南大周志华《机器学习》课程笔记

> Introduction：最近自学机器学习课程，注意到了南京大学周志华老师的课程。我是在学堂在线平台观看的，注意到b站上也有相应视频，但b站上并未获得授权，随时有消失的可能。
>
> 周志华老师的网络教学视频中，与其西瓜书相比确实少了一些内容。但幸运的是，缺失的内容实际上对于初学者来说并不会产生太大影响。目前这一笔记也遵循视频内容，相比西瓜书中也会有一些缺失，敬请谅解。可能以后如果有机会和时间，我会再阅读周志华老师的书籍将缺失内容补全。
>
> 一切内容敬请关注我的个人Page页面。
>
> 全系列笔记请见：click here
> **About Me:**点击进入我的Personal Page

第四章 决策树

决策树基本流程

> 决策树是机器学习中最早变得重要的一个模型，可以说它是使得机器学习能够变成一门学科的模型，而其本身较为简单。

决策树基于“树”结构进行决策

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

学习过程：通过对训练样本的分析来确定“划分属性”（即内部结点所对应的属性）

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

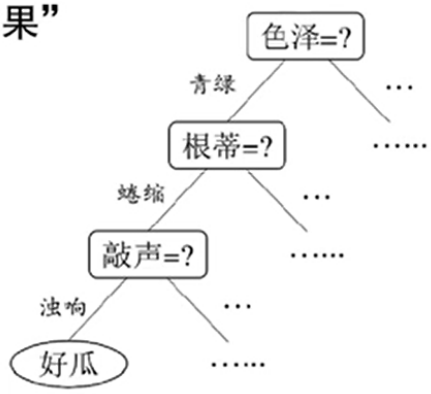


图 4.1 西瓜问题的一棵决策树

基本流程

关于递归：先看每一个节点如何划分，再考虑下一个节点。而下一个节点划分过程与根节点划分是相同的，这是十分适合计算机进行计算和迭代的一个操作。

策略：“分而治之”(divide-and-conquer)

自根至叶的递归过程

在每个中间结点寻找一个“划分”(split or test)属性

三种停止条件：

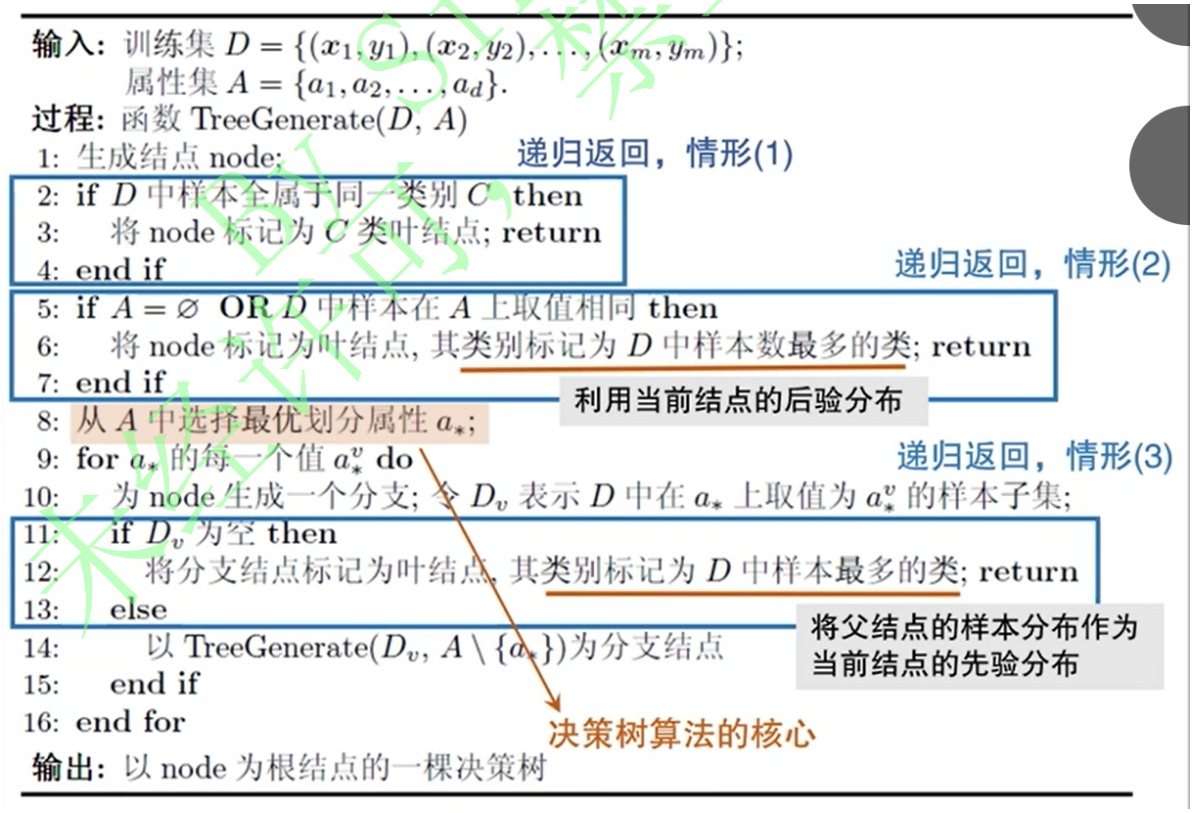
(1) 当前结点包含的样本全属于同一类别，无需划分;

