第四章：决策树

主讲：连德富 特任教授 | 博士生导师

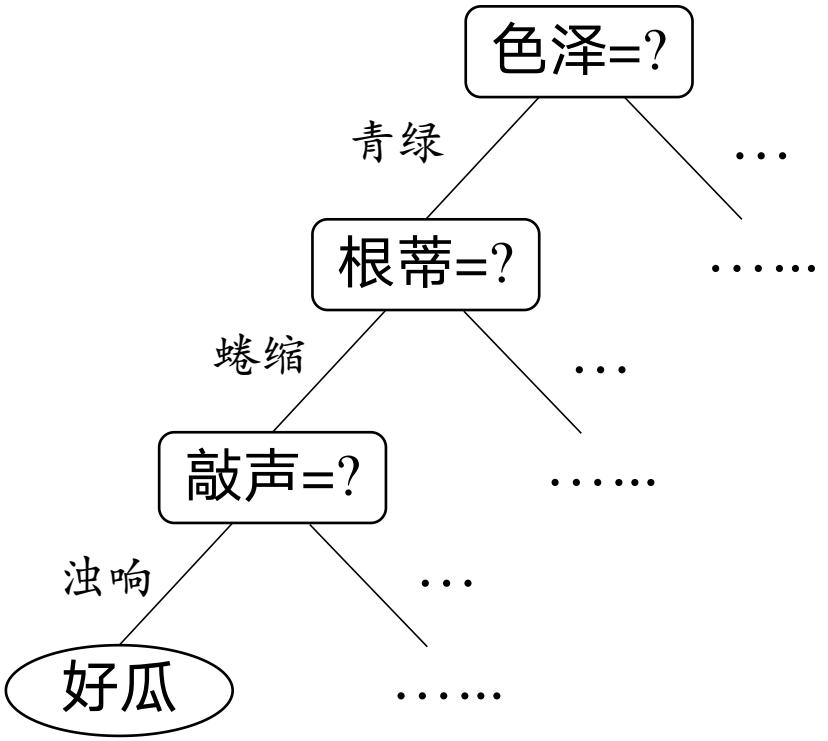
邮箱：liandefu@ustc.edu.cn

手机：13739227137

主页：http://staff.ustc.edu.cn/~liandefu

基本流程

- 决策树基于树结构来进行预测



- 决策过程的最终结论对应了我们所希望的判定结果（叶子节点）
- 决策过程中提出的每个判定问题都是对某个属性的“测试”（内部节点）
- 每个测试的结果或是导出最终结论，或者导出进一步的判定问题，其考虑范围是在上次决策结果的限定范围之内
- 每个节点包含的样本集合根据属性测试划分到子节点中（根节点包括样本全集）
- 从根结点到每个叶结点的路径对应了一个判定测试序列

决策树学习的目的是为了产生一棵泛化能力强，即处理未见示例能力强的决策树

基本流程

决策树的生成是一个递归过程

Algorithm 1 决策树学习基本算法

输入:

- 训练集 $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$;
- 属性集 $A = \{a_1, \ldots, a_d\}$.

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 生成每一个分枝; 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:        将分枝结点标记为叶结点, 其类别标记为 D 中样本最多的类; return
13:    else
14:        以 TreeGenerate(D_v, A − {a_*}) 为分枝结点
15:    end if
16: end for
```

输出: 以 node 为根结点的一棵决策树

下述三种情况递归返回

（1）当前结点包含的样本全部属于同一类别

（2）当前属性集为空，或所有样本在所有属性上取值相同

（3）当前结点包含的样本集合为空

划分选择

- 决策树学习的关键在于如何选择最优划分属性

一般而言，随着划分过程不断进行，我们希望决策树的分支结点所包含的样本尽可能属于同一类别，即结点的“纯度”(purity)越来越高

经典的属性划分方法：

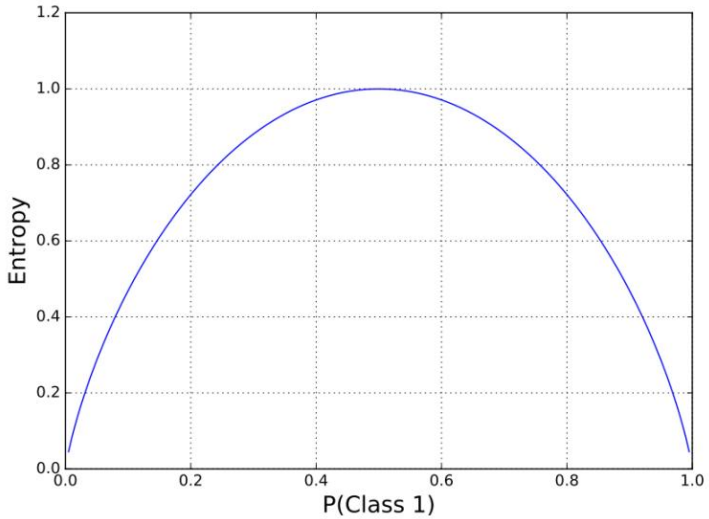
信息增益　增益率　基尼指数

划分选择—信息增益

- “信息熵”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$p_k(\mathrm{k}=1,2,\cdots,|\mathcal{Y}|)$，则$D$的信息熵定义为

$$Ent(D) = -\sum_k p_k \log_2 p_k$$



二分类：

$$Ent(D) = -p_1 \log p_1 - (1-p_1)\log(1-p_1)$$

划分选择—信息增益

- 对于任意一个离散随机变量Y，信息熵 $H(Y) = -\sum_i P(Y=i)\log P(Y=i)$
- 信息熵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$p_k(k = 1,2,\cdots,|\mathcal{Y}|)$，则$D$的信息熵定义为

$$Ent(D) = -\sum_k p_k \log_2 p_k$$

$Ent(D)$的值越小，则 D的纯度越高

若f为凹函数，则 $\sum_i a_i f(x_i) \leq f(\sum_i a_i x_i)$

易证：$$Ent(D) = \sum_k p_k \log_2 \frac{1}{p_k} \leq \log_2 \sum_k p_k \frac{1}{p_k} = \log_2 |\mathcal{Y}|$$

均匀分布的熵最大

计算信息熵时约定：若 $p=0$，则 $p\log p = 0$

划分选择—信息增益

- 信息增益是是变量间相互依赖性的度量，是联合分布和边缘分布乘积的相似程度度量

$$I(X,Y) = \sum_{y}\sum_{x} P(X=x, Y=y)\log\frac{P(X=x, Y=y)}{P(X=x)P(Y=y)}$$

$$= H(Y) - \sum_{x} P(X=x)H(Y|X=x)$$

$$= H(Y) - H(Y|X)$$

$$= H(X) - H(X|Y)$$

划分选择—信息增益

- 假设离散属性a有V个可能的取值$\{a^1, a^2, \cdots, a^V\}$
- 若用a来进行划分，则会产生V个分支结点；第v个分支结点包含D中所有在属性a上取值为a^v的样本，记为D^v

用属性a对样本集D进行划分所获得的“信息增益”：

