机器学习导论

第四讲 决策树

机器学习模型引线

- 基础机器学习模型：线性模型
  - 高级模型：支持向量机、神经网络等
  - 高级模型：统计学习方法和理论

- 基础机器学习模型：决策树
  - 高级模型：Adaboost，随机森林，GBDT等

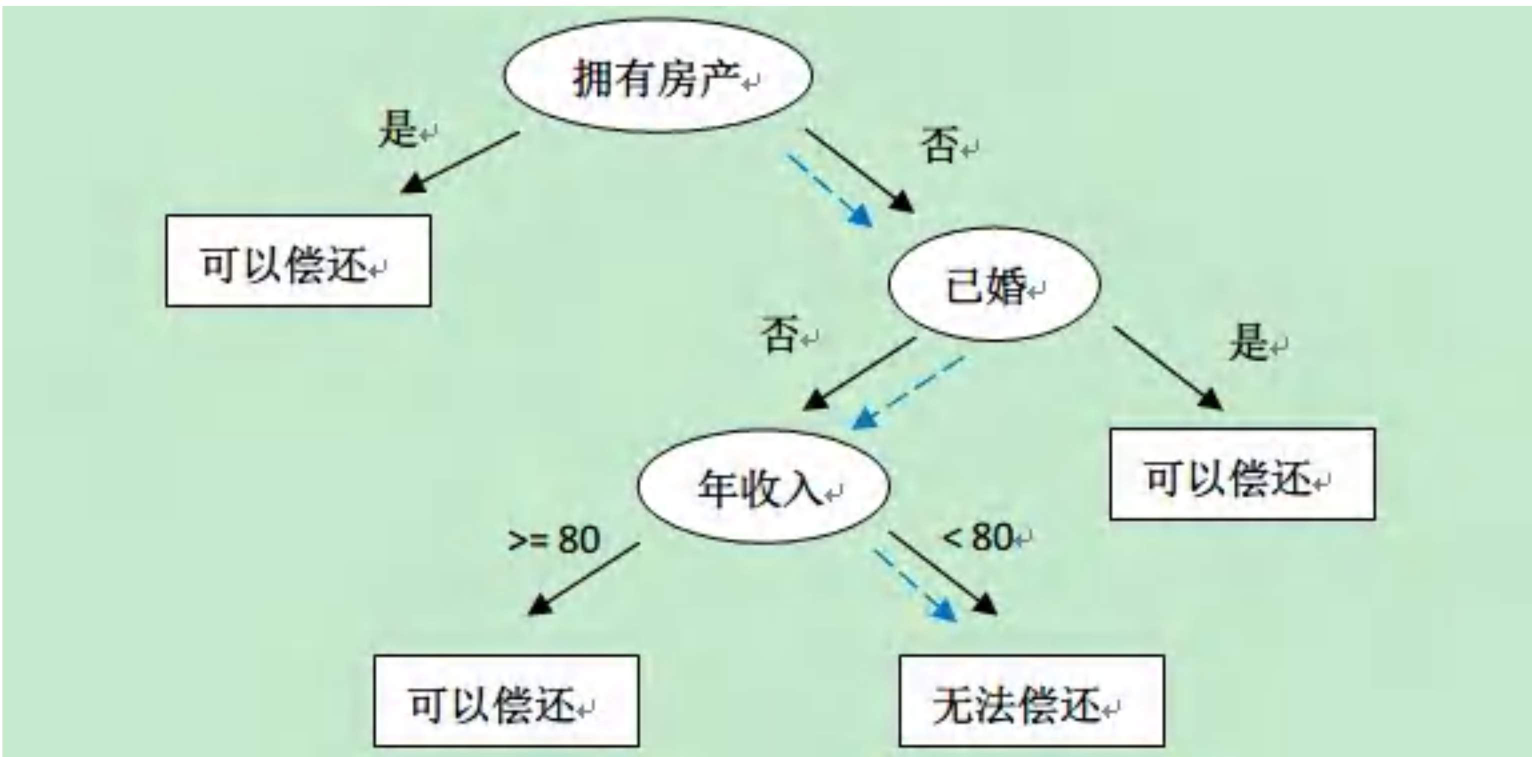
- 基础机器学习模型：贝叶斯模型
  - 高级模型：图模型等

大纲

- 决策树简介（基本流程）
- 决策树算法的关键：划分选择
- 克服过拟合的问题：剪枝处理
- 处理多种类型数据：连续与缺失值
- 决策树的变体：多变量决策树

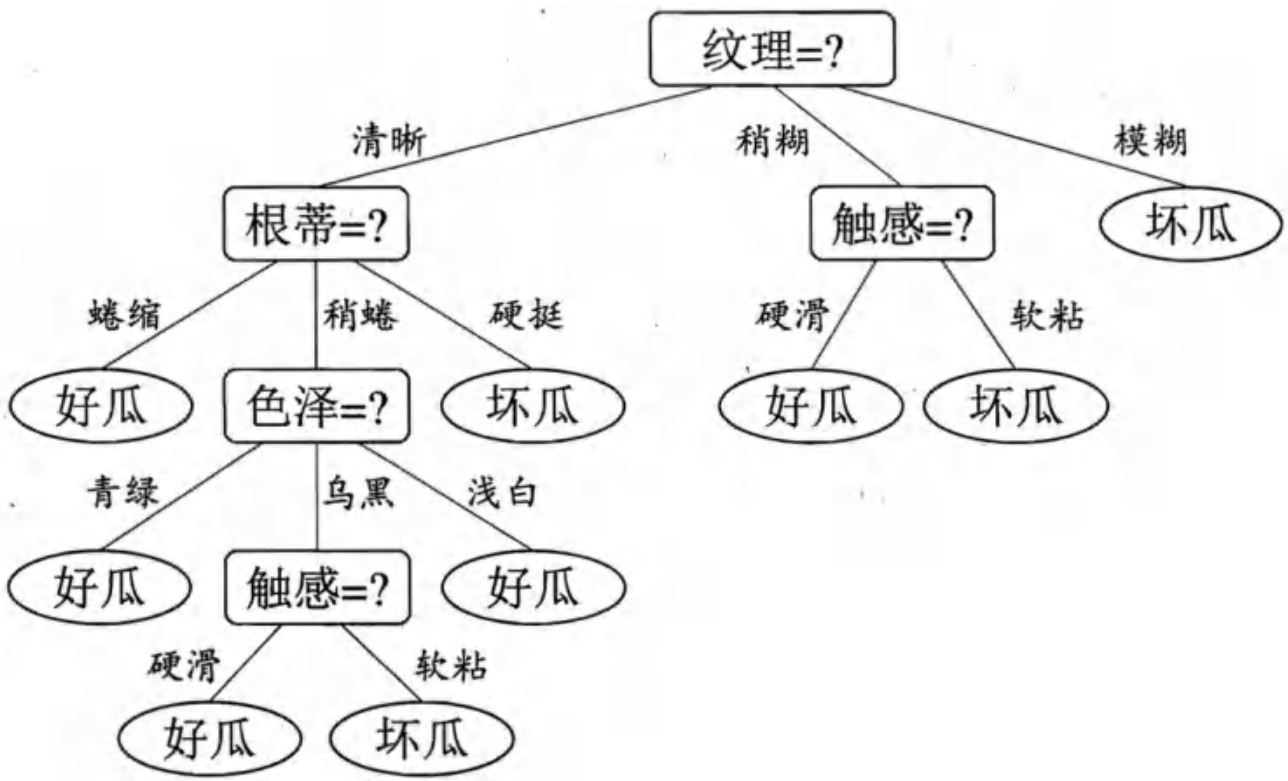
基本流程

- 例子



基本流程

决策树基于树结构来进行预测



基本流程

- 决策过程中每个判定问题都是对某个属性的“测试”
- 决策过程的最终结论对应了我们所希望的判定结果
- 每个测试的结果或是导出最终结论，或者导出进一步的判定问题；嵌套关系——当前节点考虑范围是在上次决策结果的限定范围之内
- 从根结点到每个叶结点的路径对应了一个判定测试序列

决策树学习的目的是为了产生一棵泛化能力强，
即处理未见示例能力强的决策树

基本流程 – 递归过程

Algorithm 1 决策树学习基本算法

输入:

- 训练集 $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$;
- 属性集 $A = \{a_1, \ldots, a_d\}$.

过程: **函数** TreeGenerate(D, A)

生成结点 node;
if D 中样本全属于同一类别 C **then**
    将 node 标记为 C 类叶结点; **return**
end if
if $A = \emptyset$ OR D 中样本在 A 上取值相同 **then**
    将 node 标记叶结点，其类别标记为 D 中样本数最多的类; **return**
end if
从 A 中选择最优划分属性 a_*;
for a_* 的每一个值 a_*^v **do**
    为 node 生成每一个分枝; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
