第6讲：决策树

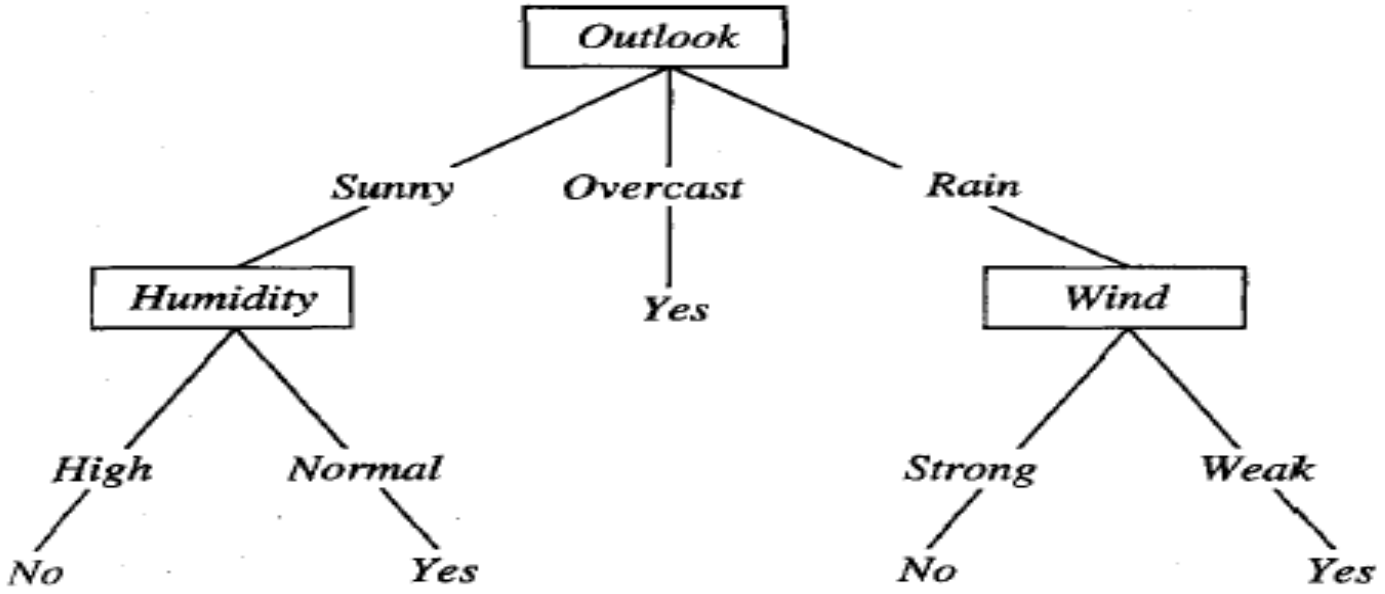
Lecture 6: Decision Tree

张永飞

2023年10月17日

引　例

- 引例1：天气是否适合打网球?



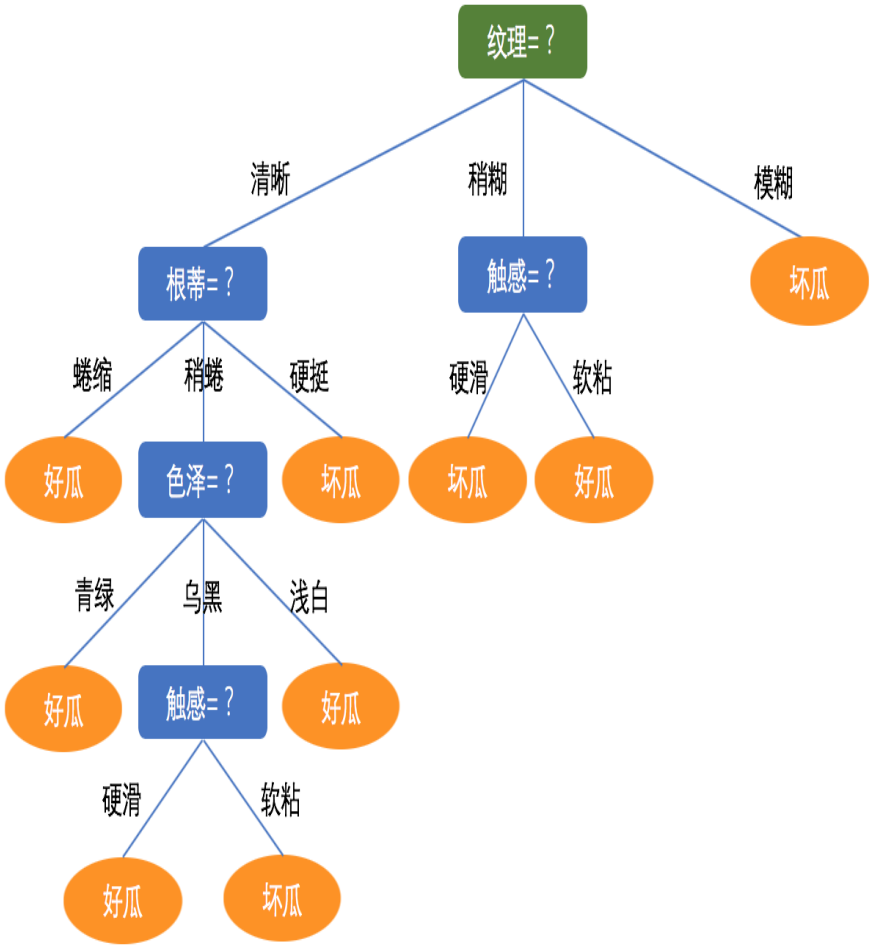
$$(Outlook = Sunny \wedge Humidity = Normal)$$
$$\vee \quad (Outlook = Overcast)$$
$$\vee \quad (Outlook = Rain \wedge Wind = Weak)$$

P38 机器学习， T.M. Mitchell著，曾华军译

引　例

- 引例2：西瓜分类

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	0.46	是
2	乌黑	-	沉闷	-	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	-	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	-	清脆	模糊	平坦	硬滑	0.245	-	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	-	清晰	稍凹	软粘	0.36	0.37	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

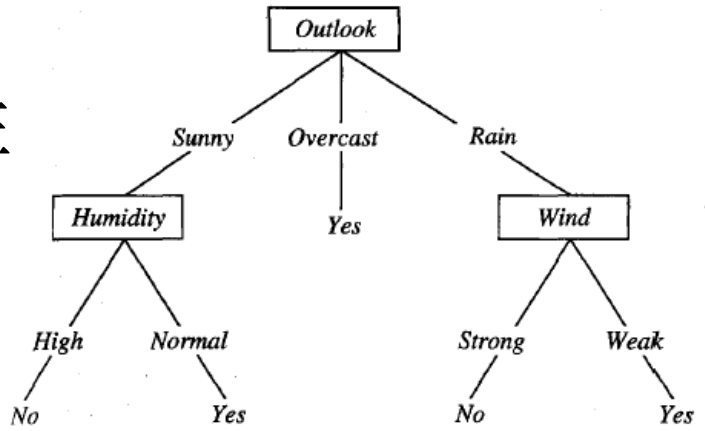


引　例

- **问题举例**
  - 根据症状或检查结果分类患者
  - 根据起因或现象分类设备故障
  - …
- **适用问题的特征**
  - 实例由“属性-值”对表示
  - 属性可以是连续值或离散值
  - 具有离散的输出值
  - 训练数据可以包含缺少属性值的实例
- **分类问题**
  - 核心任务是把样例分类到各可能的离散值对应的类别

决策树算法

- **基本思想：**采用自顶向下的递归方法，（以信息熵为度量）构造一棵（熵值下降最快的）树，（到叶子节点处的熵值为零）此时每个叶节点中的实例都属于同一类
- 决策树是一种树型结构，由结点和有向边组成
  - 节点
    - 内部结点表示一个属性或特征
    - 叶结点代表一种类别
  - 有向边/分支
  - 分支代表一个测试输出
- 可看成一个if-then的规则集合
  - 将特征空间划分为不相交的区域(region)



决策树算法

- **决策树分类：分为两步**
  - **第1步-决策树生成/学习、训练**：利用训练集建立(并精化)一棵决策树，建立决策树模型。这个过程实际上是一个从数据中获取知识，进行机器学习的过程
  - **第2步-分类/测试**：利用生成的决策树对输入数据进行分类。对输入的记录，从根结点依次测试记录的属性值，直到到达某个叶结点，从而找到该记录所在的类

决策树生成/学习算法

- step 1：选取一个属性作为决策树的根结点，然后就这个属性所有的取值创建树的分支
- step 2：用这棵树来对训练数据集进行分类：
  - 如果一个叶结点的所有实例都属于同一类，则以该类为标记标识此叶结点
  - 如果所有的叶结点都有类标记，则算法终止
- step 3：否则，选取一个从该结点到根路径中没有出现过的属性为标记标识该结点，然后就这个属性所有的取值继续创建树的分支；重复算法步骤step 2

主要算法

- 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

  示例-高考报志愿：好学校？好专业？好位置？...

  找对象：年龄？学历？兴趣爱好？身高？...