(2) 当前属性集为空，或是所有样本在所有属性上取值相同，无法划分;

(3) 当前结点包含的样本集合为空，不能划分.

↗ 对于第一种停止条件，可见其为我们想要的叶节点。↗ 对于第二种停止条件，虽然当前的样本划分中存在多类的样本，但已经没有属性可以进行划分了，因此划分中哪一类的节点数量多就预测这样一个划分中节点为对应类别(例如：正类或负类)。↗ 对于第三类停止条件，可知这一属性值对应数据在样本中并未出现，因此以父节点中哪一类的节点数量多就预测这样一个划分中节点为对应类别(例如：正类或负类)。

下为决策树基本算法伪代码：



可知决策树最重要的在于选择属性进行划分时选择先选择哪一个属性来进行划分。下为一种最著名的选择方法：

信息增商(Information Gain)划分

决策树的提出很大程度上是受到了信息论的启发，因而其中很多是以信息论中的准则在作为判断依据的。而在信息论中最重要的一个量就是"熵"，其定义为$P\log_2 P$

信息熵 (entropy) 是度量样本集合“纯度”最常用的一种指标

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

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

计算信息熵时约定：若 $p=0$，则 $p\log_2 p = 0$.

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

$\mathrm{Ent}(D)$的值越小，则 D的纯度越高

信息增益直接以信息熵为基础，计算当前划分对信息熵所造成的变化

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

D^v: D 中在 a 上取值 $= a^v$ 的样本集合

以属性 a 对 数据集 D 进行划分所获得的信息增益为：

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

划分前的信息熵

划分后的信息熵

第 v 个分支的权重，样本越多越重要

ID3算法中使用

下为一个案例：

表 4.1 西瓜数据集 2.0

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

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

根结点的信息熵为

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

假设以色泽对其进行划分，可知色泽有(青绿、乌黑、浅白)，因此原始数据集D就划分成了三个数据集：

表 4.1 西瓜数据集 2.0

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

以属性“色泽”为例，其对应的3个子集分别为：

D^1(色泽=青绿)

D^2(色泽=乌黑)

D^3(色泽=浅白)