    if D_v 为空 **then**
        将分枝结点标记为叶结点，其类别标记为 D 中样本最多的类; **return**
    else
        以 TreeGenerate$(D_v, A \setminus \{a_*\})$ 为分枝结点
    end if
end for

输出: 以 node 为根结点的一棵决策树

遇到如下三个停止条件之一停止递归

（1）当前结点**包含的样本全部属于同一类别**

（2）当前属性集为空，或所有样本在所有属性上取值相同

（3）当前结点包含的样本集合为空

大纲

- 决策树简介（基本流程）
- 决策树算法的关键：划分选择
- 克服过拟合的问题：剪枝处理
- 处理多种类型数据：连续与缺失值
- 决策树的变体：多变量决策树

划分选择

- 决策树学习的关键在于如何选择最优划分属性。
- 一般而言，随着划分过程不断进行，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，即结点的“纯度”(purity) 越来越高
- 从启发式想法到算法——经典的属性划分方法：
  - 信息增益
  - 增益率
  - 基尼指数

划分选择-信息增益

“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.

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 $\log_2|\mathcal{Y}|$.

划分选择-信息增益

假定离散属性 a 有 V 个可能的取值 $\{a^1, a^2, \ldots, a^V\}$，若使用 a 来对样本集 D 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，记为 D^v．我们可根据式(4.1) 计算出 D^v 的信息熵，再考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重 $|D^v|/|D|$，即样本数越多的分支结点的影响越大，于是可计算出用属性 a 对样本集 D 进行划分所获得的"信息增益"(information gain)

