日期:　　/

决策树

一、基本概念

1. 分类流程

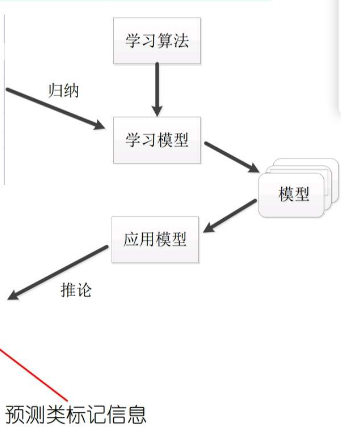
表 4.2

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

未见样本集

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	?
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	?
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	?
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	?
11	浅白	硬挺	清脆	模糊	平坦	硬滑	?
12	浅白	蜷缩	浊响	模糊	平坦	软粘	?
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	?



2. 决策树是一种进行分类的树型数据结构

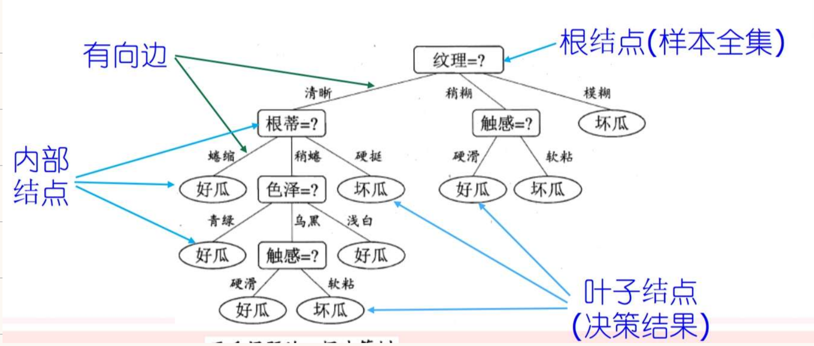
① 根结点 (root node)：没有入边，但有零或多条出边，包含全部样本

② 内部结点 (internal node)：恰有一条入边和2条或多条出边，表示一个属性/特征

③ 叶子结点 (leaf node)：恰有一条入边，无出边，表示决策结果(类标记)

④ 决策树一般包含结点(3个)和有向边、

⑤ 决策规则：从树的根结点到叶子结点的一条路径代表一条决策规则



⑥ 决策树目的：产生一颗泛化能力强的决策树，即处理未见样本能力强.

二、决策树构造.

1、信息与信息熵.

① 信息：用来消除某种随机不确定性

② 信息熵(Ent)：衡量随机变量的不确定性的某种度量

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

D: $[x_1, x_2, \cdots, x_m]$　　p: $[p_1, p_2, \ldots, p_m]$

D代表当前样本(状态)集合，p代表概率集合.

假定D中第k类样本占比例为p_k (|y|表示类别数目

<1> 是度量样本集合纯度最常用的一种指标

日期:　　/

<2> Ent(D) 值越小，则 D的纯度越高，样本确定性越高

③ 信息增益

假设事物未获得某条信息之前的状态集与概率集为

$$X:[x_1,x_2\cdots,x_n],\ P:[p_1,p_2\cdots,p_n]$$

而获得了某条信息之后的状态与概率集合为

$$X':[x_1,x_2\cdots,x_m],\ P':[p_1,p_2\cdots,p_m]$$

$$I = H(X) - H(X') = -\sum_{i=1}^{n} p_i \log p_i - \left(-\sum_{i=1}^{m} p_i \log p_i\right)$$

2、ID3 算法（信息论角度）

① 条件属性：特征 ｝ 希望找到带给我们最
决策属性：标记 ｝ 多信息的条件属性

[信息增益最大]

② 划分选择：信息增益.

离散属性 a 有 V 个可能的取值 $\{a^1, a^2, \cdots, a^V\}$，用 a 来进行划分，则会产生 V 个分支结点，其中第 v 个分支结点包含了 D 中所有在属性 a 上取值为 a^v 的样本 → a^v 的样本个数.

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

样本数越多的分支结点的影响越大.

