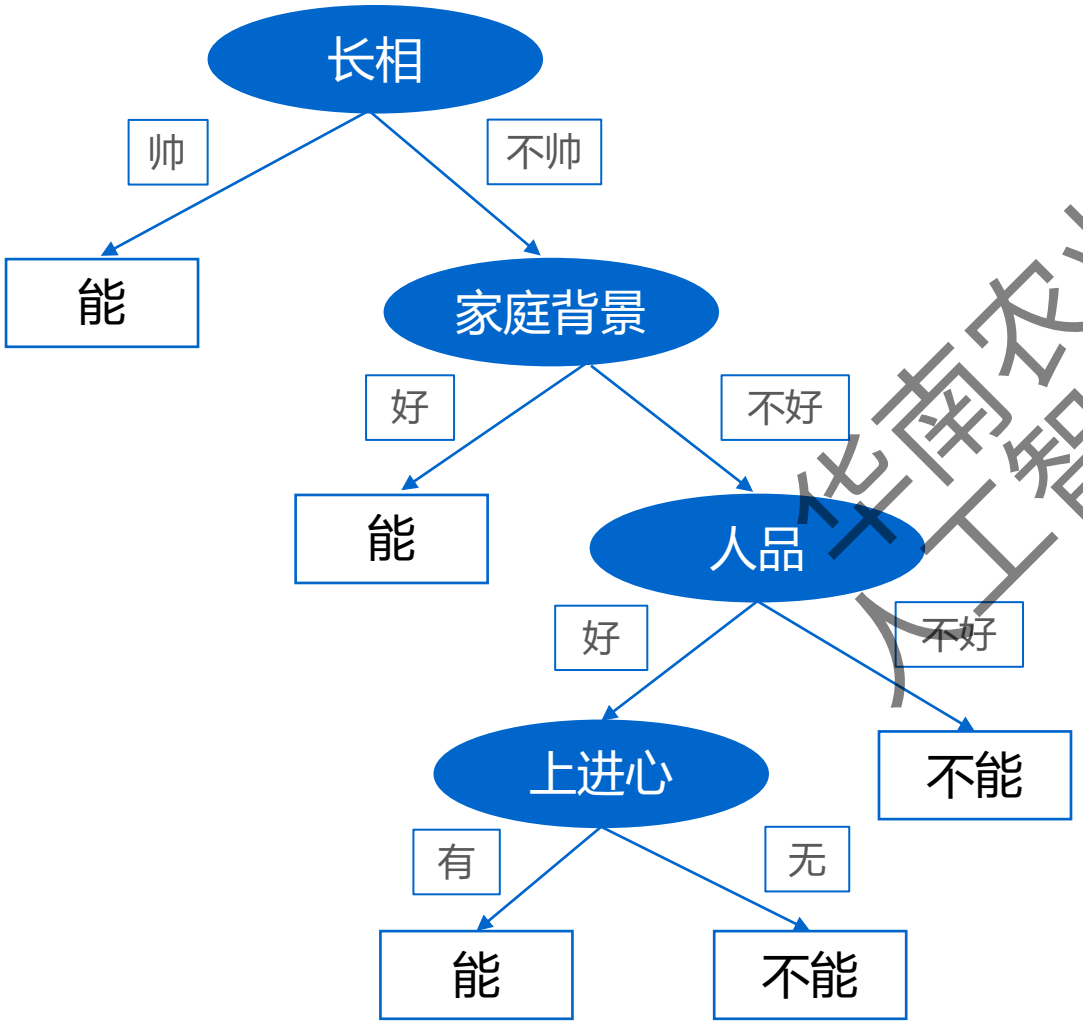
7.4 缺失值处理

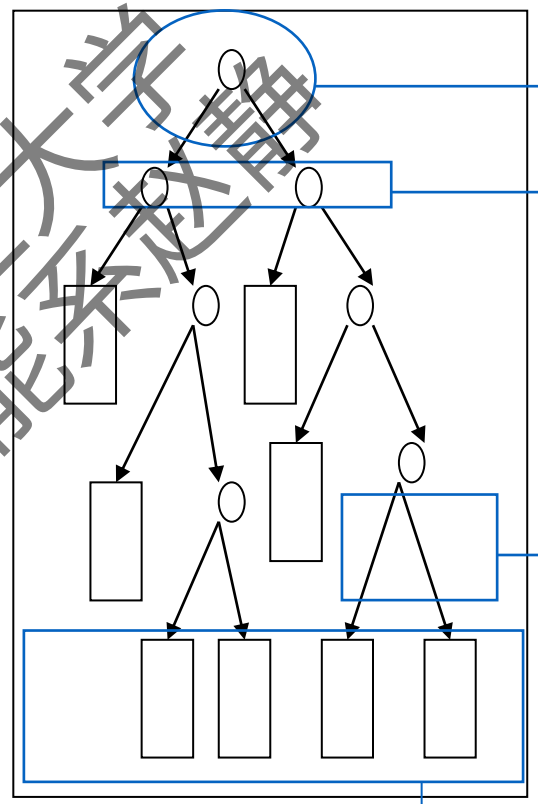
算法的发展过程

- 1979年, J.R. Quinlan 给出ID3算法，并在1983年和1986年对ID3 进行了总结和简化，使其成为决策树学习算法的典型。
- 1993年，Quinlan 进一步发展了ID3算法，改进成C4.5算法。
- 另一类决策树算法为CART，CART的决策树由二元逻辑问题生成，每个树节点只有两个分枝，分别包括学习实例的正例与反例。

7.1 基本原理

7.1.1 决策树的表示





根节点 (root node)

非叶子节点 (non-leaf node)
(代表测试条件，对数据属性的测试)

分支 (branches)
(代表测试结果)

叶节点 (leaf node)
(代表分类后所获得的分类标记)

7.1.2 决策树的特点

- 决策树算法属于监督学习方法
- 决策树算法采用贪心算法
- 基本策略：“分而治之”(divide-and-conquer)，属于判别模型
- 基本流程：自根至叶的递归过程，在每个中间结点寻找一个划分属性
- 在决策树的生成过程中，划分属性选择是关键

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.

7.1.3 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)

```
1: 生成结点 node;
2: if D 中样本全属于同一类别 C then
3:     将 node 标记为 C 类叶结点; return
4: end if
5: if A = ∅ OR D 中样本在 A 上取值相同 then
6:     将 node 标记为叶结点, 其类别标记为 D 中样本数最多的类; return
7: end if
8: 从 A 中选择最优划分属性 a_*;
9: for a_* 的每一个值 a_*^v do
10:    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;
11:    if D_v 为空 then
12:        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
13:    else
14:        以 TreeGenerate(D_v, A \ {a_*})为分支结点
15:    end if
16: end for
```

输出： 以 node 为根结点的一棵决策树

递归返回，情形(1)