- **目标：**每个分支节点的样本尽可能属于同一类别，即节点的“纯度”(purity)越来越高；最具区分性的属性！
- 根据不同目标函数，建立决策树主要有以下三种算法
  - **ID3：　信息增益**
  - C4.5：　信息增益率
  - CART：基尼指数

ID3算法

- **ID3 (Iterative Dichotomiser 3)算法**
  - ID3算法是一种最经典的决策树学习算法，由Ross Quinlan于1975年提出
  - **基本思想**：以信息熵为度量，用于决策树节点的属性选择，每次优先选取信息增益最大的属性，亦即能使熵值变为最小的属性，以构造一颗熵值下降最快的决策树，到叶子节点处的熵值为0。此时，每个叶子节点对应的实例集中的实例属于同一类
  - 熵值下降 → 无序变有序

信息论基础

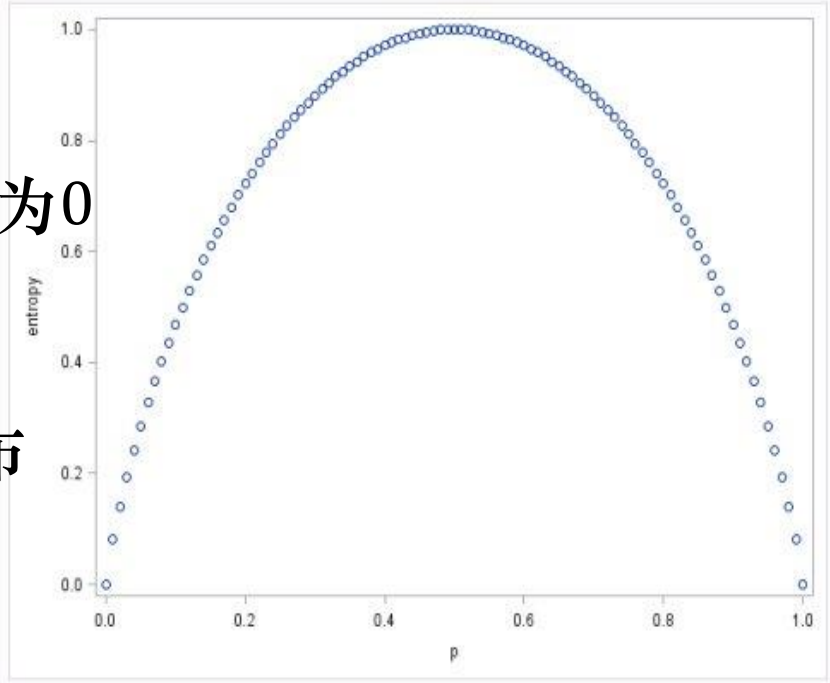
- 信息论与概率统计中，熵表示随机变量不确定性的大小，是度量样本集合纯度最常用的一种指标
- **信息量：** 具有确定概率事件的信息的定量度量
  - 定义：$I(x) = -\log_2 p(x)$

    其中 $p(x)$ 为事件x发生的概率
- **信息熵：** 事件集合的信息量的平均值
  - 定义：$H(x) = \sum_i h(x_i) = \sum_i p(x_i) I(x_i) = -\sum_i p(x_i) \log_2 p(x_i)$
  - 若 X 为连续随机变量，则概率分布变成概率密度函数，求和符号变成积分符号即可

信息论基础

- **信息熵**
  - 熵定义了一个函数(概率密度函数pdf)到一个值(信息熵)的映射　　$p(x) \rightarrow H$（函数→数值）
  - 熵是随机变量**不确定性**的度量：
    - 不确定性越大，熵值越大
    - 若随机变量退化成定值，熵为0

    示例：明天晴天？明天下雪？
  - 均匀分布是“最不确定”的分布



ID3算法

- 经验(信息)熵
  - 假设当前样本集合D中第c（c=1,2,...,C）类样本所占比例为p_c（c=1,2,...,C），则D的经验信息熵(简称经验熵)定义为

$$H(D) = -\sum_{c=1}^{C} p_c \log_2 p_c$$

$$= -\sum_{c=1}^{C} \frac{|D_c|}{|D|} \log_2 \frac{|D_c|}{|D|}$$

  - $H(D)$的值越小，则D的纯度越高

ID3算法

- **条件熵（conditional entropy）**
  - 对随机变量(*X*, *Y*)，联合分布为：$p(X = x_i, Y = y_i) = p_{ij}$
  - 条件熵 ***H(Y |X)*** 表示在已知随机变量X 的条件下，随机变量Y的不确定性：

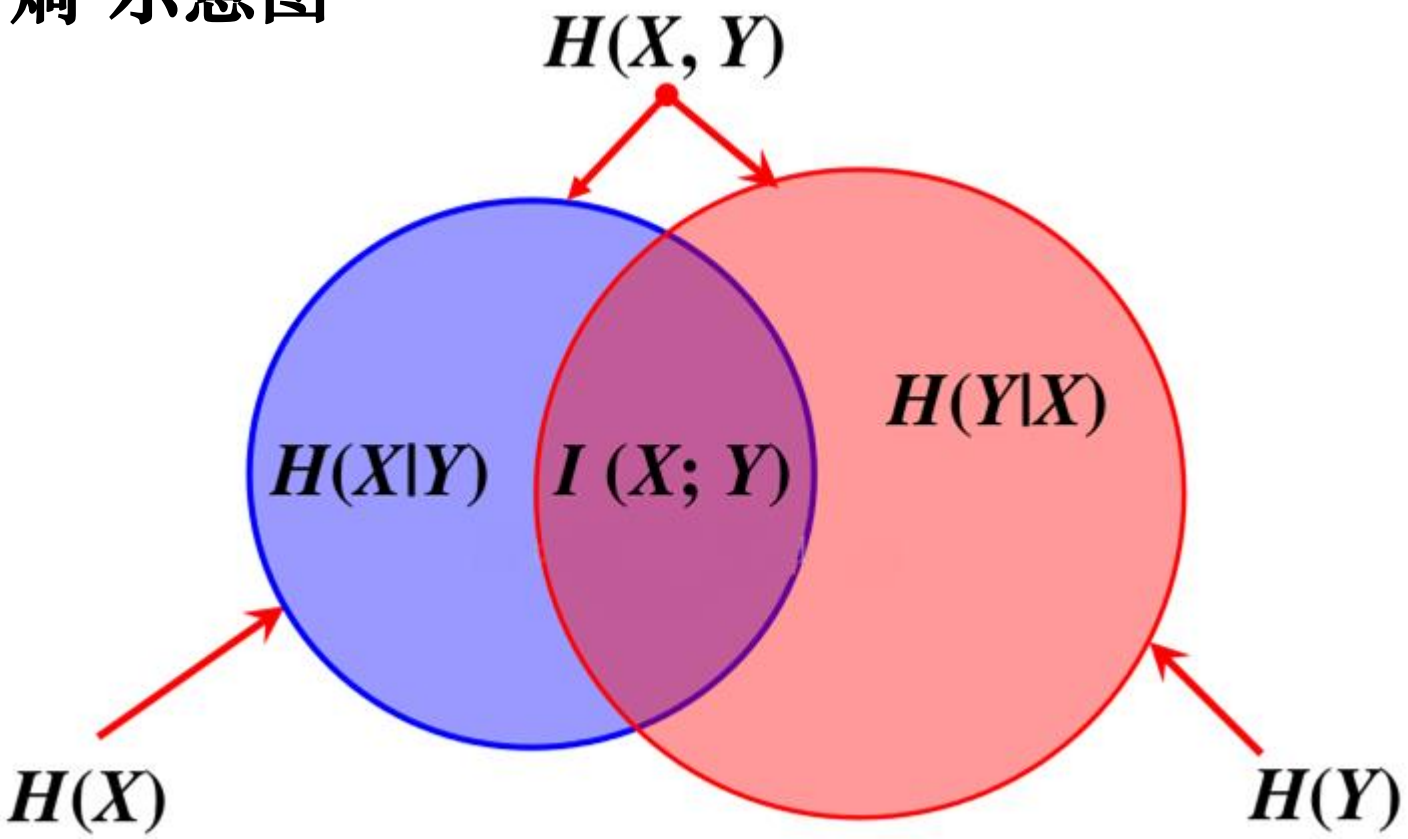
$$H(Y \mid X) = -\sum_{i=1}^{n} p_i H(Y \mid X = x_i)$$

  - 条件熵H(Y|X)相当于联合熵$H(X,Y)$减去单独的熵H(X)，即

$$H(Y \mid X) = H(X, Y) - H(X)$$

ID3算法

- 条件熵-示意图



Venn图

ID3算法

- **条件熵-推导**

$$H(Y \mid X) = H(X,Y) - H(X)$$

$$= -\sum_{x,y} p(x,y)\log_2\ p(x,y) + \sum_{x} p(x)\log_2\ p(x)$$

$$= -\sum_{x,y} p(x,y)\log_2\ p(x,y) + \sum_{y}(\sum_{x} p(x,y))\log_2\ p(x)$$

$$= -\sum_{x,y} p(x,y)\log_2\ p(x,y) + \sum_{x,y} p(x,y)\log_2\ p(x)$$

$$= -\sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)}$$

$$= -\sum_{x,y} p(x,y)\log_2\ p(y \mid x)$$

$$H(Y \mid X) = \sum_{x} p(x)H(Y \mid X = x)$$

$$= -\sum_{x} p(x)\sum_{y} p(y \mid x)\log_2\ p(y \mid x)$$

$$= -\sum_{x,y} p(x,y)\log_2\ p(y \mid x)$$

ID3算法

- **经验条件熵**
  - 假设当前样本集合D共有C类，每一类有D_c个样本，属性a $(a \in A)$有不同的取值$\{a_1, a_2, ..., a_N\}$，每一类中属性为i 的样本数为$D_c{}^n$，则D的经验条件熵定义为

$$H(D|a) = -\sum_{n,c} p(D_c, a_n) \log_2 p(D_c | a_n)$$

$$= -\sum_{n=1}^{N} \frac{|D^n|}{|D|} \sum_{c=1}^{C} \frac{|D_c^n|}{|D^n|} \log_2 \frac{|D_c^n|}{|D^n|}$$

$$= \sum_{n=1}^{N} \frac{|D^n|}{|D|} H(D^n)$$

  - 即特征a的信息对样本D的信息的不确定性减少的程度

ID3算法

- **信息增益（information gain）：**
  - 特征 ***a*** 对训练数据集 ***D*** 的信息增益 ***G(D, a)*** ，定义为集合D 的经验熵 ***H(D)*** 与特征 ***a*** 给定条件下 ***D*** 的经验条件熵 ***H(D | a)*** 之差，即

$$G(D,a) = H(D) - H(D \mid a)$$

$$= H(D) - \sum_{n=1}^{N} \frac{\left|D^n\right|}{\left|D\right|} H(D^n)$$

  - ID3算法即是以此信息增益为准则，对每次递归的节点属性进行选择的

ID3算法

- 决策树的生成算法：

输入：训练数据集D，特征集A，阈值ε

输出：决策树T

（1）若D中所有实例属于同一类C_k，则T为单结点树，并将类C_k作为该结点的类标记，返回T;