日期: /

信息增益越大，则纯度提升越大

ID3：决策树学习算法以信息增益为准。

举例：

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

若按纹理划分，则

纹理
- 清晰 $\{1,2,3,4,5,6,8,10,15\}$
- 稍糊 $\{7,9,13,14,17\}$
- 模糊 $\{11,12,16\}$

$D=\{1,\dots,17\}$ 包括17个训练样本 $|y|=2$，在决策树开始学习时，根结点包括正例 $p_1=\frac{8}{17}$，$p_2=\frac{9}{17}$

$$Ent(D) = -\left(\frac{8}{17}\log\frac{8}{17}+\frac{9}{17}\log\frac{9}{17}\right)$$

$$=0.998$$

- 以属性“色泽”为例，即使用“色泽”属性对D进行划分，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)
- 子集D^1 包含编号为$\{1,4,6,10,13,17\}$ 的6个样例，其中正例占 $p_1=\frac{3}{6}$ ，反例占 $p_2=\frac{3}{6}$ ，第一个分支结点的信息熵为：

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

- 子集 D^2 包含编号为$\{2,3,7,8,9,15\}$的6个样例，其中正例占p_1= 4/6，反例占p_2= 2/6，第二个分支结点的信息熵为：

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

- 子集 D^3 包含编号为$\{5,11,12,14,16\}$的5个样例，其中正例占p_1= 1/5，反例占p_2= 4/5，第三个分支结点的信息熵为：

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

- 属性“色泽”的信息增益为

$$\begin{aligned}\mathrm{Gain}(D,\text{色泽})&=\mathrm{Ent}(D)-\sum_{v=1}^{3}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\\&=0.998-(\tfrac{6}{17}\times1.000+\tfrac{6}{17}\times0.918+\tfrac{5}{17}\times0.722)\\&=0.109\end{aligned}$$

按照相同的方法，可以计算出根蒂、敲声、纹理、脐部等的信息增益如下：

$$\mathrm{Gain}(D,\text{根蒂})=0.143 \qquad \mathrm{Gain}(D,\text{敲声})=0.141$$
$$\mathrm{Gain}(D,\text{纹理})=0.381 \qquad \mathrm{Gain}(D,\text{脐部})=0.289$$
$$\mathrm{Gain}(D,\text{触感})=0.006$$

最后可以发现，纹理的信息增益最大。因此，纹理属性帮助我们进行西瓜判断最有用，选择纹理属性作为划分属性，即作为决策树的根结点，得到如下决策树。

③ ID3 算法流程:

算法总结（ID3 算法）：

输入: 训练集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$

属性集 $A=\{a_1,a_2,\cdots,a_d\}$

输出: 决策树 T : $\mathrm{TreeGenerate}(D,A)$

STEP 1: 若 D 中所有样本都属于同一类 C, 则 T 为单结点树, 并将 C 作为该结点的类标记, 返回 T, 否则转 **STEP 2**;

STEP 2: 依据决策属性计算信息熵 $\mathrm{Ent}(D)$, 令 $k=1$,

1: 选择 a_k, 假设 a_k 具有 v_k 个可能的取值, $D^{a_k^i}$ 为属性 $a_k=v_k^i$ 的样本集合, 计算条件信息熵 $\mathrm{Ent}(D|a_k)=\sum_{i=1}^{v_k}\frac{|D^{a_k^i}|}{|D|}\mathrm{Ent}(D^{a_k^i})$

2: 计算 a_k 属性的信息增益, $\mathrm{Gain}(D,a_k)=\mathrm{Ent}(D)-\mathrm{Ent}(D|a_k)$;

3: $k=k+1$, 若 $k<m$, 则跳转到 1;

决策树算法核心

STEP 3: 选择信息增益最大的属性 a_p 设为根结点, 根据 a_p 将数据集分成 v_p 个子集 $\{D^{a_p^1},D^{a_p^2},\cdots,D^{a_p^{v_p}}\}$;

STEP 4: 令 $D=D^{a_p^j}$, $A=A-a_p$, 转 **STEP 1**.

④ 缺点: 信息增益对可取值数目较多的属性有所偏好.

3. C4.5 算法 (增益率

① 增益率: 对可取值数目较少的属性有所偏好

$$\mathrm{Gain\_ratio}(D,a)=\frac{\mathrm{Gain}(D,a)}{IV(a)}$$

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

属性a的固有值, a的可能取值数越多, IV(a)的值通常越大.

