

Lecture 4: 决策树

决策树模型

决策树基础概念

概念：决策树模型

- 概念：
 - 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”或“判断”；
 - 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)，即判断结果的输出；
 - 每个“叶结点”对应于一个“预测结果”，即一种分类的结果；
 - 即：决策树模型即告诉你可能的结果，也告诉你为什么会有这个结果（决策）；
- 概念：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)
- 概念：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；

决策树的划分过程概述

概念：决策树划分过程

- 策略：
 - “分而治之”(divide-and-conquer)
- 概念：
 - 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；
 - 直到**子节点中的数据都属于同一类别**，当应用了足够多的决策条件后，会使得每个子节点的内部数据都尽可能的保持在同一类别，从而使得整个数据被这个决策树做了分类；
 - 当有了新的样本后，只知道它的特征信息，不知道它的类别，此时就可以把这个特征信息按照树的结果、从上到下，依次带入决策条件，得到这个测试样本的类别；
- 特点：
 - 优点：
 - (1) 直观，可视化；
 - (2) 易于追溯和倒推；
 - 缺点：
 - (1) 需要手动设定树最多需要预测多少次决策条件，即树有多深；
 - (2) 如果树太浅，可能会欠拟合，而如果树太深，则会过拟合；
- 问题：什么时候写递归的终止条件
 - 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；

- 假设两种属性：A、B；
- 假设当前属性全都是：is A、is B，则属性集已经空了、其属性都已经一样了，划分完成，此时必须停止；
- (3) 当前结点包含的样本集合为空，不能划分；
 - 假设用颜色划分，但是当前拿到的训练集里面，对应的属性样本为空；
- 问题：使用什么样的决策条件，来进行划分？
 - 基于信息熵和信息增益来进行划分；

概念：决策树的基本算法过程

输入：训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$;
 属性集 $A = \{a_1, a_2, \dots, a_d\}$.
 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 $A = \emptyset$ 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($D_v, A \setminus \{a_*\}$) 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