对D^1(色泽=青绿)，正例3/6，反例3/6

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

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

D^2(色泽=乌黑)，正例4/6，反例2/6

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

D^3(色泽=浅白)，正例1/5，反例4/5

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

于是，属性“色泽”的信息增益为

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

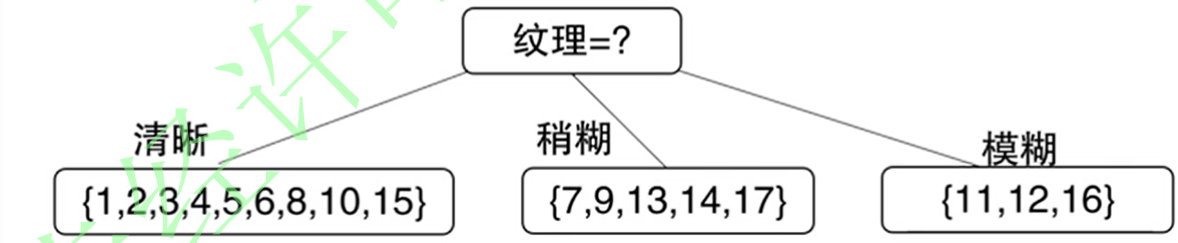
类似的，其他属性的信息增益为

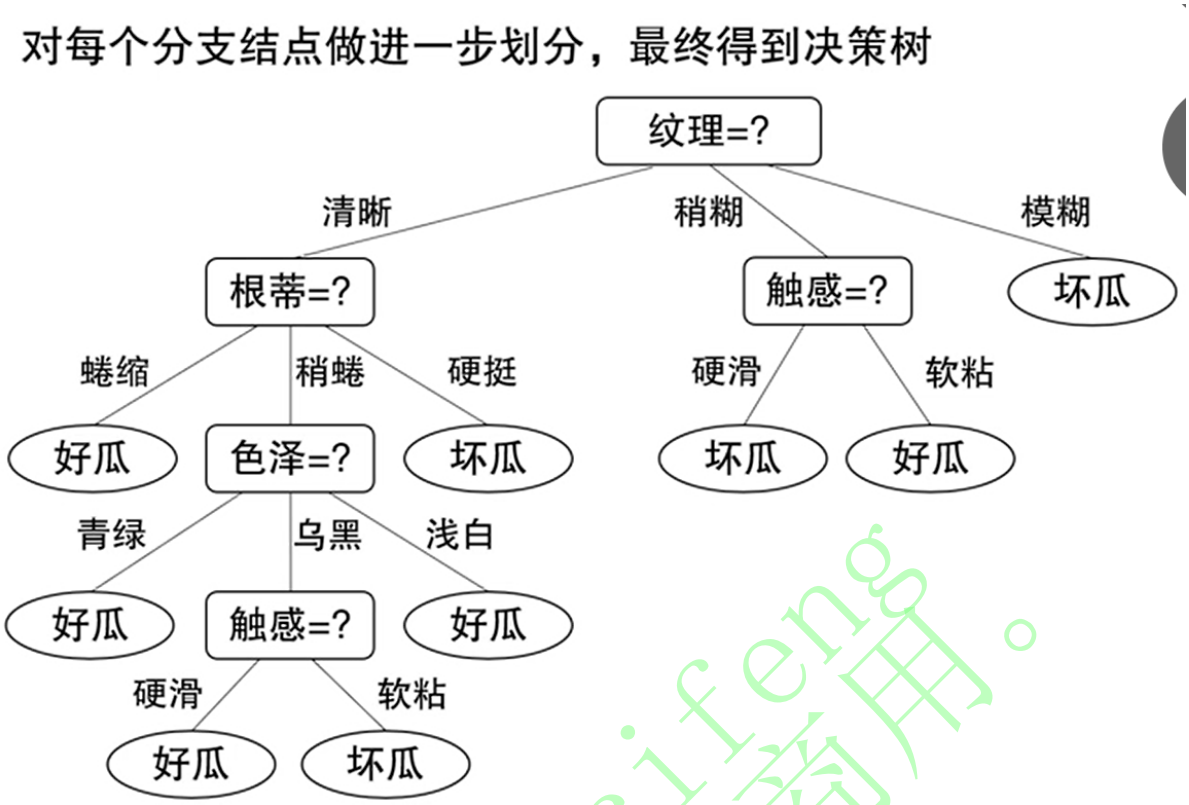
$\mathrm{Gain}(D, \text{根蒂}) = 0.143$　　$\mathrm{Gain}(D, \text{敲声}) = 0.141$