② C4.5 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

4. CART：基尼系数：
① 原因：ID3，C4.5 会涉及大量的对数运算
② 基尼系数：代表模型的不纯度，基尼系数越小，则不纯度越低，特征越好。
(1) 基尼值：数据集 D 的纯度：

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

p_k：第 k 类样本在 D 中占比.
Gini(D) 越小，数据集越纯，D 纯度越高
(2) 基尼系数：在样本集合中随机选中的样本被分错的概率

$$\mathrm{Gini\_index}(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Gini}(D^v)$$

选 基尼系数最小的属性作为划分属性

对比：
① Gini 计算速度比熵快，
② Gini 趋向于孤立数据集中数量多的类，将

它们会训一个树中；倾向于构建一颗平衡的树

三. 剪枝处理.

1. 目的：处理过拟合.

　　通过剪枝来避免因决策分支过多以致于把训练集自身一些特点当作所有数据一般性质而导致过拟合

2. 基本策略：预剪枝，后剪枝.

3. 泛化性能提升判定：留出法

　　预留法将一部分数据用作验证集，评估性能

4. 预剪枝：

① 实现：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

② 举例：

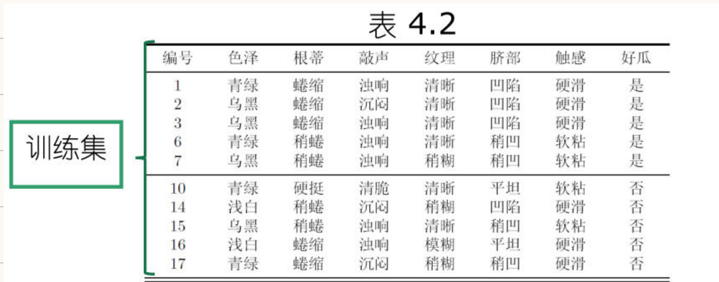


表 4.2

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

首先计算各属性信息增益.

对于训练集 D, 其中正例占 $p_1=\frac{5}{10}=\frac{1}{2}$, 反例占 $p_2=\frac{5}{10}=\frac{1}{2}$, 因此, 信息熵:

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

下面根据表 4.2 计算各个属性的信息增益值.

色泽: $D^1\{色泽=青绿\}:(1,6,10,17)$, $D^2\{色泽=乌黑\}:(2,3,7,15)$, $D^3\{色泽=浅白\}:(14,16)$, 则

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

$$\mathrm{Ent}(D^2)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{3}{4}\log_2\frac{3}{4}+\frac{1}{4}\log_2\frac{1}{4}\right)=0.811,$$

$$\mathrm{Ent}(D^3)=-\sum_{k=1}^{2}p_k\log_2 p_k=-(0+1\log_2 1)=0,$$

$$\mathrm{Gain}(D,色泽)=1-\left(\frac{4}{10}*1+\frac{4}{10}*0.811+0\right)=0.276.$$

纹理: $D^1\{纹理=清晰\}:(1,2,3,6,10,15)$, $D^2\{纹理=稍糊\}:(7,14,17)$, $D^3\{纹理=模糊\}:(16)$, 则

$$\mathrm{Ent}(D^1)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{4}{6}\log_2\frac{4}{6}+\frac{2}{6}\log_2\frac{2}{6}\right)=0.918,$$

$$\mathrm{Ent}(D^2)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right)=0.918,$$

$$\mathrm{Ent}(D^3)=-\sum_{k=1}^{2}p_k\log_2 p_k=-(0+1\log_2 1)=0.$$

$$\mathrm{Gain}(D,纹理)=1-\left(\frac{6}{10}*0.971+\frac{3}{10}*0.918+\frac{1}{10}*0\right)=0.174.$$

根蒂: $D^1\{根蒂=蜷缩\}:(1,2,3,16,17)$, $D^2\{根蒂=稍蜷\}:(6,7,14,15)$, $D^3\{根蒂=硬挺\}:(10)$, 则

$$\mathrm{Ent}(D^1)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log_2\frac{2}{5}\right)=0.971,$$