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

$I(a,Y)$　$H(Y)$　$P(a^v)$　$H(Y|a^v)$

为分支结点权重，样本数越多的分支结点的影响越大

信息增益越大，则意味着使用属性a来进行划分所获得的“纯度提升”越大

- ID3决策树算法[Quinlan, 1986]以信息增益为准则来选择划分属性

划分选择—信息增益

信息增益实例

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

数据集包含17个训练样本，$|\mathcal{Y}| = 2$

正例占$p_1 = \frac{8}{17}$，反例占$p_2 = \frac{9}{17}$

计算得到根结点的信息熵为

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

划分选择—信息增益

信息增益实例

以属性“色泽”为例，其对应的3个数据子集分别为 D^1(色泽=青绿)， D^2(色泽=乌黑)，D^3 (色泽=浅白)

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

子集D^1包含编号为{1,4,6,10,13,17} 6个样例，正例$p_1 = \frac{3}{6}$，反例$p_2 = \frac{3}{6}$

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

子集D^2包含编号为{2,3,7,8,9,15 } 6个样例，正例$p_1 = \frac{4}{6}$，反例$p_2 = \frac{2}{6}$

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

子集D^3包含编号为{5,11,12,14,16} 5个样例，正例$p_1 = \frac{1}{5}$，反例$p_2 = \frac{4}{5}$

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

属性“色泽”的信息增益为

$$\text{Gain}(D, \text{色泽}) = Ent(D) - \sum_v \frac{|D^v|}{|D|} Ent(D^v)$$

$$= 0.998 - (\frac{6}{17} \times 1.000 + \frac{6}{17} \times 0.918 + \frac{5}{17} \times 0.772) = 0.109$$

划分选择—信息增益

信息增益实例

			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

$\text{Gain}(D, \text{色泽}) = 0.109$

$\text{Gain}(D, \text{根蒂}) = 0.143$

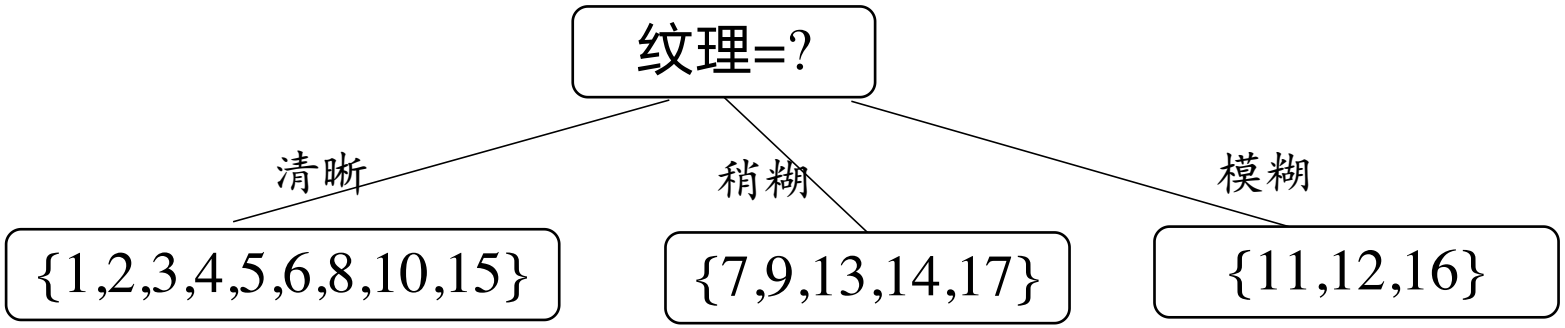
$\text{Gain}(D, \text{敲声}) = 0.141$

$\text{Gain}(D, \text{纹理}) = 0.381$

$\text{Gain}(D, \text{脐部}) = 0.289$

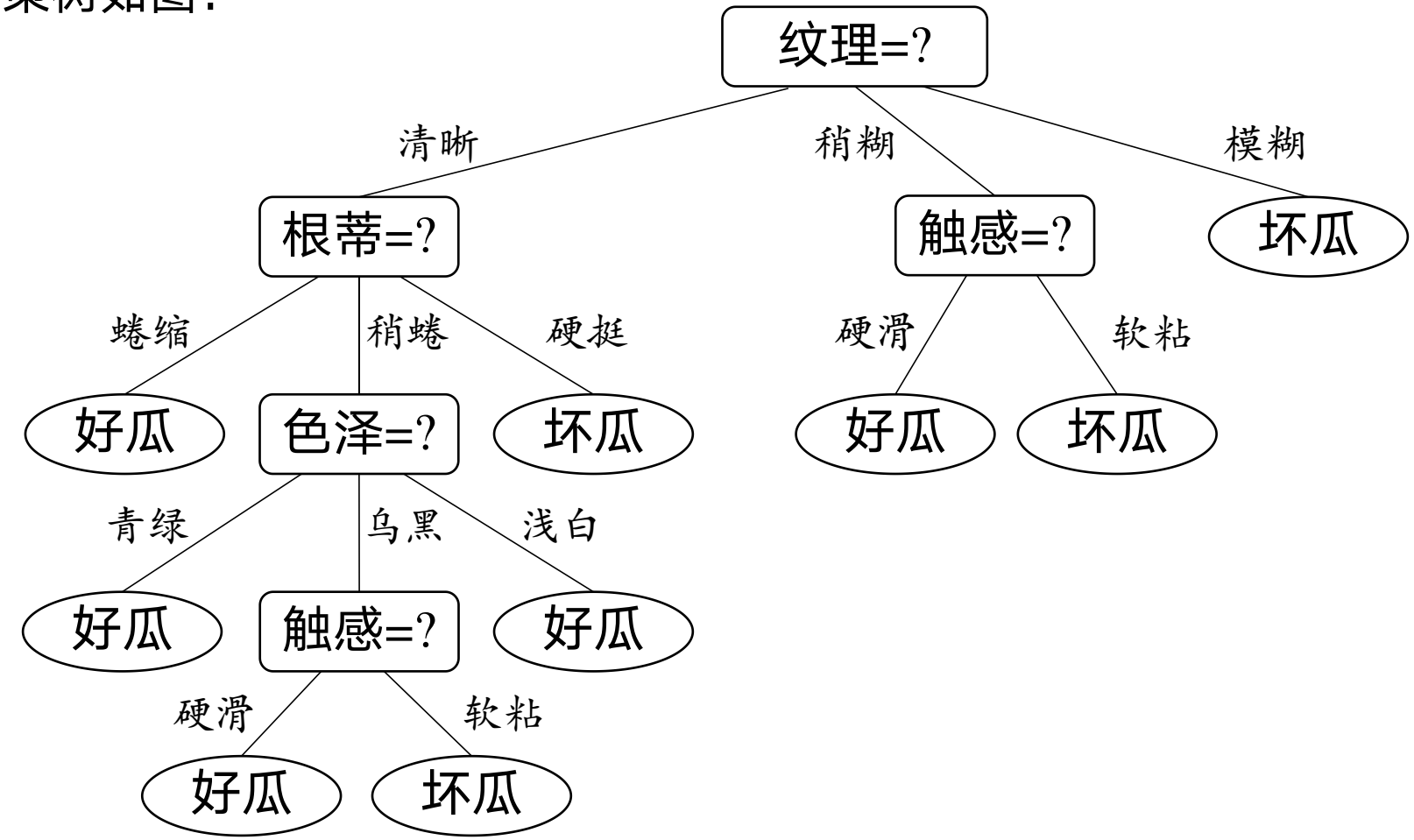
$\text{Gain}(D, \text{触感}) = 0.006$

属性“纹理”的信息增益最大，其被选为划分属性



划分选择—信息增益

决策树学习算法将对每个分支结点做进一步划分，最终得到的决策树如图：



划分选择—信息增益

存在的问题

若把“编号”也作为一个候选划分属性，则其信息增益一般远大于其他属性。显然，这样的决策树不具有泛化能力，无法对新样本进行有效预测

信息增益对可取值数目较多的属性有所偏好

划分选择—增益率

- 增益率定义：

$$\text{Gain}-\text{ratio}(D,a)=\frac{\text{Gain}(D,a)}{\text{IV}(a)} \qquad \text{其中 } \text{IV}(a)=-\sum_v \frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$$

称为属性a的“固有值”[Quinlan, 1993]，属性a的可能取值数目越多（即V越大），则$IV(a)$的值通常就越大

- 存在的问题

增益率准则对可取值数目较少的属性有所偏好

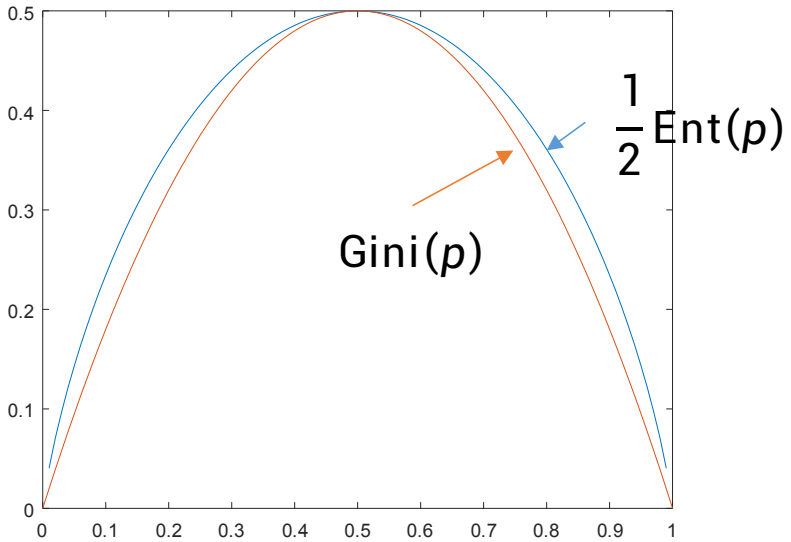
- C4.5 [Quinlan, 1993]使用了一个启发式：先从候选划分属性中找出信息增益高于平均水平的属性，再从中选取增益率最高的