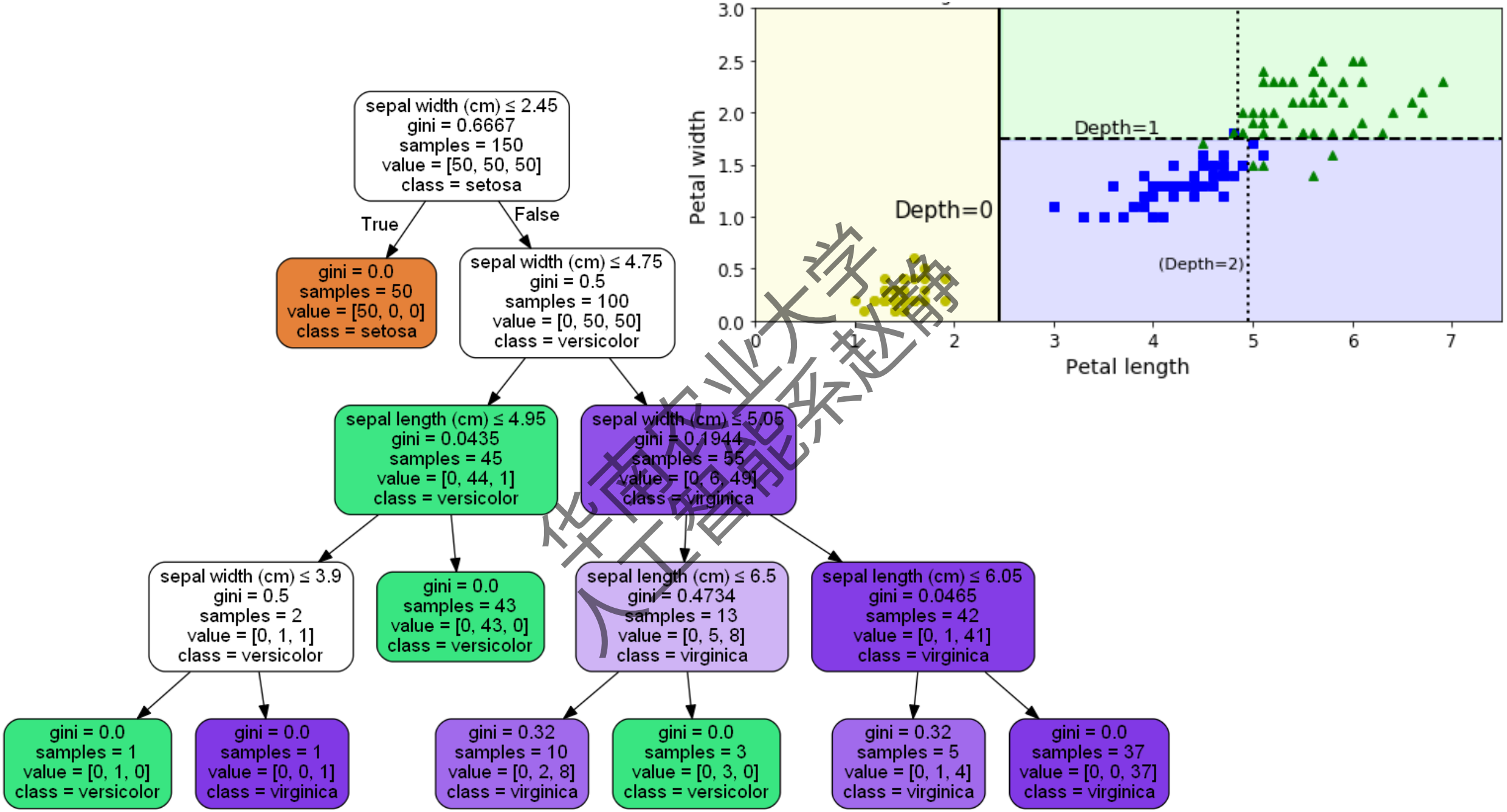
递归返回，情形(2)

利用当前结点的后验分布

递归返回，情形(3)

将父结点的样本分布作为当前结点的先验分布

决策树算法的核心



7.2 划分属性选择

7.2.1 信息熵 (information entropy)

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

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

计算信息熵时约定：若 $p = 0$，则 $p \log_2 p = 0$.

$\mathrm{Ent}(D)$ 的值越小，则 D 的纯度越高

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

信息熵（Information Entropy）

$$Ent(D) = -\sum p_i \log p_i, i = 1, 2, \ldots, n$$

例:

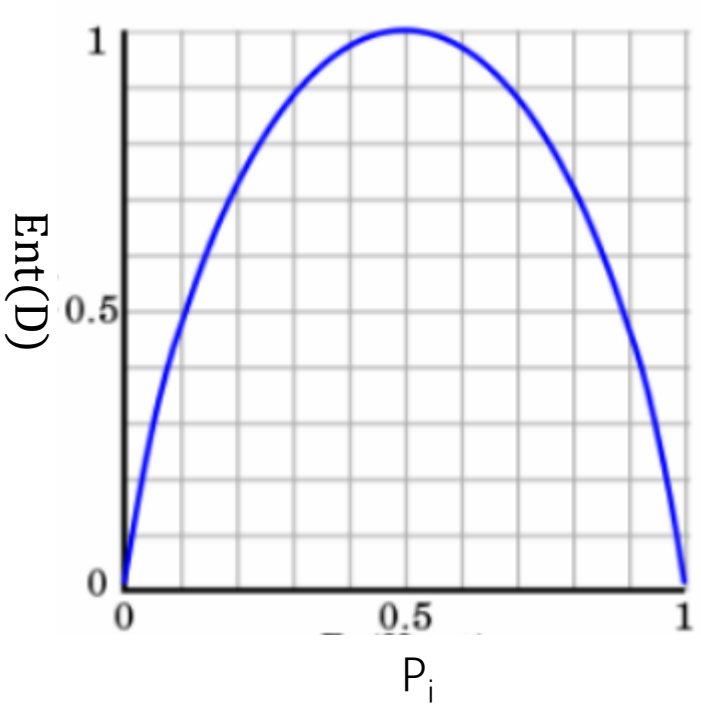
A集合y=[1,1,1,1,1,1,1,1,2,2]

B集合y=[1,2,3,4,5,6,6,5,3,1]

二分类问题中：

$$P(X=1)=p, \quad P(X=0)=1-p, \quad 0 \leqslant p \leqslant 1$$

$$Ent(D) = -\ p\log_2 p - (1-p)\log_2(1-p)$$



二分类熵值与概率的关系

7.2.2 信息增益(Information Gain) ——ID3算法

令当前节点的样本集合为$\mathcal{D}$,分裂后第v个子集样本集合为$\mathcal{D}_v$

- 用样本的比例估计概率分布：$p(Y=\mathrm{c})=\frac{|N_c|}{|N|}$
- 分裂之前的熵：$Ent(\mathcal{D})=-\sum_{\mathrm{c}=1}^{\mathrm{C}}p(Y=\mathrm{c})\log p(Y=\mathrm{c})$
- 分裂成V个子集后的熵：$Ent(\mathcal{D}|X)=\sum_{v=1}^{V}\frac{|\mathcal{D}_v|}{|\mathcal{D}|}Ent(\mathcal{D}_v)$
- **信息增益**：$gain_X(\mathcal{D})=Ent(\mathcal{D})-Ent(\mathcal{D}|X)$

第 v 个分支的权重，样本越多越重要

一个例子

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 $p_2 = \frac{9}{17}$

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

根结点的信息熵为

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

一个例子 (续)

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

D^2(色泽=乌黑)，
正例4/6，反例2/6

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

D^3(色泽=浅白)，
正例1/5，反例4/5

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

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

$$\mathrm{Gain}(D, \text{色泽}) = \mathrm{Ent}(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$

$$= 0.998 - (\frac{6}{17}\times 1.000 + \frac{6}{17}\times 0.918 + \frac{5}{17}\times 0.722)$$

$$= 0.109$$

一个例子 (续)