（2）若$A=\emptyset$，则T为单结点树，并将D中实例数最大的类C_k作为该结点类标记，返回T;

（3）否则，计算A中各特征对D的信息增益，选择信息增益最大的特征A_g

（4）如果A_g的信息增益小于阈值ε，则置T为单结点树，并将D中样本数最大的类C_k作为该结点的类标记，返回T

（5）否则，对A_g的每一个可能值a_i，依$A_g=a_i$，奖D分割为若干非空子集D_i，将D_i中实例数最大的类作为标记，构建子结点，由结点及其子结点构成树T，返回T

（6）对第i个子结点，以D_i为训练集，以$A-\{A_g\}$为特征集，递归的调用第（1）~（5）步，得到子树T_i，返回T_i。

ID3算法-示例

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

ID3算法-示例

- 计算信息熵-以属性色泽为例

$$H(D)=-\sum_{c=1}^{C}p_c\log_2 p_c=-\left(\frac{8}{17}\log_2\frac{8}{17}+\frac{9}{17}\log_2\frac{9}{17}\right)=0.998$$

编号	色泽	好瓜
1	青绿	是
2	乌黑	是
3	乌黑	是
4	青绿	是
5	浅白	是
6	青绿	是
7	乌黑	是
8	乌黑	是
9	乌黑	否
10	青绿	否
11	浅白	否
12	浅白	否
13	青绿	否
14	浅白	否
15	乌黑	否
16	浅白	否
17	青绿	否

$$H\left(D^{\text{青绿}}\right)=-\left(\frac{3}{6}\log_2\frac{3}{6}+\frac{3}{6}\log_2\frac{3}{6}\right)=1.000$$

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

$$H\left(D^{\text{浅白}}\right)=-\left(\frac{1}{5}\log_2\frac{1}{5}+\frac{4}{5}\log_2\frac{4}{5}\right)=0.722$$

■参照教材《机器学习》-周志华P75-77示例

ID3算法-示例

- 计算信息增益-以属性色泽为例

$$G(D,\text{色泽})=H(D)-\sum_{n=1}^{3}\frac{|D^n|}{|D|}H(D^n)$$

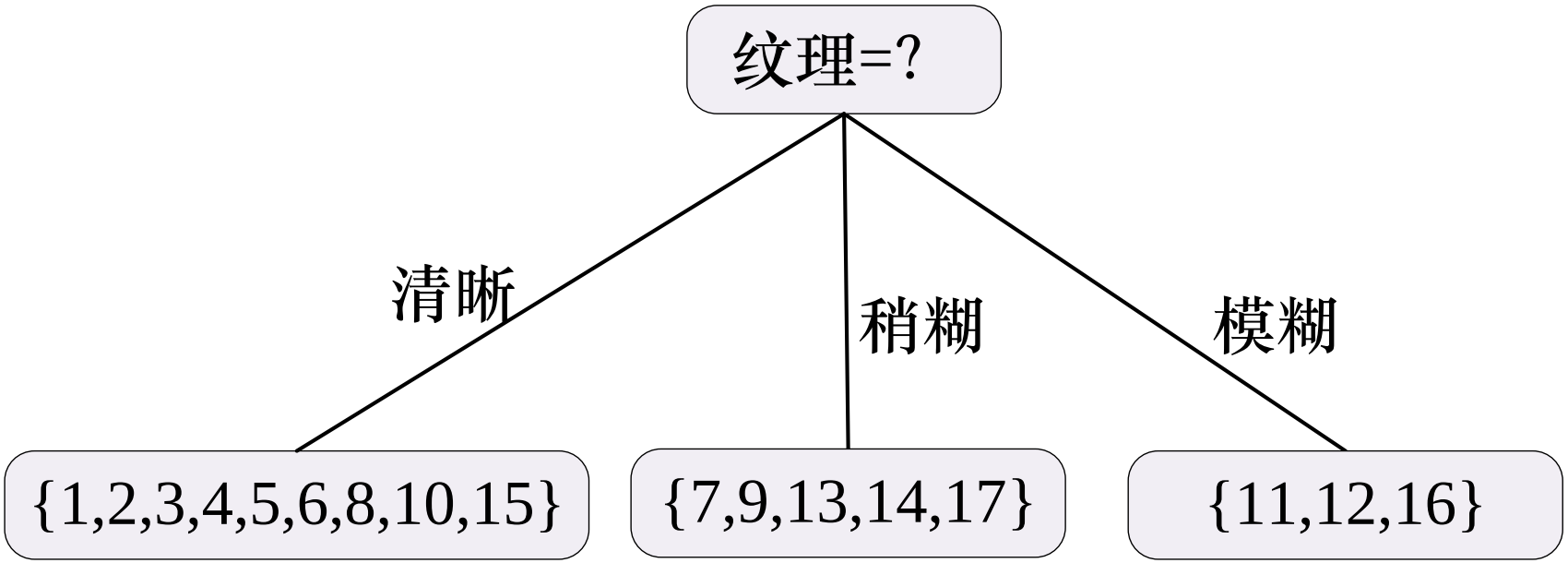
$$=0.998-\left(\frac{6}{17}\times 1.000+\frac{6}{17}\times 0.918+\frac{5}{17}\times 0.722\right)$$

$$=0.109$$

- $G(D,\text{色泽})=0.109$
- $G(D,\text{敲声})=0.141$
- $G(D,\text{脐部})=0.289$
- $G(D,\text{根蒂})=0.143$
- $G(D,\text{纹理})=0.381$
- $G(D,\text{触感})=0.006$

ID3算法-示例

- 基于属性“纹理”对根节点进行划分



ID3算法-示例

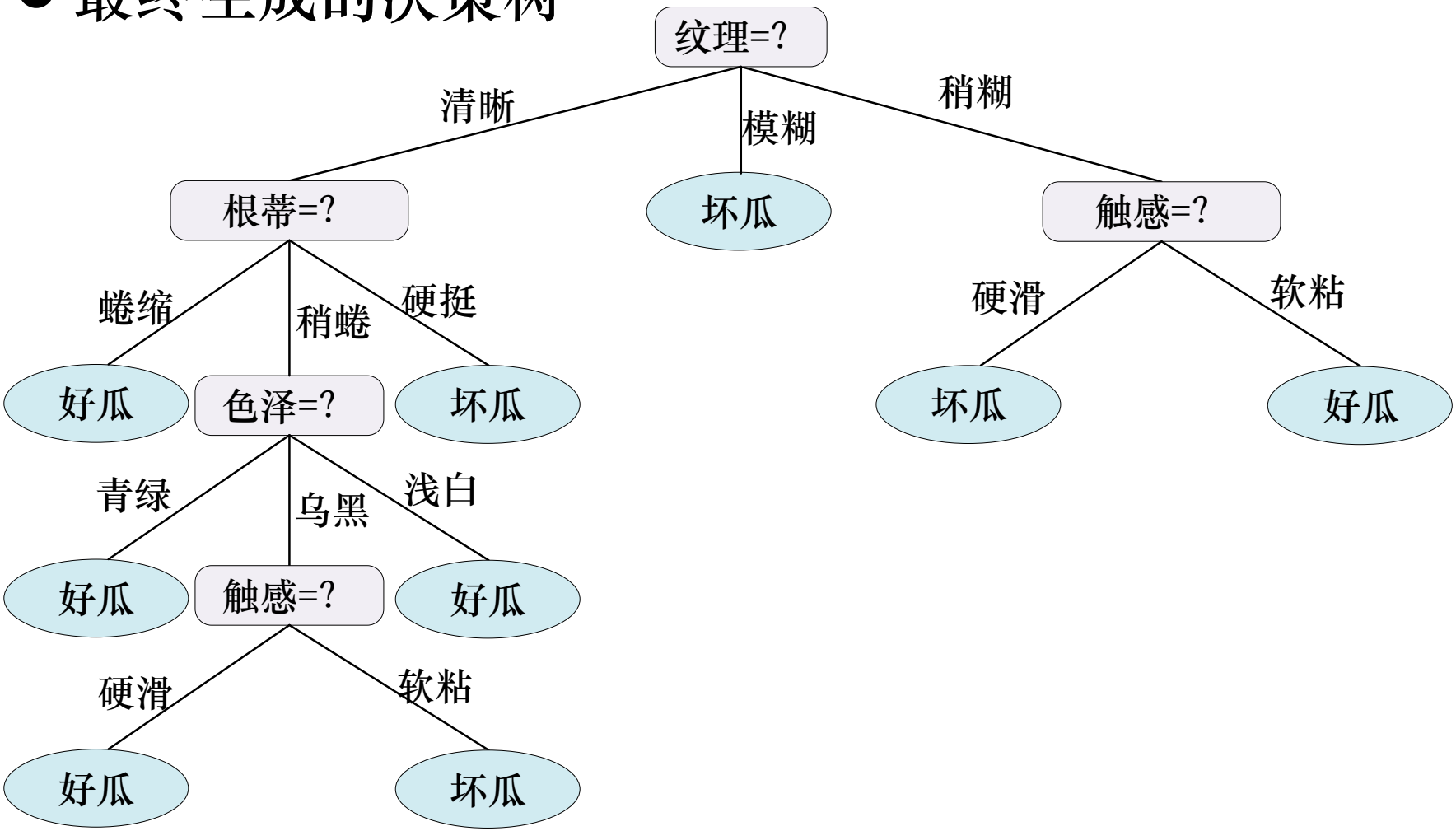
- 继续进行划分-以“纹理=清晰”分支为例
  - “纹理=清晰”分支:

    样本 {1,2,3,4,5,6,8,10,15}

  - 计算信息增益
    - $G(D^{清晰}, 色泽) = 0.043$
    - $G(D^{清晰}, 敲声) = 0.331$
    - $G(D^{清晰}, 触感) = 0.458$
    - $G(D^{清晰}, 根蒂) = 0.458$
    - $G(D^{清晰}, 脐部) = 0.458$

ID3算法-示例

- 最终生成的决策树



算法特点

- 最大优点是，它可以自学习：在学习的过程中，不需要使用者了解过多背景知识，只需要对训练实例进行较好的标注，就能够进行学习
- 决策树的分类模型是树状结构，简单直观，比较符合人类的理解方式
- 可将决策树中到达每个叶节点的路径转换为IF—THEN形式的分类规则，这种形式更有利于理解
- 从一类无序、无规则的事物(概念)中推理出决策树表示的分类规则
- 显然，属于有监督学习

ID3算法的问题

- 信息增益偏好取值多的属性(分散，极限趋近于均匀分布)
  - **属性筛选度量标准**
- 可能会受噪声或小样本影响，易出现过拟合问题
  - **剪枝处理**
- 无法处理连续值的属性
  - **连续值处理**
- 无法处理属性值不完整的训练数据
  - **缺失值处理**
- 无法处理不同代价的属性
  - **不同代价属性的处理**
- **针对这些问题的改进，ID3被扩展成C4.5(同时处理前4个问题)等算法**

属性筛选度量标准

- **信息增益的问题：**

$$G(D,a) = H(D) - \sum_{n=1}^{N} \frac{|D^n|}{|D|} H(D^n)$$

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

  - 信息增益准则对可取值数目 N 较多的属性有所偏好
  - 取值更多的属性更容易使得数据更“纯”，其信息增益更大，决策树会首先挑选这个属性作为树的顶/节点
  - 以序号为划分属性，信息增益最大!
  - 结果训练出来的形状是一棵庞大且深度很浅的树，这样的划分是极为不合理的

属性筛选度量标准

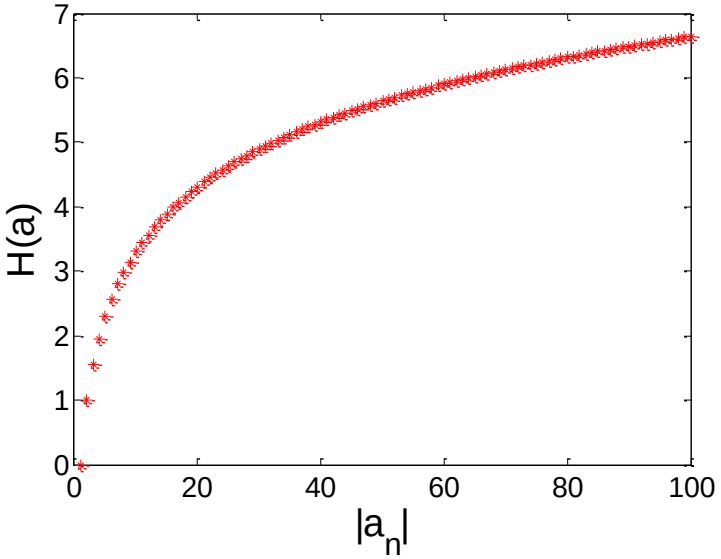
- **增益率（gain ratio）：**

$$G_{ratio}(D,a)=\frac{G(D,a)}{H(a)}$$

其中

$$H(a)=-\sum_{n=1}^{N}\frac{|D_n|}{|D|}\log_2\frac{|D_n|}{|D|}$$

称为属性a的固有值



- N 越大，$H(a)$ 通常也越大；因此，采用增益率，可缓解信息增益准则对可取值数目较多的属性的偏好

- **C4.5算法[Quinlan 1993]就基于信息增益（初选）和增益率（精选）替代了ID3 算法的信息增益**

属性筛选度量标准

- **基尼指数（Gini index）：**

$$Gini(D)=\sum_{c=1}^{C}\sum_{c'\neq c}p_c p_{c'}=1-\sum_{c=1}^{C}p_c^2=1-\sum_{c=1}^{C}\left(\frac{|D_c|}{|D|}\right)^2$$

  - 直观反映了从数据集中随机抽取两个样本，其类别不一致的概率；因此，Gini(D)越小，则数据集D的纯度越高

- 属性A的基尼指数：

$$Gini(D,a)=\sum_{n=1}^{N}\frac{|D^n|}{|D|}Gini(D^n)$$

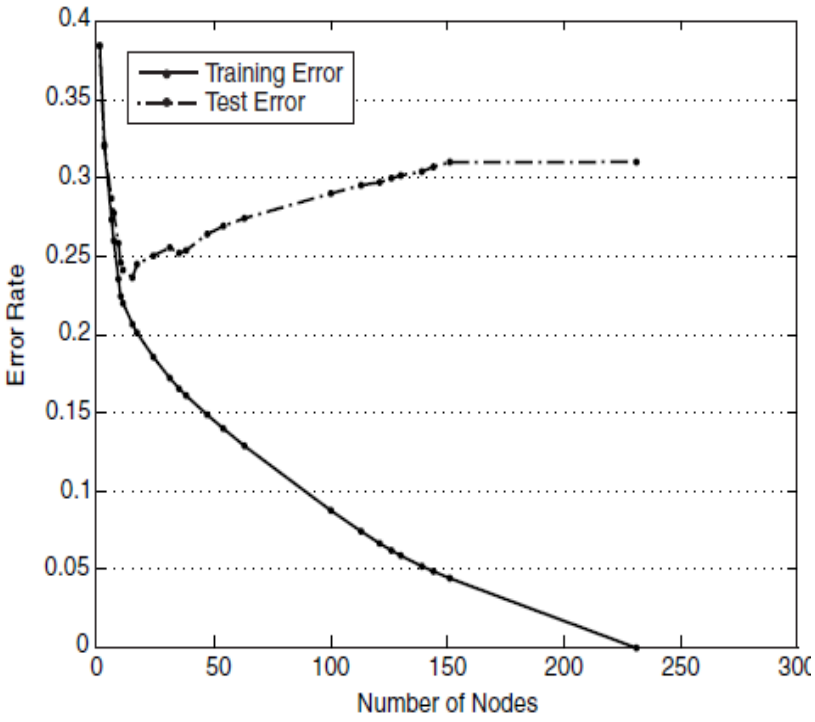
- 最优属性选择：

$$a^*=\arg\min_{a\in A}Gini(D,a)$$

- **CART算法(Breman 1984)就采用基尼指数替代了ID3 算法的信息增益**

剪枝处理（Pruning）

- 问题-过拟合
  - 决策树对训练数据有很好的分类能力，但对未知的测试数据未必有好的分类能力，泛化能力弱，即可能发生过拟合现象



- 可能原因
  - 训练数据有噪声，对训练数据拟合的同时也对噪音进行拟合，影响了分类效果
  - 叶节点样本太少，易出现耦合的规律性，使一些属性恰巧可以很好地分类，但却与实际的目标函数并无关系

剪枝处理（Pruning）

- **解决办法**
  - **剪枝**是决策树学习算法中对付“过拟合”的主要手段
- **基本策略**
  - **预剪枝策略（pre-pruning）**：决策树生成过程中，对每个节点在划分前进行估计，若划分不能带来决策树泛化性能提升，则停止划分，并将该节点设为叶节点
  - **后剪枝策略（post-pruning）**：先利用训练集生成决策树，自底向上对非叶节点进行考察，若将该叶节点对应子树替换为叶节点能带来泛化性能提升，则将该子树替换为叶节点

剪枝处理（Pruning）

训练样本

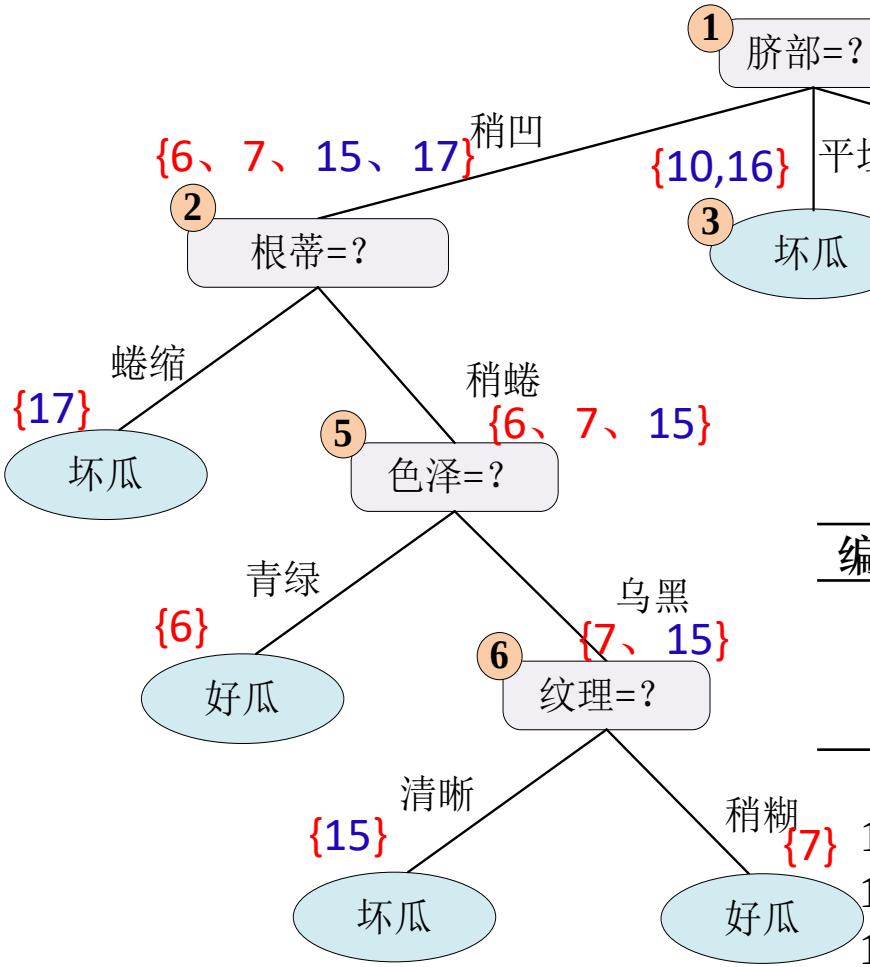
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