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

$$\mathrm{Ent}(D^3)=-\sum_{k=1}^{2}p_k\log_2 p_k=-(0+1\log_2 1)=0.$$

$$\mathrm{Gain}(D,根蒂)=1-\left(\frac{1}{2}*0.971+\frac{2}{5}*1+\frac{1}{10}*0\right)=0.115.$$

脐部: $D^1\{脐部=凹陷\}:(1,2,3,14)$, $D^2\{脐部=稍凹\}:(6,7,15,17)$, $D^3\{脐部=平坦\}:(10,16)$, 则

$$\mathrm{Ent}(D^1)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{3}{4}\log_2\frac{3}{4}+\frac{1}{4}\log_2\frac{1}{4}\right)=0.811,$$

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

$$\mathrm{Ent}(D^3)=-\sum_{k=1}^{2}p_k\log_2 p_k=-(0+1\log_2 1)=0,$$

$$\mathrm{Gain}(D,脐部)=1-\left(\frac{4}{10}*0.811+\frac{4}{10}*1+\frac{1}{10}*0\right)=0.276.$$

敲声: $D^1\{敲声=浊响\}:(1,3,6,7,15,16)$, $D^2\{敲声=沉闷\}:(2,14,17)$, $D^3\{敲声=清脆\}:(10)$, 则

$$\mathrm{Ent}(D^1)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{4}{6}\log_2\frac{4}{6}+\frac{2}{6}\log_2\frac{2}{6}\right)=0.918,$$

$$\mathrm{Ent}(D^2)=-\sum_{k=1}^{2}p_k\log_2 p_k=-\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right)=0.918,$$

$$\mathrm{Ent}(D^3)=-\sum_{k=1}^{2}p_k\log_2 p_k=-(0+1\log_2 1)=0.$$

$$\mathrm{Gain}(D,敲声)=1-\left(\frac{6}{10}*0.971+\frac{3}{10}*0.918+\frac{1}{10}*0\right)=0.174.$$

触感: $D^1\{触感=硬滑\}:(1,2,3,16,17)$, $D^2\{触感=软粘\}:(6,7,10,15)$, 则

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

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

$$\mathrm{Ent}(D^3)=-\sum_{k=1}^{2}p_k\log_2 p_k=-(0+1\log_2 1)=0,$$

$$\mathrm{Gain}(D,触感)=1-\left(\frac{6}{10}*1+\frac{4}{10}*1\right)=0.$$

脐部和色泽最大，在这里选脐部.
若不按照脐部划分，则将其标为叶结点.