划分选择-基尼指数

- 数据集D的纯度可用“基尼值”来度量　　$\log x \approx x - 1$

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

- 反映了从D中随机抽取两个样本，其类别标记不一致的概率
- Gini(D)越小，数据集D的纯度越高



划分选择-基尼指数

- 数据集D的纯度可用“基尼值”来度量

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

- 属性a的基尼指数定义为：

$$\text{Gini} - \text{index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Gini}(D^v)$$

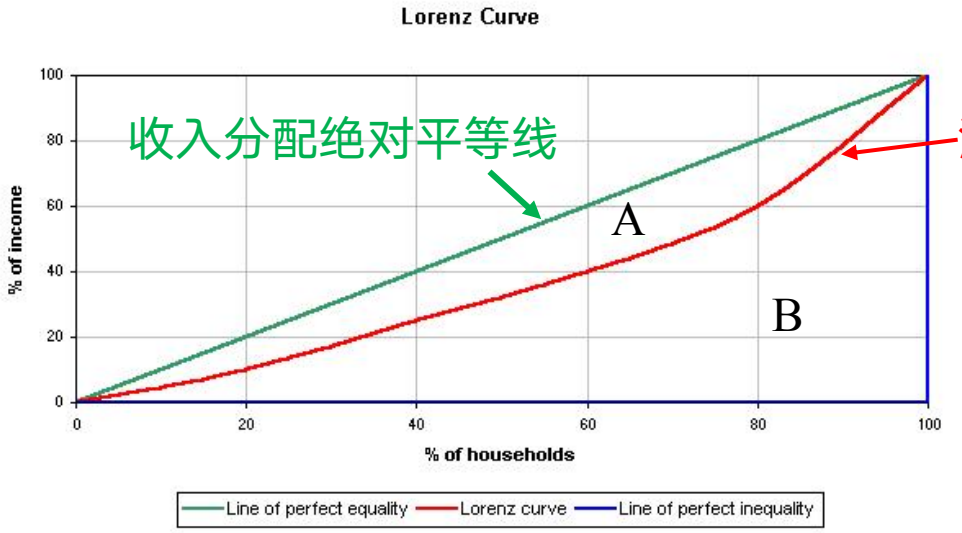
- 选择那个使划分后基尼指数最小的属性作为最优划分属性，即

$$a^\star = \arg\min_{a \in A} \text{Gini} - \text{index}(D, a)$$

- CART [Breiman et al., 1984]采用“基尼指数”来选择划分属性

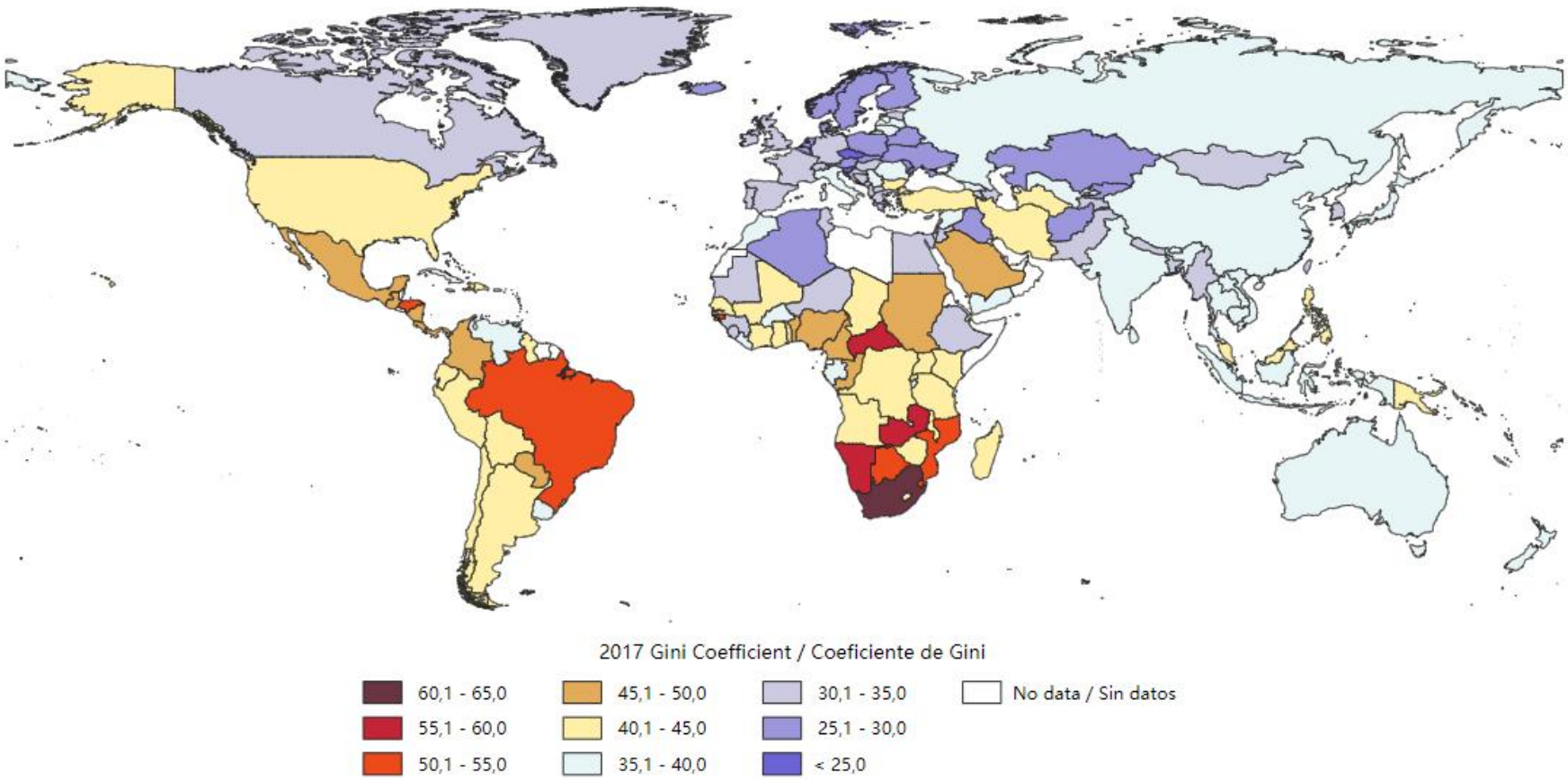
基尼系数

- 洛伦兹曲线是在过往财富分配数据上建立的累积分布函数所对应的曲线
- x%代表一部分（收入相似）家庭占整个社会家庭的比例，以y%代表该部分家庭的收入占整个社会收入的比例
- 基尼系数是根据洛伦兹曲线所定义的判断年收入分配公平程度的指标



基尼系数 $\frac{A}{A+B}$

基尼系数



2017年世界银行基尼系数世界地图。

基尼系数越小收入分配越平均，基尼系数越大收入分配越不平均。

剪枝处理

- 为什么剪枝
  - “剪枝”是决策树学习算法对付“过拟合”的主要手段
  - 可通过“剪枝”来一定程度避免因决策分支过多，以致于把训练集自身的一些特点当做所有数据都具有的一般性质而导致的过拟合

- 剪枝的基本策略
  - 预剪枝
  - 后剪枝

- 判断决策树泛化性能是否提升的方法
  - 留出法：预留一部分数据用作“验证集”以进行性能评估

剪枝处理

- 数据集

训练集

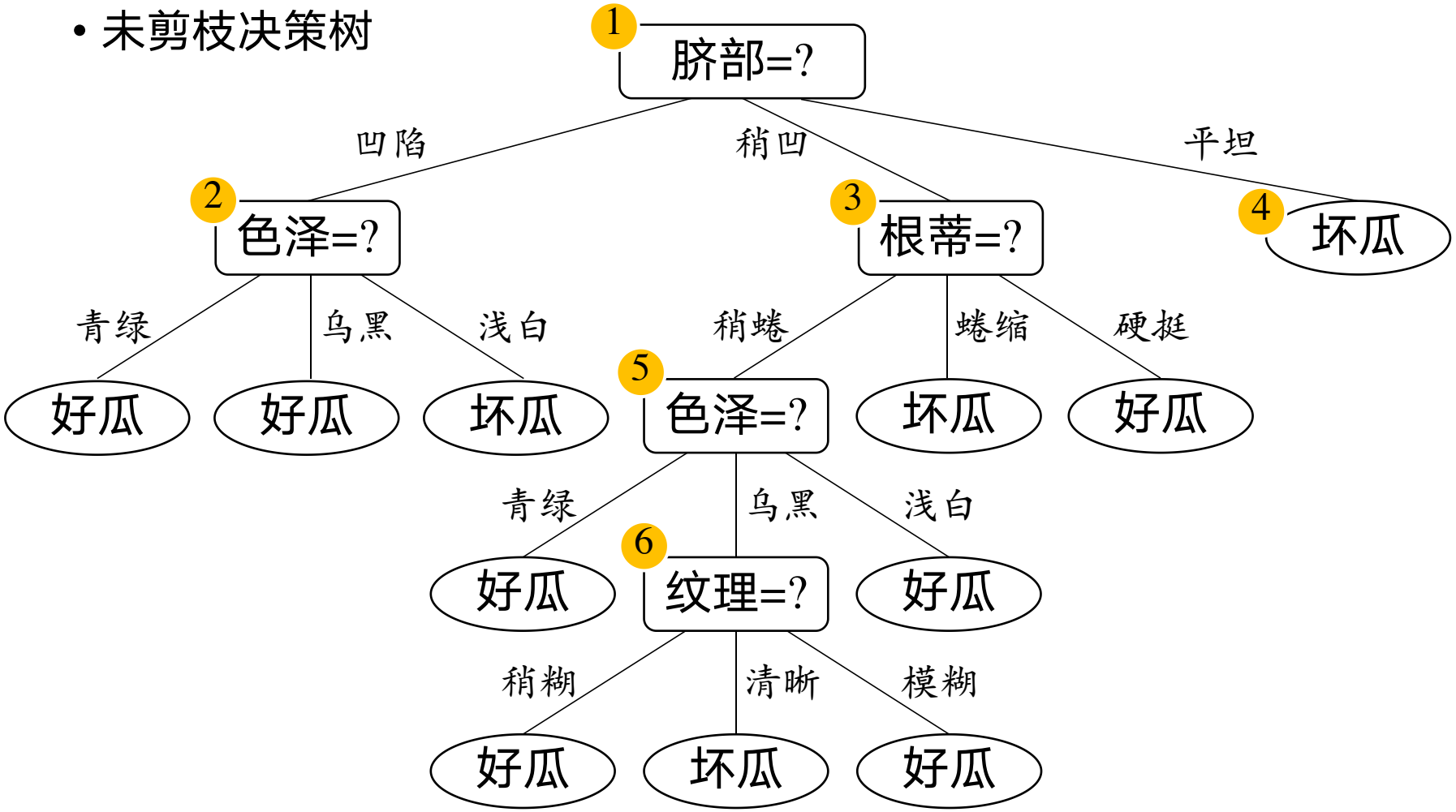
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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

剪枝处理

- 未剪枝决策树


1
脐部=?
凹陷
稍凹
平坦
2
色泽=?
3
根蒂=?
4
坏瓜
青绿
乌黑
浅白
稍蜷
蜷缩
硬挺
好瓜
好瓜
坏瓜
5
色泽=?
坏瓜
好瓜
青绿
乌黑
浅白
6
好瓜
纹理=?
好瓜
稍糊
清晰
模糊
好瓜
坏瓜
好瓜

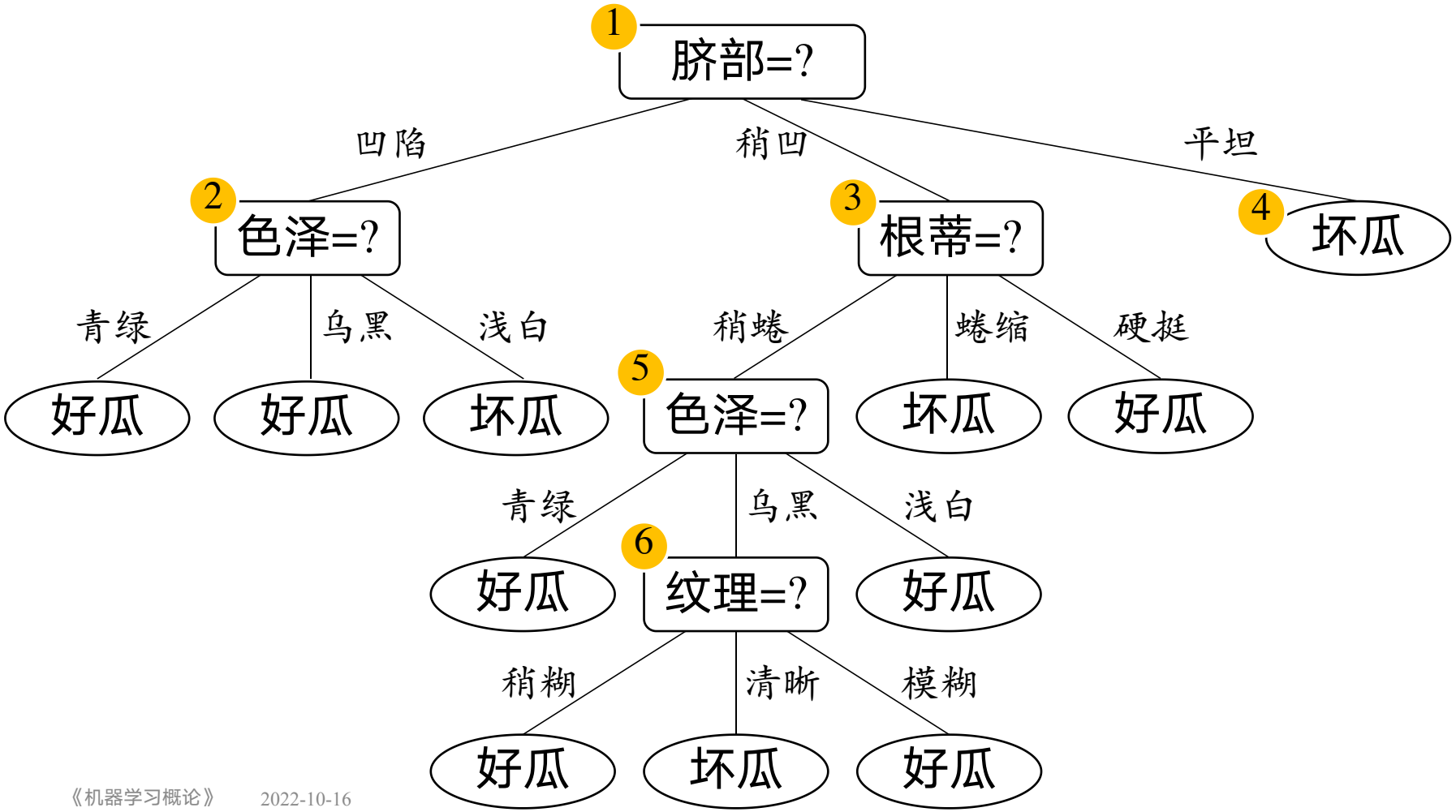
剪枝处理-预剪枝

- 决策树生成过程中，对每个结点在划分前先进行估计

- 若当前结点的划分不能带来决策树泛化性能提升，则停止划分并将当前结点记为叶结点，其类别标记为训练样例数最多的类别

剪枝处理

- 基于信息增益准则，选取属性“脐部”划分训练集
- 分别计算划分前及划分后的验证集精度，判断是否需要划分



剪枝处理—预剪枝

训练集

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

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

1 脐部=?

结点1：若不划分，则将其标记为叶结点，类别标记为训练样例中最多的类别，即好瓜。验证集中{4,5,8} 被分类正确，得到验证集精度为$\frac{3}{7} \times 100\% = 42.9\%$

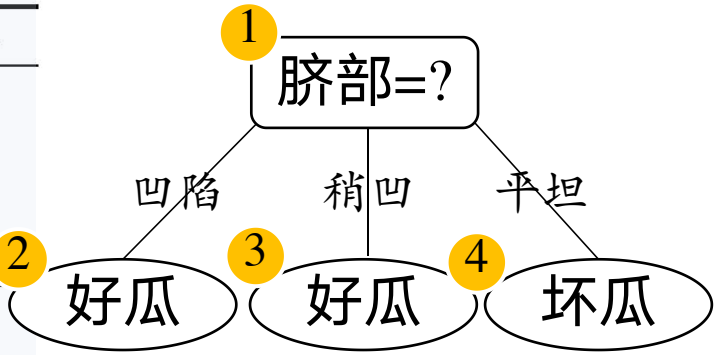
剪枝处理—预剪枝

训练集

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

验证集

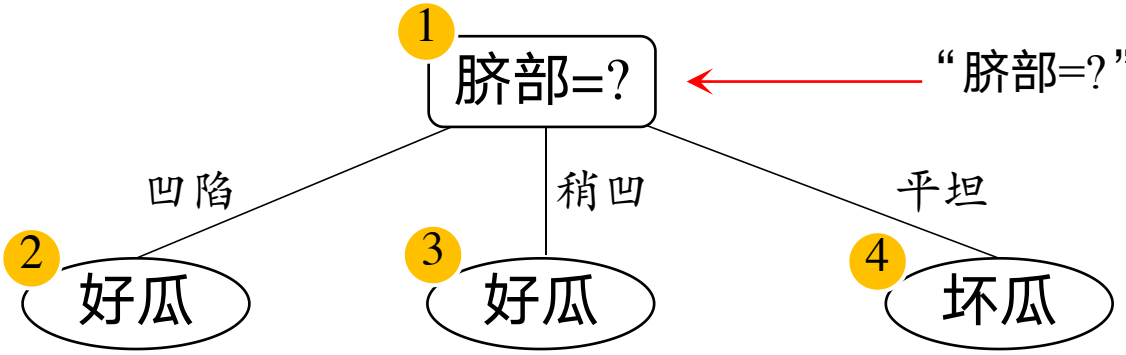
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否



结点1：若划分，根据结点 2 3 4 的训练样例，将这 3个结点分别标记为“好瓜”、“好瓜”、“坏瓜”。此时，验证集中编号 {4,5,8,11,12} 样例被划分正确，

验证集精度 $\frac{5}{7} \times 100\% = 71.4\%$

剪枝处理—预剪枝



“脐部=?”

