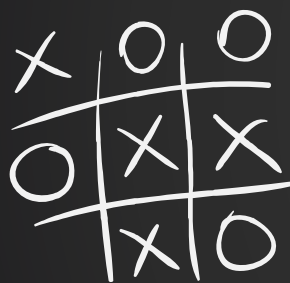


决策树



2016

PRESENTED
BY
唐汉林





决策树原理以及构成




决策树是一种利用树的结构，以分治思想为原理的机器学习方法。

由根节点，内部节点，叶节点构成的一个预测模型。树中的每个节点表示某个对象，每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。每个决策树有且仅有一个输出。





决策树适用范围

- 
- 1; 训练数据用一系列的“属性-值”来描述。
 - 2; 最终目标是对目标类进行离散性质的分割。
 - 3; 训练数据允许包含错误或可以包含缺少属性值的实例。





决策树学习的关键



决策树学习的关键是如何选择最优划分属性，随着划分的过程不断进行，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，即该结点的纯度越来越高。





信息熵与信息增益



从物理学中描述物体混乱程度的“熵”中引出，我们将信息熵定义为**离散**随机事件出现的概率，一个系统越是有序，信息熵就越低，反之一个系统越是混乱，它的信息熵就越高。所以信息熵可以被认为是系统有序化程度的一个度量。

信息熵公式：

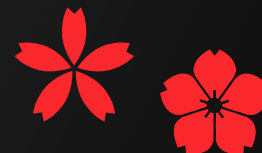
$$\text{Ent}(D) = - \sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k .$$

一个有效变量的变化情况可能越多，或者一个系统的分类可能越多，那么它携带的信息量就越大。

信息增益是针对某一特征而言的，对于一个特征，系统有它和没有它时的信息量各是多少，两者的差值就是这个特征给系统带来的信息量，即信息增益。

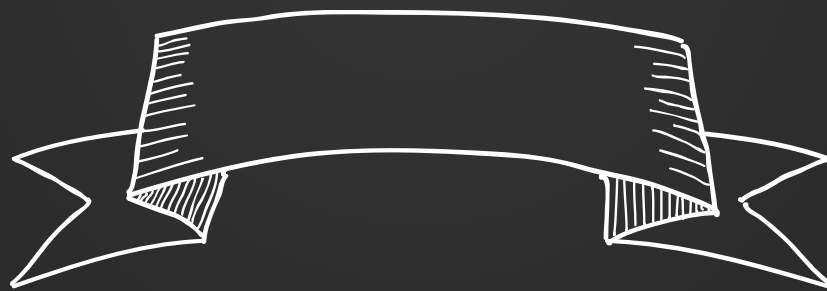
信息增益公式


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





实例下构造决策树的 ID3算法





ID3算法原理

一般而言，对于属性 a 的信息增益越大，则意味着用 a 进行划分所获得的“纯度提升”越大，ID3 (Iterative Dichotomiser) 也正是以信息增益为准则来选择划分属性。

在ID3算法构造的决策树的过程中，以尽量用较少的东西做更多的事。



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

根据所给的训练样例，用以学习一棵能预测未剖开的西瓜是不是好瓜的决策树，有 $y = 2$ ，正例占 $P1 = 8 / 17$ ，反例占 $P2 = 9 / 17$ ，于是得出根节点的信息熵：

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

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

我们应计算出当前属性集合（色泽，根蒂，敲声……）中每个属性的信息增益，以色泽为例，它有三个可能的取值{青绿，乌黑，浅白}。若用该属性对D进行划分。可以得到三个子集，再根据三个子集中正例与反例的占比可以计算出用“色泽”划分后的信息熵与信息增益：

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

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

$$\text{Ent}(D^3) = - \left(\frac{1}{5} \log_2 \frac{1}{5} + \frac{4}{5} \log_2 \frac{4}{5} \right) = 0.722 ,$$