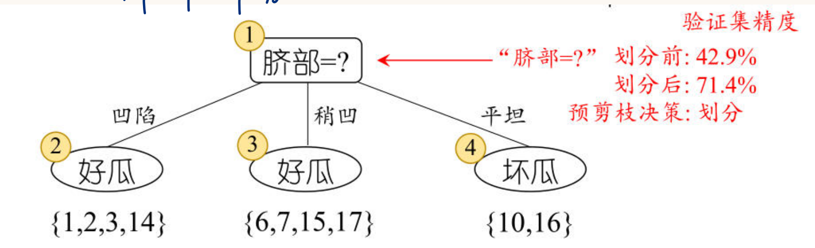
日期: /

类别标记为训练样例中最多类别 好瓜(好瓜

[优选]和坏瓜相同，选择了好瓜)，在验证集中{4,5,8}

被分类正确，验证集精度 $\frac{3}{7} \times 100\% = 42.9\%$

若按照脐部划分



③ 优缺点：

优点

- 预剪枝让决策树的很多分支没有展开，降低了过拟合风险
- 显著减少训练时间和测试时间开销

缺点

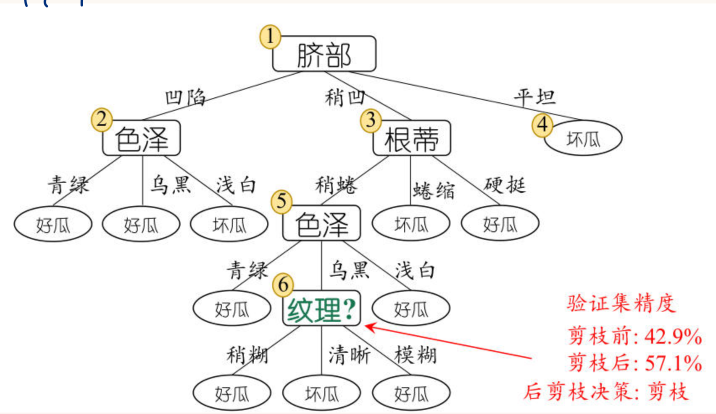
- 欠拟合风险：有些分支的当前划分虽然不能提升泛化性能，但在其基础上进行的后续划分却有可能导致性能显著提高。预剪枝基于"贪心"本质禁止这些分支展开，带来了欠拟合风险

5. 后剪枝：

① 方法：先从训练集中生成一棵完整的决策树，然后自底向上对非叶结点考察，若将该结点对应子树替换为叶结点能带来决策树泛

化性能提升，则子树 → 叶结点

② 举例



③ 优缺点

- 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

四. 连续与缺失值

1. 连续值处理方法：离散化(二分法).

① 实现 (2步).

(1) 连续属性a在样本集D上出现n个不同的

日期:　　/

取值，从小到大排列，记为 $a^1, a^2, \dots, a^n$，基于划分点 t，可将 D 分为子集 D_t^-，D_t^+，其中 D_t^- 表示在属性 a 上取值不大于 t 的样本。考虑包含 n-1 个元素的候选划分点集合。

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

即把 $[a^i, a^{i+1})$ 中位点作为候选划分点。

(2)

- 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

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

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

② 举例:

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

对于数据集中的属性“密度”，决策树开始学习时，根结点包含的17个训练样本在该属性上取值均不同。我们先把“密度”这些值从小到大排序：

$$\{0.243, 0.245, 0.343, 0.360, 0.403, 0.437, 0.481, 0.556, 0.593, 0.608, 0.634, 0.639, 0.657, 0.666, 0.697, 0.719, 0.774\}$$

根据上面计算 T_a 的公式，可得16个候选划分点集合：

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

因此，需要16个信息增益的值，选取信息增益值最大时对应的划分点作为最终划分点

计算 t 取不同值时的信息增益：

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

当 $t = 0.244$ 时：

$D_t^- = \{0.243\}$，$D_t^+ = \{0.245, 0.343, 0.360, 0.403, ..., 0.657, 0.666, 0.697, 0.719, 0.774\}$

$Ent(D_t^-) = -(0 \cdot \log_2 0 + 1 \cdot \log_2 1) = 0$，

$Ent(D_t^+) = -\left(\frac{8}{16} \cdot \log_2 \frac{8}{16} + \frac{8}{16} \cdot \log_2 \frac{8}{16}\right) = 1$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.244) = 0.998 - \left(\frac{1}{17} * 0 + \frac{16}{17} * 1\right) = 0.057$

当 $t = 0.294$ 时：

$D_t^- = \{0.243, 0.245\}$，$D_t^+ = \{0.343, 0.360, 0.403, ..., 0.657, 0.666, 0.697, 0.719, 0.774\}$

$Ent(D_t^-) = -\left(0 \cdot \log_2 0 + \frac{2}{2} \cdot \log_2 \frac{2}{2}\right) = 0$，

$Ent(D_t^+) = -\left(\frac{8}{15} \cdot \log_2 \frac{8}{15} + \frac{7}{15} \cdot \log_2 \frac{7}{15}\right) = 0.997$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.294) = 0.998 - \left(\frac{2}{17} * 0 + \frac{15}{17} * 0.997\right) = 0.118$

当 $t = 0.351$ 时：

$D_t^- = \{0.243, 0.245, 0.343\}$，$D_t^+ = \{0.360, 0.403, ..., 0.657, 0.666, 0.697, 0.719, 0.774\}$

$Ent(D_t^-) = -\left(0 \cdot \log_2 0 + \frac{3}{3} \cdot \log_2 \frac{3}{3}\right) = 0$，

$Ent(D_t^+) = -\left(\frac{8}{14} \cdot \log_2 \frac{8}{14} + \frac{6}{14} \cdot \log_2 \frac{6}{14}\right) = 0.985$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.351) = 0.998 - \left(\frac{3}{17} * 0 + \frac{14}{17} * 0.985\right) = 0.187$

