第六讲 决策树

高级机器学习

基本流程

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

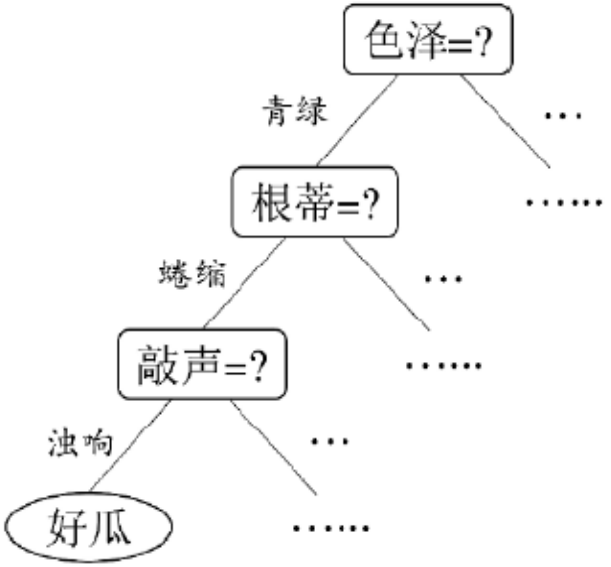


图 4.1 西瓜问题的一棵决策树

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.

划分选择

决策树学习的关键在于选择最优划分标准

- 直觉上，决策树的分支结点所包含的样本尽可能属于同一类别，即结点的“纯度”(purity) 越高越好

- 经典划分方法：
  - 信息增益
  - 增益率

划分选择-信息增益

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

“信息熵”(information 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 $(k=1,2,\ldots,|\mathcal{Y}|)$, 则 D 的信息熵定义为

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

$\mathrm{Ent}(D)$ 的值越小, 则 D 的纯度越高.

划分选择-信息增益

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

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. 因此, 我们可用信息增益来进行决策树的划分属性选择, 即在图 4.2 算法第 8 行选择属性 $a_*=\arg\max_{a\in A}\mathrm{Gain}(D,a)$. 著名的 ID3 决策树学习算法 [Quinlan, 1986] 就是以信息增益为准则来选择划分属性.

信息增益偏好取值数多的属性

划分选择-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 [Quinlan, 1993] 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. 采用与式(4.2)相同的符号表示，增益率定义为

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

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

相当于给信息增益做了一个规范化

增益率准则偏好取值数较少的属性

划分选择-信息增益

信息增益实例

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

该数据集包含17个训练样本，$|\mathcal{Y}| = 2$，其中正例占 $p_1 = \frac{8}{17}$，反例占 $p_2 = \frac{9}{17}$，计算得到根结点的信息熵为

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

划分选择-信息增益

- 以属性“色泽”为例，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 $p_2 = \frac{3}{6}$; D^2 包含编号为 $\{2, 3, 7, 8, 9, 15\}$ 的 6 个样例, 其中正、反例分别占 $p_1 = \frac{4}{6}$, $p_2 = \frac{2}{6}$; D^3 包含编号为 $\{5, 11, 12, 14, 16\}$ 的 5 个样例, 其中正、反例分别占 $p_1 = \frac{1}{5}$, $p_2 = \frac{4}{5}$. 根据式(4.1)可计算出用“色泽”划分之后所获得的 3 个分支结点的信息熵为