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

为分支结点权重，样本数越多的分支结点的影响越大

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。

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

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

划分选择-信息增益

- 类似的，其他属性的信息增益为

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

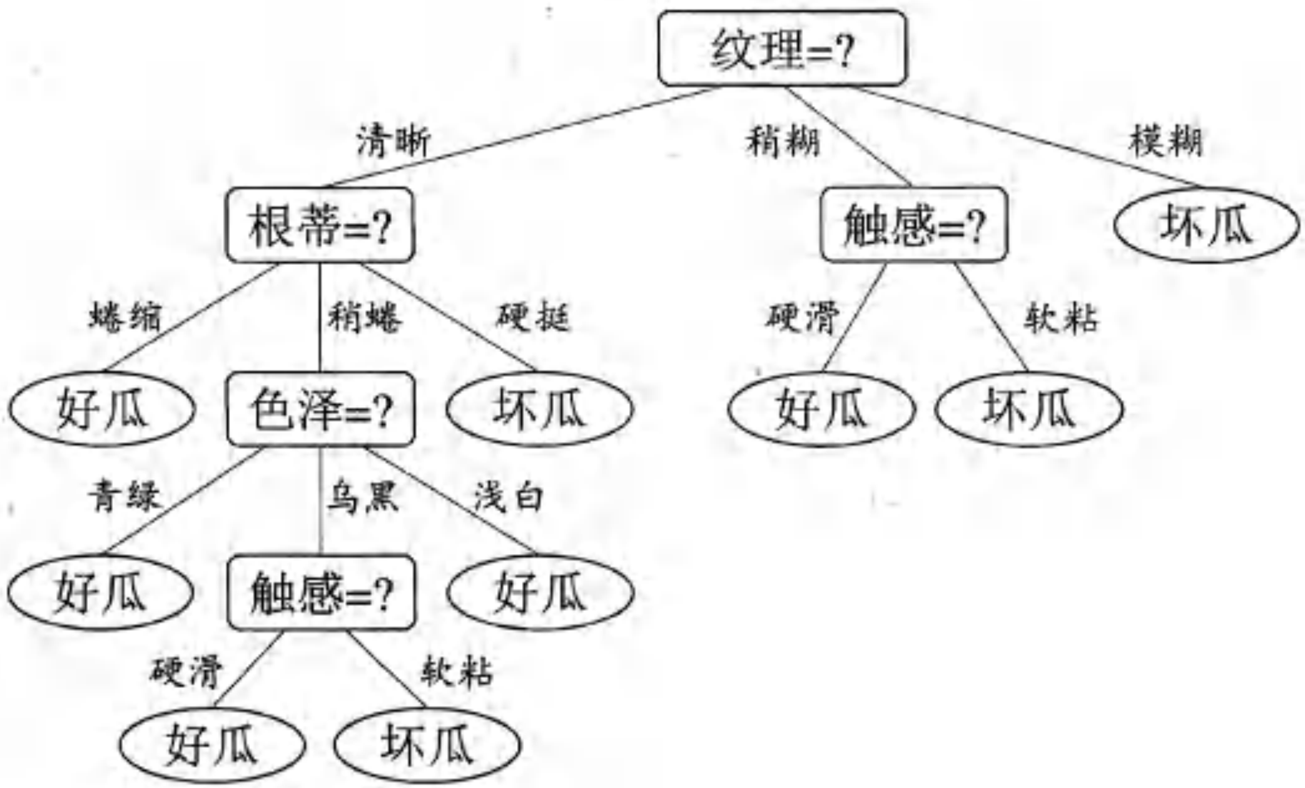
$$\text{Gain}(D,\text{纹理}) = 0.381 \qquad \text{Gain}(D,\text{脐部}) = 0.289$$

$$\text{Gain}(D,\text{触感}) = 0.006$$

- 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：



划分选择-信息增益

不足：

- 若把“编号”也作为一个候选划分属性，则其信息增益一般远大于其他属性。

信息增益指标 偏好 取值数目较多的属性

- 显然，这样的决策树不具有泛化能力，无法对新样本进行有效预测

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

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

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

相当于给信息增益做了一个规范化

称为属性 a 的"固有值"(intrinsic value) [Quinlan, 1993]. 属性 a 的可能取值数目越多(即 V 越大)，则 $\mathrm{IV}(a)$ 的值通常会越大. 例如，对表 4.1 的西瓜数据集 2.0，有 IV(触感) = 0.874 ($V = 2$)，IV(色泽) = 1.580 ($V = 3$)，IV(编号) = 4.088 ($V = 17$).

划分选择-增益率

- 存在的问题

增益率准则 偏好 取值数目较少的属性

- 取值多好，还是少好呢? 少走极端 “不多不少最合适”
- 经典实现C4.5
  - 先找出 信息增益 高于平均水平的属性
  - 从中选 增益率 最高的属性

划分选择-基尼指数

CART 决策树 [Breiman et al., 1984] 使用"基尼指数"(Gini index)来选择划分属性. 采用与式(4.1) 相同的符号, 数据集 D 的纯度可用基尼值来度量:

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

反映了从数据集中随机抽取两个样本，其类别标记不一致的概率；

基尼指数越小，说明越纯

采用与式(4.2)相同的符号表示, 属性 a 的基尼指数定义为

$$\mathrm{Gini\_index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Gini}(D^v) . \qquad (4.6)$$

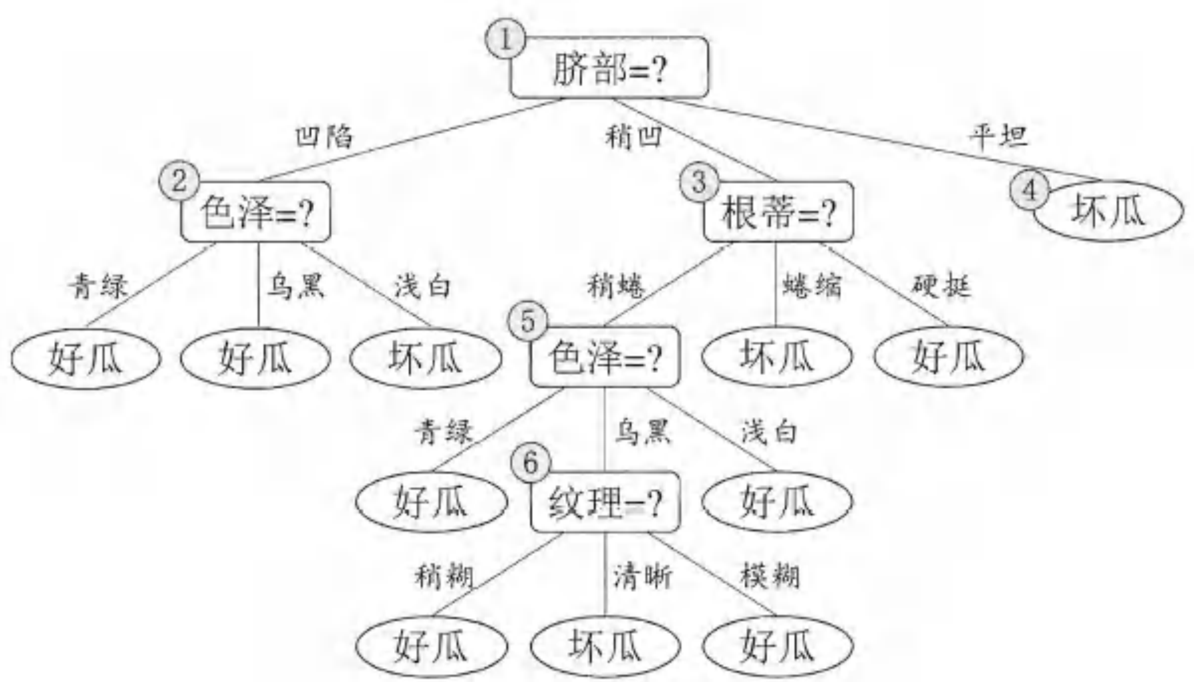
于是, 我们在候选属性集合 A 中, 选择那个使得划分后基尼指数最小的属性作为最优划分属性, 即 $a_* = \underset{a \in A}{\arg\min}\ \mathrm{Gini\_index}(D, a)$.