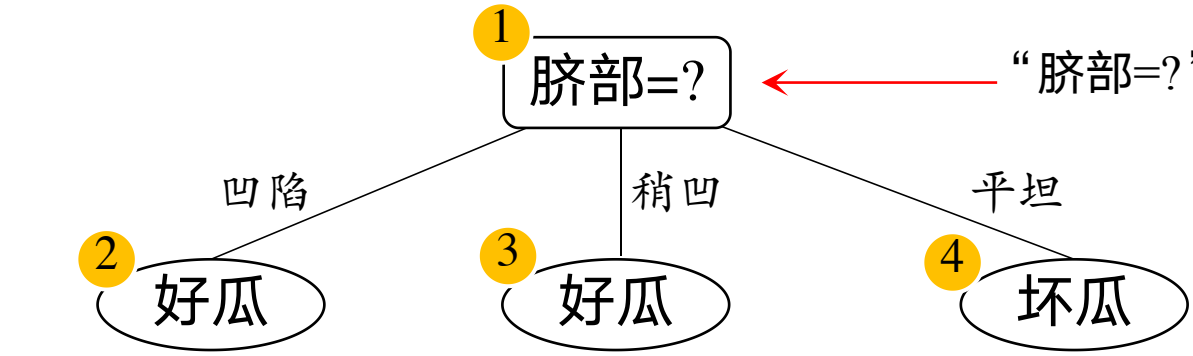
验证集精度

划分前: 42.9%

划分后: 71.4%

预剪枝决策: 划分

剪枝处理—预剪枝



验证集精度

“脐部=?” 划分前: 42.9%
划分后: 71.4%

预剪枝决策: 划分

“色泽=?”

验证集精度

划分前: 71.4%
划分后: 57.1%

预剪枝决策: 禁止划分

“根蒂=?”

验证集精度

划分前: 71.4%
划分后: 71.4%

预剪枝决策: 禁止划分

最终得到仅有一层划分的决策树，称为“**决策树桩**”

剪枝处理—预剪枝

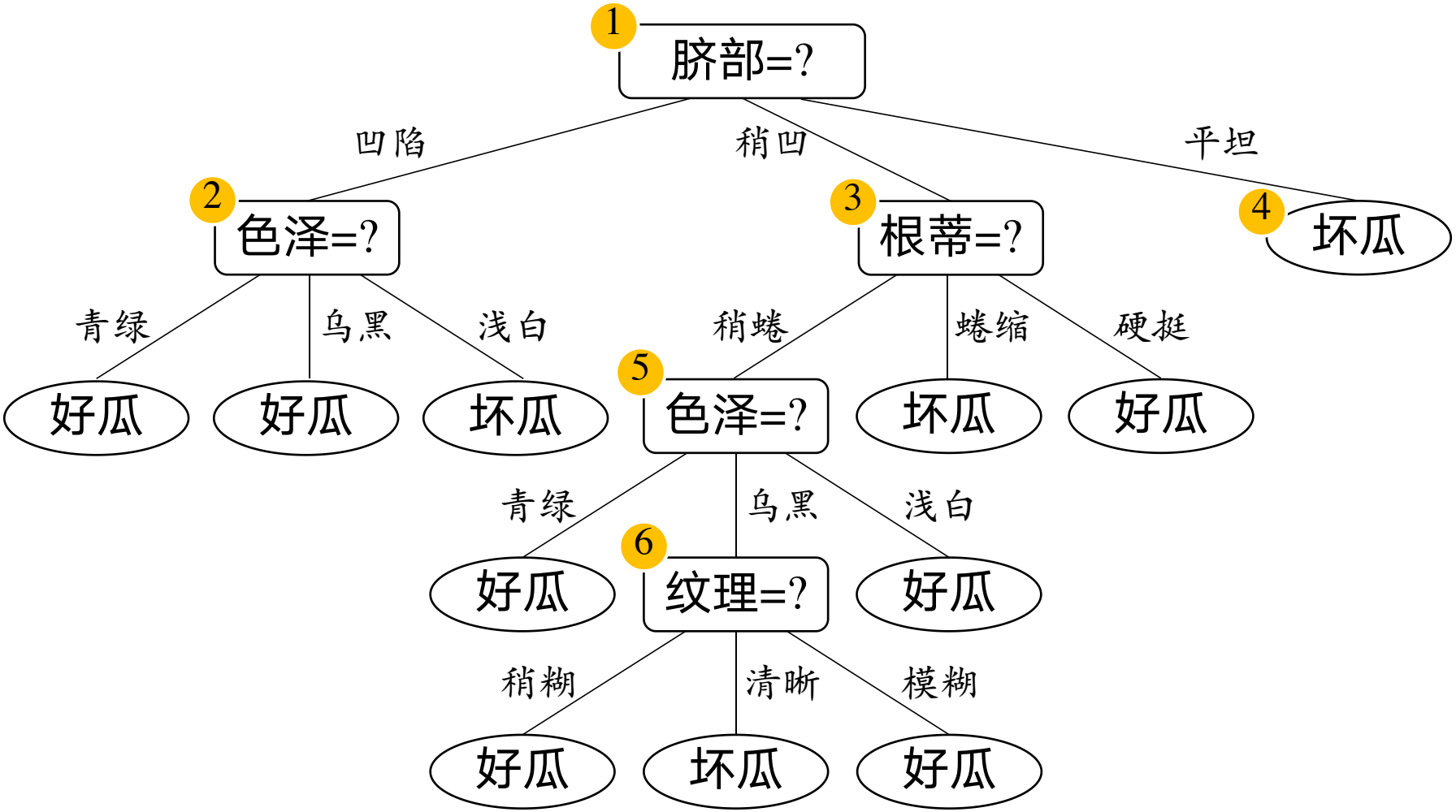
优点

- 降低过拟合风险
- 显著减少训练时间和测试时间开销

缺点

- 欠拟合风险：有些分支的当前划分虽然不能提升泛化性能，但在其基础上进行的后续划分却有可能导致性能显著提高。预剪枝基于“贪心”本质禁止这些分支展开，带来了欠拟合风险

剪枝处理—后剪枝

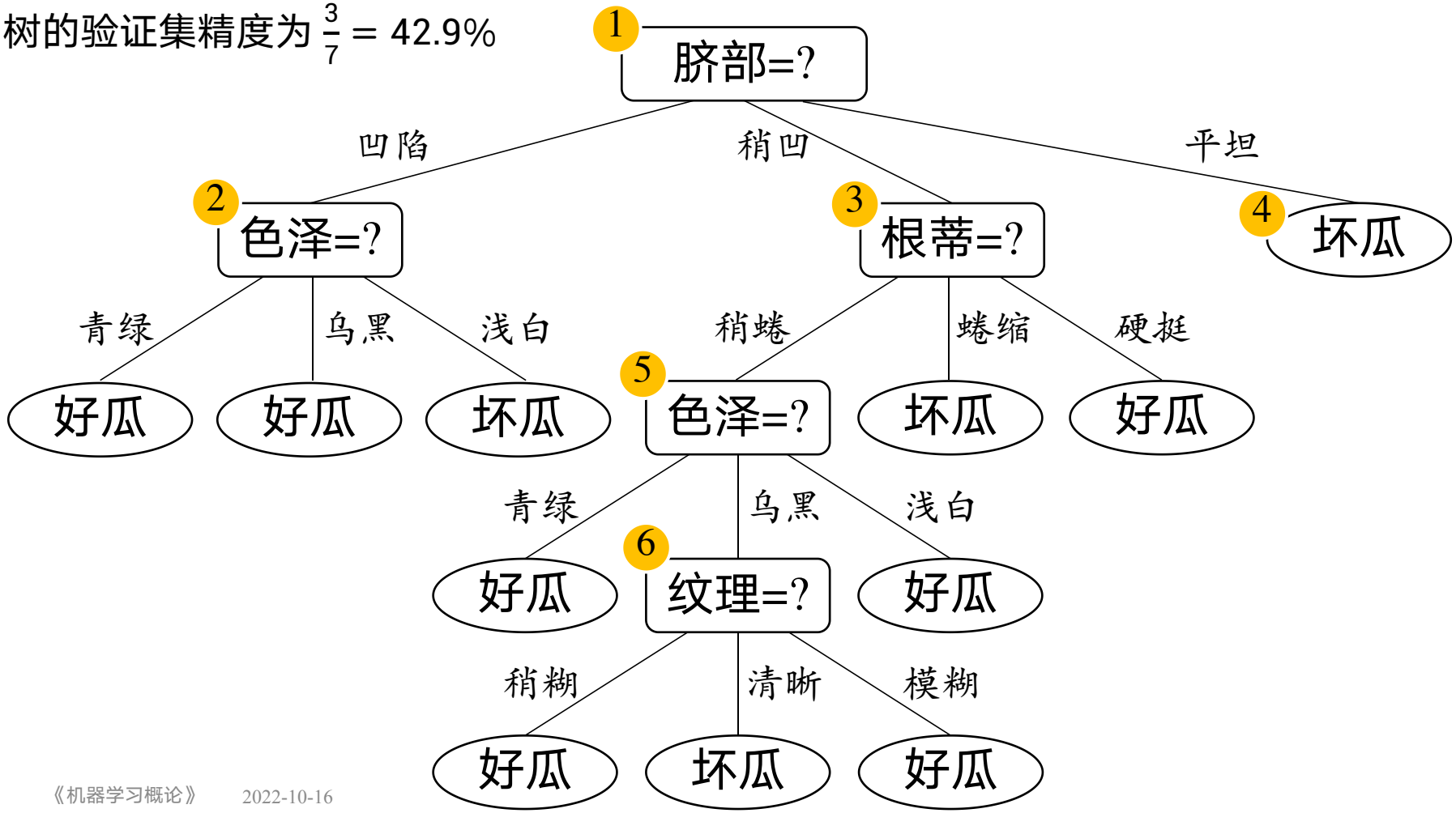


先从训练集生成一棵完整的决策树，然后自底向上地对非叶结点进行考察，若将该结点对应的子树替换为叶结点能带来决策树泛化性能提升，则将该子树替换为叶结点

剪枝处理—后剪枝

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

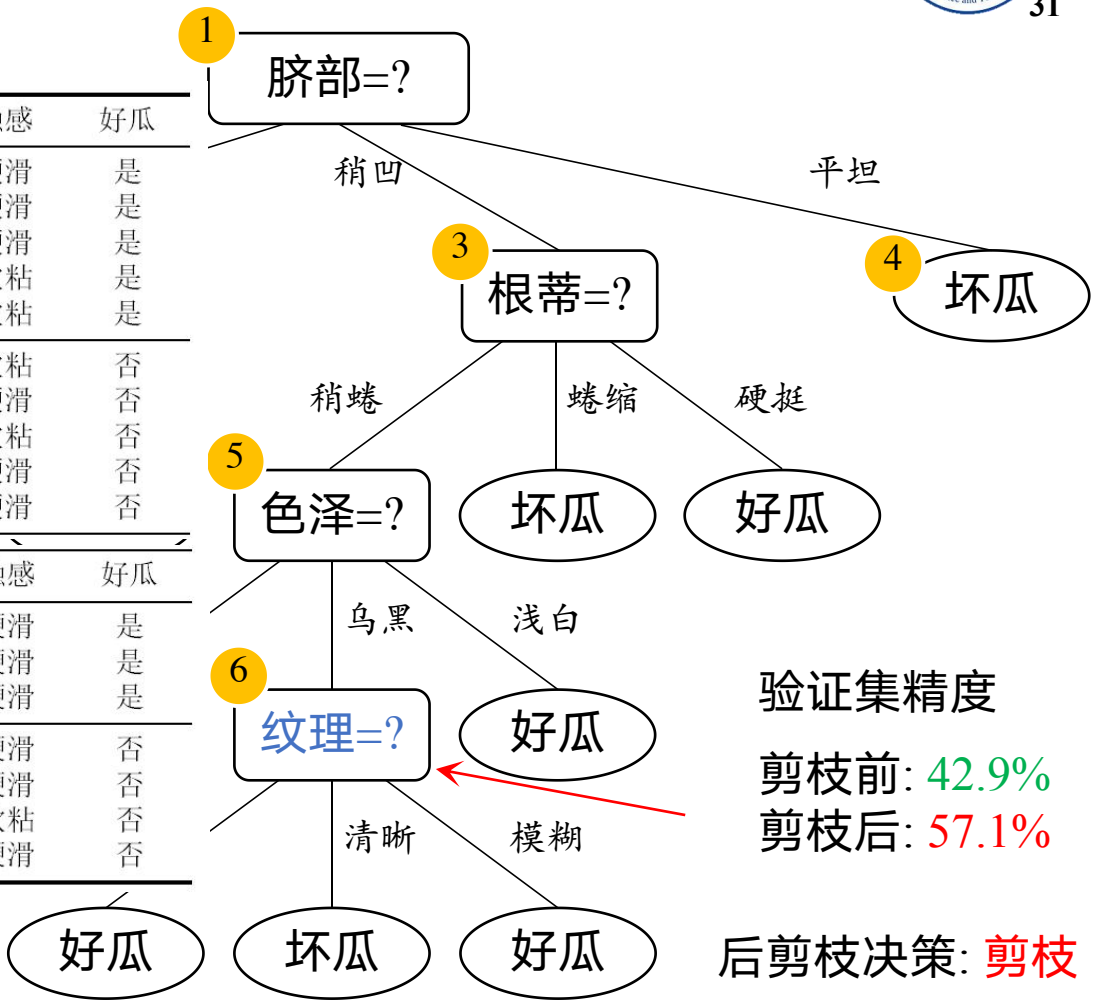
首先生成一棵完整的决策树，验证集{4,11,12 }判断正确，该决策树的验证集精度为 $\frac{3}{7} = 42.9\%$



剪枝处理—后剪枝

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



验证集精度

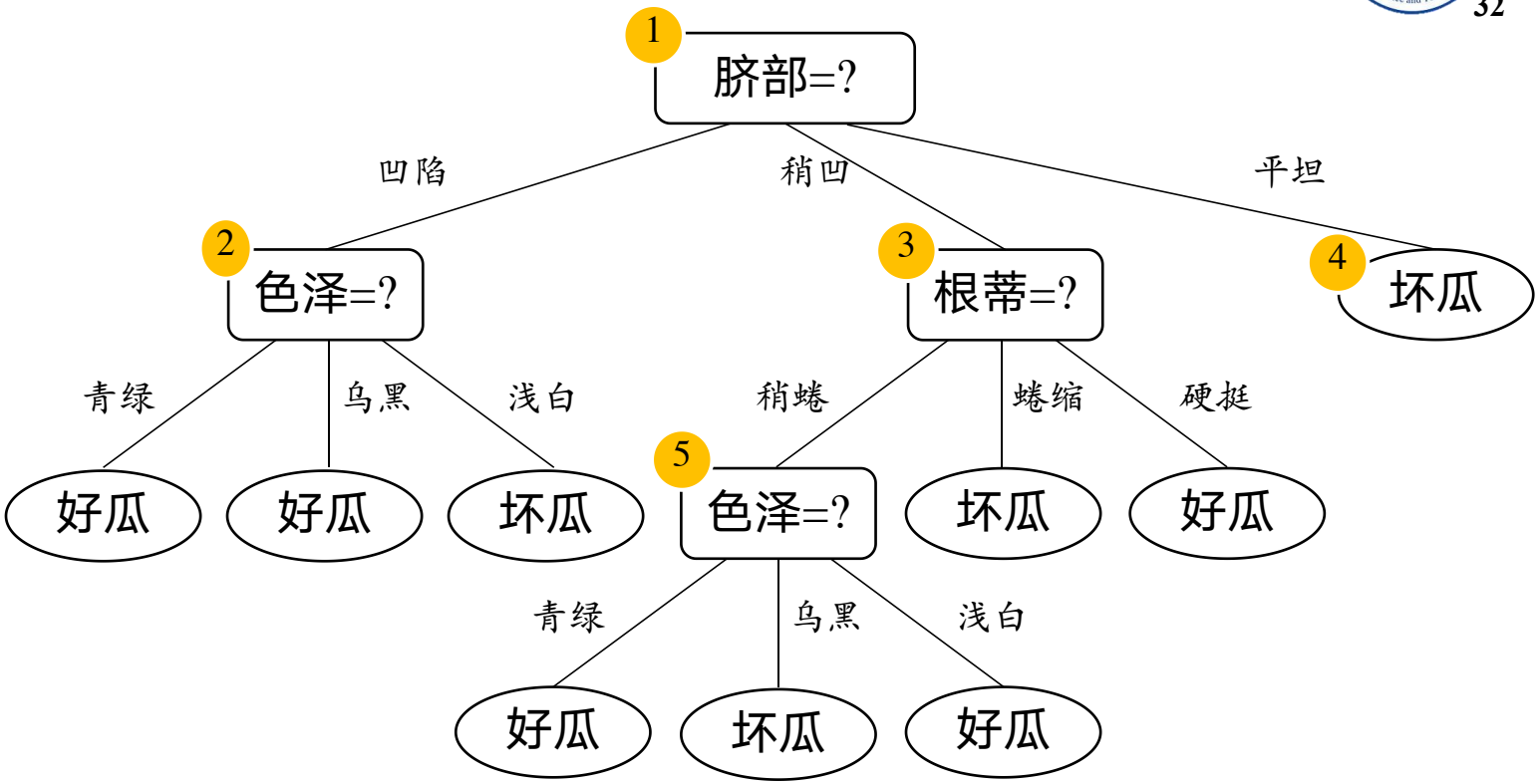
剪枝前: 42.9%

剪枝后: 57.1%

后剪枝决策: 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本{7,15}将其标记为“好瓜”。 验证集{4,8,11,12}判断正确，精度提高至57.1% 。

剪枝处理—后剪枝

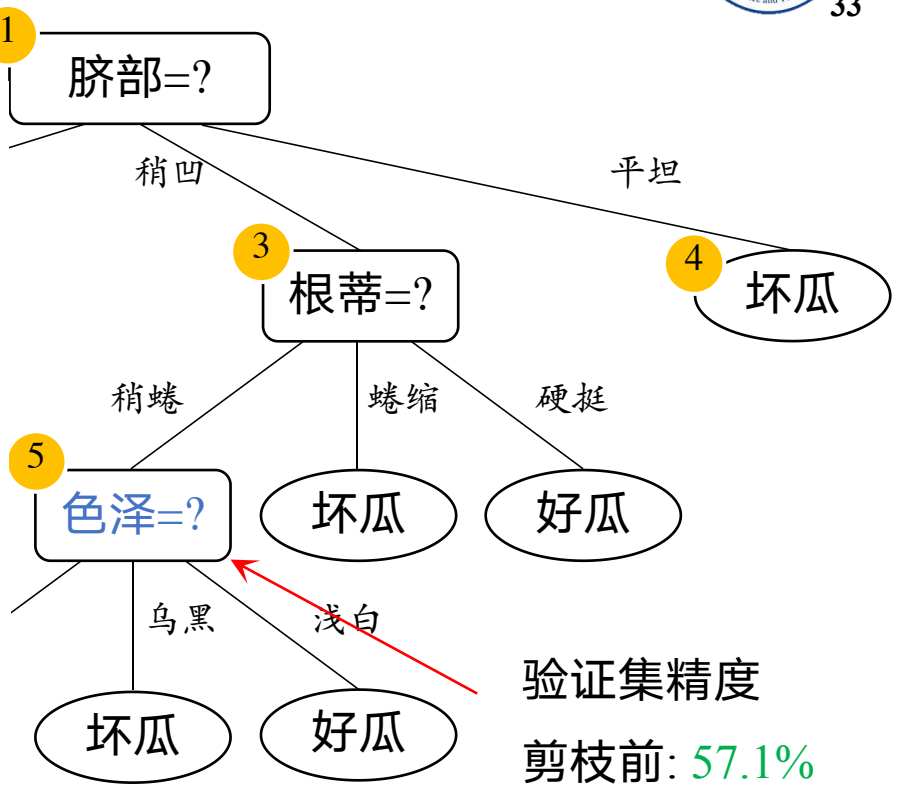


首先考虑结点 6，若将其替换为叶结点，根据落在其上的训练样本{7,15}将其标记为“好瓜”。 验证集{4,8,11,12}判断正确，精度提高至57.1% 。

剪枝处理—后剪枝

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



验证集精度

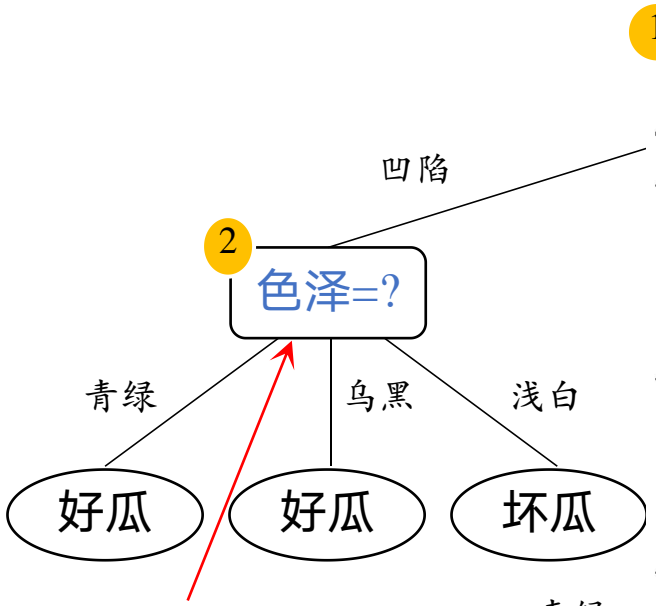
剪枝前: 57.1%

剪枝后: 57.1%

后剪枝决策: 不剪枝

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本{6,7,15}将其标记为“好瓜”。验证集{4,8,11,12}判断正确，精度仍然为57.1% 。

剪枝处理—后剪枝



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

验证集精度

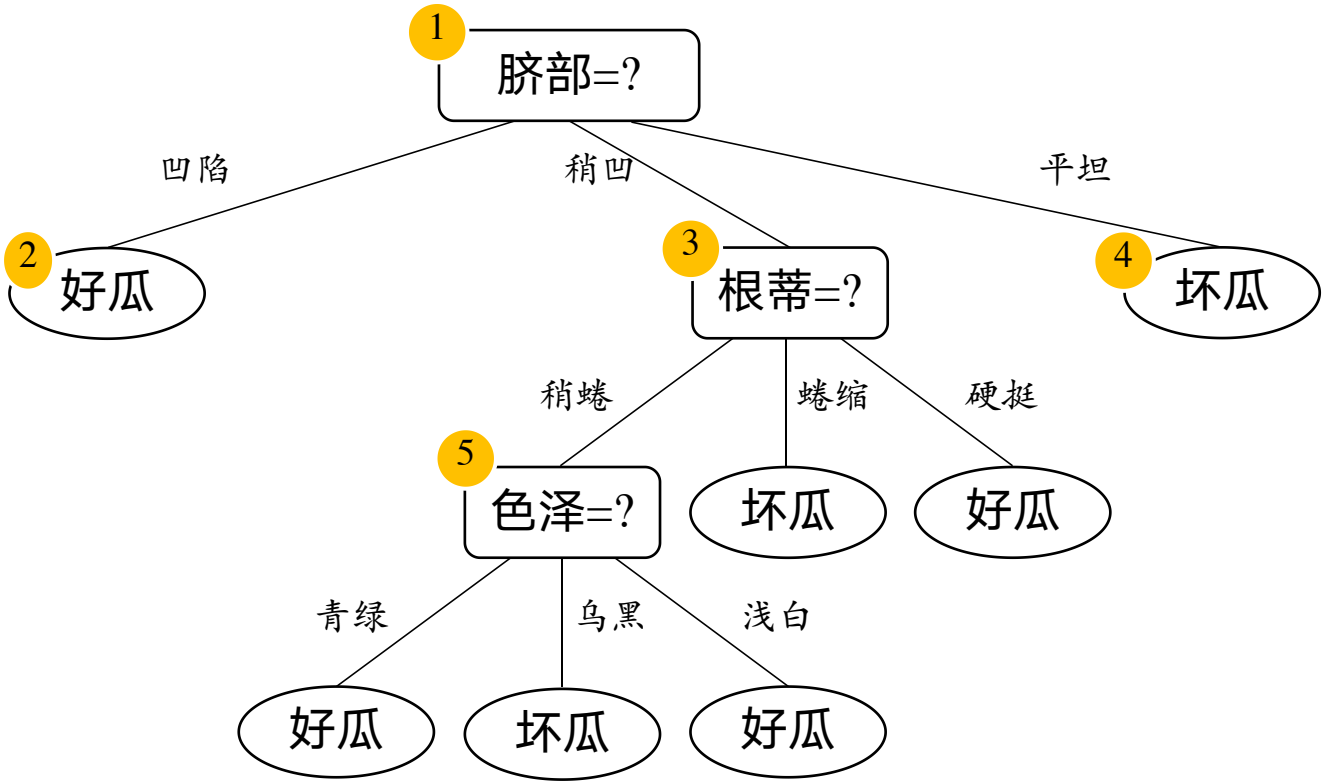
剪枝前: 57.1%

剪枝后: 71.4%

后剪枝决策: 剪枝

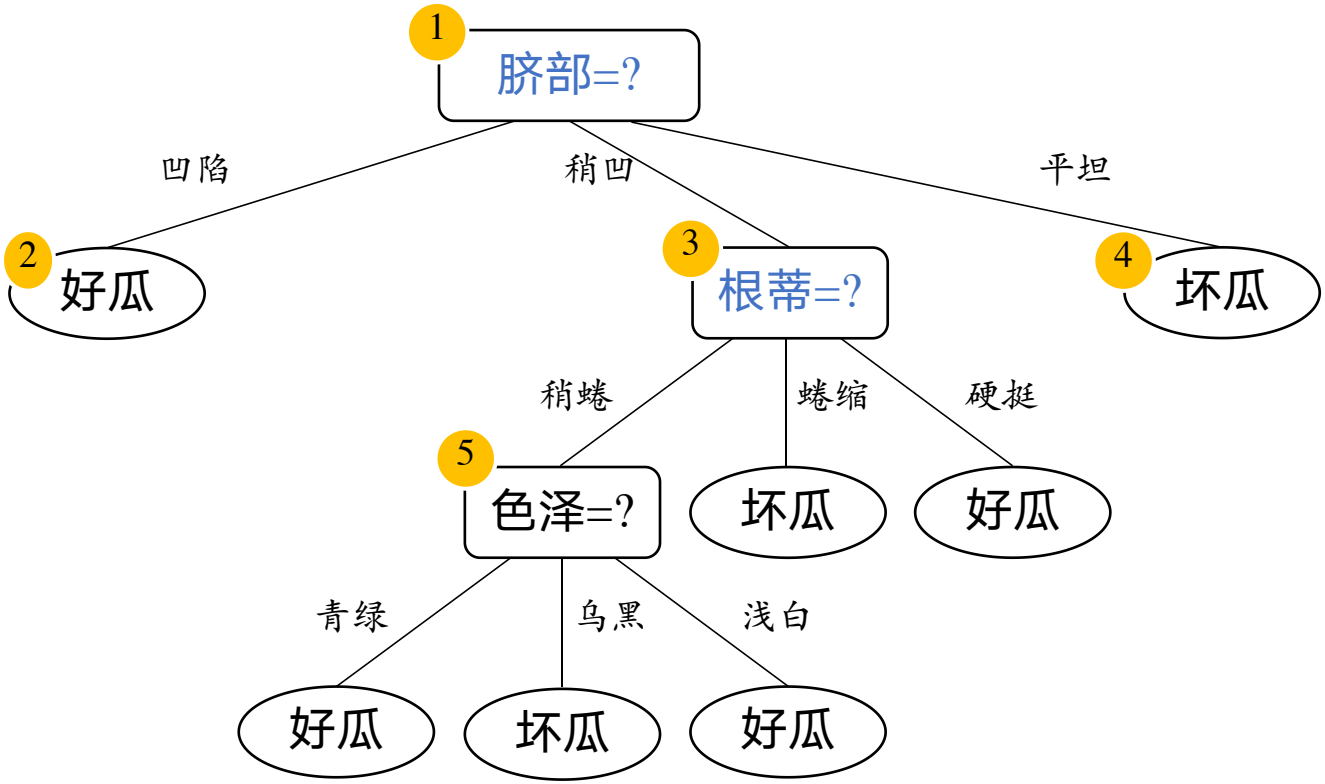