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

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

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

划分选择-信息增益

- 属性“色泽”的信息增益为

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

划分选择-信息增益

- 类似的，其他属性的信息增益为

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

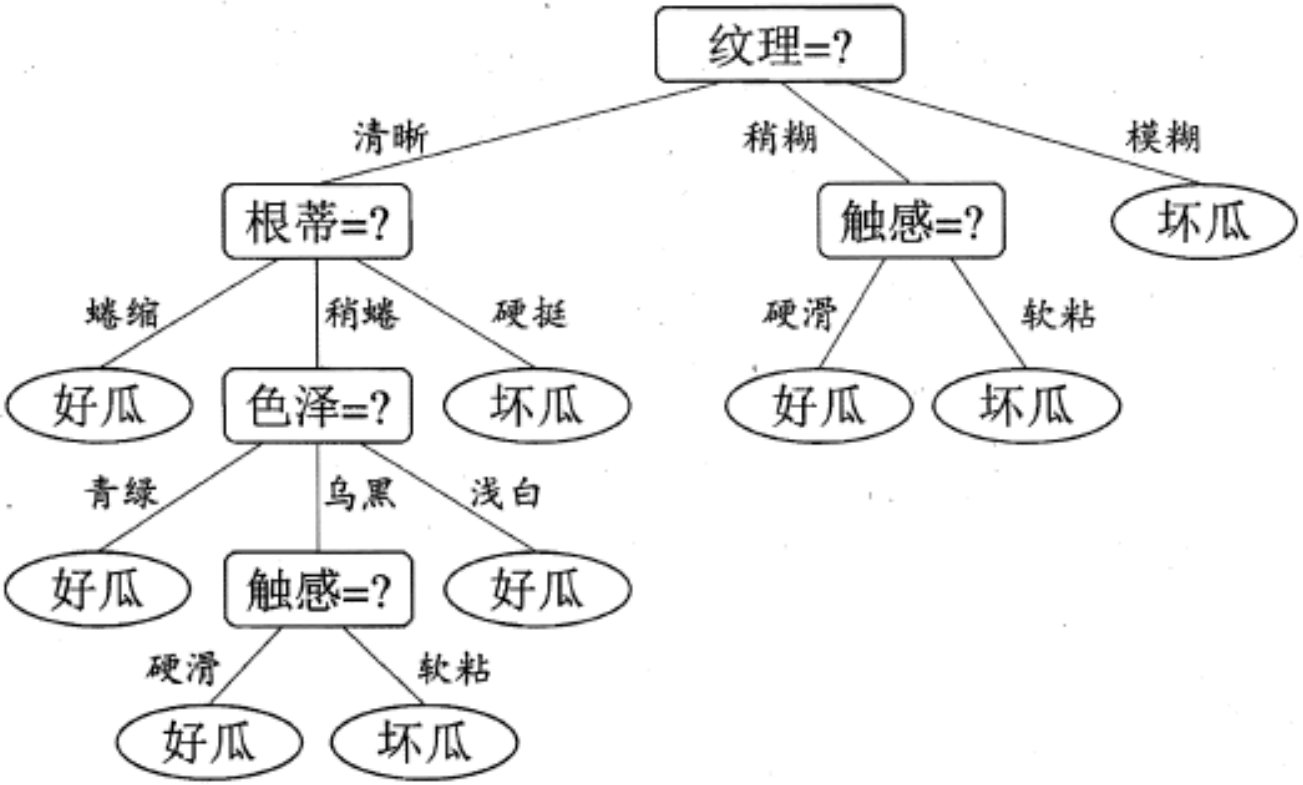
$$\text{Gain}(D, \text{纹理}) = 0.381 \qquad \text{Gain}(D, \text{脐部}) = 0.289$$

$$\text{Gain}(D, \text{触感}) = 0.006$$

- 显然，属性“纹理”的信息增益最大，其被选为划分属性

划分选择-信息增益

- 决策树学习算法将对每个分支结点做进一步划分，最终得到的决策树如图：



过拟合和剪枝

- 决策树的不足，过拟合
  - 决策树决策分支过多，以致于把训练集自身的一些特点当做所有数据都具有的一般性质而导致的过拟合
- 剪枝的基本策略
  - 预剪枝：边建树，边剪枝
  - 后剪枝：先建树，后剪枝
- 判断决策树泛化性能是否提升的方法
  - **留出法：预留一部分数据用作“验证集”以进行性能评估**

剪枝处理

数据集

训练集

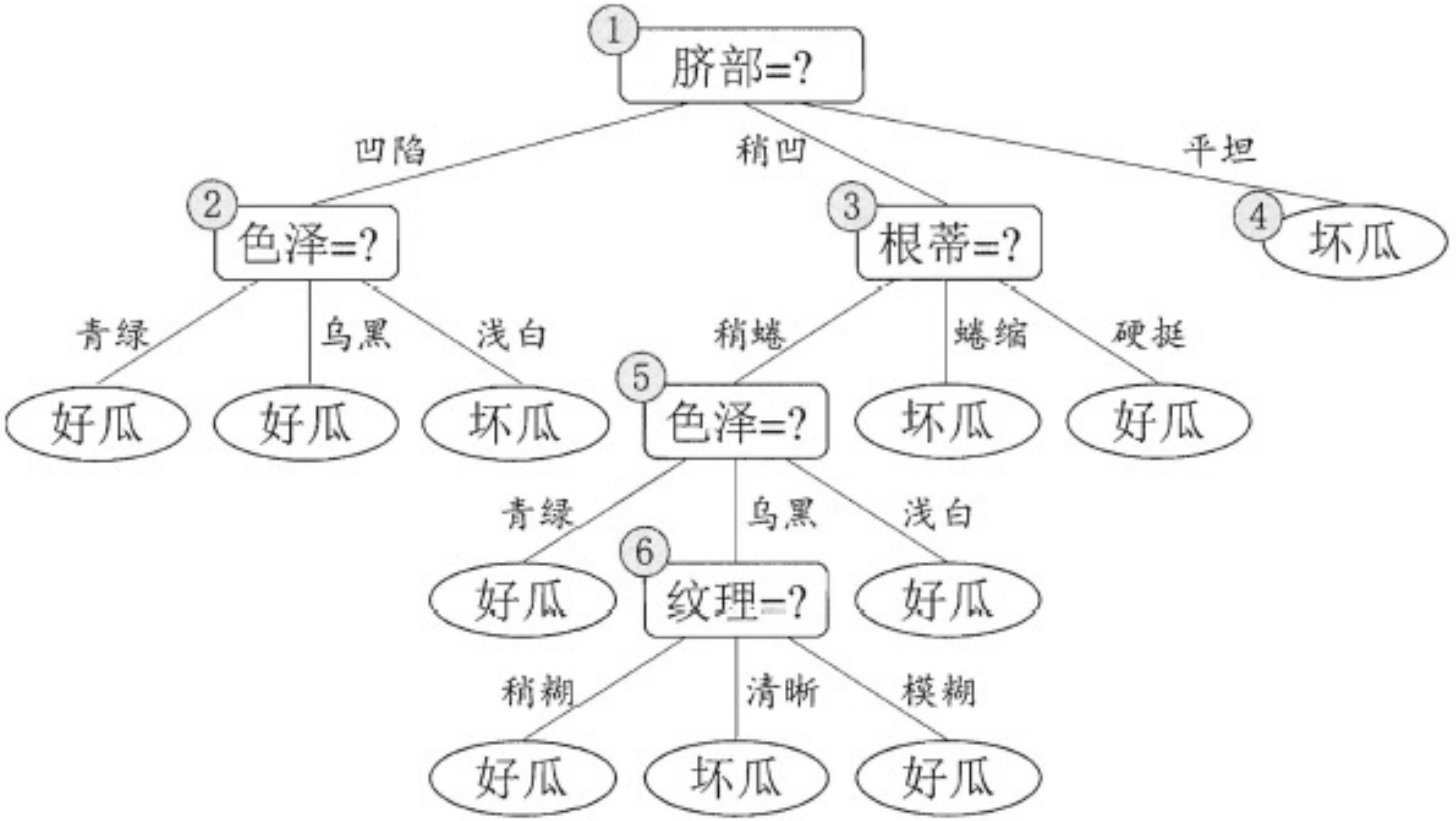
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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