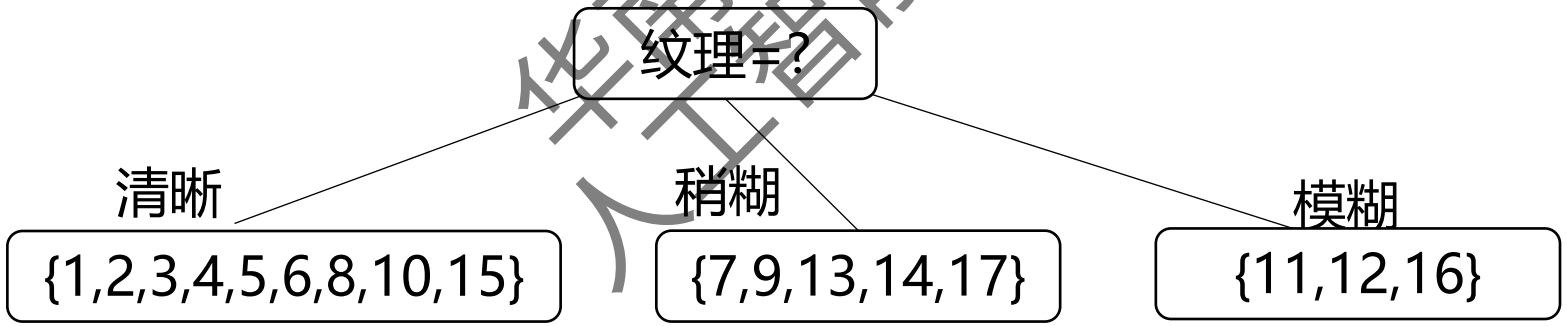
类似的，其他属性的信息增益为

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

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

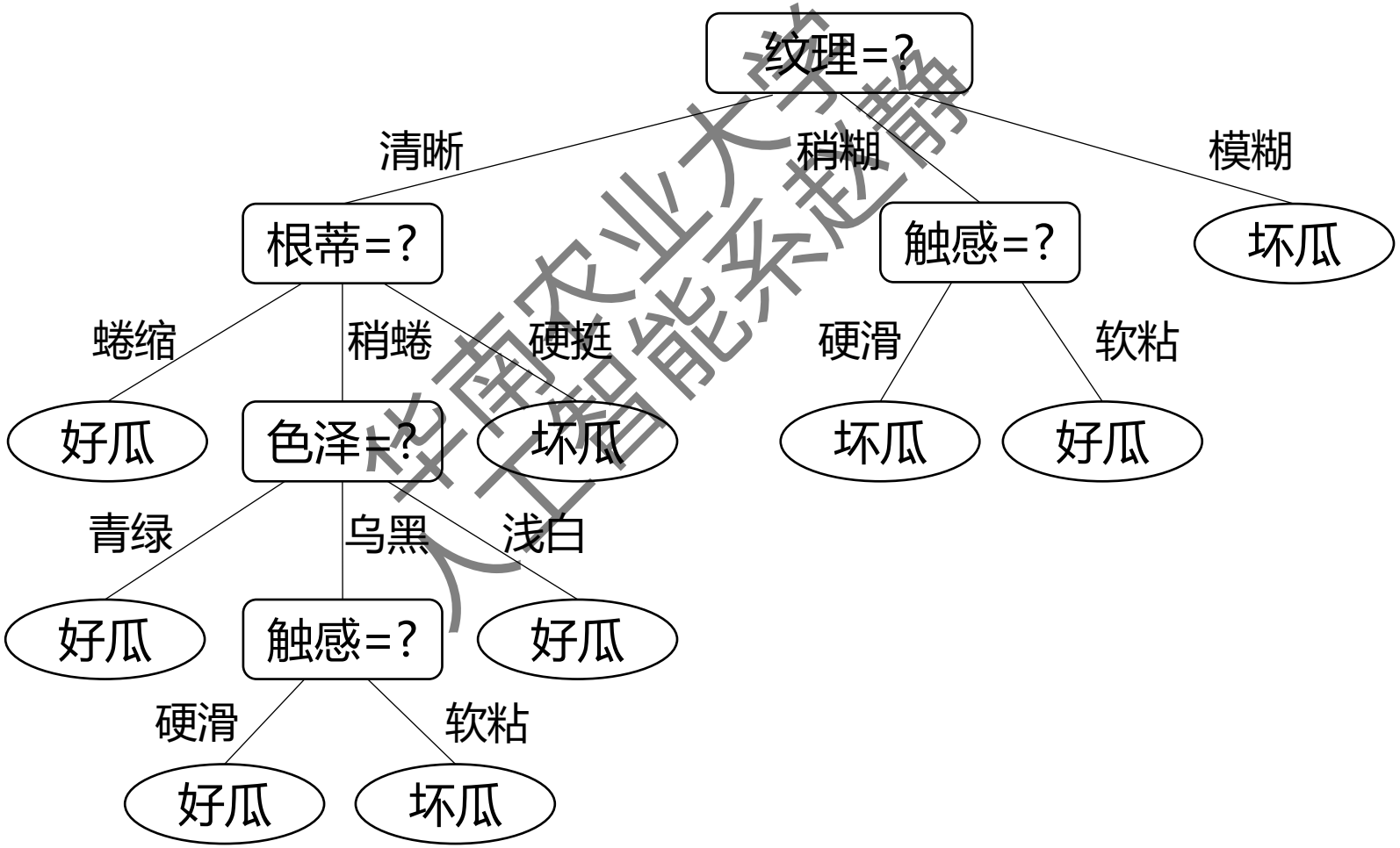
$$\text{Gain}(D, \text{触感}) = 0.006$$

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



7.2.3 信息增益率——C4.5算法

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

属性固有值： $$\text{IV}(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$$

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

C4.5：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

例:就业因素调查

学号	性别	学生干部	综合成绩	毕业论文	就业情况
1	男	是	70-79	优	已
2	女	是	80-89	中	已
3	男	不是	60-69	不及格	未
4	男	是	60-69	良	已
5	男	是	70-79	中	已
6	男	不是	70-79	良	未
7	女	是	60-69	良	已
8	男	是	60-69	良	已
9	女	是	70-79	中	未
10	男	不是	60-69	及格	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
13	男	不是	70-79	及格	未
14	男	不是	60-69	及格	已
15	男	是	70-79	良	已
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
18	女	是	70-79	良	已
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
21	女	是	60-69	优	已
22	男	是	60-69	良	已

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
6	男	不是	70-79	良	未
9	女	是	70-79	中	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
1	男	是	70-79	优	已
2	女	是	80-89	中	已
4	男	是	60-69	良	已
5	男	是	70-79	中	已
7	女	是	60-69	良	已
8	男	是	60-69	良	已
10	男	不是	60-69	及格	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
14	男	不是	60-69	及格	已
15	男	是	70-79	良	已
18	女	是	70-79	良	已
21	女	是	60-69	优	已
22	男	是	60-69	良	已

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
6	男	不是	70-79	良	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
1	男	是	70-79	优	已
4	男	是	60-69	良	已
5	男	是	70-79	中	已
8	男	是	60-69	良	已
10	男	不是	60-69	及格	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
14	男	不是	60-69	及格	已
15	男	是	70-79	良	已
22	男	是	60-69	良	已
9	女	是	70-79	中	未
2	女	是	80-89	中	已
7	女	是	60-69	良	已
18	女	是	70-79	良	已
21	女	是	60-69	优	已

$$\text{Entropy}(\text{男}) = -\frac{10}{17}\log_2\frac{10}{17} - \frac{7}{17}\log_2\frac{7}{17} = 0.97742$$

$$\text{Entropy}(\text{女}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.72193$$

$$\text{Entropy}(\text{性别}) = \frac{17}{22} * 0.97742 + \frac{5}{22} * 0.72193 = 0.91935$$

$$\text{Gain}(\text{性别}) = 0.94566 - 0.91935 = 0.02631$$

$$Ent_A(\text{性别}) = -\frac{17}{22}\log_2\frac{17}{22} - \frac{5}{22}\log_2\frac{5}{22} = 0.77323$$

$$\text{Gain\_Ratio}(\text{性别}) = 0.02631/0.77323 = 0.03403$$

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
6	男	不是	70-79	良	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
17	男	不是	80-89	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
10	男	不是	60-69	及格	已
14	男	不是	60-69	及格	已
1	男	是	70-79	优	已
4	男	是	60-69	良	已
5	男	是	70-79	中	已
8	男	是	60-69	良	已
11	男	是	80-89	及格	已
12	男	是	70-79	良	已
15	男	是	70-79	良	已
22	男	是	60-69	良	已
9	女	是	70-79	中	未
2	女	是	80-89	中	已
7	女	是	60-69	良	已
18	女	是	70-79	良	已
21	女	是	60-69	优	已

$$\text{Gain}(\text{学生干部}) = 0.94566 - 0.54382 = 0.40184$$

$$Ent_A(\text{学生干部}) = -\frac{13}{22}\log_2\frac{13}{22} - \frac{9}{22}\log_2\frac{9}{22} = 0.97602$$

$$\text{Gain\_Ratio}(\text{学生干部}) = 0.40184/0.97602 = 0.41171$$

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
10	男	不是	60-69	及格	已
14	男	不是	60-69	及格	已
4	男	是	60-69	良	已
8	男	是	60-69	良	已
22	男	是	60-69	良	已
7	女	是	60-69	良	已
21	女	是	60-69	优	已
6	男	不是	70-79	良	未
13	男	不是	70-79	及格	未
16	男	不是	70-79	良	未
19	男	不是	70-79	不及格	未
20	男	不是	70-79	良	未
1	男	是	70-79	优	已
5	男	是	70-79	中	已
12	男	是	70-79	良	已
15	男	是	70-79	良	已
9	女	是	70-79	中	未
18	女	是	70-79	良	已
17	男	不是	80-89	良	未
11	男	是	80-89	及格	已
2	女	是	80-89	中	已

$$\text{Gain}(\text{综合成绩}) = 0.94566 - 0.819897 = 0.125763$$

$$Ent_A(\text{综合成绩}) = -\frac{8}{22}\log_2\frac{8}{22} - \frac{11}{22}\log_2\frac{11}{22} - \frac{3}{22}\log_2\frac{3}{22} = 1.422675$$

$$\text{Gain\_Ratio}(\text{综合成绩}) = 0.125763/1.422675 = 0.088391$$

$$\text{Entropy}(\text{就业情况}) = -\frac{14}{22}\log_2\frac{14}{22} - \frac{8}{22}\log_2\frac{8}{22} = 0.94566$$