然后考虑结点 2 ，若将其替换为叶结点，根据落在其上的训练样本{1,2,3,14}将其标记为“好瓜”。验证集{4,5,8,11,12}判断正确，精度提高至71.4%。

剪枝处理—后剪枝



然后考虑结点 2 ，若将其替换为叶结点，根据落在其上的训练样本{1,2,3,14}将其标记为“好瓜”。验证集{4,5,8,11,12}判断正确，精度提高至71.4% 。

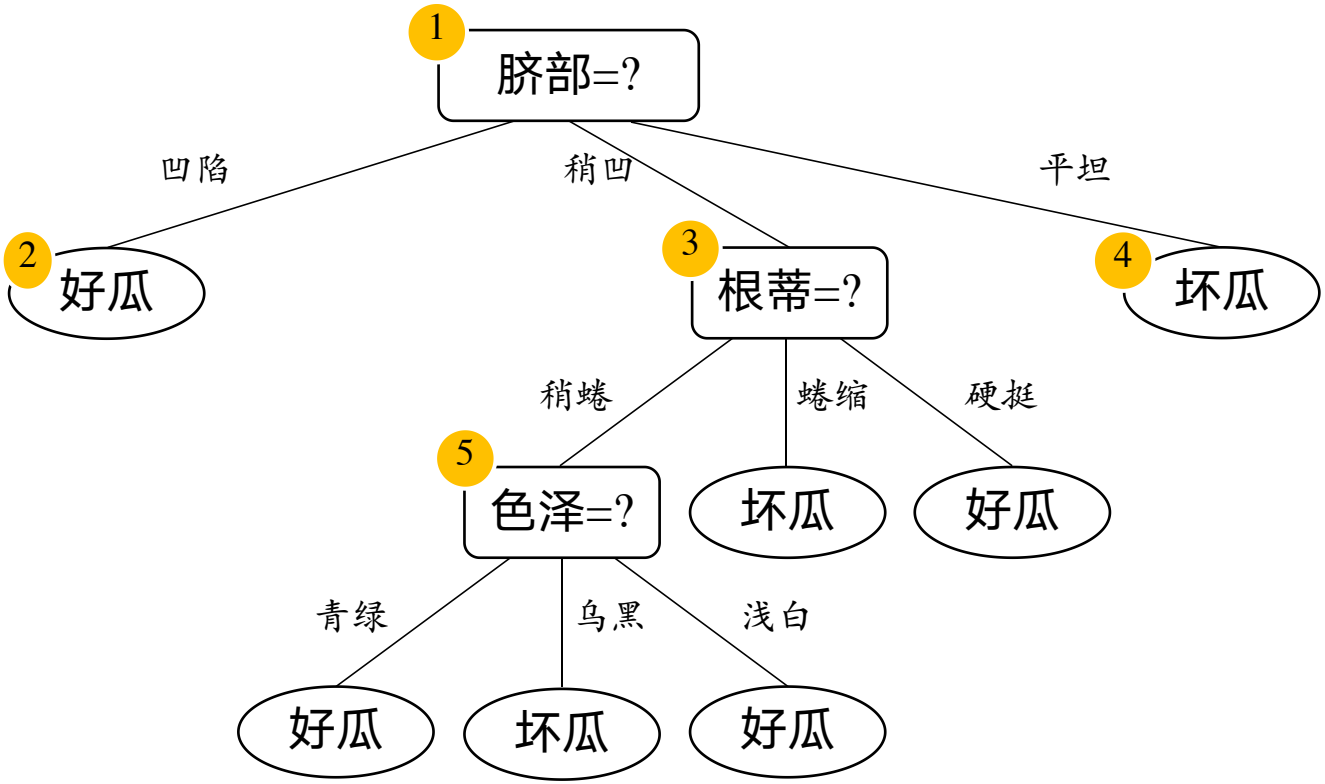
剪枝处理—后剪枝



考虑结点 1 3，先后替换为叶结点，验证集精度均未提升，则分支得到保留

剪枝处理—后剪枝

最终基于后剪枝策略得到的决策树如图所示



剪枝处理—后剪枝

优点

- 后剪枝比预剪枝保留了更多的分支，欠拟合风险小，泛化性能往往优于预剪枝决策树

缺点

- 训练时间开销大：后剪枝过程是在生成完全决策树之后进行的，需要自底向上对所有非叶结点逐一考察

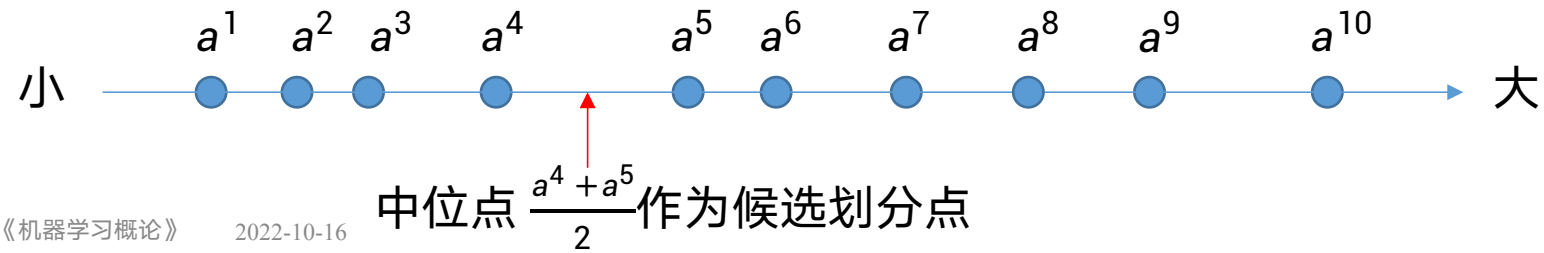
连续与缺失值—连续值处理

- 连续属性离散化(二分法)

第一步：
- 假定连续属性a在样本集 D上出现 n个不同的取值，从小到大排列，记为 $a^1, a^2, \cdots, a^n$ 。
- 基于划分点 t，可将 D 分为子集D_t^- 和D_t^+，
- D_t^-包含那些在属性a上取值不大于t的样本
- D_t^+包含那些在属性a上取值大于t的样本

考虑包含$n-1$ 个元素的候选划分点集合

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



连续与缺失值—连续值处理

- 连续属性离散化(二分法)

第二步：
- 采用离散属性值方法，考察这些划分点，选取最优的划分点进行样本集合的划分

$$\text{Gain}(D, a) = \max_{t \in T_a} \text{Gain}(D, a, t)$$

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

其中$\text{Gain}(D, a, t)$是样本集D基于划分点t二分后的信息增益
于是可选择使 $\text{Gain}(D, a, t)$ 最大化的划分点

连续与缺失值—连续值处理

- 连续值处理实例

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17个候选值：

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

属性“密度”划分点为0.381

对属性“含糖量”进行同样处理

属性“含糖量”划分点为0.126，信息增益为0.349

$$\mathrm{Gain}(D,a,0.244)=0.998-\frac{16}{17}\times\left(-\frac{8}{16}\log_2\frac{8}{16}-\frac{8}{16}\log_2\frac{8}{16}\right)=0.0568$$

$$\vdots$$

$$\mathrm{Gain}(D,a,0.381)=0.998-\frac{13}{17}\times\left(-\frac{8}{13}\log_2\frac{8}{13}-\frac{5}{13}\log_2\frac{5}{13}\right)=0.2629$$

$$\vdots$$

连续与缺失值—连续值处理

- 连续值处理实例

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

$\text{Gain}(D, \text{色泽}) = 0.109$

$\text{Gain}(D, \text{根蒂}) = 0.143$