大纲

- 决策树简介（基本流程）
- 决策树算法的关键：划分选择
- 克服过拟合的问题：剪枝处理
- 处理多种类型数据：连续与缺失值
- 决策树的变体：多变量决策树

剪枝处理

- 决策树有什么不足？为什么要剪枝？
  - 决策树很容易**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



剪枝处理

- 剪枝 是决策树学习算法对付“过拟合”的主要手段

- 剪枝的基本策略
  - 预剪枝
  - 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

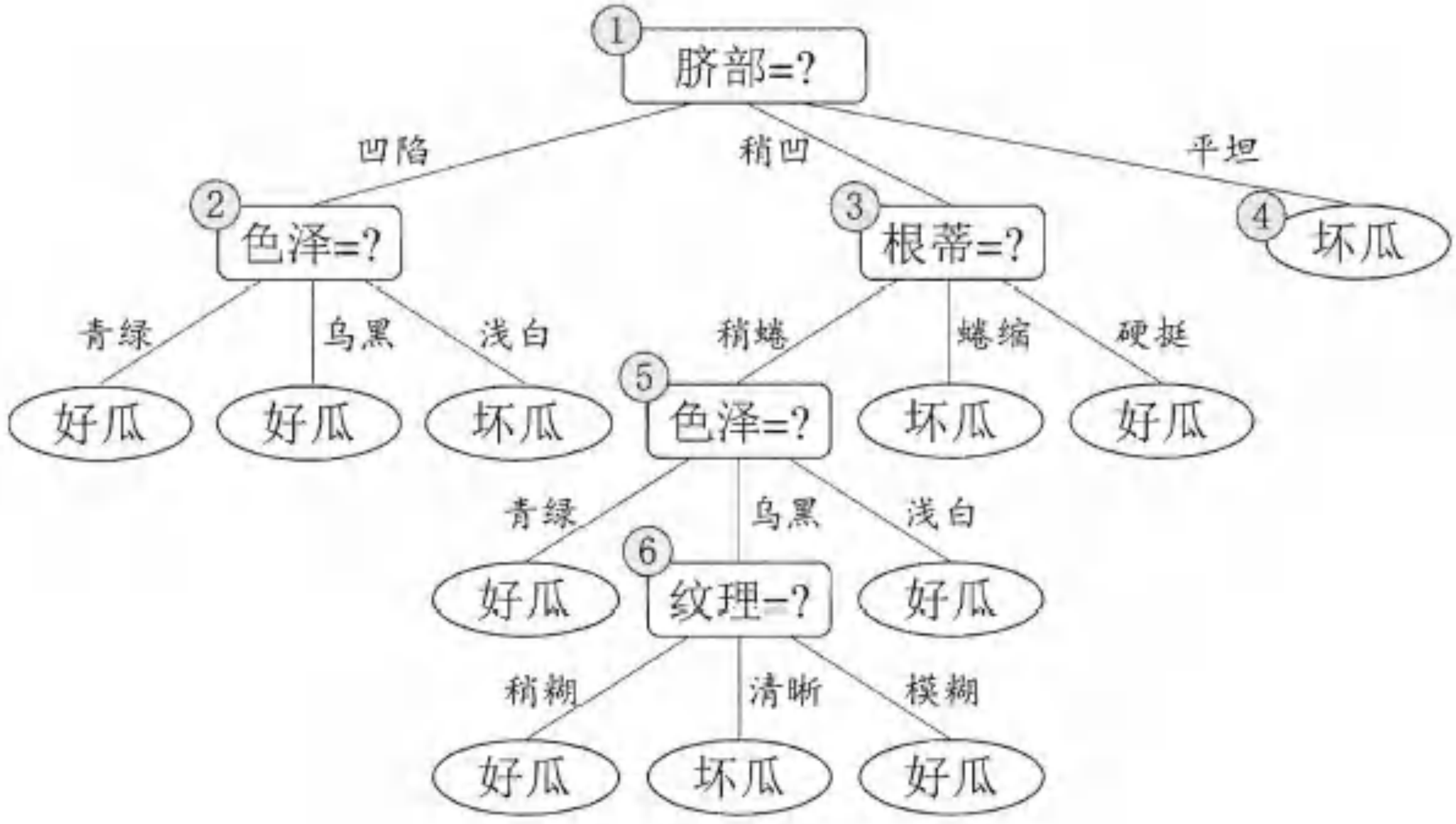
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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
好瓜
6
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：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