学号	性别	学生干部	综合成绩	毕业论文	就业情况
3	男	不是	60-69	不及格	未
19	男	不是	70-79	不及格	未
10	男	不是	60-69	及格	已
14	男	不是	60-69	及格	已
13	男	不是	70-79	及格	未
11	男	是	80-89	及格	已
4	男	是	60-69	良	已
8	男	是	60-69	良	已
22	男	是	60-69	良	已
7	女	是	60-69	良	已
6	男	不是	70-79	良	未
16	男	不是	70-79	良	未
20	男	不是	70-79	良	未
12	男	是	70-79	良	已
15	男	是	70-79	良	已
18	女	是	70-79	良	已
17	男	不是	80-89	良	未
21	女	是	60-69	优	已
1	男	是	70-79	优	已
5	男	是	70-79	中	已
9	女	是	70-79	中	未
2	女	是	80-89	中	已

$$\text{Gain}(\text{毕业论文}) = 0.94566 - 0.745557 = 0.200103$$

$$Ent_A(\text{毕业论文}) = -\frac{2}{22}\log_2\frac{2}{22} - \frac{4}{22}\log_2\frac{4}{22} - \frac{11}{22}\log_2\frac{11}{22} - \frac{2}{22}\log_2\frac{2}{22} - \frac{3}{22}\log_2\frac{3}{22} = 2.100806$$

$$\text{Gain\_Ratio}(\text{毕业论文}) = 0.200103/2.00103 = 0.10167158$$

	性别	学生干部	综合成绩	论文
Gain	0.02631	0.40184	0.12576	0.200103
Gain_ratio	0.03403	0.41171	0.08839	0.10167158

填空题 0.5分

设置

在二分类任务中，若当前样本集合的正类和负类的数量刚好各一半，此时信息熵为 [填空1] （保留1位小数）

作答

7.2.4 基尼指数 (Gini Index)——CART树

$$\mathrm{Gini}(D) = \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2$$

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

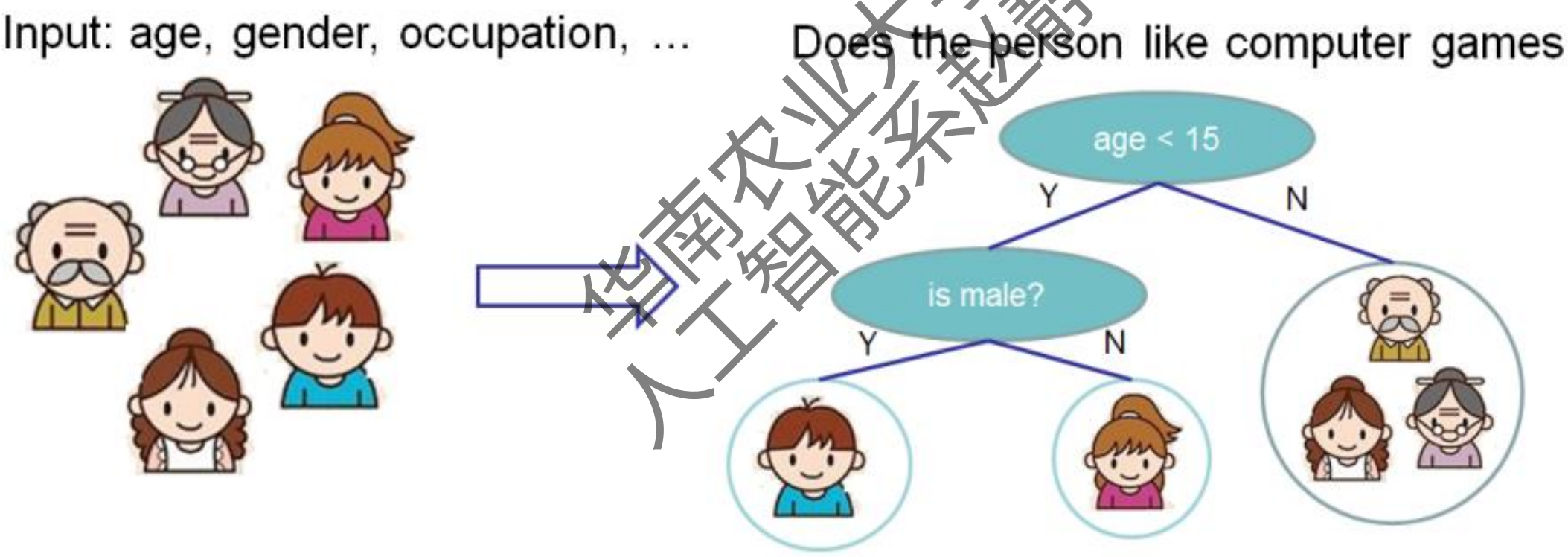
Gini(D) 越小，数据集 D 的纯度越高

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，选取使划分后基尼指数最小的属性

CART树是二叉树：二分递归划分，将当前样本集合划分为两个子集为两个子节点，使得生成的每个非叶子结点都有两个分支



例：贷款审批

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

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

$$Gini(D,A_1=\text{青年})=\frac{5}{15}\times\left(2\times\frac{2}{5}\times\left(1-\frac{2}{5}\right)\right)+\frac{10}{15}\times\left(2\times\frac{7}{10}\times\left(1-\frac{7}{10}\right)\right)=0.44$$

$Gini(D,A_1=\text{中年})=0.48$

$Gini(D,A_1=\text{老年})=0.44$

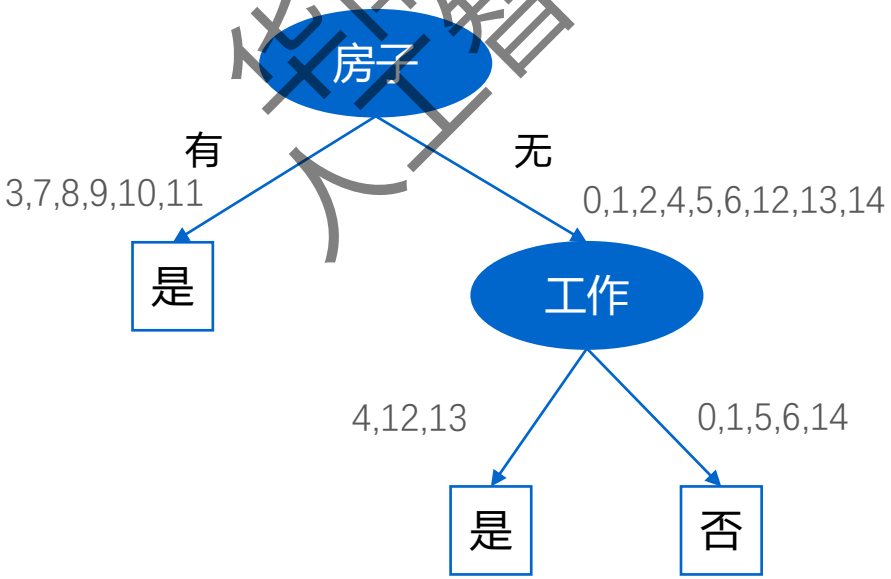
$Gini(D,A_2=\text{是})=0.32$

$Gini(D,A_3=\text{是})=0.27$

$Gini(D,A_4=\text{非常好})=0.36$

$Gini(D,A_4=\text{好})=0.47$

$Gini(D,A_4=\text{一般})=0.32$



	年龄	有工作	有房子	信用	类别
0	青年	否	否	一般	否
1	青年	否	否	好	否
2	青年	是	否	好	是
3	青年	是	是	一般	是
4	青年	否	否	一般	否
5	中年	否	否	一般	否
6	中年	否	否	好	否
7	中年	是	是	好	是
8	中年	否	是	非常好	是
9	中年	否	是	非常好	是
10	老年	否	是	非常好	是
11	老年	否	是	好	是
12	老年	是	否	好	是
13	老年	是	否	非常好	是
14	老年	否	否	一般	否