$\text{Gain}(D, \text{敲声}) = 0.141$

$\text{Gain}(D, \text{纹理}) = 0.381$

$\text{Gain}(D, \text{脐部}) = 0.289$

$\text{Gain}(D, \text{触感}) = 0.006$

$\text{Gain}(D, \text{密度}) = 0.262$

$\text{Gain}(D, \text{含糖率}) = 0.349$

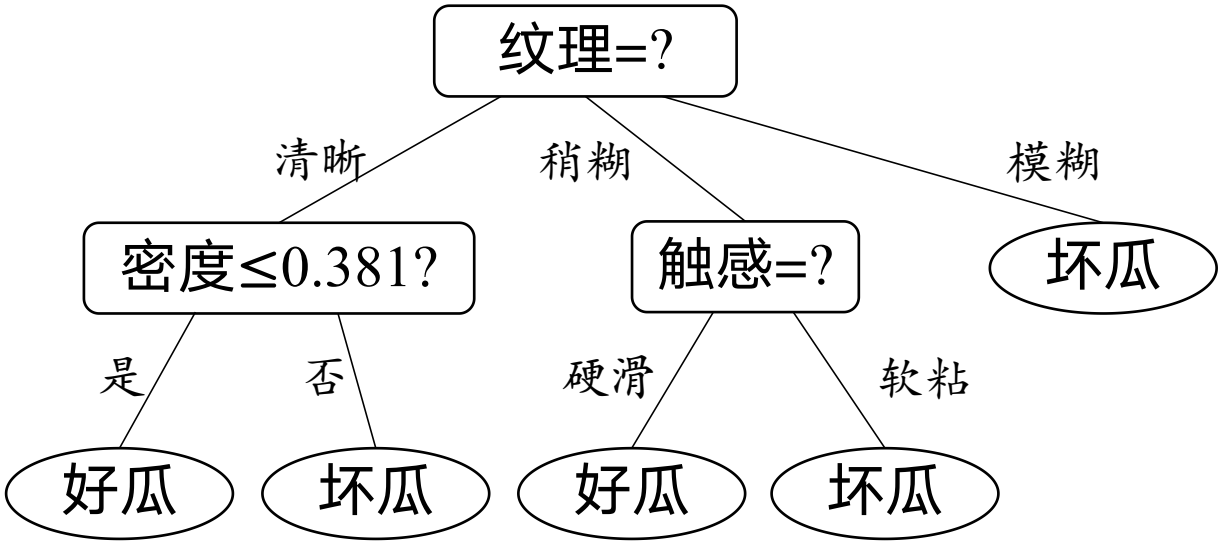
连续与缺失值—连续值处理

- 连续值处理实例

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						
4	青绿	蜷缩	沉闷						
5	浅白	蜷缩	浊响						
6	青绿	稍蜷	浊响						
7	乌黑	稍蜷	浊响						
8	乌黑	稍蜷	浊响						
9	乌黑	稍蜷	沉闷						
10	青绿	硬挺	清脆						
11	浅白	硬挺	清脆						
12	浅白	蜷缩	浊响						
13	青绿	稍蜷	浊响						
14	浅白	稍蜷	沉闷						
15	乌黑	稍蜷	浊响						
16	浅白	蜷缩	浊响						
17	青绿	蜷缩	沉闷						

$\text{Gain}(D, \text{色泽}) = 0.109$

$\text{Gain}(D, \text{根蒂}) = 0.143$



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

连续与缺失值—缺失值处理

- 不完整样本，即样本的属性值缺失
- 仅使用无缺失的样本进行学习?

对数据信息极大的浪费

- 使用有缺失值的样本，需要解决哪些问题？

Q1：如何在属性缺失的情况下进行划分属性选择？

Q2：给定划分属性,若样本在该属性上的值缺失，如何对样本进行划分？

连续与缺失值—缺失值处理

Q1：如何在属性缺失的情况下进行划分属性选择？

- $\tilde{D}$ 表示D中在属性a上没有缺失值的样本子集，$\tilde{D}^v$表示$\tilde{D}$中在属性a上取值为a^v的样本子集，$\tilde{D}_k$表示$\tilde{D}$中属于第k类的样本子集

$$\begin{aligned}\mathrm{Gain}(D,a) &= \rho\,\mathrm{Gain}(\tilde{D},a)\\ &= \rho\left(\mathrm{Ent}(\tilde{D}) - \sum_{v=1}^{V}\tilde{r}_v\,\mathrm{Ent}(\tilde{D}^v)\right)\end{aligned}$$

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

$$\mathrm{Ent}(\tilde{D}) = -\sum_i \tilde{p}_k \log_2 \tilde{p}_k$$

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

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

w_x为每个样本x的权重

连续与缺失值—缺失值处理

Q2：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？

- 若样本x在划分属性a上的取值已知，则将x划入与其取值对应的子结点，且样本权值在子结点中保持为w_x