测试样本

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

预剪枝算法

- **ID3算法生成的决策树**



● 泛化性能：

{4,11,12}被正确划分

3/7= 42.9%

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

预剪枝算法

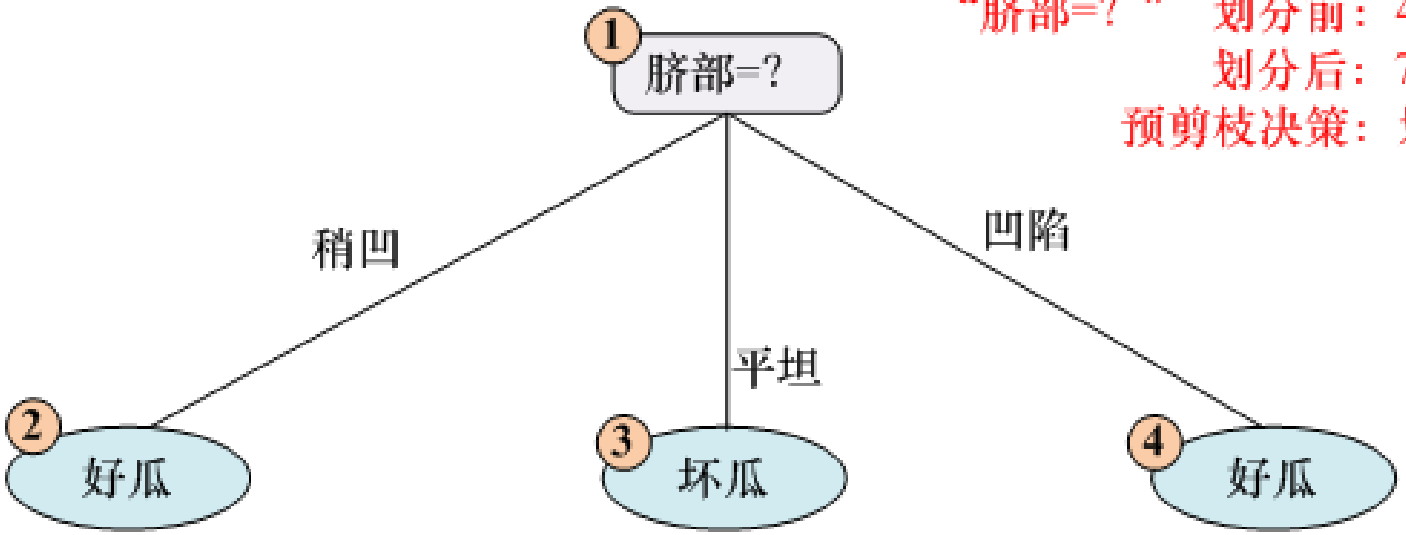
- **第一步：评估节点1**
  - **属性选择：** 基于信息增益准则，选择属性“脐部”
  - **不划分：**
    - 标记为训练样例数最多的类别，如“好瓜”
    - 泛化性能：{4,5,8}被正确分类， **3/7= 42.9%**
  - **划　分：**
    - 节点2：稍凹{6,7,15,17} “好瓜”
    - 节点3：平坦{10,16} “坏瓜”
    - 节点4：凹陷{1,2,3,14} “好瓜”
    - 泛化性能：{4,5,8,11,12}被正确分类， **5/7= 71.4%**
  - **评估结果/预剪枝决策： 划分**

预剪枝算法

- **第二步：评估节点2：训练样本{6,7,15,17}**
  - **属性选择：**基于信息增益准则，选择属性“根蒂”
  - **不划分：** {4,5,8,11,12}被正确分类 **5/7= 71.4%**
  - **划　分：** {4,5,8,11,12}被正确分类 **5/7= 71.4%**
  - 评估结果/预剪枝决策： **不划分**
- **第三步：评估节点4：训练样本{1, 2, 3, 14}**
  - **属性选择：**基于信息增益准则，选择属性“色泽”
  - **不划分：** {4,5,8,11,12}被正确分类 **5/7= 71.4%**
  - **划　分：** {4,~~5,~~8,11,12}被正确分类 **4/7= 57.1%**
  - 评估结果/预剪枝决策：**不划分**

预剪枝算法

- 最终生成的决策树



测试集精度
“脐部=？” 划分前：42.9%
划分后：71.4%
预剪枝决策：划分

测试集精度
“脐部=？” 划分前：71.4%
划分后：57.1%
预剪枝决策：不划分

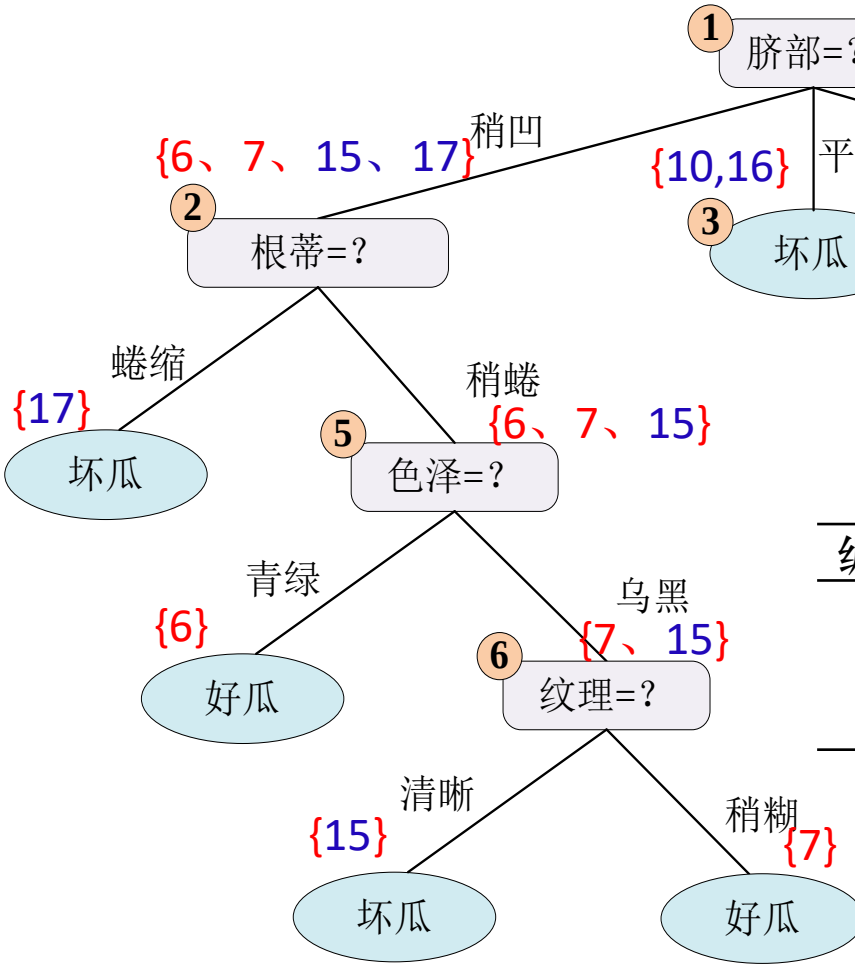
测试集精度
“色泽=？” 划分前：71.4%
划分后：71.4%
预剪枝决策：不划分

预剪枝算法

- 结论
  - **优势**：预剪枝“剪掉了”很多没必要展开的分支，降低了过拟合的风险，并且显著减少了决策树的训练时间开销和测试时间开销
  - **劣势**：有些分支的当前划分有可能不能提高甚至降低泛化性能，但后续划分有可能提高泛化性能；预剪枝禁止这些后续分支的展开，可能会导致欠拟合