CART回归

✓ 集合$\mathcal{D}$的不纯净性度量：

$$H(\mathcal{D}) = \sum_{i \in \mathcal{D}} (\bar{y} - y_i)^2$$

$$\bar{y} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} y_i$$

集合中样本的 y 值越接近越纯净

相当于损失函数取L2损失，选择最小L2损失的分裂

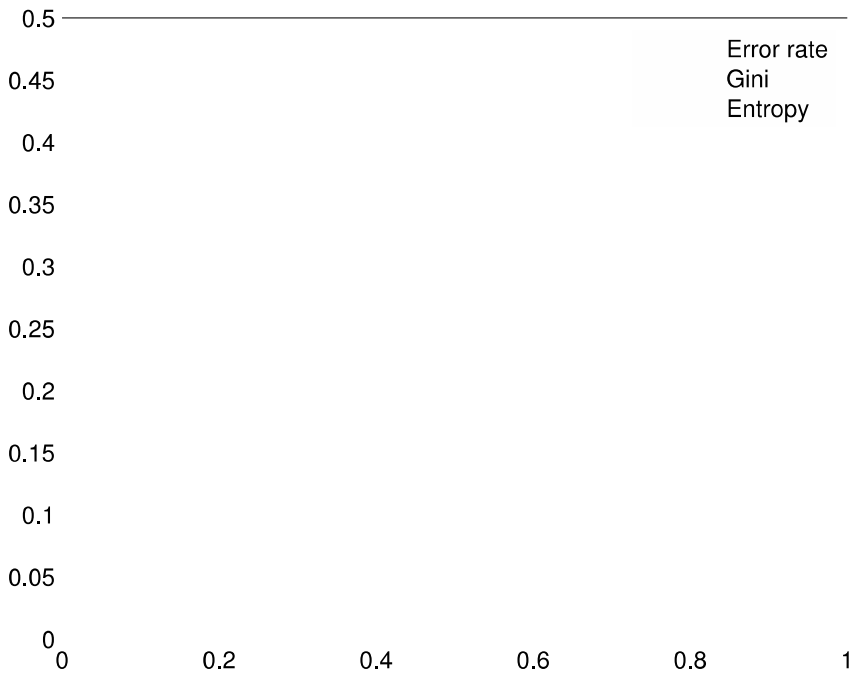
L2损失：$L(\hat{y}(\boldsymbol{\theta}), y) = (\hat{y}(\boldsymbol{\theta}) - y)^2 = (\bar{y} - y)^2$

预测值$\hat{y}$为样本均值 $\bar{y} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} y_i$ 时L2损失最小

总结：常用的不纯净度量

- 错误率 $H(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\mathbb{I}(y_i \neq \hat{y}) = 1 - p_k$
- 熵（ID3/C4.5） $\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$
- Gini指数（无需log，计算更快）

$$Gini(\mathcal{D}) = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2$$



两类分类的不纯度度量

7.3 剪枝(pruning)处理

剪枝 (pruning) 是决策树对付“过拟合”的主要手段

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

数据集

表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

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

训练集

验证集

7.3.1 预剪枝

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

验证集

结点1：若不划分，则根结点为叶结点，类别标记为训练样例最多的类别，若选“好瓜”，则验证集中{4,5,8} 被分类正确，验证集精度为 3/7 x 100% = 42.9%

验证集精度
划分前: 42.9%

预剪枝

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

验证集

结点1：若不划分，则根结点为叶结点，类别标记为训练样例最多的类别，若选“好瓜”，则验证集中{4,5,8} 被分类正确，验证集精度为 3/7 x 100% = 42.9%

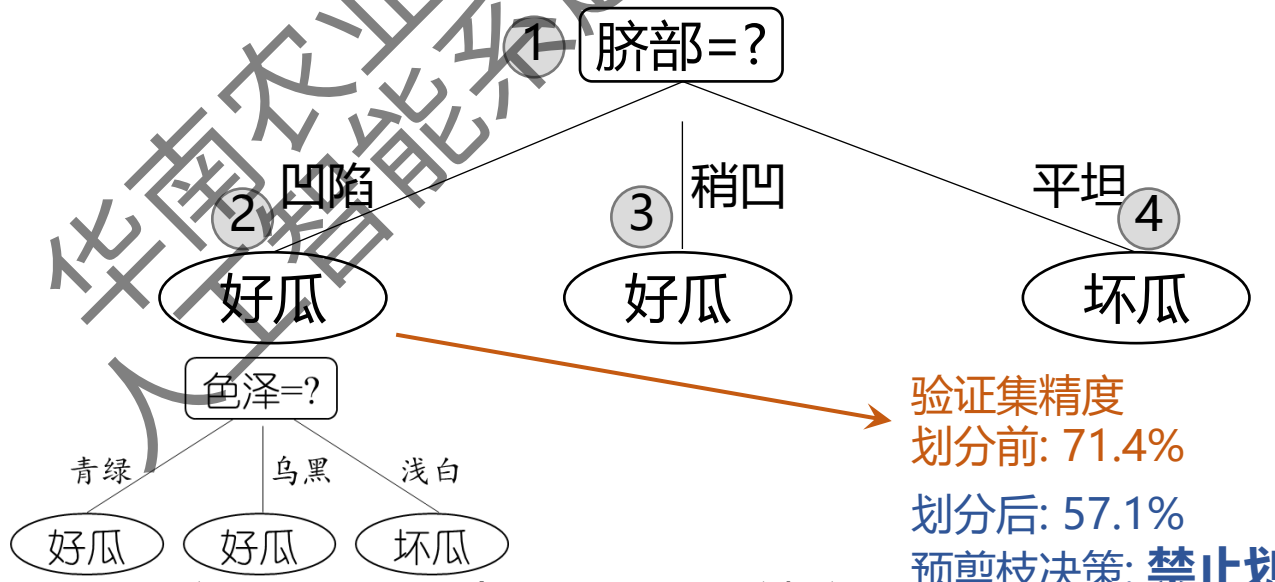
验证集精度
划分前: 42.9%
划分后: 71.4%
预剪枝决策: **划分**

结点1若划分，则根据划分后结点②③④ 的训练样例，它们将分别标记为“好瓜” “好瓜” “坏瓜”。此时，验证集中编号为 {4,5,8,11,12}的样例被划分正确，验证集精度为 5/7 x 100% = 71.4%