信息增益划分

信息熵

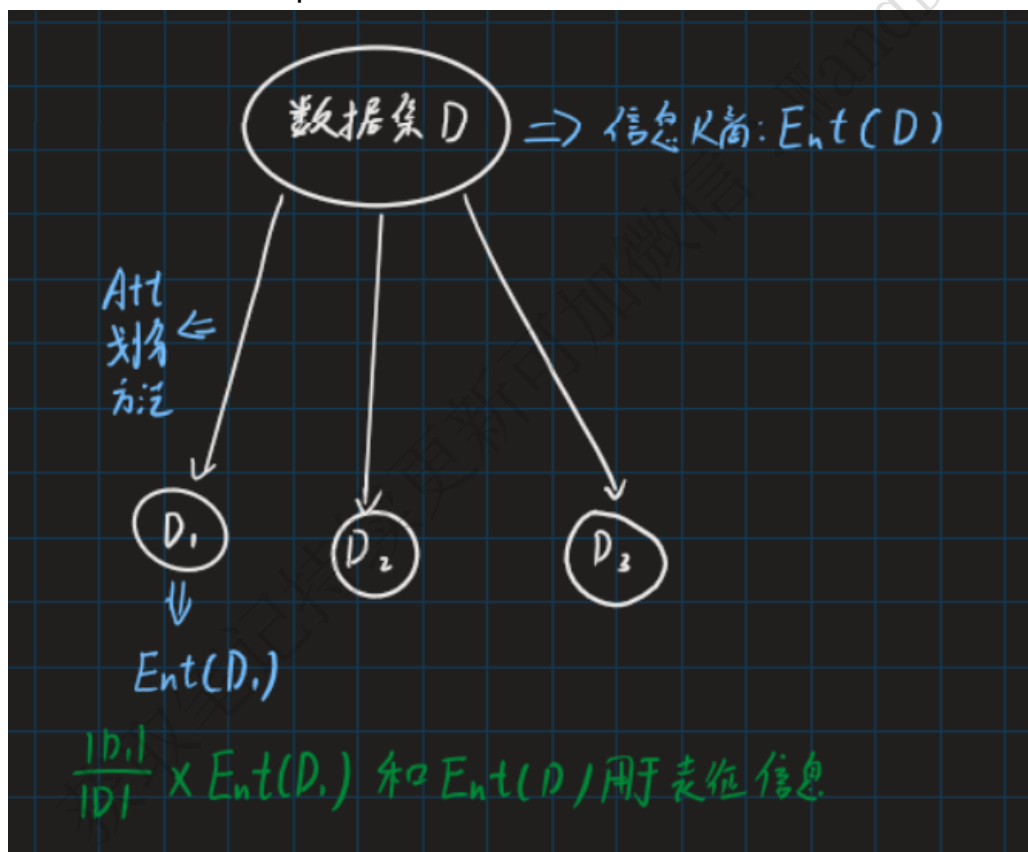
概念：基于信息增益划分

- 在进行决策树划分的过程当中，选择什么属性来对决策树中的结点中的样本进行划分？
- 让信息增益最大，也就是让每一次划分的结果只需要最少的信息、就可以将其分干净；
- 此时就需要基于信息熵进行判断；

概念：信息熵

- 概念：
 - 信息熵 (entropy) 是度量样本集合“纯度”最常用的一种指标，即衡量一个节点内的不确定性；
 - 如果一个系统的信息熵越高，则这个系统的不确定性就越大；

- 信息熵越大，样本分类越平均，信息熵越小，样本越倾向于某一类；
 - 因为决策树划分是为了划分的熵尽可能的小；
 - 定义：
 - 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：
- $$Ent(D) = - \sum_{k=1}^{|Y|} p_k \log_2 p_k$$
- 比如：假设其中一个类比的概率为 1，即当前样本全是同一类的，则此时其信息熵为 0（因为 $\log_2 1 = 0$ ），所以最小化了。当每一类都差不多时，则熵大，比较糟糕；
 - $Ent(D)$ 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；
 - 含义：
 - 为了把一个东西描述清楚，你还需要给我多少信息？
 - 如果信息熵为 0，即 $p_k = 1$ ，表示这些样本都是一类东西，也就不需要更多信息了；



信息增益

概念：信息增益

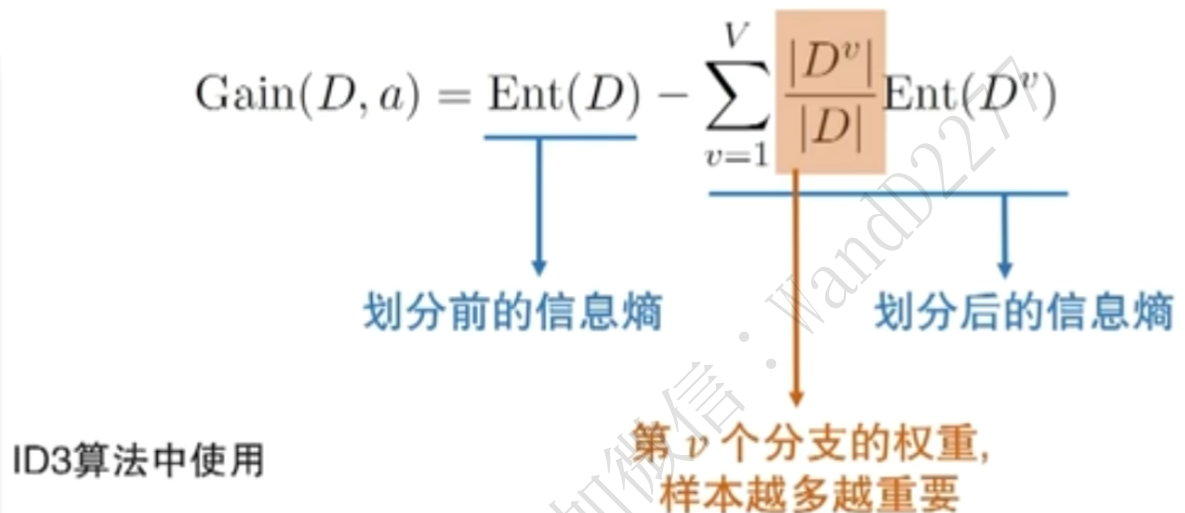
- 概念：
 - 熵的增益数学上表示为熵的变化量；
 - 在决策树模型中，信息增益表示当前：信息增益 = 根节点的熵 - 当前决策条件下所有子节点熵的和
 - 一般而言，信息增益越大，则意味着使用属性 a 来进行划分所获得的“纯度提升”越大；
- 定义：