剪枝处理

未剪枝决策树


1
脐部=?
凹陷
稍凹
平坦
2
色泽=?
3
根蒂=?
4
坏瓜
青绿
乌黑
浅白
稍蜷
蜷缩
硬挺
好瓜
好瓜
坏瓜
5
色泽=?
坏瓜
好瓜
青绿
乌黑
浅白
6
好瓜
纹理=?
好瓜
稍糊
清晰
模糊
好瓜
坏瓜
好瓜

剪枝处理-预剪枝

思路：边建树，边剪枝

- 思路：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记为叶结点，其类别标记为训练样例数最多的类别

剪枝处理-预剪枝

(1) 针对上述数据集，基于信息增益准则，选取属性“脐部”划分训练集。

(2) 分别计算划分前（即直接将该结点作为叶结点）及划分后的验证集精度，判断是否需要划分。

(3) 若划分后能提高验证集精度，则划分，对划分后的属性，执行同样判断；否则，不划分

剪枝处理-预剪枝

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

结点1：若不划分，则将其标记为叶结点，类别标记为训练样例中最多的类别，即好瓜。验证集中，$\{4, 5, 8\}$被分类正确，得到验证集精度为$\frac{3}{7} \times 100\% = 42.9\%$

验证集精度

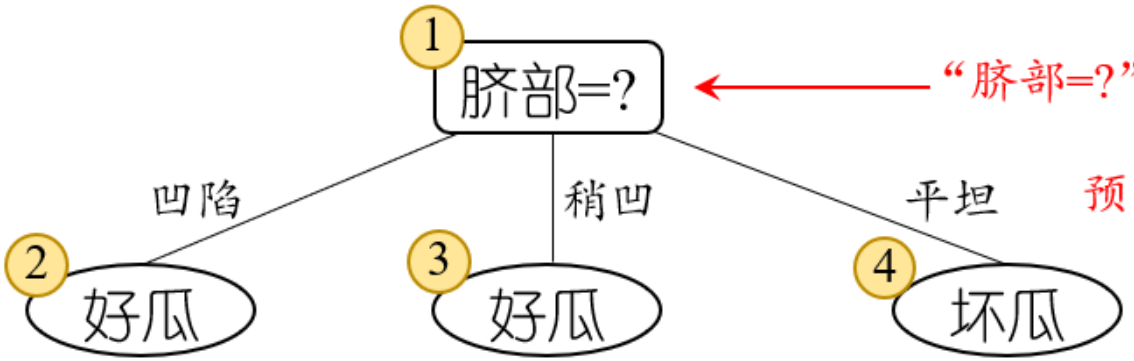
剪枝处理-预剪枝

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

结点1：若划分，根据结点②，③，④的训练样例，将这3个结点分别标记为“好瓜”、“好瓜”、“坏瓜”。此时，验证集中编号为$\{4, 5, 8, 11, 12\}$的样例被划分正确，验证集精度为$\frac{5}{7} \times 100\% = 71.4\%$