预剪枝

验证集

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



验证集精度
划分前: 71.4%

划分后: 57.1%

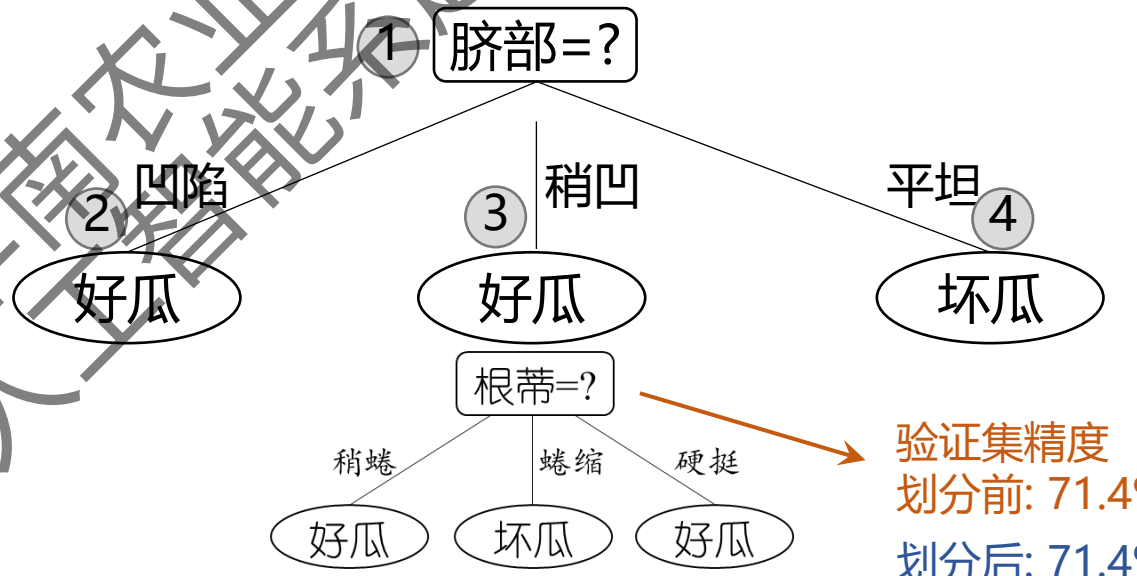
预剪枝决策: **禁止划分**

结点2：若划分，则验证集中{4,8,11,12} 被分类正确，验证集精度为 4/7 x100% = 57.1%

预剪枝

验证集

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



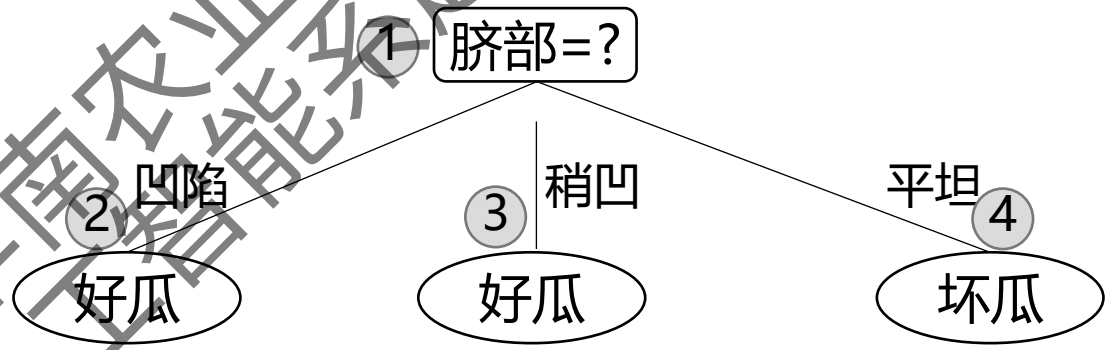
验证集精度
划分前: 71.4%
划分后: 71.4%
预剪枝决策: **禁止划分**

结点3：若划分，则验证集中{4,5,8,11,12} 被分类正确，验证集精度为 5/7 x100% = 71.4%

预剪枝

验证集

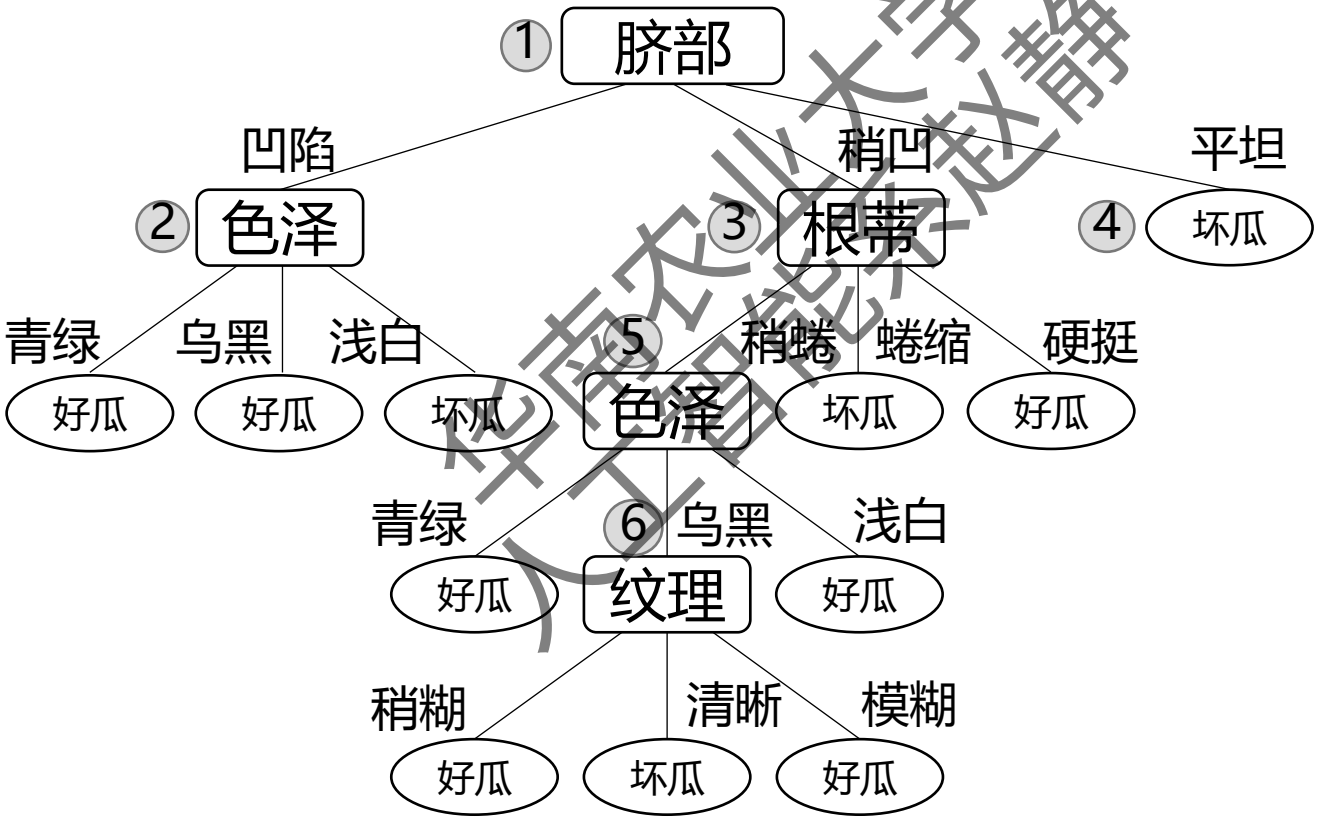
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否