- $$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v).$$
- 因此，可用信息增益来进行决策树的划分属性选择；

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

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：



- 注意：
 - 这一种 ID3 算法主要用于分类问题；

分析：决策树划分距离

- 图示：

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

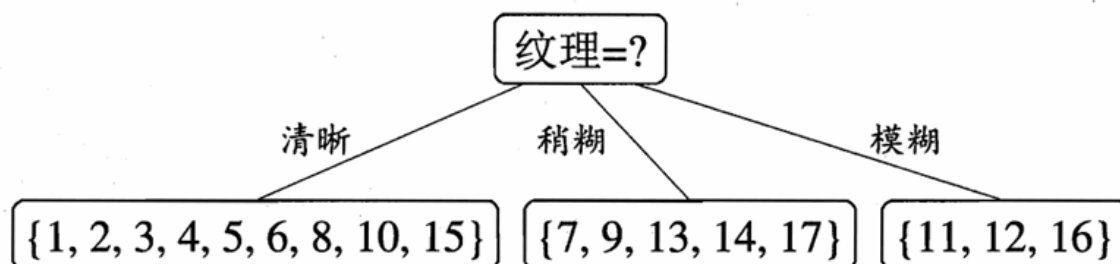
• 计算:

$$\bullet \text{Ent}(D^1) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.000, \text{Ent}(D^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918, E$$

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

$$= 0.998 - \left(\frac{6}{17} \times 1.000 + \frac{6}{17} \times 0.918 + \frac{5}{17} \times 0.722\right) = 0.109.$$

- 如果属性“纹理”的信息增益最大, 于是它被选为划分属性;
- 图给出了基于“纹理”对根结点进行划分的结果, 各分支结点所包含的样例子集显示在结点中;



其他属性划分准则

增益率

概念：增益率

- 引入：

- 假设将上述新内容的编号也作为一个候选划分属性，则可计算出它的信息增益为 0.998，远大于其他候选划分属性；
- 这很容易理解：“编号”将产生 17 个分支，每个分支结点仅包含一个样本，这些分支结点的纯度已达最大；
- 然而，这样的决策树显然不具有泛化能力，无法对新样本进行有效预测；
- 即：使用信息增益来进行划分，天然对可取值数目较多的属性有所偏好；
- 为减少这种偏好可能带来的不利影响，著名的 C4.5 决策树算法不直接使用信息增益，而是使用“增益率”(gain ratio)来选择最优划分属性；

- 定义：

$$\text{Gain_ratio}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)},$$

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

- 概念：

- $\text{IV}(a)$ 实际上就是利用计算信息熵的公式、来计算当前每个分类的数目的信息熵；
- 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；
- 增益率就是在信息增益和 IV 综合考虑下计算的；

- 实现：C4.5 算法

- 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，再从中选取增益率最高的；

概念：规范化

- 将之前不可以比较的属性、规范化后，可以进行直接比较；
- 对数值变为 0、1 后，成为归一化；

基尼指数

概念：基尼指数

- 概念：

- CART 决策树使用“基尼指数”(Gini index)来选择划分属性；
- 数据集 D 的纯度可用基尼值来度量：

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

- 属性 a 的基尼指数: $\text{Gini index}(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v)$

- 意义:

- 直观来说, $\text{Gini}(D)$ 反映了从数据集 D 中随机抽取两个样本, 其类别标记不一致的概率, 即不确定程度;
- 因此, $\text{Gini}(D)$ 越小, 则数据集 D 的纯度越高;
- 基尼指数就是用一个概率, 比如一个袋子里面抓两个球, 两个球不一样的概率小了、就表示更纯净了;

- 举例:

- 假设一个二分类问题的结果有批准和拒绝, 则: $\text{Gini} = 1 - p(\text{批准})^2 - p(\text{拒绝})^2$

剪纸处理

概念: 剪纸处理

- 引入:

- 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。剪枝方法和程度对决策树泛化性能的影响更为显著, 在数据带噪时甚至可能将泛化性能提升 25%;