验证集精度



“脐部=?” 划分前: 42.9%

划分后: 71.4%

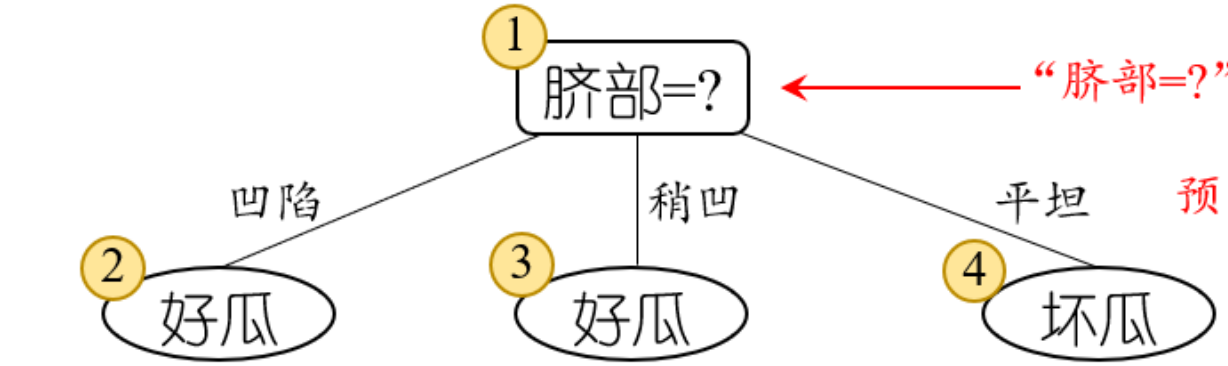
预剪枝决策: 划分

剪枝处理-预剪枝

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

对结点②，③，④分别进行剪枝判断，结点②，③都禁止划分，结点④本身为叶子结点。最终得到仅有一层划分的决策树，称为“决策树桩”



验证集精度

“脐部=?” 划分前: 42.9%

划分后: 71.4%

预剪枝决策: 划分

“色泽=?”

验证集精度

划分前: 71.4%

划分后: 57.1%

预剪枝决策: 禁止划分

“根蒂=?”