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

$\mathrm{Gain}(D, \text{触感}) = 0.006$

属性“纹理”的信息增益最大，被选为划分属性



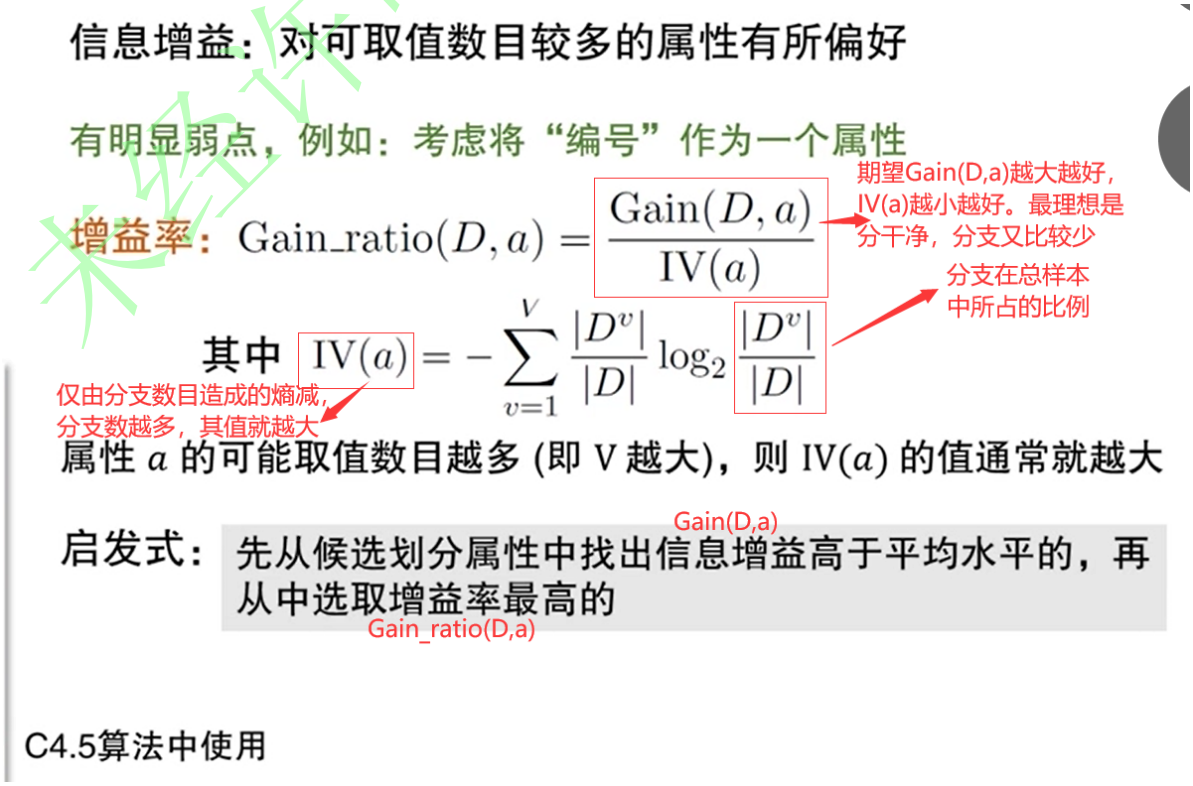


总结：信息增益最大，即每一步划分结果需要最少的信息就可以将其划分干净。最少的信息就是其已经干净了，因此信息增益是在每一步都在追求划分结果尽可能干净，这是一种贪心的策略。

其他属性划分准则

然而，信息增益采用贪心思想存在一些问题。例如，我们会对数据集中数据进行编号，假如采用编号进行划分，那么可以得到一个以编号划分的树，其每个划分中都只有一个节点。当然，实际中我们不会使用编号进行划分，但假如属性为电话呢？如果这样划分，得到的结果泛化能力必然十分糟糕。

因此，信息增益只考虑信息量的获取，其一定程度上偏好了分支多的属性。而这可能对其可能是不利的。因此C4.5对id3进行了改造，其不再采用信息增益作为划分准则，而是采用增益率(Gain Ratio)。



目前已经看了id3和C4.5两种决策树，再看一下CART决策树中的方式——基尼指数(Gini Index)

基尼指数思路是概率。假设从袋子中抓两个球，如果两个球同色，认为其较为纯净，若异色则认为其不太纯净。重复$|y|$次，即可得到Gini(D)值。

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

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，选取那个使划分后基尼指数最小的属性

CART算法中使用

> 注：k和k'分别表示好瓜和坏瓜，而不是不同的属性值。

不同划分标准下的决策树算法

可以看到，我们可以设计出非常多的划分准则，其中的关键在于如何衡量经过某个操作，后者变得更加“纯净”。信息增益是通过“信息熵”的计算来判定什么是“更加纯净”，而基尼指数中则采用概率来衡量什么是“更加纯净”。因此，我们可以通过很多种方法得到很多衡量，从而获得很多不同的决策树算法。