当 $t = 0.381$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360\}$，$D_t^+ = \{0.403, ..., 0.657, 0.666, 0.697, 0.719, 0.774\}$

$Ent(D_t^-) = -\left(0 \cdot \log_2 0 + \frac{4}{4} \cdot \log_2 \frac{4}{4}\right) = 0$，

$Ent(D_t^+) = -\left(\frac{8}{13} \cdot \log_2 \frac{8}{13} + \frac{5}{13} \cdot \log_2 \frac{5}{13}\right) = 0.961$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.381) = 0.998 - \left(\frac{4}{17} * 0 + \frac{13}{17} * 0.961\right) = 0.263$

当 $t = 0.420$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403\}$，$D_t^+ = \{0.437, ..., 0.657, 0.666, 0.697, 0.719, 0.774\}$

$Ent(D_t^-) = -\left(\frac{1}{5} \cdot \log_2 \frac{1}{5} + \frac{4}{5} \cdot \log_2 \frac{4}{5}\right) = 0.722$，

$Ent(D_t^+) = -\left(\frac{7}{12} \cdot \log_2 \frac{7}{12} + \frac{5}{12} \cdot \log_2 \frac{5}{12}\right) = 0.980$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.420) = 0.998 - \left(\frac{5}{17} * 0.722 + \frac{12}{17} * 0.980\right) = 0.094$

当 $t = 0.459$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403, 0.437\}$，$D_t^+ = \{0.481, ..., 0.666, 0.697, 0.719, 0.774\}$

$Ent(D_t^-) = -\left(\frac{2}{6} \cdot \log_2 \frac{2}{6} + \frac{4}{6} \cdot \log_2 \frac{4}{6}\right) = 0.918$，

$Ent(D_t^+) = -\left(\frac{6}{11} \cdot \log_2 \frac{6}{11} + \frac{5}{11} \cdot \log_2 \frac{5}{11}\right) = 0.994$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.459) = 0.998 - \left(\frac{6}{17} * 0.918 + \frac{11}{17} * 0.994\right) = 0.03$

当 $t = 0.518$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403, 0.437, 0.481\}$，$D_t^+ = \{0.556, ..., 0.697, 0.719, 0.774\}$

$Ent(D_t^-) = -\left(\frac{3}{7} \cdot \log_2 \frac{3}{7} + \frac{4}{7} \cdot \log_2 \frac{4}{7}\right) = 0.985$，

$Ent(D_t^+) = -\left(\frac{5}{10} \cdot \log_2 \frac{5}{10} + \frac{5}{10} \cdot \log_2 \frac{5}{10}\right) = 1$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.518) = 0.998 - \left(\frac{7}{17} * 0.985 + \frac{10}{17} * 1\right) = 0.004$

当 $t = 0.574$ 时：

$D_t^- = \{0.243, 0.245, 0.343, 0.360, 0.403, 0.437, 0.481, 0.574\}$，$D_t^+ = \{0.593, ...,, 0.719, 0.774\}$

$Ent(D_t^-) = -\left(\frac{4}{8} \cdot \log_2 \frac{4}{8} + \frac{4}{8} \cdot \log_2 \frac{4}{8}\right) = 1$，

$Ent(D_t^+) = -\left(\frac{4}{9} \cdot \log_2 \frac{4}{9} + \frac{5}{9} \cdot \log_2 \frac{5}{9}\right) = 0.991$，

$\therefore Gain(D,a,t) = Gain(D, 密度, 0.574) = 0.998 - \left(\frac{8}{17} * 1 + \frac{9}{17} * 0.991\right) = 0.002$

当 $t = 0.600$，t=0.621，t=0.636，t=0.648，.....就不在展示详细的计算过程了。

比较能够发现，当 $t = 0.381$ 时，Gain（D，a，t）最大为 0.263。因此选择该划分点。

对于属性“含糖率”，按照同样的方法能够计算出，$t = 0.126$，Gain(D，a，t)=0.349。

③：连续属性在根结点用了一次，后代可以接着用
2次划分点不同

2. 缺失值：

日期:　　/

① 缺失样本数量较少，把不完备的样本删掉

② ：属性值缺失情况如何进行划分属性选择？

对于**问题1**，我们根据 $\tilde{D}$ (即在该属性上没有缺失的样本集) 来计算某个属性 a 的信息增益或者其它指标。我们再给根据 $\tilde{D}$ 计算出来的值一个权重，就可以表示训练集 D 中属性 a 的优劣。