验证集精度

划分前: 71.4%

划分后: 71.4%

预剪枝决策: 禁止划分

剪枝处理-预剪枝

预剪枝的优缺点

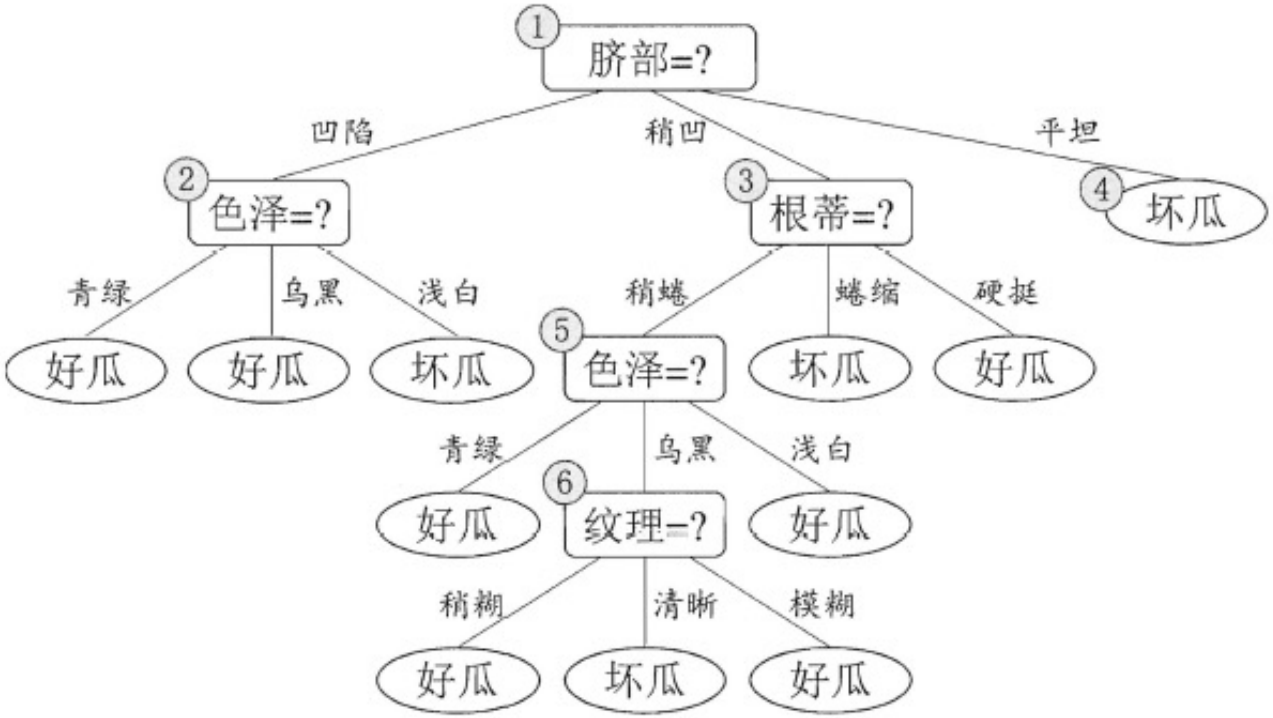
- 优点
  - 降低过拟合风险
  - 显著减少训练时间和测试时间开销

- 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

剪枝处理-后剪枝

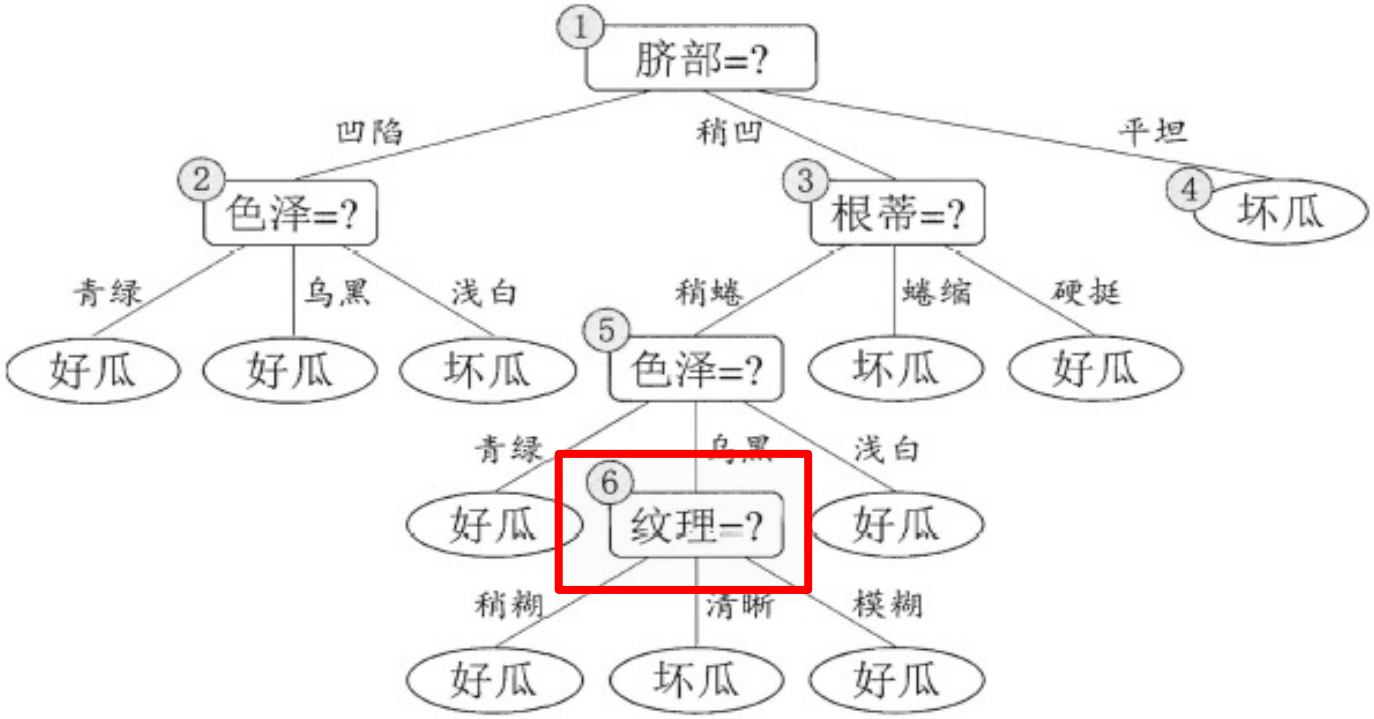
思路：先建树，后剪枝

首先生成一棵完整的决策树，该决策树的验证集精度为 42.9%



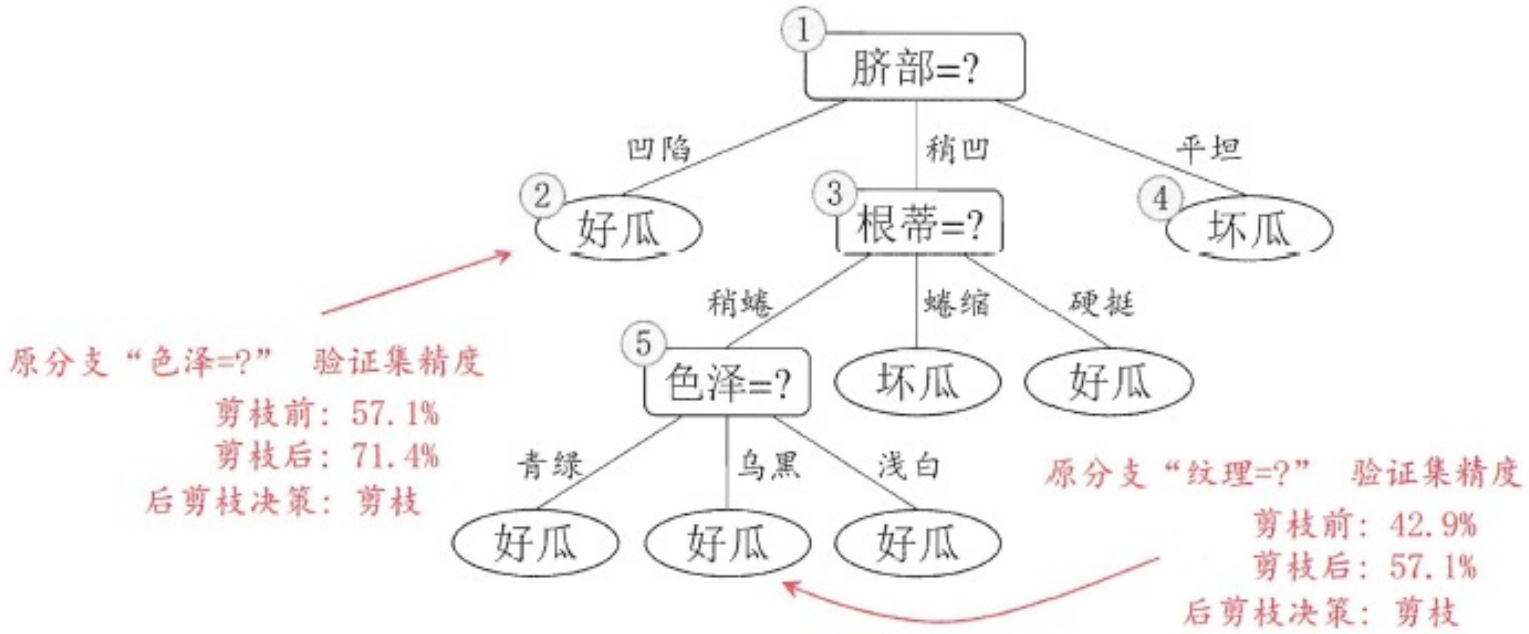