边建树，边剪枝

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

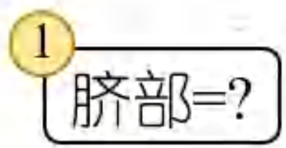
剪枝处理-预剪枝

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



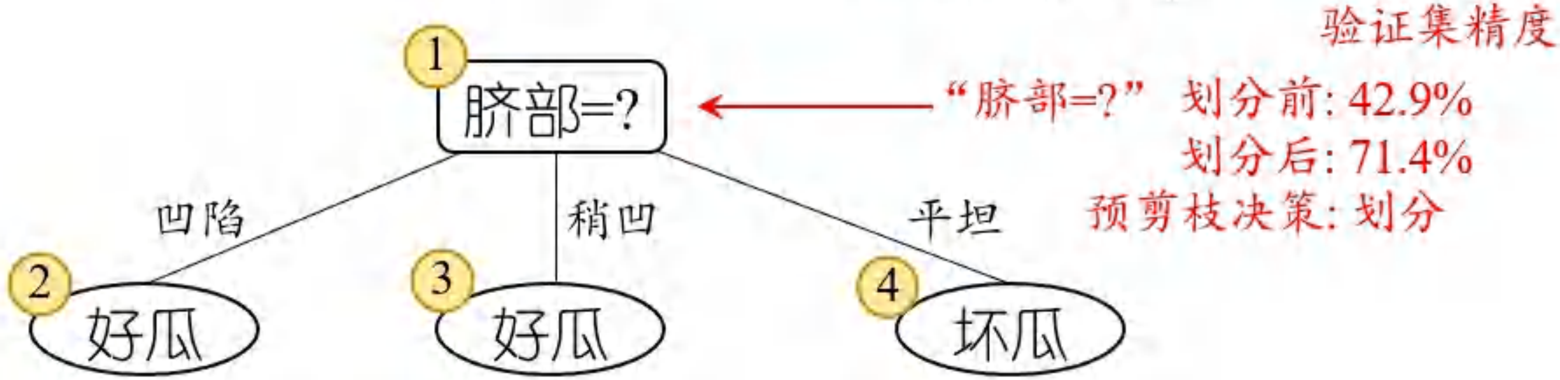
“脐部=?” 划分前: 42.9%

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$\frac{5}{7} \times 100\% = 71.4\%$