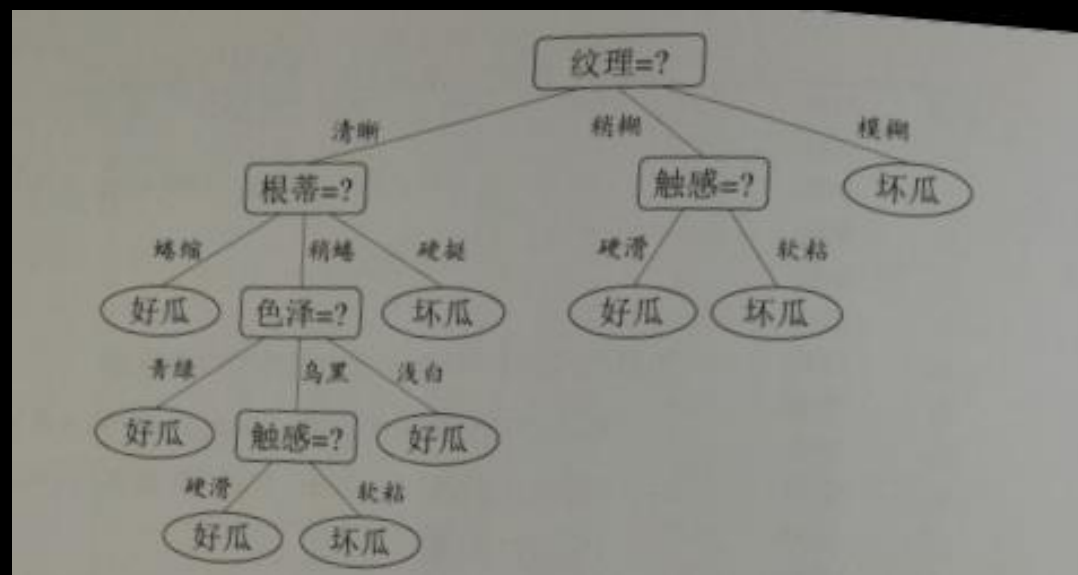
$$\begin{aligned} \text{Gain}(D, \text{色泽}) &= \text{Ent}(D) - \sum_{v=1}^3 \frac{|D^v|}{|D|} \text{Ent}(D^v) \\ &= 0.998 - \left(\frac{6}{17} \times 1.000 + \frac{6}{17} \times 0.918 + \frac{5}{17} \times 0.722 \right) \\ &= 0.109 . \end{aligned}$$


类似的我们可以计算出其他属性的信息增益

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

显然, 属性“纹理”的信息增益最大, 于是选择纹理作为当前的划分属性, 同时我们得到了三个内部节点, 对于每个结点再次进行以上操作, 求出除了“纹理”以外的各属性的信息增益。进而得到进一步的结点

最终得到决策树:






ID3小结

由最终ID3算法构造的决策树来看，信息增益最大的分类属性将会被首先判定，所以ID3算法是寻求当前最优解的贪婪算法，采用自顶向下的贪婪搜索遍历可能的决策空间。





ID3小结

- 
- 1; 根据常识, 编号与西瓜的甜度无关系, 所以对于信息增益为0.998的“编号”属性将在决策树的构造过程中直接忽略。因此并不能单纯的根据信息增益的高低来决定能否都成为划分属性。
 - 2; 例子中的训练样例的属性均为离散的, 在决策数据中对于连续性属性的使用ID3存在盲点。
 - 3; 对于缺失值的处理存在缺陷。
 - 4; 当训练集增加时, ID3的算法的决策树会随之变化, 不便于渐进学习。
 - 5; 每个节点仅含一个特征, 特征之间的相关性强调不够。