剪枝处理-后剪枝

后剪枝首先考察图 4.5 中的结点⑥. 若将其领衔的分支剪除, 则相当于把⑥ 替换为叶结点. 替换后的叶结点包含编号为 $\{7, 15\}$ 的训练样本, 于是, 该叶结点的类别标记为“好瓜”, 此时决策树的验证集精度提高至 57.1%. 于是, 后剪枝策略决定剪枝, 如图 4.7 所示.



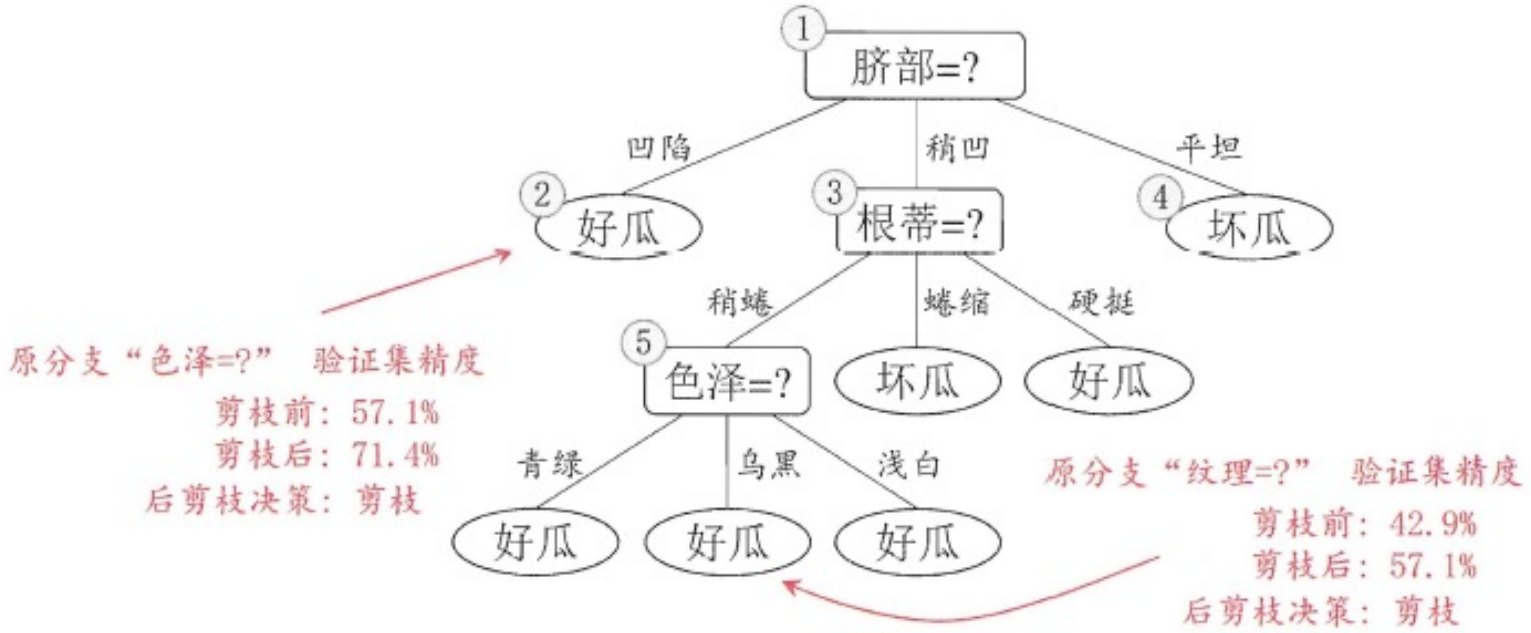
剪枝处理-后剪枝

然后考察结点⑤, 若将其领衔的子树替换为叶结点, 则替换后的叶结点包含编号为 $\{6,7,15\}$ 的训练样例, 叶结点类别标记为“好瓜”, 此时决策树验证集精度仍为 57.1%. 于是, 可以不进行剪枝.