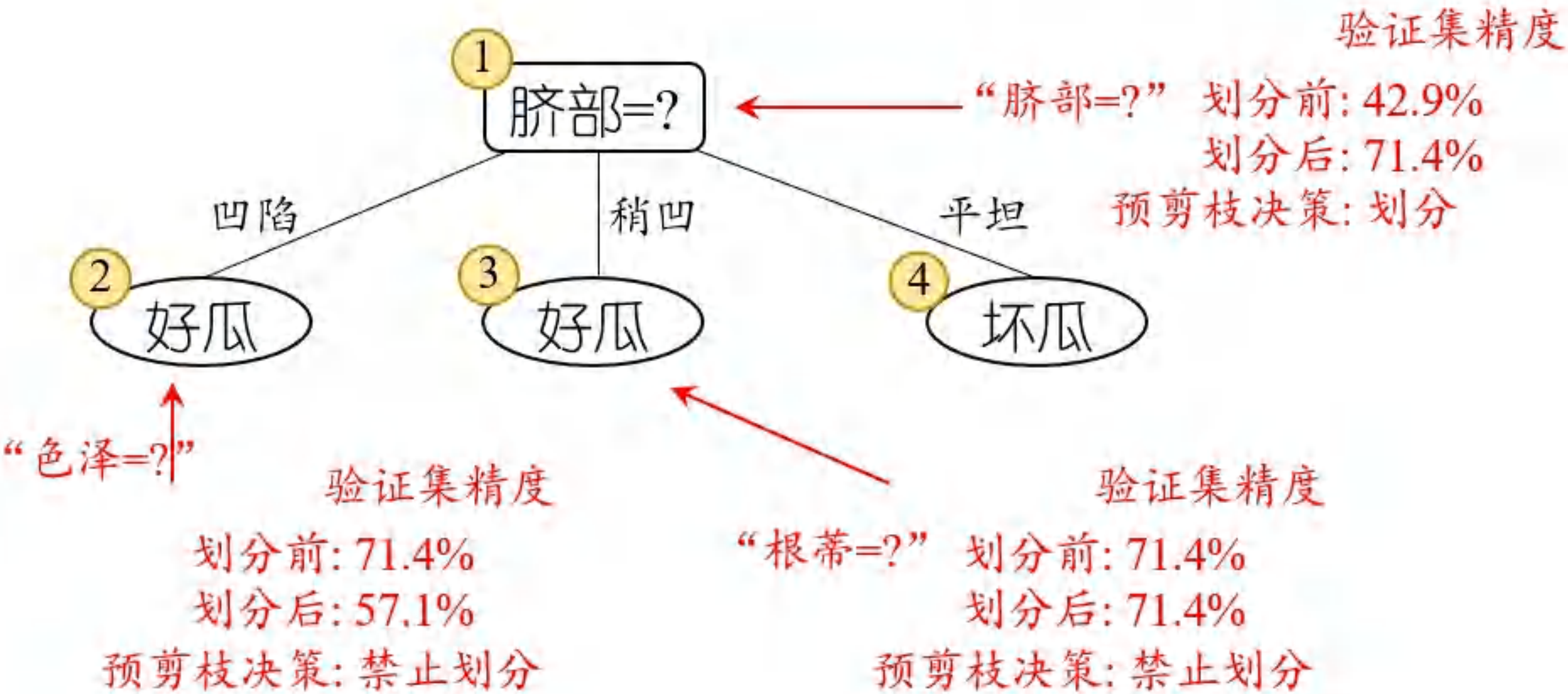


剪枝处理-预剪枝

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”**



剪枝处理-预剪枝

预剪枝的优缺点

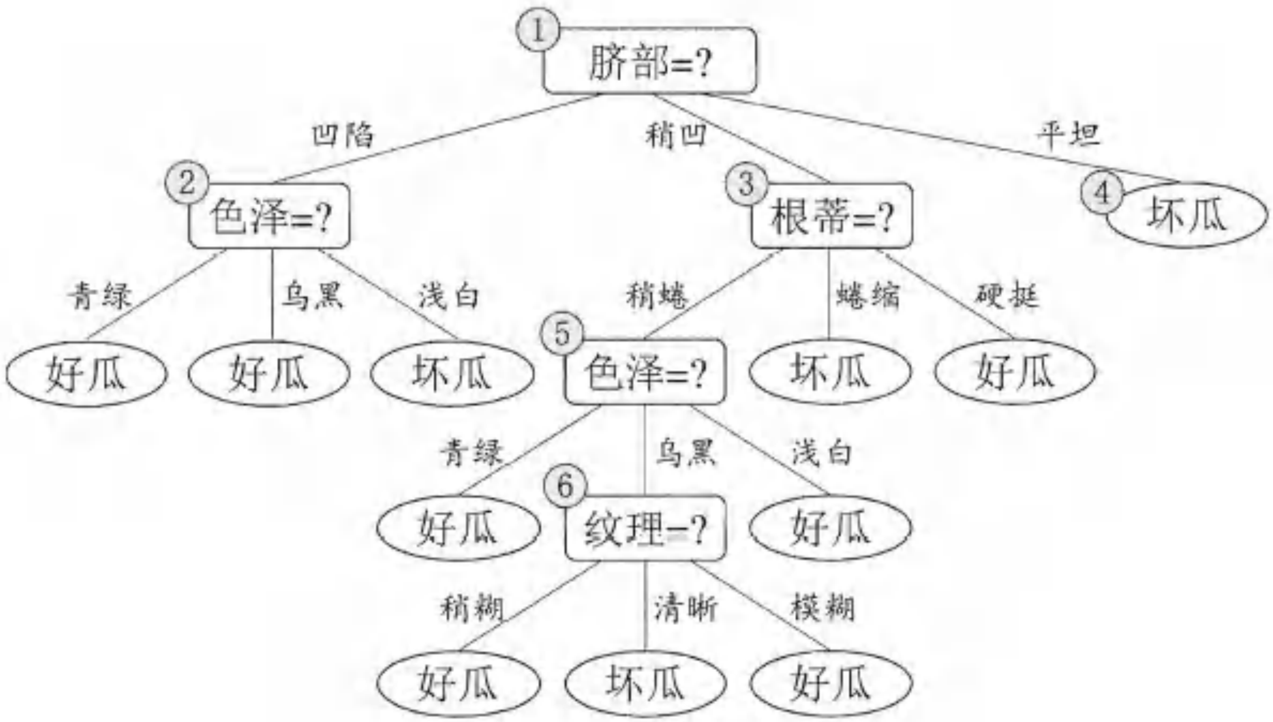
- 优点
  - 降低过拟合风险
  - 显著减少训练时间和测试时间开销

- 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

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%

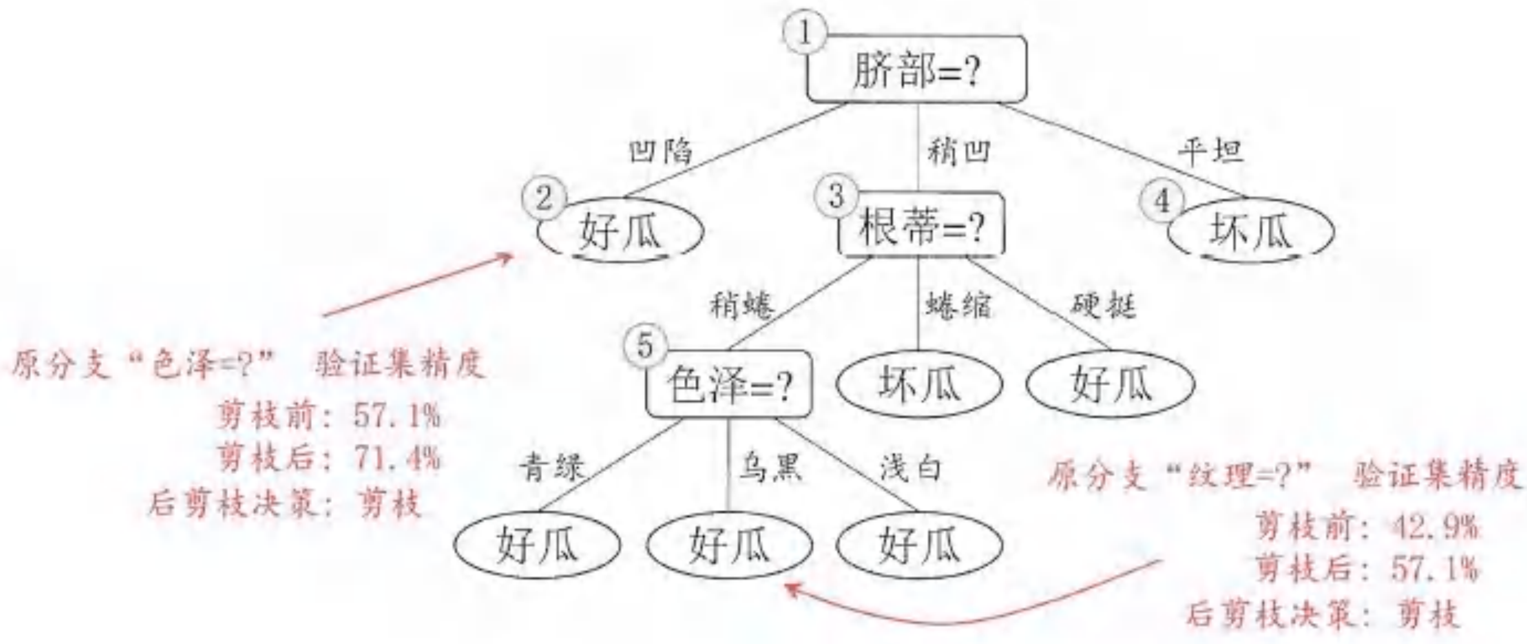


剪枝处理-后剪枝

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．

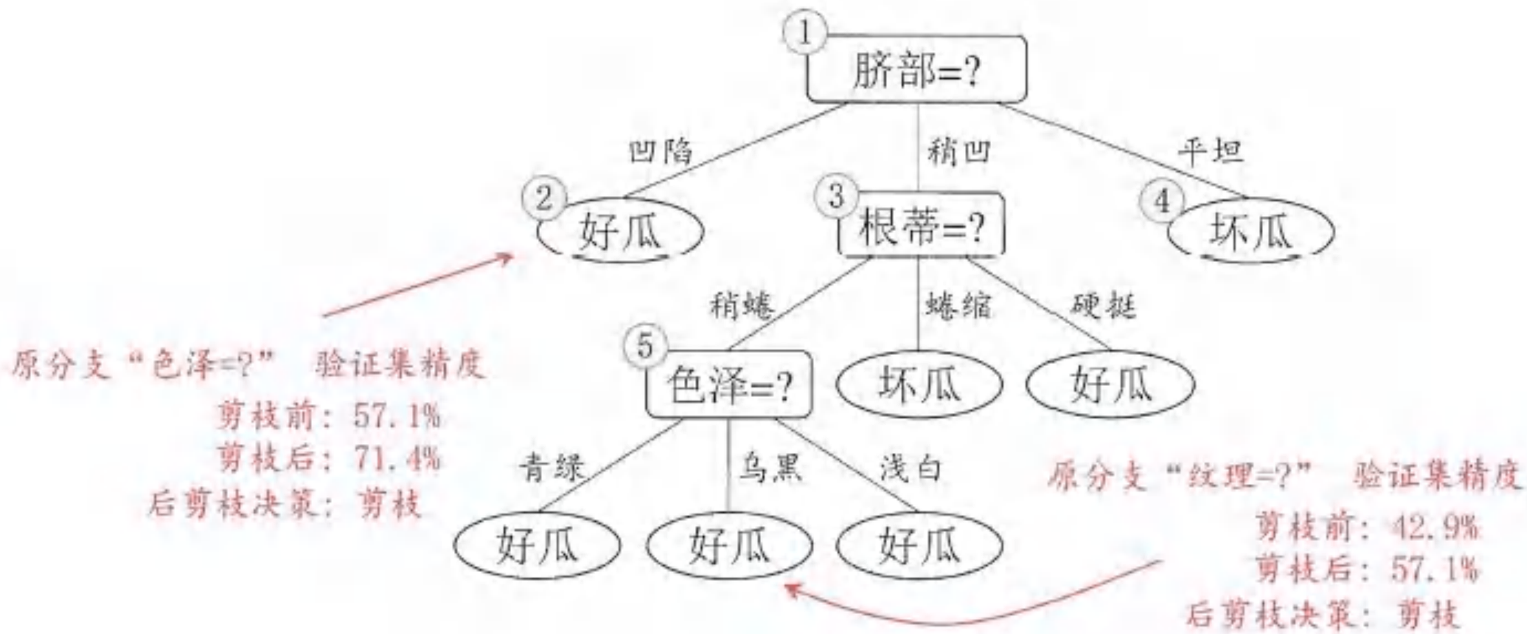
剪枝处理-后剪枝

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.



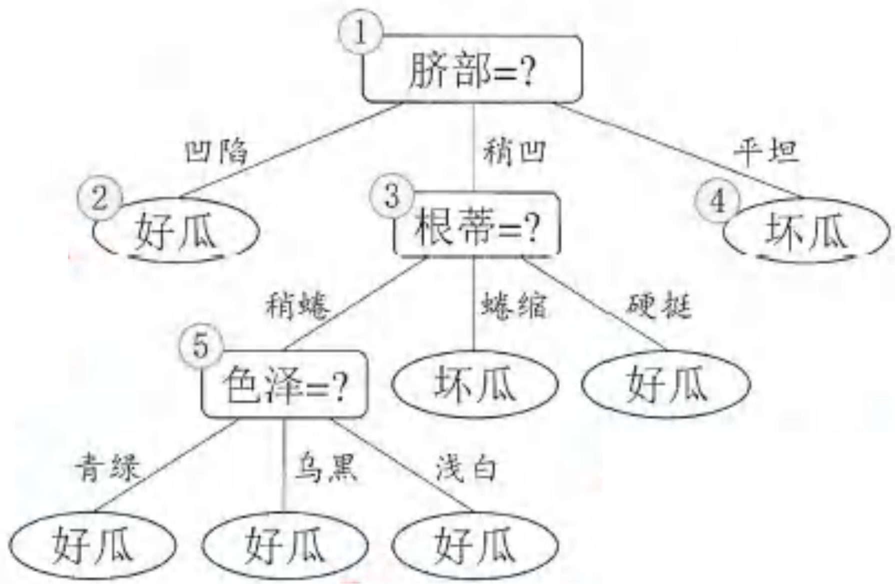
剪枝处理-后剪枝

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

大纲

- 决策树简介（基本流程）
- 决策树算法的关键：划分选择
- 克服过拟合的问题：剪枝处理
- 处理多种类型数据：连续与缺失值
- 决策树的变体：多变量决策树

连续与缺失值 – 连续值处理

- 决策树只能处理离散属性，如何处理连续属性？**离散化**

  - 连续属性离散化(二分法)
    - 第一步：假定连续属性a在样本集D上出现n个不同的取值，从小到大排列，记为 $a^1, a^2, ...a^n$，基于划分点t，可将D分为子集D_t^-和D_t^+，其中D_t^-包含那些在属性a上取值不大于t的样本，D_t^+包含那些在属性a上取值大于t的样本。考虑包含$n-1$个元素的候选划分点集合

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

即把区间$[a^i, a^{i-1})$ 的中位点 $\frac{a^i+a^{i+1}}{2}$作为候选划分点

连续与缺失值 – 连续值处理

- 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

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

其中 $\mathrm{Gain}(D,a,t)$ 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 $\mathrm{Gain}(D,a,t)$ 最大化的划分点