$\tilde{D}$ 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

假定为每个样本 $\boldsymbol{x}$ 赋予一个权重 $w_{\boldsymbol{x}}$，并定义：

(1) 无缺失值样本所占比例：$\frac{\text{无缺失}}{\text{总}}$

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

(2) 无缺失样本中第k类所占比例　$\frac{\text{第k类无缺失}}{\text{无缺失}}$

$$\tilde{p}_k = \frac{\sum_{x\in\tilde{D}_k} w_x}{\sum_{x\in\tilde{D}} w_x} \quad 1\le k\le|y|$$

(3) 无缺失样本中属性a取 a^v 的样本比例　$\frac{a^v\text{无缺失}}{\text{无缺失}}$

$$\tilde{r}_v = \frac{\sum_{x\in\tilde{D}^v} w_x}{\sum_{x\in\tilde{D}} w_x} \quad (1\le v\le V)$$

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

- 其中

$$\operatorname{Ent}(\tilde{D}) = -\sum_{k=1}^{|\mathcal{Y}|}\tilde{p}_k\log_2\tilde{p}_k$$

日期:　　/

③ 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

对于问题2

- 若样本 $\boldsymbol{x}$ 在划分属性 a 上的取值已知，则将 $\boldsymbol{x}$ 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 $\boldsymbol{x}$ 在划分属性 a 上的取值未知，则将 $\boldsymbol{x}$ 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 $\tilde{r}_v \cdot w_x$（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）

④ 举例

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

开始时，D中有17个样本，样本权值均为1

以色泽为例计算信息熵：

(14个) $\tilde{D} = \{2,3,4,6,7,8,9,10,11,12,14,15,16,17\}$

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

$$= 0.985$$

日期: /

- 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

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

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

- 因此，样本子集 $\tilde{D}$ 上属性“色泽”的信息增益为

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

- 于是，样本集 D 上属性“色泽”的信息增益为

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

同理计算其它的可得纹理增益最大

划分结果为：

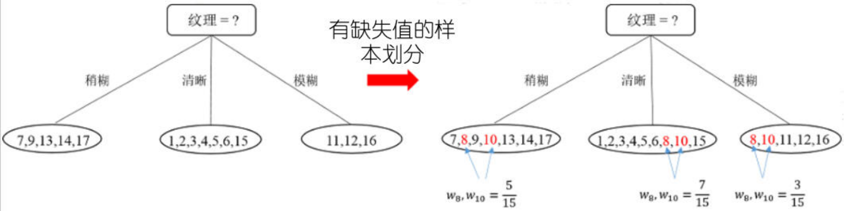
“纹理=稍糊”分支：$\{7, 9, 13, 14, 17\}$，

“纹理=清晰”分支：$\{1, 2, 3, 4, 5, 6, 15\}$，

“纹理=模糊”分支：$\{11, 12, 16\}$。

{8, 10}
缺失!!!
如何划分?

因此，经过第一次划分后的决策树如下图所示：



日期:　　　/

以稍糊结点，色泽为例继续计算

西瓜数据集 2.0α

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
$\boxed{8}$	乌黑	稍蜷	浊响	–	稍凹	硬滑	是
9	乌黑	–	沉闷	稍糊	稍凹	硬滑	否
$\boxed{10}$	青绿	硬挺	清脆	–	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
17	青绿	–	沉闷	稍糊	稍凹	硬滑	否