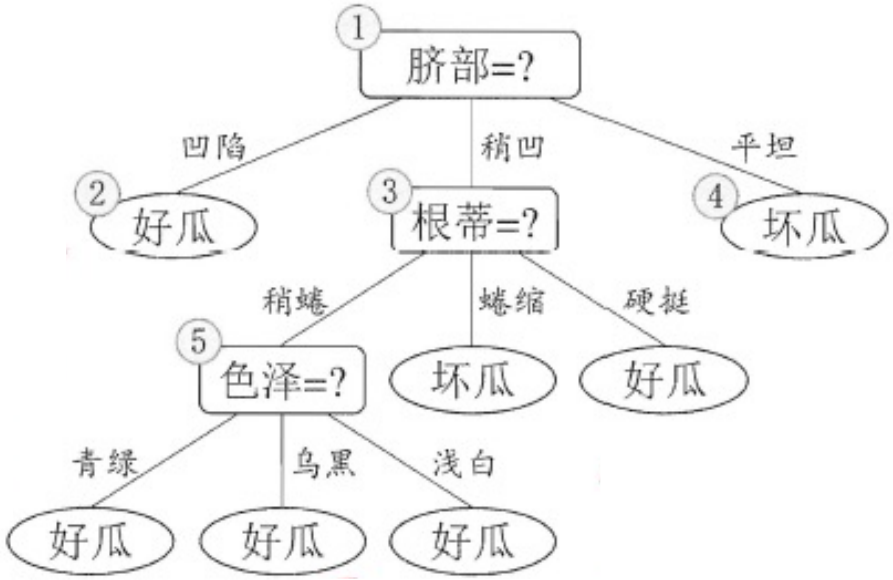
剪枝处理-后剪枝

后剪枝首先考察图 4.5 中的结点⑥. 若将其领衔的分支剪除, 则相当于把⑥ 替换为叶结点. 替换后的叶结点包含编号为 $\{7, 15\}$ 的训练样本, 于是, 该叶结点的类别标记为“好瓜”, 此时决策树的验证集精度提高至 57.1%. 于是, 后剪枝策略决定剪枝, 如图 4.7 所示.