最终，预剪枝的得到的决策树

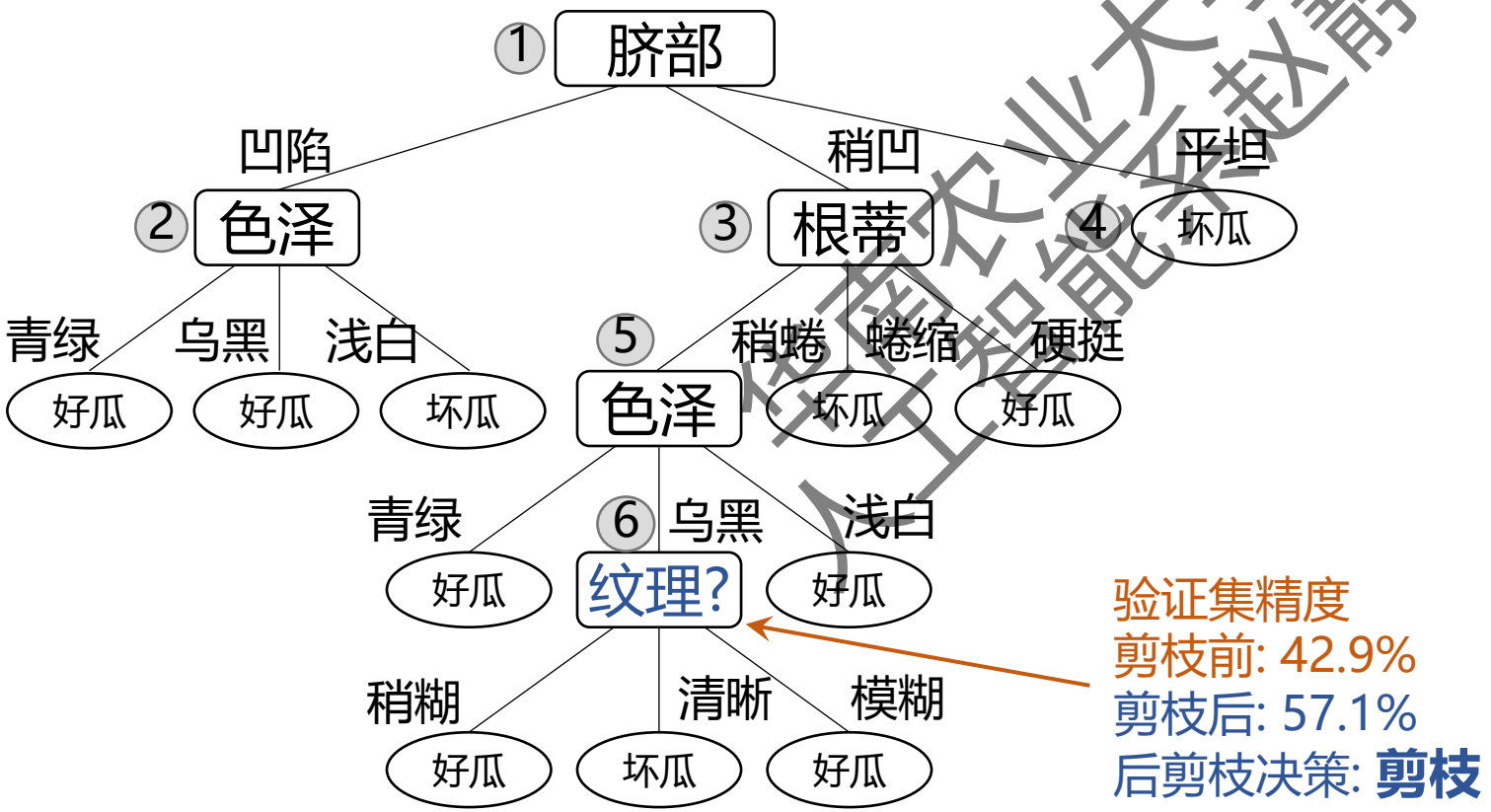
7.3.2 后剪枝

先生成一棵完整的决策树，其验证集精度测得为 42.9%



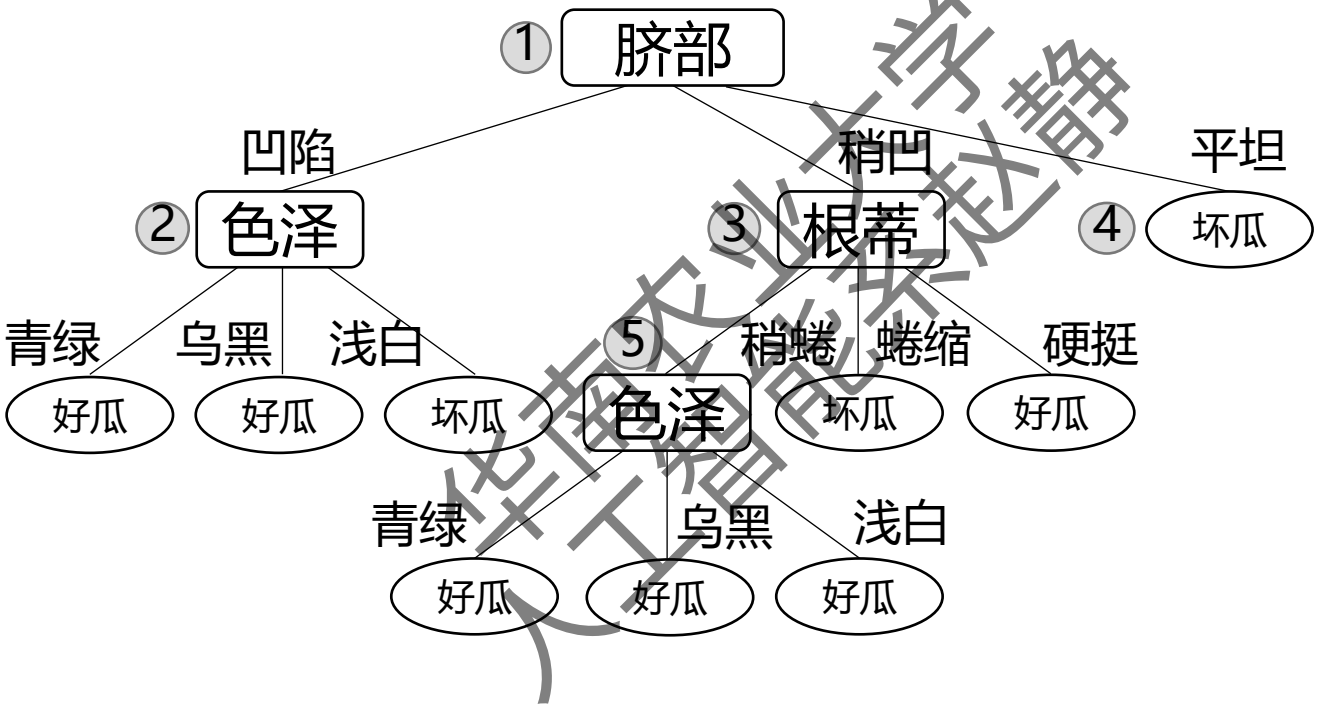
后剪枝

首先考虑结点⑥，若将其替换为叶结点，根据落在其上的训练样例 $\{7, 15\}$ 将其标记为“好瓜”，测得验证集精度提高至 57.1%，于是决定剪枝



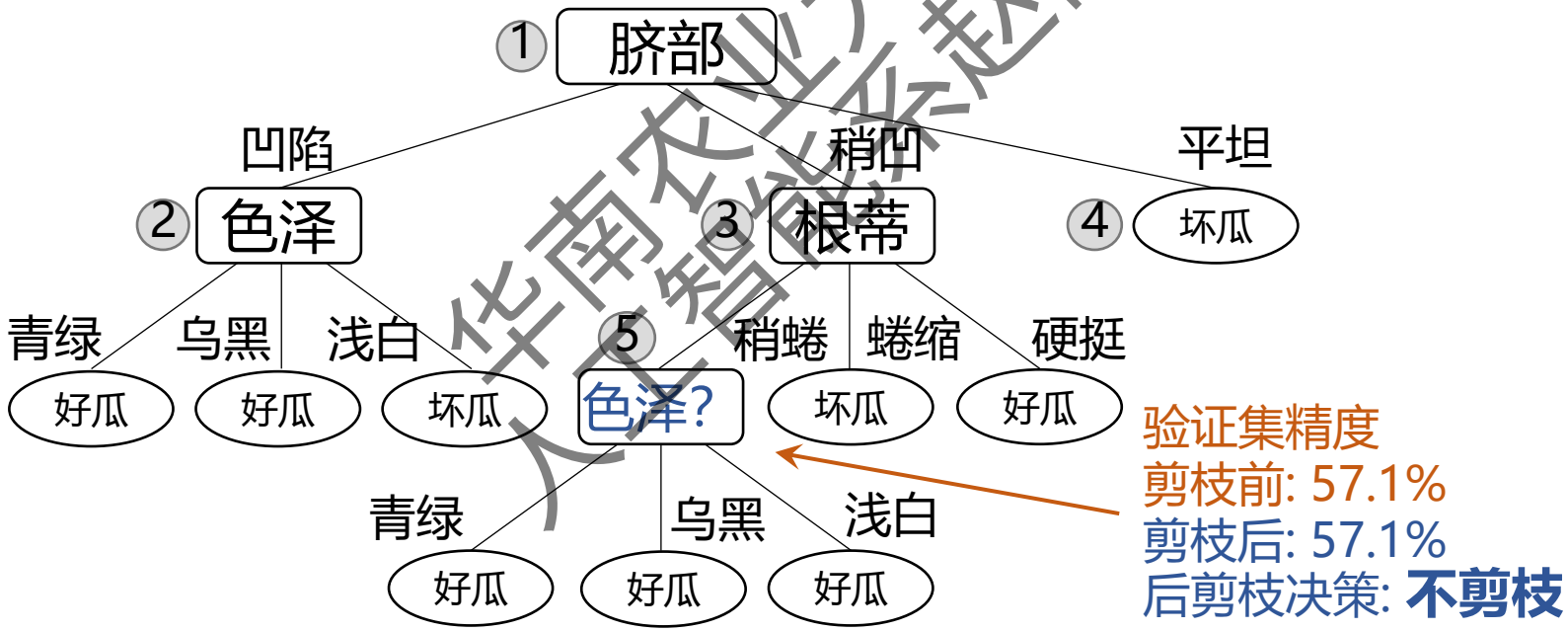
验证集精度
剪枝前: 42.9%
剪枝后: 57.1%
后剪枝决策: **剪枝**