- 若样本x在划分属性 a上的取值未知，则将x同时划入所有子结点，且样本权值在与属性值a^v对应的子结点中调整为 $r_v \times w_x$

  直观来看，相当于让同一个样本以不同概率划入不同的子结点中去

连续与缺失值—缺失值处理

- 缺失值处理实例

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

- 学习开始时，根结点包含样本集D中全部17个样例，各样例的权值均为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

$\tilde{D}$的信息熵 $\mathrm{Ent}(\tilde{D}) = -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985$

连续与缺失值—缺失值处理

- 缺失值处理实例

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

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

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

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

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

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

$\tilde{D}$上属性“色泽”的信息增益为

$$\mathrm{Gain}(\tilde{D}, \text{色泽}) = \mathrm{Ent}(\tilde{D}) - \sum_v \tilde{r}_v \mathrm{Ent}(\tilde{D}^v)$$

$$= 0.985 - \left(\frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0\right)$$

连续与缺失值—缺失值处理

- 类似地可计算出所有属性在数据集上的信息增益

Gain(D, 色泽) = 0.252	Gain(D, 敲声) = 0.145	Gain(D, 脐部) = 0.289
Gain(D, 根蒂) = 0.171	Gain(D, 纹理) = 0.424	Gain(D, 触感) = 0.006

进入“纹理=清晰”分支 7个样本

进入“纹理=稍糊”分支 5个样本

进入“纹理=模糊”分支 3个样本

样本权重在各子结点仍为1

在属性“纹理”上出现缺失值，样本8和10同时进入3个分支，调整8和10在3分分支权值分别为7/15，5/15，3/15

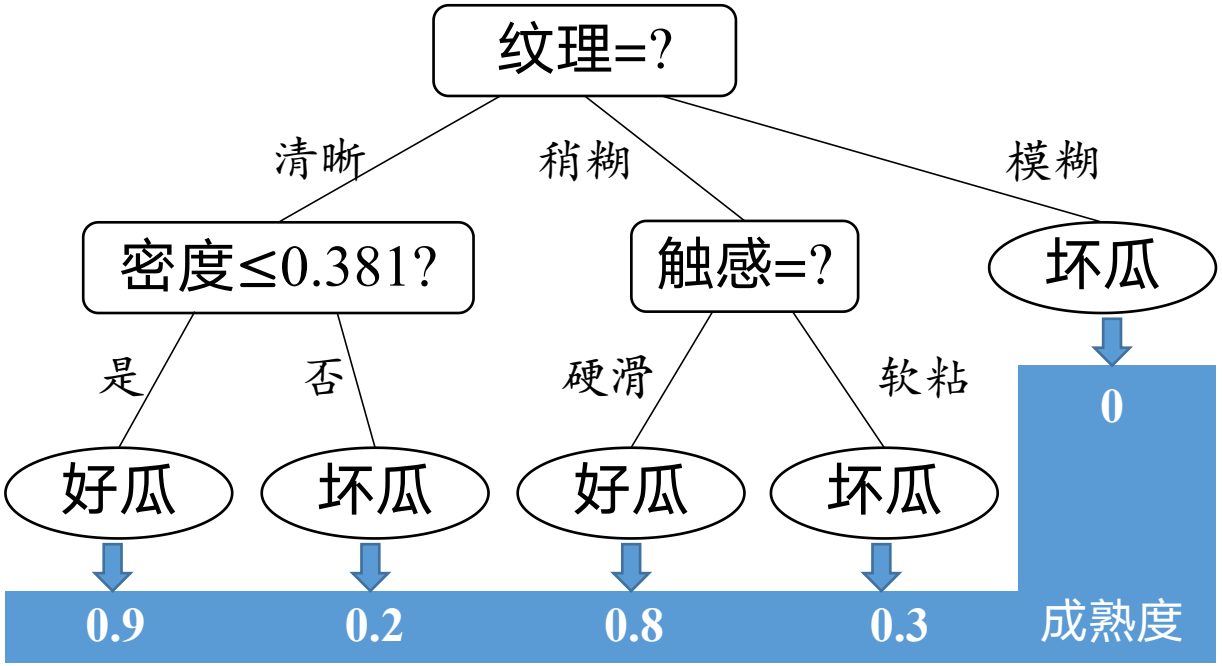
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	–	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	–	是
3	乌黑	蜷缩	–	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	–	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	–	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	–	稍凹	硬滑	是