解决方法——增益率

实际上，信息增益准则对可取值数目较多的属性有所偏好，为降低这种方法的不利影响，在ID3的基础上改进的C4.5算法使用了增益率来辅助选择最优划分属性。

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

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





解决方法——基尼指数

基尼指数为CART决策树选择划分属性的评判标准



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

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





解决方法——剪枝处理

剪枝是决策树学习算法中预防过拟合的主要手段，即将对决策树中的部分子树替换成叶节点，可以提升决策树的泛化性。





泛化性的判定

采用留出法，即随机预留出一部分数据用作验证集以进行性能评估。并根据训练集再作决策树



表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

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





剪枝处理——预剪枝

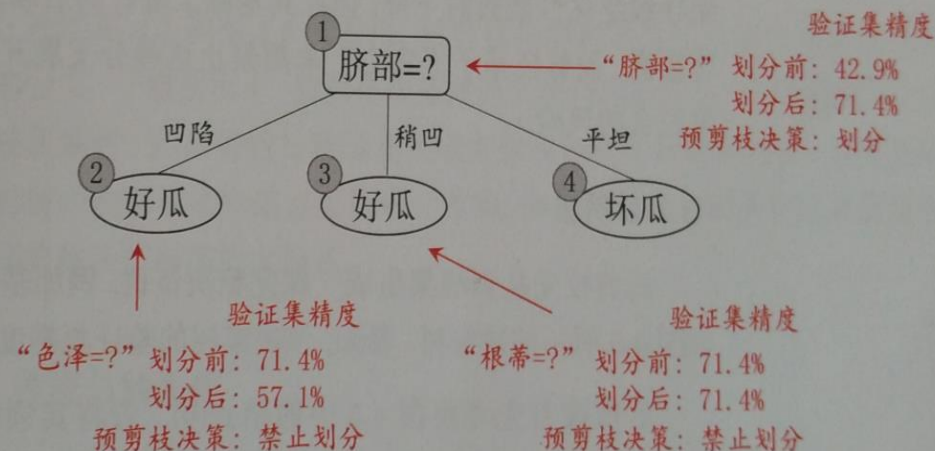
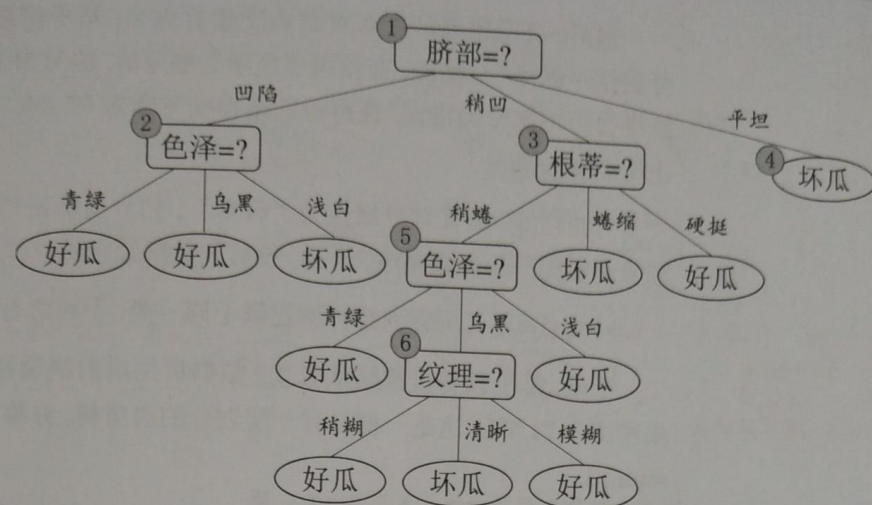
在决策树生成过程中，对每个节点在划分前先进行估计，若当前节点的划分不能带来决策树的泛化性提升，则停止划分并将当前结点标记为叶节点。



表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

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





预剪枝——小结

对比剪枝前后的决策树可以看出，预剪枝使得决策树的很多分支没有展开，这不仅降低了过拟合的风险，还显著减少了决策树的训练时间和测试时间的开销。但是预剪枝是基于局部最优的贪婪算法，无法保证全局最优，因此预剪枝给决策树带来了欠拟合的风险。