后剪枝



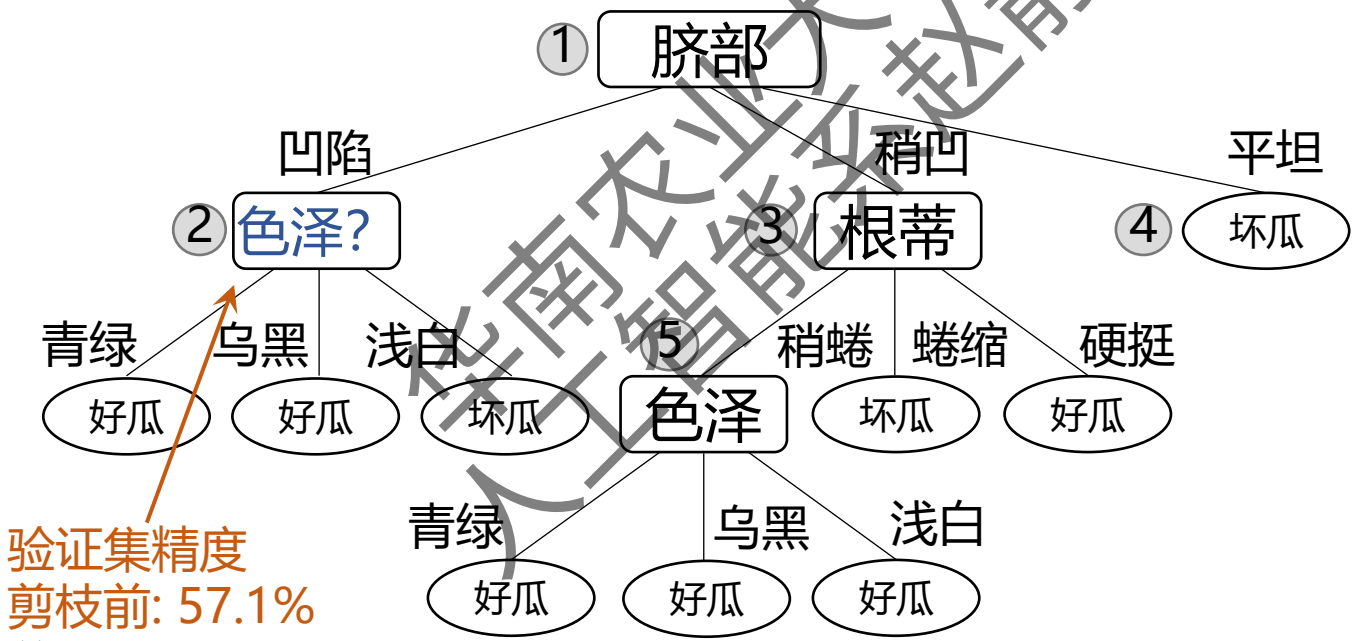
后剪枝

然后考虑结点⑤，若将其替换为叶结点，根据落在其上的训练样例 $\{6, 7, 15\}$ 将其标记为“好瓜”，测得验证集精度仍为 57.1%，可以不剪枝



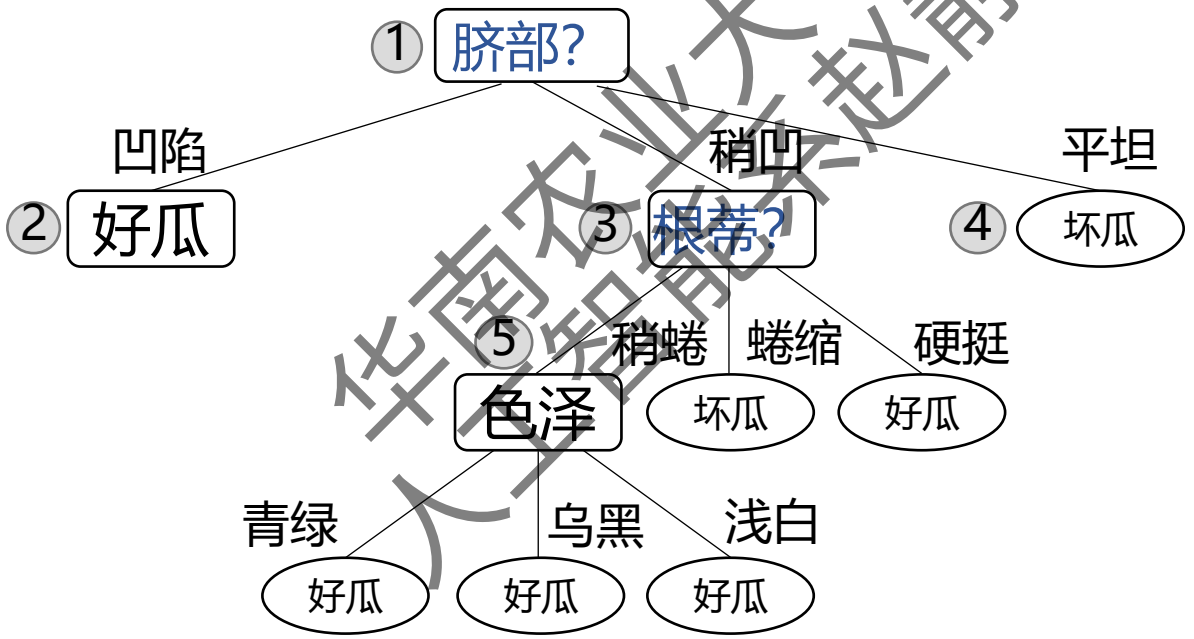
后剪枝

对结点②，若将其替换为叶结点，根据落在其上的训练样例 $\{1, 2, 3, 14\}$ ，将其标记为“好瓜”，测得验证集精度提升至 71.4%，决定剪枝



后剪枝

对结点③ 和①，先后替换为叶结点，均未测得验证集精度提升，于是不剪枝



最终，后剪枝得到的决策树

预剪枝 vs. 后剪枝

- 时间开销：
  - 预剪枝：测试时间开销降低，训练时间开销降低
  - 后剪枝：测试时间开销降低，训练时间开销增加

- 过/欠拟合风险：
  - 预剪枝：过拟合风险降低，欠拟合风险增加
  - 后剪枝：过拟合风险降低，欠拟合风险基本不变

- 泛化性能：后剪枝 通常优于 预剪枝

7.4 连续与缺失值

7.4.1 连续值处理——连续属性离散化(二分法)

1. 产生候选划分点

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

2. 选择最优划分点

$$\begin{aligned} \mathrm{Gain}(D, a) &= \max_{t \in T_a} \mathrm{Gain}(D, a, t) \\ &= \max_{t \in T_a} \mathrm{Ent}(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} \mathrm{Ent}(D_t^\lambda) \end{aligned}$$

一个例子

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

对属性“密度”，其候选划分点集合包含16 个候选值：

$$T_{\text{密度}} = \{0.244, 0.294, 0.351, 0.381, 0.420, 0.459, 0.518, 0.574, 0.600, 0.621, 0.636, 0.648, 0.661, 0.681, 0.708, 0.746\}$$

可计算其最大信息增益为0.262，对应划分点为0.381

与离散属性不同，若当前结点划分属性为连续属性，该属性还可作为其后代结点的划分属性

7.4.2 缺失值

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

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

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

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

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	[illegible]	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

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

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

$\tilde{r}_v$：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

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

ρ：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

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

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	青绿	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	浅白	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

总结

➢ 树模型的优点

- ■容易解释
- ■对特征预处理要求少,能自动进行特征选择
- ■可处理缺失数据
- ■可扩展到大数据规模

➢ 树模型的缺点

- 正确率不高：建树过程过于贪心

  可作为Boosting的弱学习器（深度不太深）
- 模型不稳定（方差大）：输入数据小的变化会带来树结构的变化

  Bagging：随机森林
- 当特征数目相对样本数目太多时，容易过拟合

决策树的三种基本类型对比

建立决策树的关键，即在当前状态下选择哪个属性作为分类依据。根据不同的目标函数，决策树主要有三种算法： ID3(Iterative Dichotomiser)、C4.5、CART(Classification And Regression Tree)。

算法	支持模型	树结构	特征选择	连续值处理	缺失值处理	剪枝	特征属性多次使用
ID3	分类	多叉树	信息增益	不支持	不支持	不支持	不支持
C4.5	分类	多叉树	信息增益率	支持	支持	支持	不支持
CART	分类 回归	二叉树	基尼指数 均方差	支持	支持	支持	支持

- ID3——信息增益
- C4.5——信息增益率
- CART——Gini指数