那么，就需要思考这些标准之间的差别有多大。在决策树经过多年发展，经历很多变体后，有人开始将目光聚焦于这一问题。

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

例如信息增益与基尼指数产生的结果，仅在约 2% 的情况下不同

决策树的剪枝

真正对决策树泛化性能影响最为显著的是剪枝而非划分属性。在数据有噪声的情况下，剪枝甚至可能将决策树的泛化性能提升25%。

Why？因为【剪枝(purning)】是决策树对付“过拟合”的主要手段。

一般在工具包中都可以选择是否进行剪枝，一般单个决策树通常是要选择进行剪枝的。但在第八章集成学习中，由于其设计观念与前述都不一样，在其中反而要选择不剪枝，具体学习到集成学习再进行介绍。但单独使用一个决策树算法，缺省情况都是需要选择进行剪枝的。

为了尽可能正确分类训练样本，有可能造成分支过多，从而产生过拟合。可以通过主动去掉一部分分支来降低过拟合的风险。其基本策略如下：

- 预剪枝(pre-pruning)：提前终止某些分支的增长
- 后剪枝(post-pruning)：生成一棵完全的决策树后，再”回头“进行剪枝

剪枝过程中需要评估剪枝前后决策树的优劣——>第2章内容

> 这里预剪枝和后剪枝的视频内容不全，但实际上看一下西瓜书80-83页的两个案例就可以很快了解预剪枝和后剪枝的流程。

缺失值的处理

在现实应用中，经常会遇到属性值“缺失”(missing)的情况。

如果仅使用无缺失的样例，则是对样例的极大浪费。而使用带缺失值的样例，需要解决下面问题：

- 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

基本思想：**样本赋权，权重划分**

下为一个案例

如果我们将下图中存在缺失值的样本扔掉，那么最终17条数据中只有4条可以用。

表 4.4　西瓜数据集 2.0α

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

因此，我们对于每个样本给予权重1(之前的决策树内容)