剪枝——后剪枝

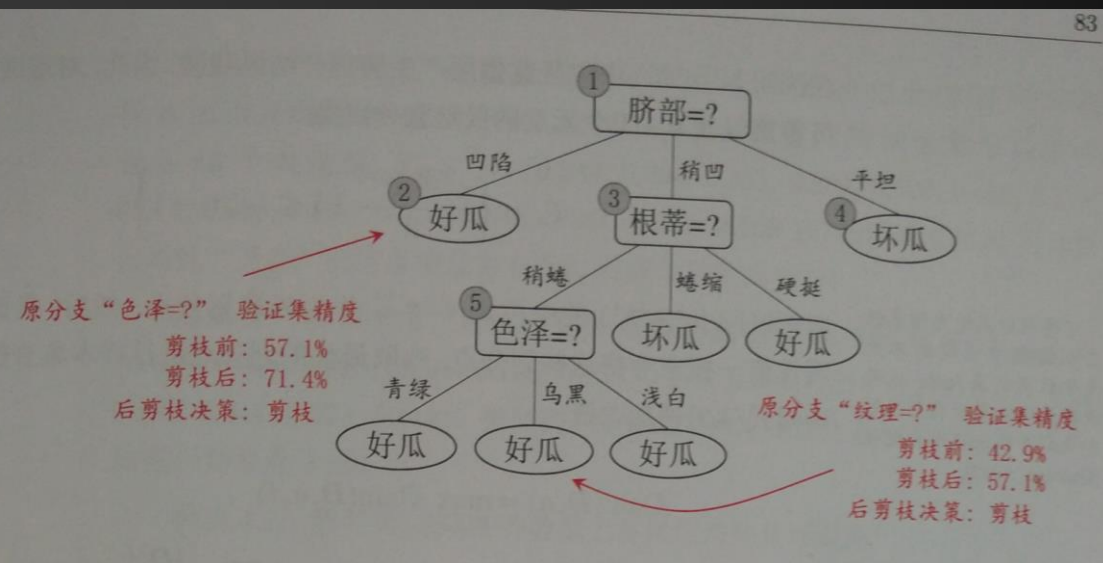
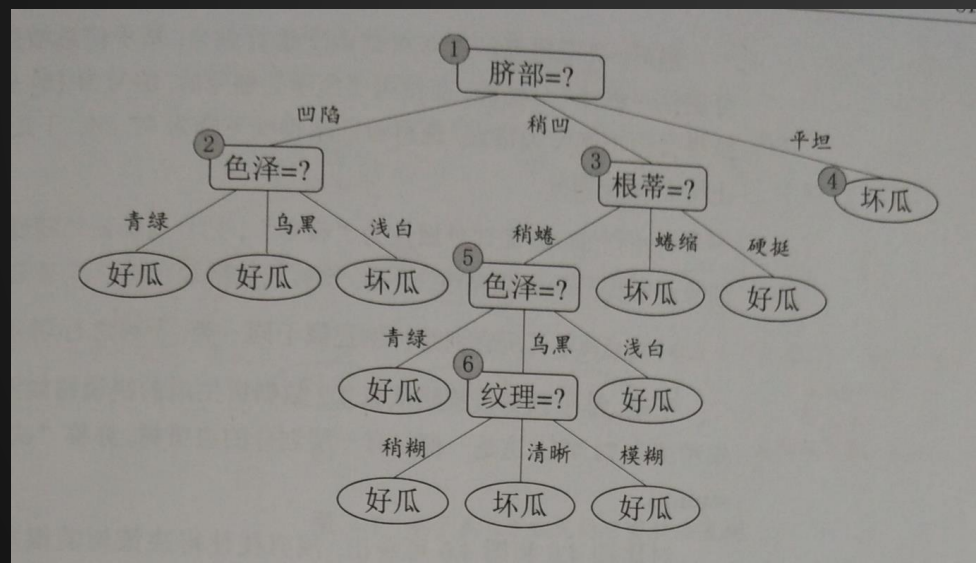
后剪枝先从训练集生成一棵完整的决策树，
然后自底向上地对非叶节点进行考察，若将
该节点对应的子树替换为叶节点能带来决策
树泛化性的提升，则进行替换



表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

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	清晰	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	硬粘	否
16	浅白	蜷缩	浊响	模糊	平坦	软滑	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





后剪枝——小结

后剪枝决策树通常比预剪枝决策树保留了更多的分支，欠拟合性风险小于预剪枝，泛化性性能优于预剪枝，但是训练时间开销比未剪枝决策树和预剪枝决策树要大很多。





解决方法——连续值处理

由于连续属性的可取值数目不再有限，因此，不能直接根据连续属性的可取值进行划分，此时，我们可以使连续属性离散化，最简单的策略是采用二分法，这也是C4.5决策树算法中采用的机制。





连续值处理——二分法

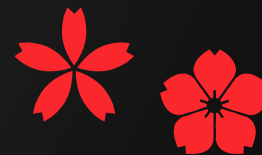


对于连续属性 a 在 D 中出现的 n 个不同的取值, 先将这些值进行从小到大的排序, 取 t 为相邻两元素的中点, 我们可以得出候选划分点的集合:

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \mid 1 \leq i \leq n-1 \right\}$$

我们就可像离散属性值一样来考察这些划分点, 进而选取最优的划分点进行样本集合的划分

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





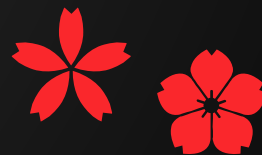
连续值处理——二分法



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，在决策树学习开始时，根节点包含的17个训练数据，因此该属性的候选划分点包含16个候选值： $T_{(\text{密度})} = \{0.244, 0.294, 0.351, 0.381, \dots\}$

根据公式得出属性“密度”的最优增益率为0.262，对应划分点0.81。

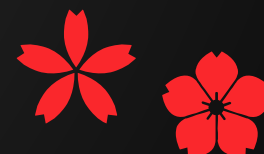




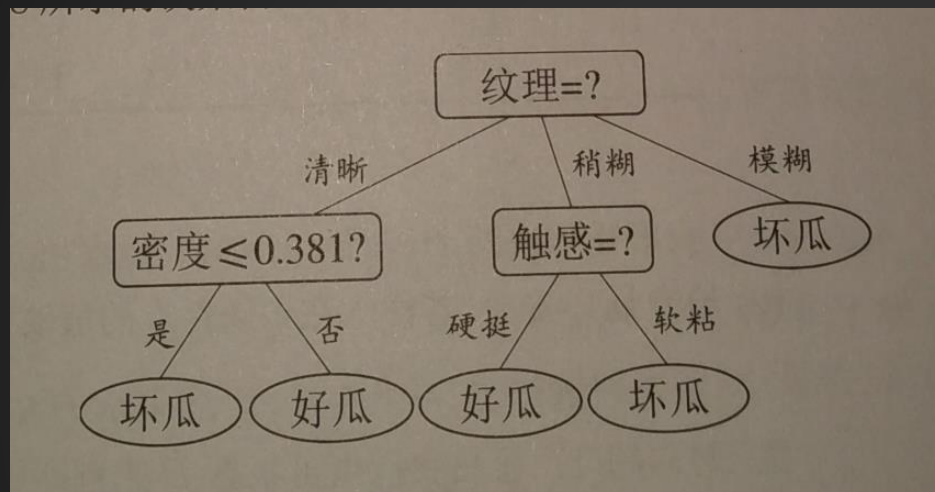
连续值处理——二分法

同理我们得出含糖率的最优信息增益以及其他属性的信息增益：

$$\begin{aligned} \text{Gain}(D, \text{色泽}) &= 0.109; & \text{Gain}(D, \text{根蒂}) &= 0.143; \\ \text{Gain}(D, \text{敲声}) &= 0.141; & \text{Gain}(D, \text{纹理}) &= 0.381; \\ \text{Gain}(D, \text{脐部}) &= 0.289; & \text{Gain}(D, \text{触感}) &= 0.006; \\ \text{Gain}(D, \text{密度}) &= 0.262; & \text{Gain}(D, \text{含糖率}) &= 0.349. \end{aligned}$$



连续值处理——二分法

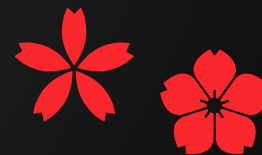


与离散属性不同的是，若当前节点划分属性为连续属性，该属性还可作为其后代节点的划分属性。



存在的问题

由于自身数学能力的限制，对于通过加权进行缺失值处理的算法的存在着不理解，希望能在今后数学能力提升之后继续对决策树的研究。





参考文献

《机器学习理论及运用》——李凡长

《机器学习》——周志华

《机器学习》——Tom M. Mitchell

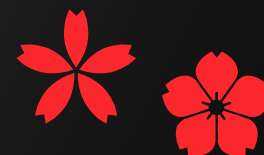
《机器学习理论与算法》——张燕平

《机器学习方法》

《机器学习导论》

《人工智能技术导论》

《机器学习与数据挖掘：方法和应用》



☀️ 加油! ٩('ω' *)و ✨