剪枝处理-后剪枝

- 最终，基于后剪枝策略得到的决策树如图所示



剪枝处理-后剪枝

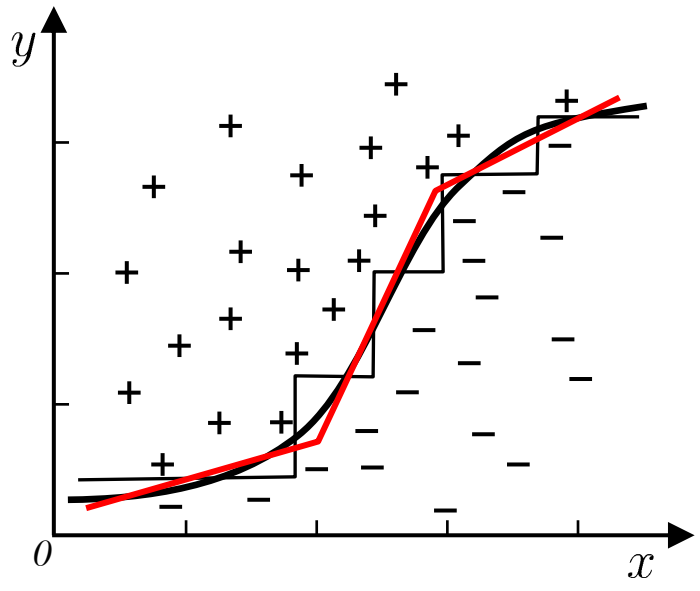
后剪枝的优缺点

- 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；其训练时间要远大于预剪枝决策树

多变量决策树

- 单变量决策树分类边界:轴平行
- 多变量决策树

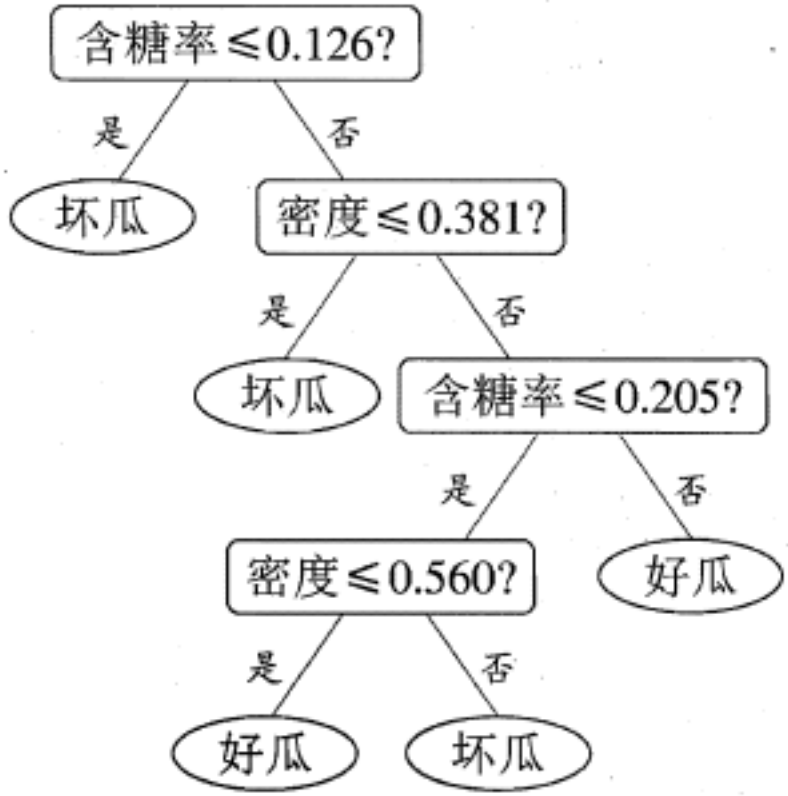


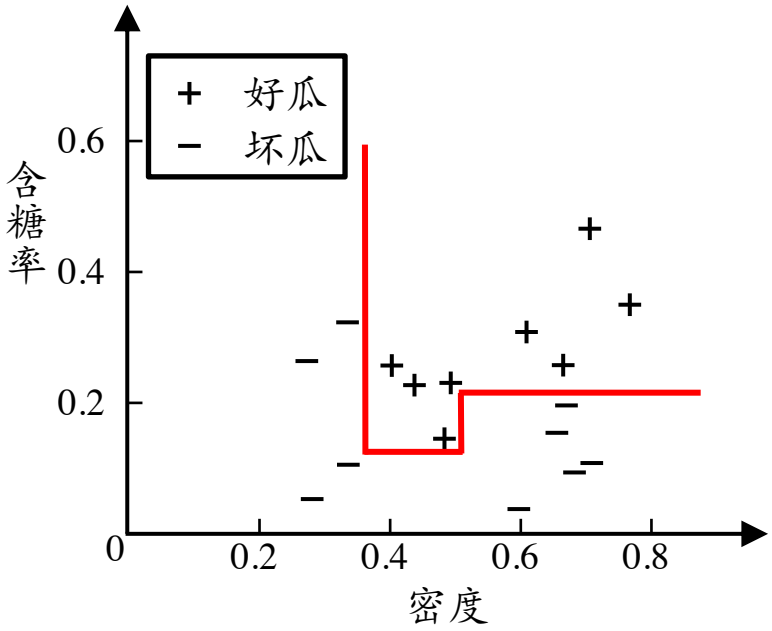
- 非叶节点不再是仅对某个属性, 而是对属性的线性组合

- 每个非叶结点是一个形如 $\sum_{i=1}^{d} w_i a_i = t$ 的线性分类器，其中 w_i 是属性 a_i 的权值，w_i 和 t 可在该结点所含的样本集和属性集上学得

多变量决策树

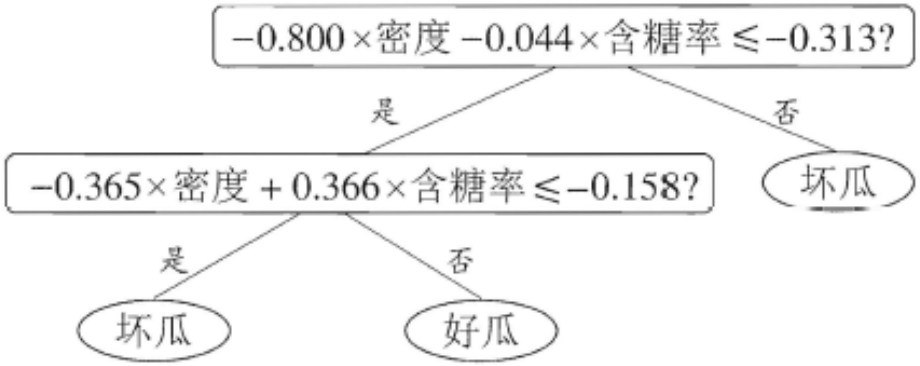
- 单变量决策树

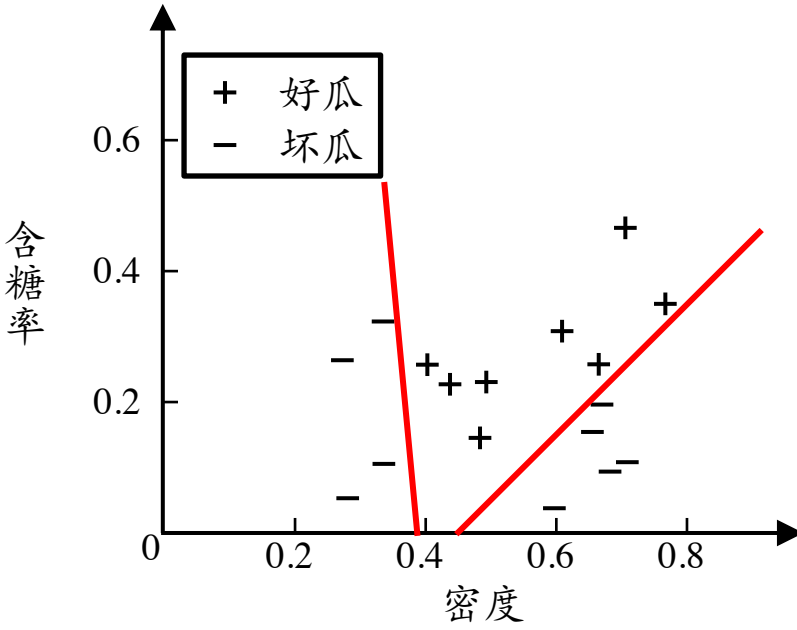




多变量决策树

- 多变量决策树





小结

- 决策树简介（基本流程）
  - 掌握决策树基本流程和原理
- 决策树算法的关键：划分选择
  - 熟悉划分准则
- 克服过拟合的问题：剪枝处理
  - 预剪枝 vs 后剪枝
- 决策树的变体：多变量决策树
  - 了解基本原理