后剪枝算法

- ID3算法生成的决策树



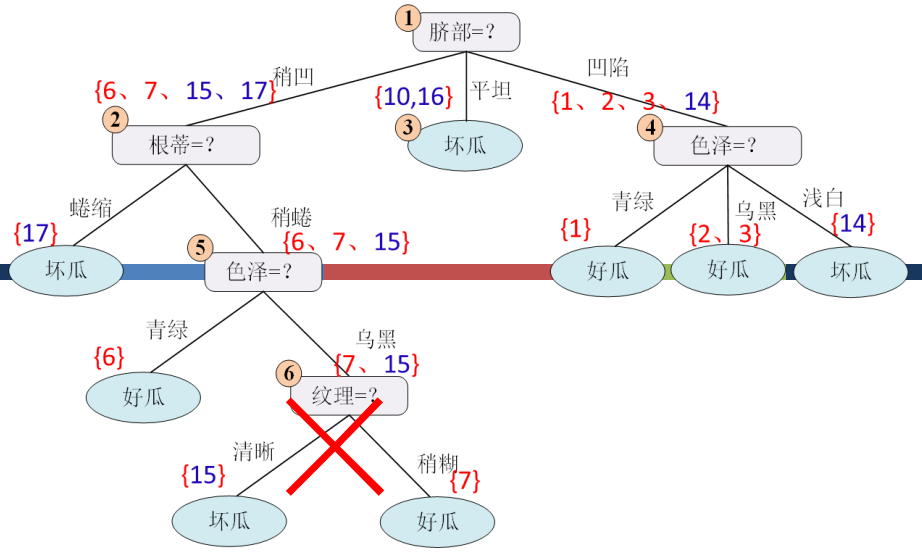
- **泛化性能：**

{4,11,12}被正确划分

3/7= 42.9%

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

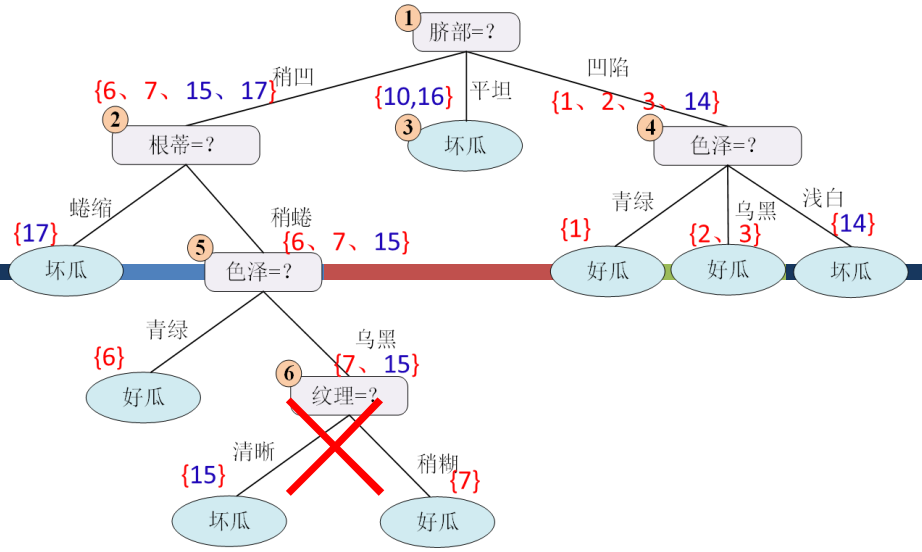
后剪枝算法



- **第一步：评估节点6**
  - **剪枝前：**
    - 属性为“纹理”；样本为{7,15}
    - 泛化性能：{4,11,12}被正确分类，　　**3/7= 42.9%**
  - **剪枝后：**
    - 把节点6替换为叶节点，　　“好/坏瓜”
    - 泛化性能　{4, **8/9**,11,12}被正确分类，　**4/7= 57.1%**

- **评估结果/后剪枝决策：　剪枝**

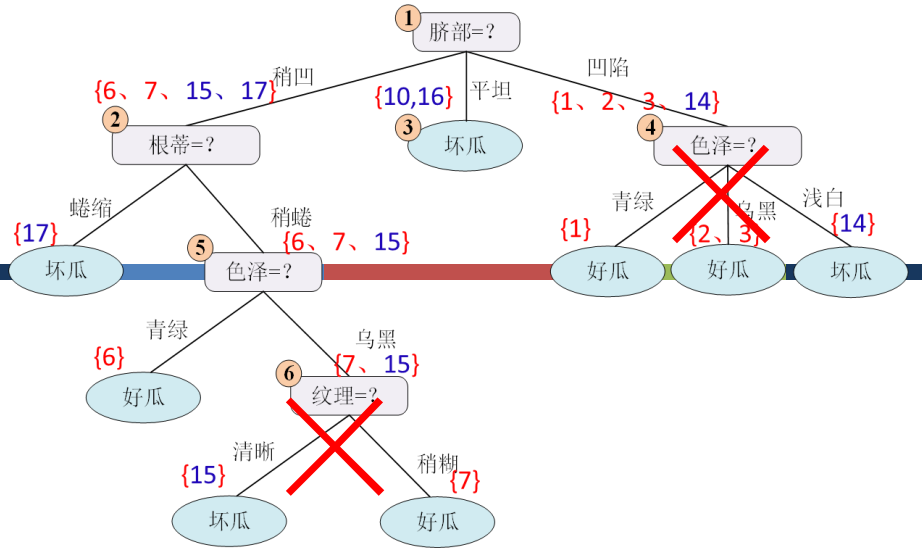
后剪枝算法



- **第二步：评估节点5**
  - **剪枝前：**
    - 属性为“色泽”，样本{6,7,15}
    - 泛化性能：同第一步　　**4/7= 57.1%**
  - **剪枝后：**
    - 把节点5替换为叶节点，　　“好瓜”
    - 泛化性能：{4, 8,11,12}被正确分类，　　**4/7= 57.1%**

- **评估结果/后剪枝决策：　不剪枝**

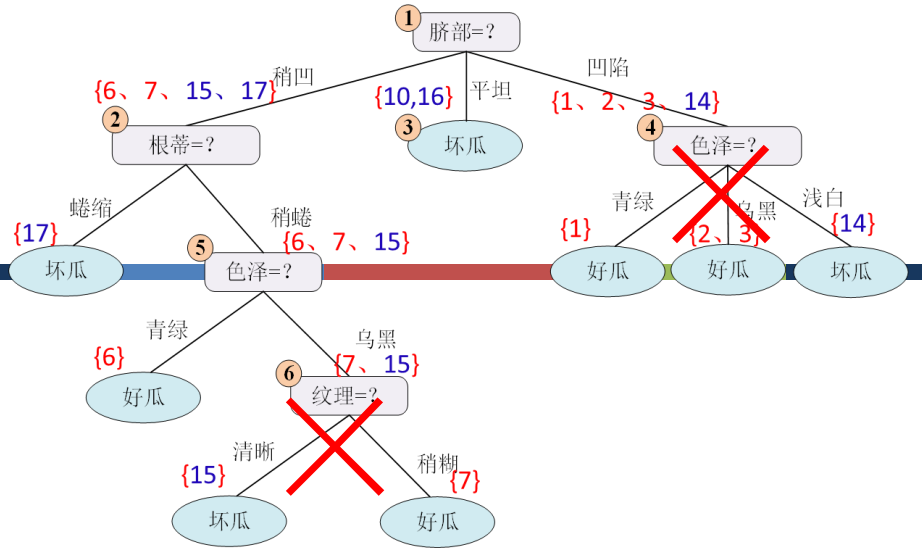
后剪枝算法



- **第三步：评估节点4**
  - **剪枝前：**
    - 属性为“色泽”，样本{1,2,3,14}
    - 泛化性能：同上一步　　**4/7= 57.1%**
  - **剪枝后：**
    - 把节点4替换为叶节点，　　“好瓜”
    - 泛化性能：{4, **5**,8,11,12}被正确分类，　　**5/7= 71.4%**

- **评估结果/后剪枝决策：　剪枝**

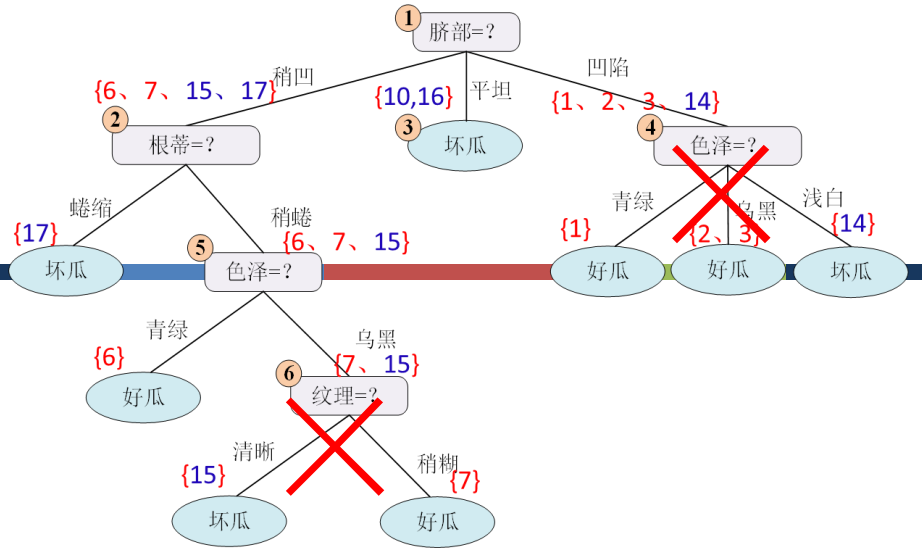
后剪枝算法



- **第四步：评估节点2**
  - **剪枝前：**
    - 属性为“根蒂”，样本{6,7,15,17}
    - 泛化性能：同上一步 **5/7= 71.4%**
  - **剪枝后：**
    - 把节点2替换为叶节点， “好/坏瓜”
    - 泛化性能： {4,5,**8/9**,11,12}被正确分类，**5/7= 71.4%**