色泽：

该属性上无缺失值的样本子集$\tilde{D}=\{7,8,9,10,14,17\}$共 6 个样本，但是样本 8 和样本 10 的权重都不再是 1，而是$\frac{1}{3}$，因此$\rho=\frac{4+\frac{2}{3}}{5+\frac{2}{3}}=\frac{14}{17}$，其中正样本比例$\tilde{p}_1=\frac{1+\frac{1}{3}}{4+\frac{2}{3}}=\frac{4}{14}$，$\tilde{p}_2=\frac{3+\frac{1}{3}}{4+\frac{2}{3}}=\frac{10}{14}$。则，

$$Ent(\tilde{D})=-\sum_{k=1}^{2}\tilde{p}_k\log_2\tilde{p}_k=-\left(\frac{4}{14}\log_2\frac{4}{14}+\frac{10}{14}\log_2\frac{10}{14}\right)=0.863$$

$\tilde{D}^1\{色泽=乌黑\}$：(7,8,9)，$\tilde{D}^2\{色泽=青绿\}$：(10,17)，$\tilde{D}^3\{色泽=浅白\}$：(14)。则，

$$\tilde{r}_1=\frac{2+\frac{1}{3}}{4+\frac{2}{3}}=\frac{7}{14},\ \tilde{r}_2=\frac{1+\frac{1}{3}}{4+\frac{2}{3}}=\frac{4}{14},\ \tilde{r}_3=\frac{1}{4+\frac{2}{3}}=\frac{3}{14}$$

$$Ent(\tilde{D}^1)=-\left(\frac{1+\frac{1}{3}}{2+\frac{1}{3}}\log_2\frac{1+\frac{1}{3}}{2+\frac{1}{3}}+\frac{1}{2+\frac{1}{3}}\log_2\frac{1}{2+\frac{1}{3}}\right)$$
$$=-\left(\frac{4}{7}\log_2\frac{4}{7}+\frac{3}{7}\log_2\frac{3}{7}\right)=0.985$$

$$Ent(\tilde{D}^2)=-\left(\frac{0}{1+\frac{1}{3}}\log_2\frac{0}{1+\frac{1}{3}}+\frac{1+\frac{1}{3}}{1+\frac{1}{3}}\log_2\frac{1+\frac{1}{3}}{1+\frac{1}{3}}\right)=0$$

$$Ent(\tilde{D}^3)=-\left(\frac{0}{1}\log_2\frac{0}{1}+\frac{1}{1}\log_2\frac{1}{1}\right)=0$$

$\Rightarrow$ $Gain(\tilde{D},色泽)=Ent(\tilde{D})-\sum_{v=1}^{3}\tilde{r}_v\,Ent(\tilde{D}^v)$

$$=0.863-\left(\frac{7}{14}*0.985+\frac{4}{14}*0+\frac{3}{14}*0\right)=0.371$$

则 $Gain(D,色泽)=\rho*Gain(\tilde{D},色泽)=\frac{14}{17}*0.371=0.305$

日期: /

缺失值问题可以从三个方面来考虑:

1. 在选择划分属性的时候，训练样本存在缺失值，如何处理？

➢ 计算划分损失减少值时，忽略特征缺失的样本，最终计算的值乘以比例（实际参与计算的样本数除以总的样本数）

假如使用ID3算法，那么选择分类属性时，就要计算所有属性的信息增益(Gain)。假设10个样本，属性是*a*,b,c。在计算a属性熵时发现，第10个样本的*a*属性缺失，那么就把第10个样本去掉，前9个样本组成新的样本集，在新样本集上按正常方法计算*a*属性的熵增。然后结果乘9/10（新样本占raw样本的比例），就是*a*属性最终的熵。

2. 分类属性选择完成，对训练样本分类，发现样本属性缺失怎么办？

➢ 将该样本分配到所有子结点中，权重由1变为具有属性*a*的样本被划分成的子集样本个数的相对比率，计算错误率的时候，需要考虑到样本权重

比如该结点是根据*a*属性划分，但是待分类样本*a*属性缺失，怎么办？假设*a*属性离散，有1,2两种取值，那么就把该样本分配到两个子结点中去，但是权重由1变为相应离散值个数占样本的比例。然后计算错误率的时候，注意，不是每个样本都是权重为1，存在分数。

3. 训练完成，给测试集样本分类，有缺失值怎么办？

➢ 分类时，如果待分类样本有缺失变量，而决策树决策过程中没有用到这些变量，则决策过程和没有缺失的数据一样；否则，如果决策要用到缺失变量，决策树也可以在当前结点做多数投票来决定（选择样本数最多的特征值方向）。