- 概念: 为什么要剪枝

- 剪枝 (pruning) 是决策树学习算法对付“过拟合”的主要手段;
- 在决策树学习中, 为了尽可能正确分类训练样本, 结点划分过程将不断重复, 有时会造成决策树分支过多, 这时就可能因训练样本学得“太好”了, 以致于把训练集自身的一些特点当作所有数据都具有的一般性质而导致过拟合;
- 因此, 剪纸就是通过主动去掉一些分支来降低过拟合的风险;

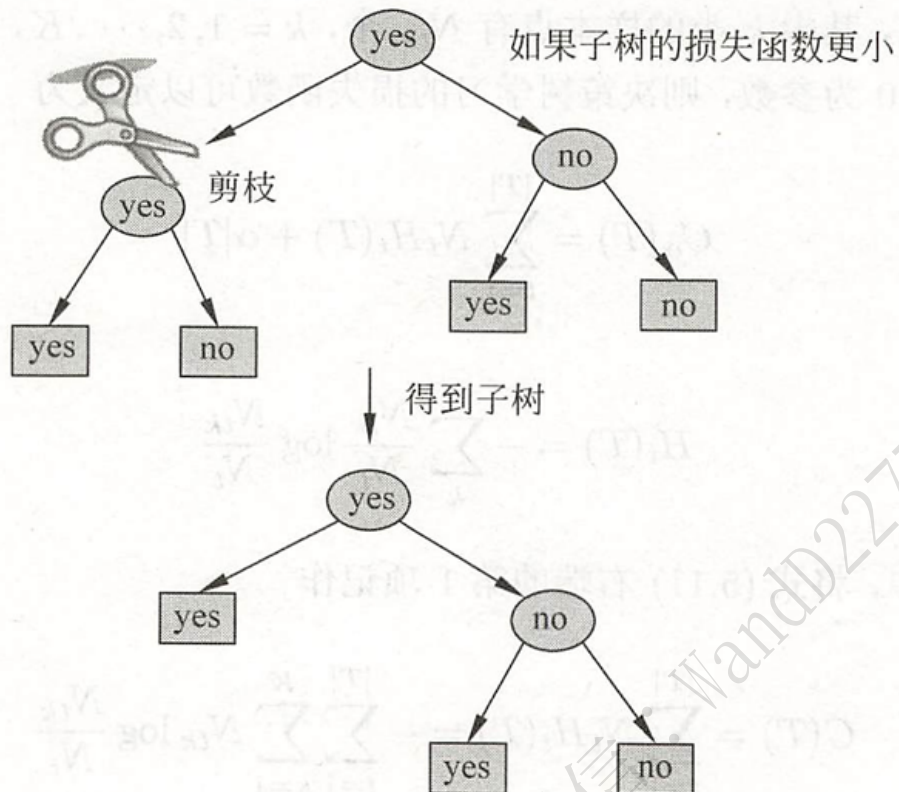
- 概念: 剪纸的策略

- 决策树剪枝的基本策略有“预剪枝”(prepruning)和“后剪枝”(post-pruning);
- 预剪枝:
 - 在决策树生成过程中, 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标记为叶结点;
 - 即: 划分前预估泛化性能
- 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;
 - 即: 划分后、对子树看是否能合并

- 使用:

- 通常, 使用单个决策树的时候, 是进行剪纸的;

- 但集成学习的时候，它的学习思路是不剪枝的；
- 图示：



缺失值

概念：缺失值

- 引入：
 - 现实应用中，经常会遇到属性值“缺失”(missing)现象；
 - 仅使用无缺失的样例是对数据的极大浪费
 - 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？
- 核心：
 - 样本赋值，权重划分；
- 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

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

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

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

因此，样本子集 $\tilde{D}$ 上属性“色泽”的信息增益为

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

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

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

无缺失值样例占比

类似地可计算出所有属性在数据集上的信息增益

$$\text{Gain}(D, \text{色泽}) = 0.252$$

$$\text{Gain}(D, \text{根蒂}) = 0.171$$

$$\text{Gain}(D, \text{敲声}) = 0.145$$

$$\text{Gain}(D, \text{纹理}) = 0.424$$

$$\text{Gain}(D, \text{脐部}) = 0.289$$

$$\text{Gain}(D, \text{触感}) = 0.006$$

 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	—	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

- 在“纹理”上出现缺失值，样本 8, 10 同时进入三个分支，三支上的权重分别为 7/15, 5/15, 3/15；

获取笔记持续更新可加微信: Wechat: 182043633