9	乌黑	–	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	–	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	–	否
12	浅白	蜷缩	–	模糊	平坦	软粘	否
13	–	稍蜷	浊响	稍糊	凹陷	硬滑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

回归树

当标记为离散值时，叶子节点标记为数量最多的类别

若标记为连续值时，叶子节点应该如何标记？（均值？中位数？）

如何选择划分节点？



回归树

- 假设待划分节点中的数据为 $D = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$
- 离散属性a有V个可能的取值$\{a^1, a^2, \cdots, a^V\}$
- 若用a来进行划分，则会产生V个分支结点；第v个分支结点包含D中所有在属性a上取值为a^v的样本，记为D^v

- 若优化均方误差，

划分前误差 $\min_c \sum_i (y_i - c)^2 \Rightarrow c = \text{avg}(\{y | (\boldsymbol{x}, y) \in D\})$

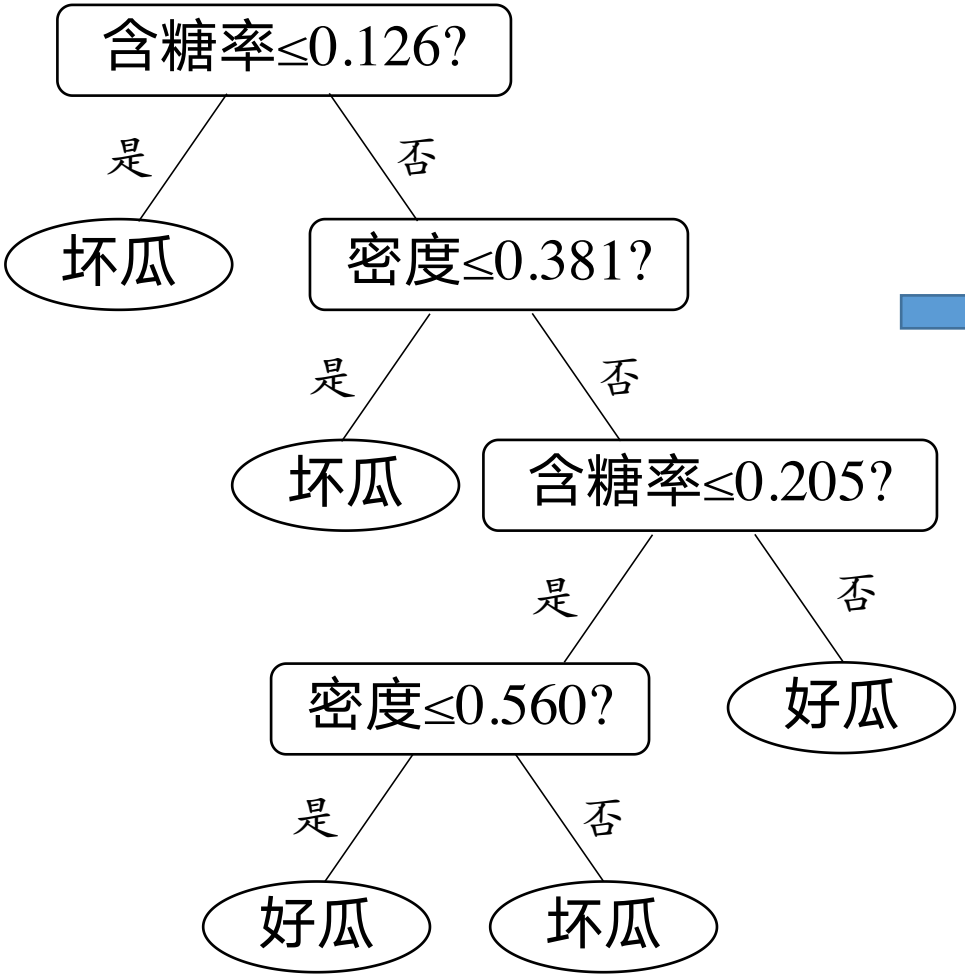
划分后误差 $\text{error}(\text{D}, \text{a}) = \sum_{v=1}^{V} \min_{c^v} \sum_{(\boldsymbol{x}_i, y_i) \in D^v} (y_i - c^v)^2 \Rightarrow c^v = \text{avg}(\{y | (\boldsymbol{x}, y) \in D^v\})$

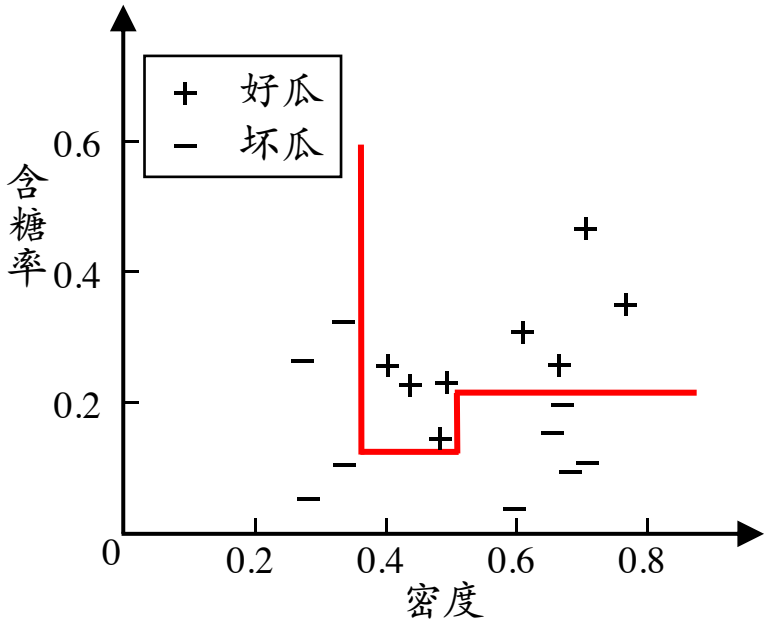
- 选择划分后误差最小的属性作为最优划分属性

$$a^\star = \arg\min_{a \in A} \text{error}(D, a)$$

多变量决策树

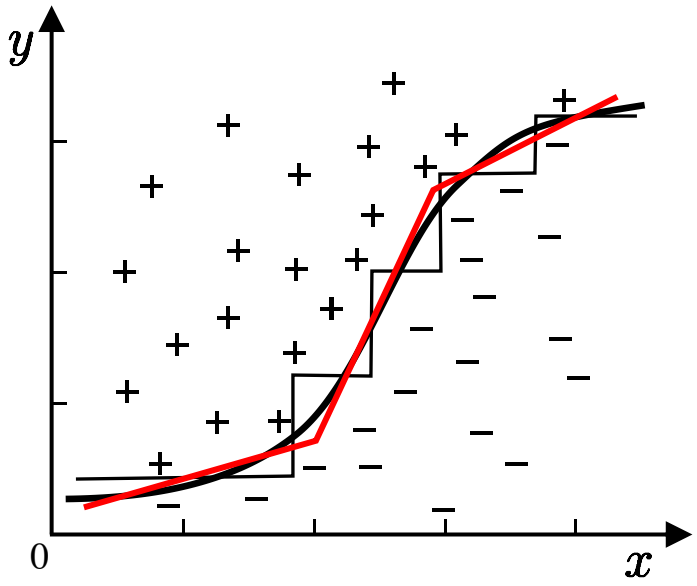
- 单变量决策树分类边界:轴平行





多变量决策树

- 单变量决策树分类边界:轴平行



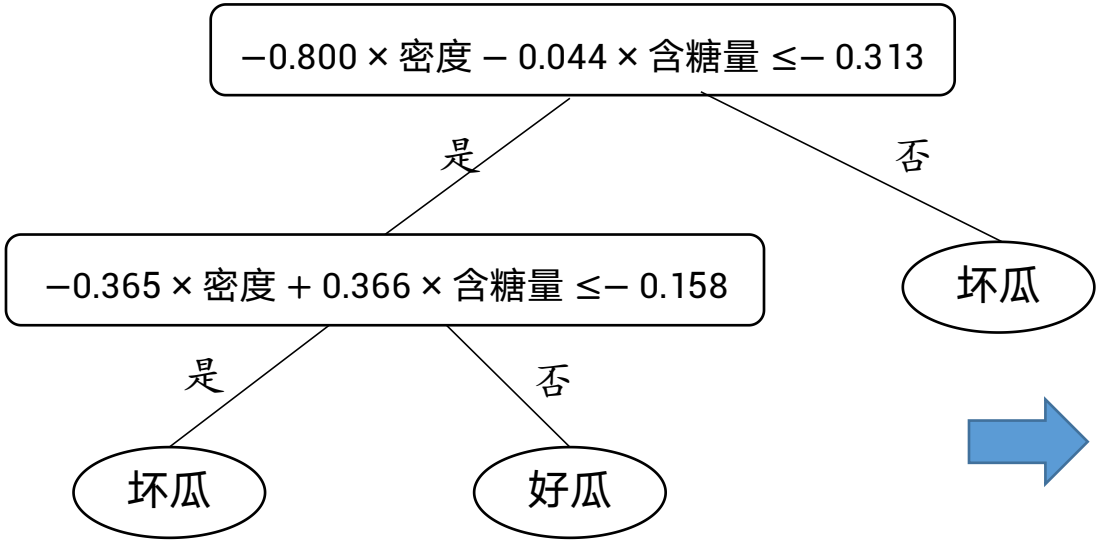
非叶节点不再是仅对某个属性，而是对属性的线性组合 进行测试

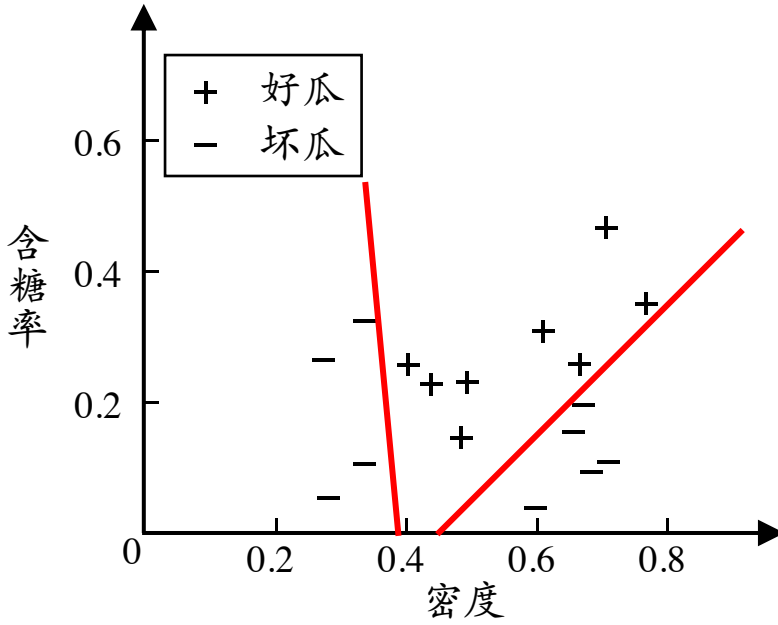
每个非叶结点是一个形如$\sum_{i=1}^{d} w_i a_i = t$的线性分类器，其中$w_i$是属性$a_i$的权值，$w_i$和$t$可在该结点所含的样本集和属性集上学得

- 多变量决策树

多变量决策树

- 多变量决策树





习题

- 4.1

- 4.9

- 假设离散随机变量$X \in \{1, \cdots, K\}$，其取值为k的概率$P(X = k) = p_k$，其熵为$H(\boldsymbol{p}) = -\sum_k p_k \log_2 p_k$，试用朗格朗日乘子法证明熵最大的分布为均匀分布

习题

- 下表表示的二分类数据集，具有三个属性A,B,C，样本标记为两类“+”，“-”。请运用你学过的知识完成如下问题：

实例	A	B	C	类别
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-
10	F	F	2.0	+

(a) 整个训练样本关于类属性的熵是多少
(b) 数据集中A，B两个属性的信息增益各是多少
(c) 对于属性C，计算所有可能划分的信息增益
(d) 根据Gini指数，A和B两个属性哪个是最优划分
(e) 采用算法C4.5，构造决策树