- **评估结果/后剪枝决策： 不剪枝**

后剪枝算法

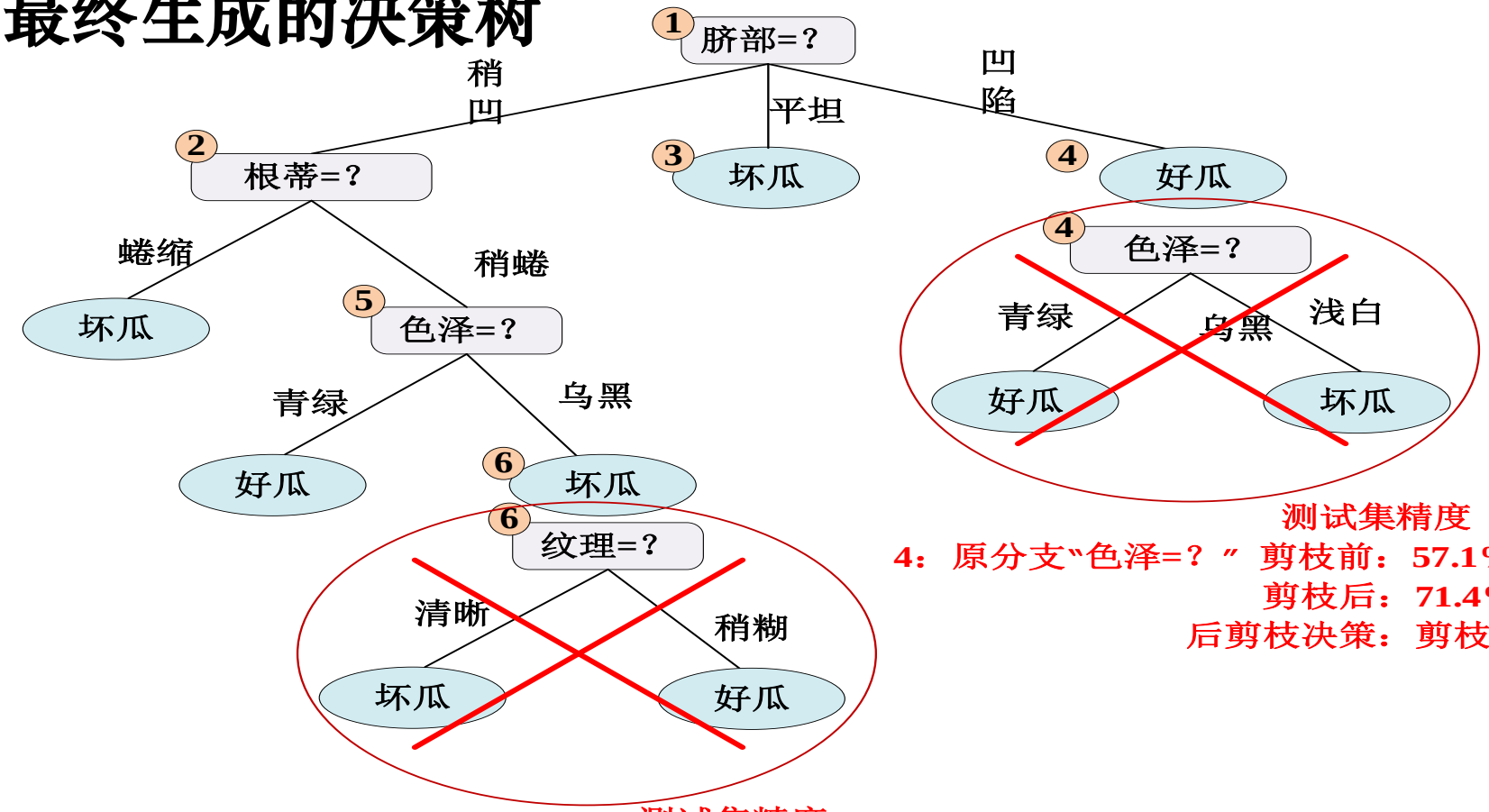


- **第五步：评估节点1**
  - **剪枝前：**
    - 泛化性能：同上一步　　**5/7= 71.4%**
  - **剪枝后：**
    - 把节点1替换为叶节点
    - 泛化性能：{4,5,8,11,12}被正确分类，　**5/7= 71.4%**

- **评估结果/后剪枝决策：　不剪枝**

后剪枝算法

- 最终生成的决策树



测试集精度

6：原分支"纹理=？"剪枝前：42.9%

剪枝后：57.1%

后剪枝决策：剪枝

测试集精度

4：原分支"色泽=？"剪枝前：57.1%

剪枝后：71.4%

后剪枝决策：剪枝

后剪枝算法

- **结论**
  - **优势**：测试了所有分支，比预剪枝决策树保留了更多分支，降低了欠拟合的风险，泛化性能一般优于预剪枝决策树
  - **劣势**：后剪枝过程在生成完全决策树后在进行，且要自底向上对所有非叶节点逐一评估；因此，决策树的训练时间开销要高于未剪枝决策树和预剪枝决策树

连续值处理

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	0.46	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	0.36	0.37	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

连续值处理

- **基本思想**：采用二分法(bi-partition)进行离散化
  - 给定样本集 D 和连续属性 $a(a \in A)$，假定a 在 D 上有N个不同取值，将这些值从大到小排序得$\{a_1, a_2, \ldots, a_N\}$
  - 基于划分点 t，可将 D 分为子集 D_t^+ 和 D_t^-，其中D_t^+ (D_t^-)包含了属性值A不小（大）于 t 的样本子集
  - t 在$[a_n, a_{n+1})$上的任意取值的划分结果都相同
  - **候选划分点集合**

$$T_a = \left\{ \frac{a_n + a_{n+1}}{2} \,\middle|\, 1 \le n \le N-1 \right\}$$

  - **C4.5算法[Quinlan 1993]就采用了二分法进行离散化**

连续值处理

- **信息增益**

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

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

其中，$G(D,A,t)$ 是样本集D基于划分点 t 二分后的信息增益，所以，我们需选择使 $G(D,A,t)$ 最大的划分点 t

连续值处理-示例

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	0.46	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	0.36	0.37	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

连续值处理-示例

- **计算候选划分点集合**

$$T_a = \left\{ \frac{a_n + a_{n+1}}{2} \,\middle|\, 1 \le n \le N-1 \right\}$$

- $T_{密度}$={0.244, 0.294, 0.351,…, 0.708, 0.746} **1*16**
- $T_{含糖率}$={0.049, 0.074, 0.095,…, 0.373, 0.126} **1*16**

■参照教材《机器学习》-周志华P84-85示例

连续值处理-示例

- 计算信息增益

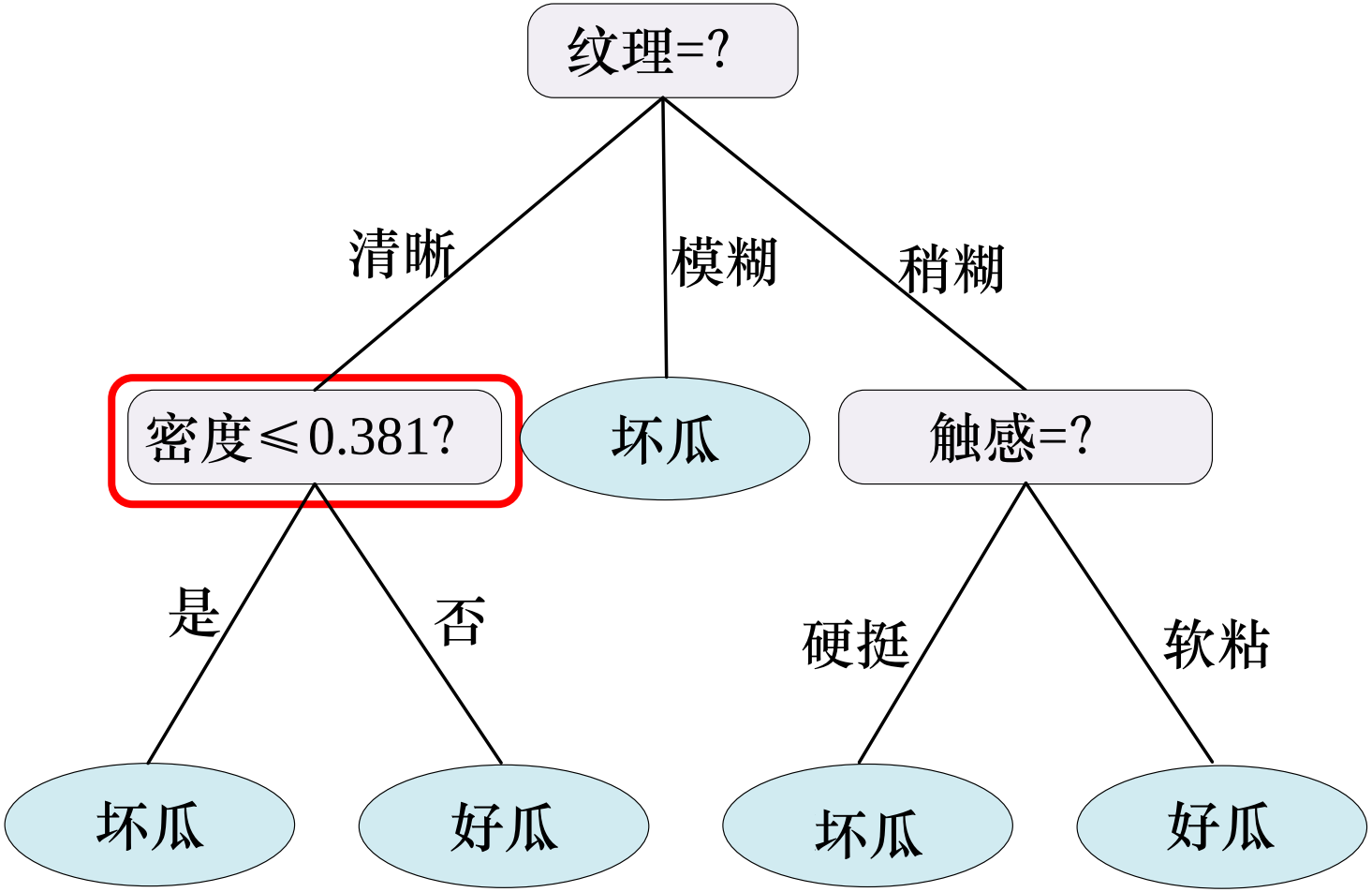
$$G(D,a)=H(D)-\sum_{n=1}^{N}\frac{|D^n|}{|D|}H(D^n) \Rightarrow G(D,a)=\max_{t\in T_a}\left(H(D)-\sum_{\lambda\in\{+,-\}}\frac{|D_t^\lambda|}{|D|}H(D_t^\lambda)\right)$$

  - $G(D,$ 密度$)=0.262$　　$T^*_{\text{密度}}=0.381$
  - $G(D,$ 含糖率$)=0.349$　　$T^*_{\text{含糖率}}=0.126$

- 已知
  - $G(D,$ 色泽$)=0.109$　　$G(D,$ 根蒂$)=0.143$
  - $G(D,$ 敲声$)=0.141$　　$G(D,$ 纹理$)=0.381$
  - $G(D,$ 脐部$)=0.289$　　$G(D,$ 触感$)=0.006$

连续值处理-示例

- 最终生成的决策树



缺失值处理

- 问题
  - 前面假设：所有样本的属性完整
  - 实际情况：存在不完整样本：即样本的某些属性缺失；特别是属性数目较多时
  - 如果简单放弃不完整样本，会导致数据信息的浪费
  - 实际中确实需要属性缺失情况下进行决策
  - 例如：医疗领域，由于诊测成本、隐私保护等问题，只有部分诊断结果

缺失值处理

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

缺失值处理

- **需要解决的两个问题**
  - 如何在属性值缺失的情况下进行划分属性选择（计算信息增益）？
  - 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?