表 4.4 西瓜数据集 2.0α

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

仅通过无缺失值的样例来判断划分属性的优劣

学习开始时，根结点包含样例集 D 中全部17个样例，权重均为 1

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

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

注意两个权重值，因为存在缺失值，所以这里的权重值的分母不再是除以17了。

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

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

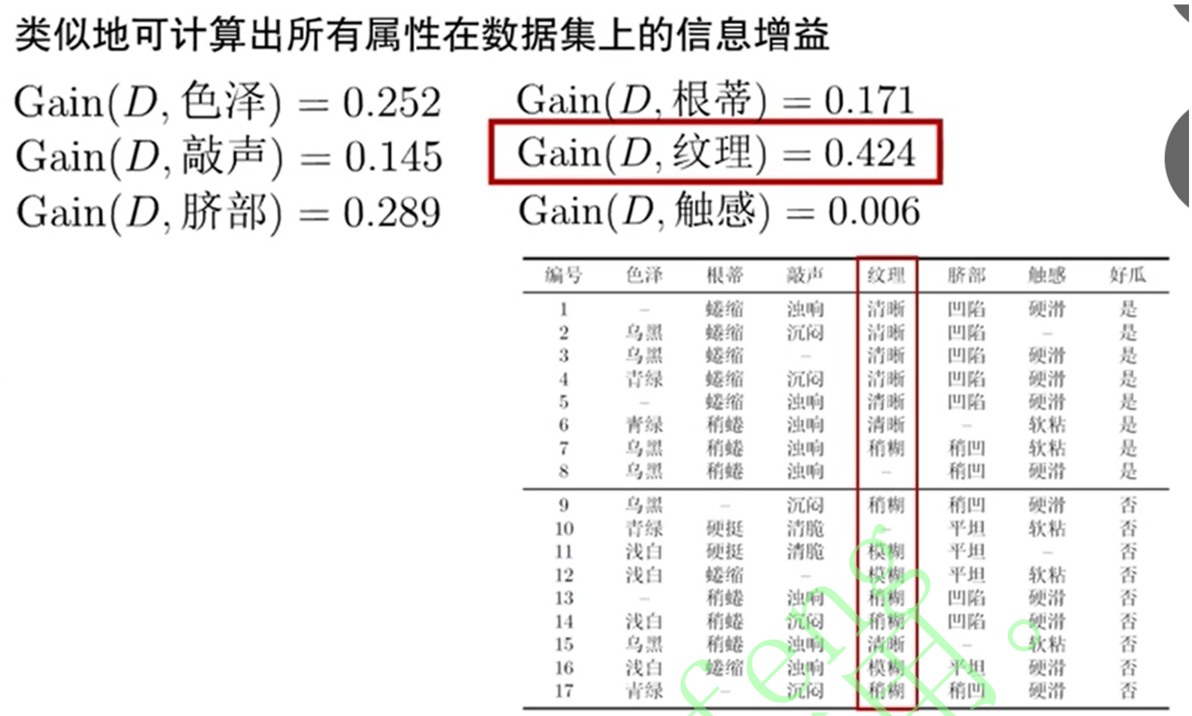
无缺失值样例中属性 a 取值为 v 的占比

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

以此类推，可以得到给定数据集上所有属性的信息增益，取最大增益属性-纹理来进行属性划分。可见纹理这一属性的属性值有{清晰、稍糊、模糊}三个值，17条数据中包含两条缺失了纹理属性值。

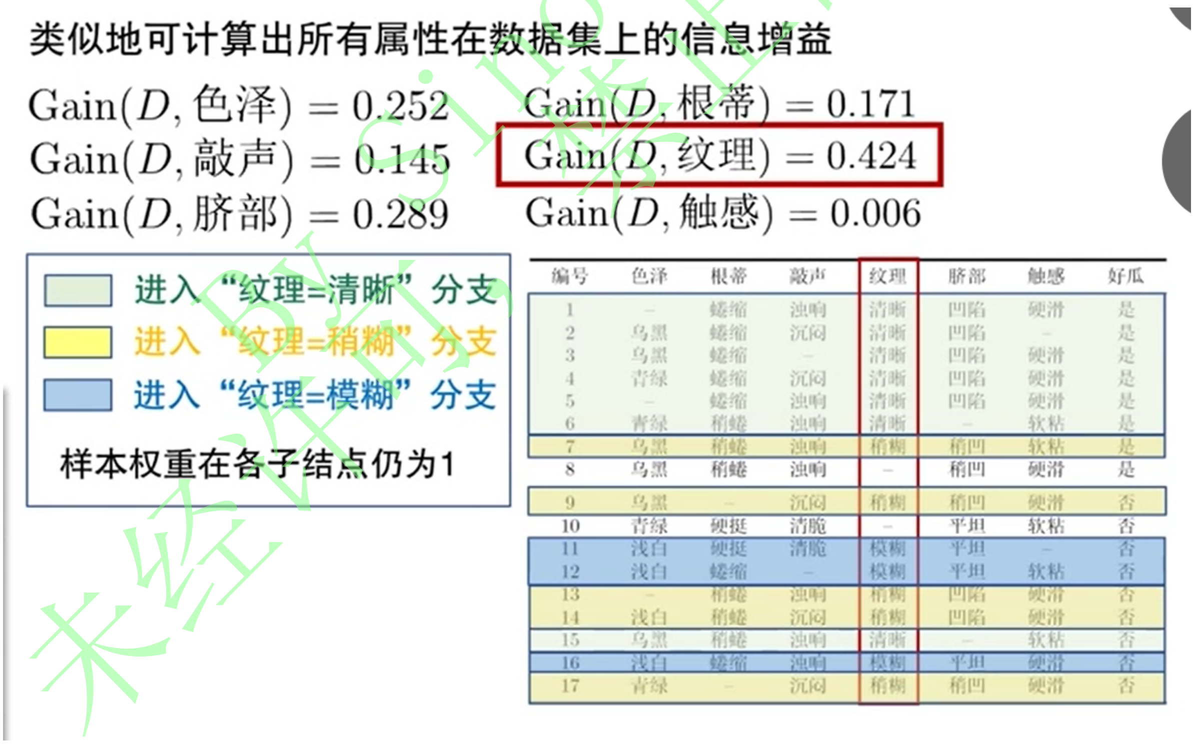


类似地可计算出所有属性在数据集上的信息增益

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

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

对“纹理”进行决策树划分，得到如下三分支划分结果，可见其中8，10号数据还没有被划分到任意一个分支中去：



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

可知，已经划分的15条数据中，清晰有7条，稍糊有5条，模糊有3条。换言之，给定一个新的西瓜，在不知道纹理的情况下，其进入这三个分支的概率分别是7/15，5/15和3/15。

因此，让样本8和10同时进入三个分支中去，其在三个分支上的权重值分别为7/15，5/15和3/15。这就是**权重划分**。

其他属性值以此类推，进行权重划分计算。