缺失值处理

- **问题1：信息增益计算**
  - 仅可利用没有属性缺失的样本
- 定义：
  - $\tilde{D}$：D中在属性a上没有缺失值的样本子集
  - 属性$a(a \in A)$ 有N个可取值$\{a_1, a_2, \ldots, a_N\}$
  - $\tilde{D}^n$：$\tilde{D}$ 中在属性a上取值为 a_n的样本子集
  - $\tilde{D}_c$：$\tilde{D}$ 中属于第$c(c \in C)$类的样本子集
  - ω_x：**样本x的权重，属性缺失时，可能以不同概率被划分**

缺失值处理

- 定义：
  - 无缺失值样本所占比例 $\rho = \frac{\sum_{x \in \tilde{D}} \omega_x}{\sum_{x \in D} \omega_x}$
  - 无缺失样本中第c类比例 $\tilde{p}_c = \frac{\sum_{x \in \tilde{D}_c} \omega_x}{\sum_{x \in \tilde{D}} \omega_x}, (1 \le c \le C), \sum_{c=1}^{C} \tilde{p}_c = 1$
  - 无缺失样本中属性a上取值为 a_n 的比例 $\tilde{r}_n = \frac{\sum_{x \in \tilde{D}^n} \omega_x}{\sum_{x \in \tilde{D}} \omega_x}, (1 \le n \le N), \sum_{n=1}^{N} \tilde{r}_n = 1$

缺失值处理

- 信息增益

$$G(D,a) = H(D) - \sum_{n=1}^{N} \frac{|D^n|}{|D|} H(D^n)$$

$$G(D,a) = \rho \times G(\tilde{D},a) = \rho \times \left(H(\tilde{D}) - \sum_{n=1}^{N} \tilde{r}_n H(\tilde{D}_n) \right)$$

其中 $$H(\tilde{D}) = -\sum_{c=1}^{C} \tilde{p}_c \log_2 \tilde{p}_c$$

缺失值处理

- **问题2：含有缺失属性的样本的划分**

- 原则：让样本根据属性情况以不同概率划分到不同子节点

  - 若样本x在属性a上的取值已知：则将x划入与其取值对应的子节点，且样本权值保持为ω_x

  - 若样本x在属性a上的取值未知：则将x划入所有子节点，且其在与属性值对应的子节点中的权值根据属性a上已知的样本的比例调整为 $\tilde{r}_n \cdot \omega_x$

- **C4.5算法[Quinlan 1993]就采用了上述方法处理缺失值**

缺失值处理-示例

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

缺失值处理-示例

- 计算信息熵-以属性色泽为例

$$H\left(\tilde{D}\right)=-\sum_{c=1}^{C}\tilde{p}_c\log_2\tilde{p}_c=-\left(\frac{6}{14}\log_2\frac{6}{14}+\frac{8}{14}\log_2\frac{8}{14}\right)=0.985$$

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

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

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

■参照教材《机器学习》-周志华P86-88示例

缺失值处理-示例

- 计算信息增益-以属性色泽为例

$$G\left(\tilde{D}, \text{色泽}\right)=H\left(\tilde{D}\right)-\sum_{n=1}^{3}\tilde{r}_n H(\tilde{D}_n)$$

$$=0.985-\left(\frac{4}{14}\times 1.000+\frac{6}{14}\times 0.918+\frac{4}{14}\times 0.000\right)$$

$$=0.306$$

$$G\left(D, \text{色泽}\right)=\rho\times G\left(\tilde{D}, \text{色泽}\right)=\frac{14}{17}\times 0.306=0.252$$

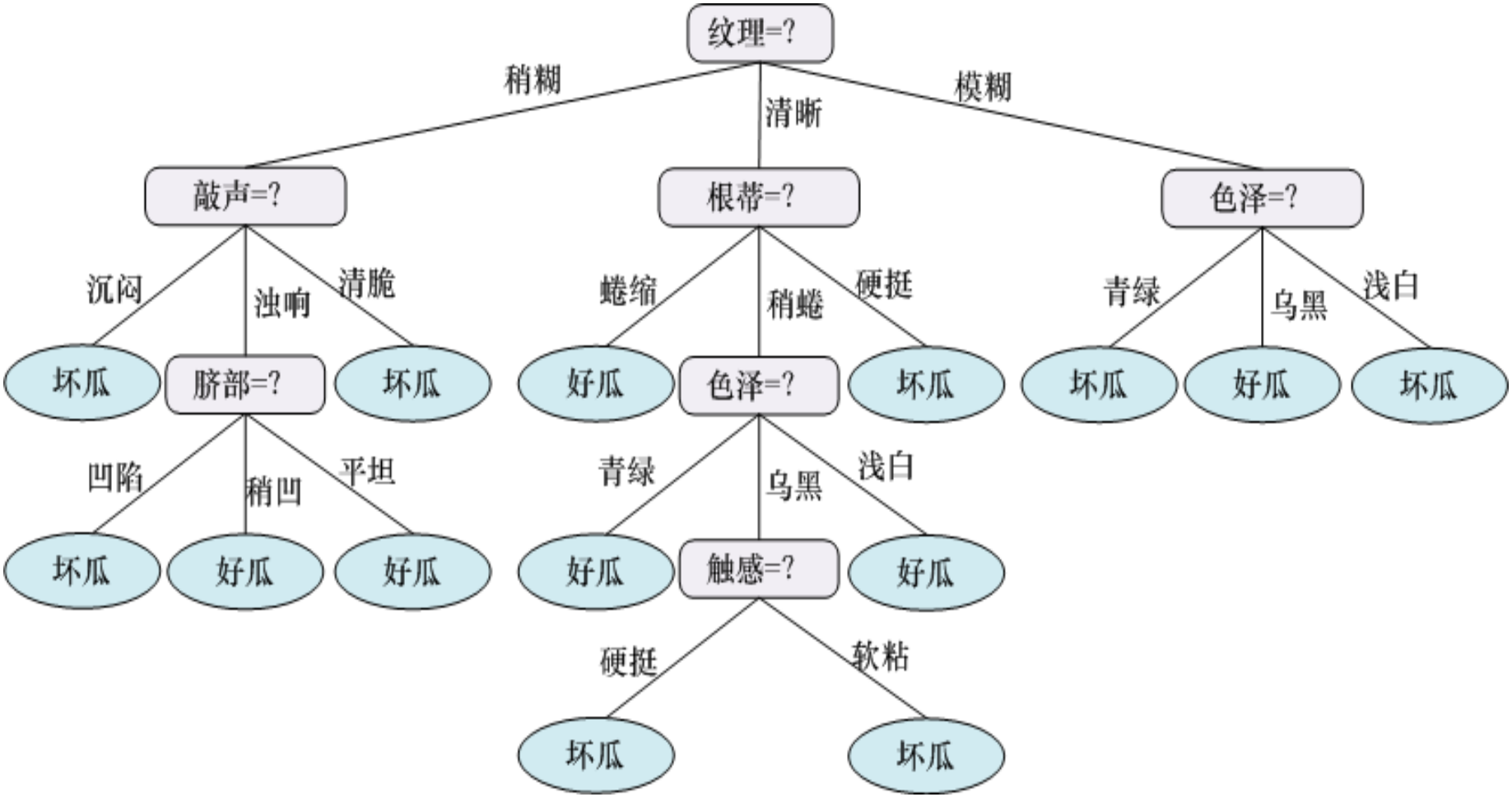
- $G(D, \text{色泽}) = 0.252$
- $G(D, \text{敲声}) = 0.145$
- $G(D, \text{脐部}) = 0.289$
- $G(D, \text{根蒂}) = 0.171$
- $G(D, \text{纹理}) = 0.424$
- $G(D, \text{触感}) = 0.006$

缺失值处理-示例

- 因此选择属性“纹理”用于对根结点划分
  - 属性不缺失样本
    - “纹理=清晰”分支：样本 {1, 2, 3, 4, 5, 6, 15}
    - “纹理=稍糊”分支：样本 {7, 9, 13, 14, 17}
    - “纹理=模糊”分支：样本 {11, 12, 16}
    - 且样本在各节点中的权重ω保持为1
  - 属性缺失样本
    - 样本{8,10}:同时进入三个分支，权重ω分别为$\frac{7}{15}, \frac{5}{15}, \frac{3}{15}$

缺失值处理-示例

- 最终生成的决策树



不同代价属性的处理

- **问题**
  - 不同的属性测量具有不同的代价
  - 医疗诊断为例：
    - 属　　性：体温、血压、血常规、活检
    - 代价不同：所需时间、费用、友好性等
- **解决思路**
  - 在属性筛选度量标准中考虑属性的不同代价
  - 优先选择低代价属性的决策树
  - 必要时才依赖高代价属性

不同代价属性的处理

- **属性筛选度量标准1**[Tan et al., 1990, 1993]

$$G_{Cost}(D,a)=\frac{G(D,a)}{Cost(a)}$$

- **属性筛选度量标准2**[Nunez et al., 1988, 1991]

$$G_{Cost}(D,a)=\frac{2^{G(D,a)}-1}{(Cost(a)+1)^{\omega}}$$

其中$Cost(\boldsymbol{a})$为属性 $\boldsymbol{a}$ 的代价

$\omega \in [0,1]$为一常数，决定代价对于信息增益的相对重要性

延伸阅读

- **概念学习系统(Concept Learning System, CLS)**
  - 1966年由Hunt等人提出，奠定决策树算法发展的基础
- **CART(Classification And Regression Tree)算法：**
  - 即分类回归树算法，简称CART算法，是一种二分递归分割技术， 1984年由Breiman等人提出
- **J. R. Quinlan**
  - **1975年：ID3算法； 1993年：C4.5算法**
- **多变量决策树**
  - **OC1[Murthy et al., 1994]等**