连续与缺失值 – 连续值处理

连续值处理实例

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

对属性“密度”，其候选划分点集合包含16 个候选值：

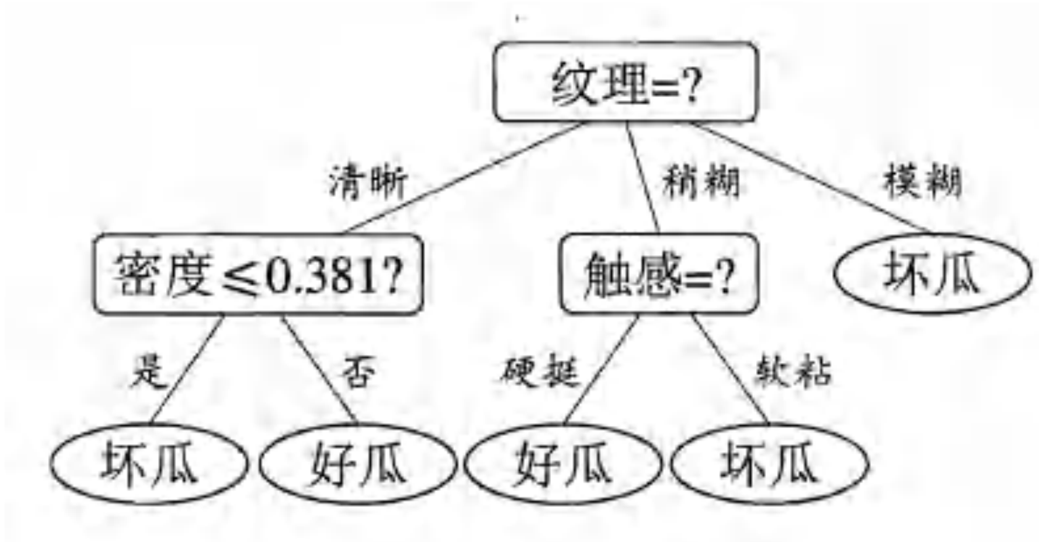
$$T_{密度} = \{0.244, 0.294, 0.351, 0.381, 0.420, 0.459, 0.518, 0.574, 0.600, 0.621, 0.636, 0.648, 0.661, 0.681, 0.708, 0.746\}$$

可计算其信息增益为0.262，对应划分点为0.381

对属性“含糖量”进行同样处理

连续与缺失值 – 连续值处理

连续值下的决策树模型



与离散属性不同，若当前结点划分属性为连续属性，该属性还可作为其后代结点的划分属性

例如在父结点上使用了“密度≤0.381”，不会禁止在子结点上使用“密度≤0.294”.

连续与缺失值 – 缺失值处理

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

连续与缺失值 – 缺失值处理

- 决策树要求所有样本的属性都完整
- 现实应用经常出现不完整样本，即样本的属性值缺失
- 仅使用无缺失的样本进行学习?

对数据信息极大的浪费

- 使用有缺失值的样本，需要解决哪些问题?

Q1：在属性缺失的情况下，如何选择划分属性？

Q2：给定划分属性，若样本在该属性上的值缺失，如何对该样本进行划分？

连续与缺失值 – 缺失值处理

给定训练集 D 和属性 a，令 $\tilde{D}$ 表示 D 中在属性 a 上没有缺失值的样本子集. 对问题(1)，显然我们仅可根据 $\tilde{D}$ 来判断属性 a 的优劣. 假定属性 a 有 V 个可取值 $\{a^1, a^2, \ldots, a^V\}$，令 $\tilde{D}^v$ 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 $(k = 1, 2, \ldots, |\mathcal{Y}|)$ 的样本子集，则显然有 $\tilde{D} = \bigcup_{k=1}^{|\mathcal{Y}|} \tilde{D}_k$，$\tilde{D} = \bigcup_{v=1}^{V} \tilde{D}^v$. 假定我们为每个样本 $\boldsymbol{x}$ 赋予一个权重 $w_{\boldsymbol{x}}$，并定义

为每个样本$\boldsymbol{x}$赋予一个权重w_x，并定义：

- 无缺失值样本所占的比例

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

- 无缺失值样本中第k类所占比例

$$\tilde{p}_k = \frac{\sum_{x \in \tilde{D}_k} w_x}{\sum_{x \in \tilde{D}} w_x} \quad (1 \le k \le |\mathcal{Y}|)$$

- 无缺失值样本中在属性a上取值a^v 的样本所占比例

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

在决策树学习开始阶段，根结点中各样本的权重初始化为 1.

连续与缺失值 – 缺失值处理

- 对于Q1问题：

基于上述定义，我们可将信息增益的计算式(4.2)推广为

$$
\begin{aligned}
\mathrm{Gain}(D,a) &= \rho \times \mathrm{Gain}(\tilde{D},a) \\
&= \rho \times \left(\mathrm{Ent}\left(\tilde{D}\right) - \sum_{v=1}^{V} \tilde{r}_v \, \mathrm{Ent}\left(\tilde{D}^v\right) \right) ,
\end{aligned}
\tag{4.12}
$$

其中由式(4.1)，有

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

- 跟传统决策树一致，只不过仅在有属性值的子集上计算信息增益，不考虑无属性值的样本

连续与缺失值 – 缺失值处理

- 对于Q2问题：

对问题(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 $w_{\boldsymbol{x}}$. 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 $\tilde{r}_v \cdot w_{\boldsymbol{x}}$；直观地看，这就是让同一个样本以不同的概率划入到不同的子结点中去.

- 直观地看，就是让同一个样本以不同的概率划入到不同的子节点中去

连续与缺失值 – 缺失值处理

缺失值处理实例

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

- 学习开始时，根结点包含样本集D中全部17个样例，各样例的权值均为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，$\tilde{D}$ 的信息熵为

$$\mathrm{Ent}(\tilde{D}) = -\sum_{k=1}^{2} \tilde{p}_k \log_2 \tilde{p}_k$$

$$= -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985$$

连续与缺失值 – 缺失值处理

令 $\tilde{D}^1$, $\tilde{D}^2$ 与 $\tilde{D}^3$ 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

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

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

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

因此, 样本子集 $\tilde{D}$ 上属性 “色泽” 的信息增益为

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

连续与缺失值 – 缺失值处理

于是，样本集 D 上属性"色泽"的信息增益为

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

- 类似地可计算出所有属性在数据集上的信息增益

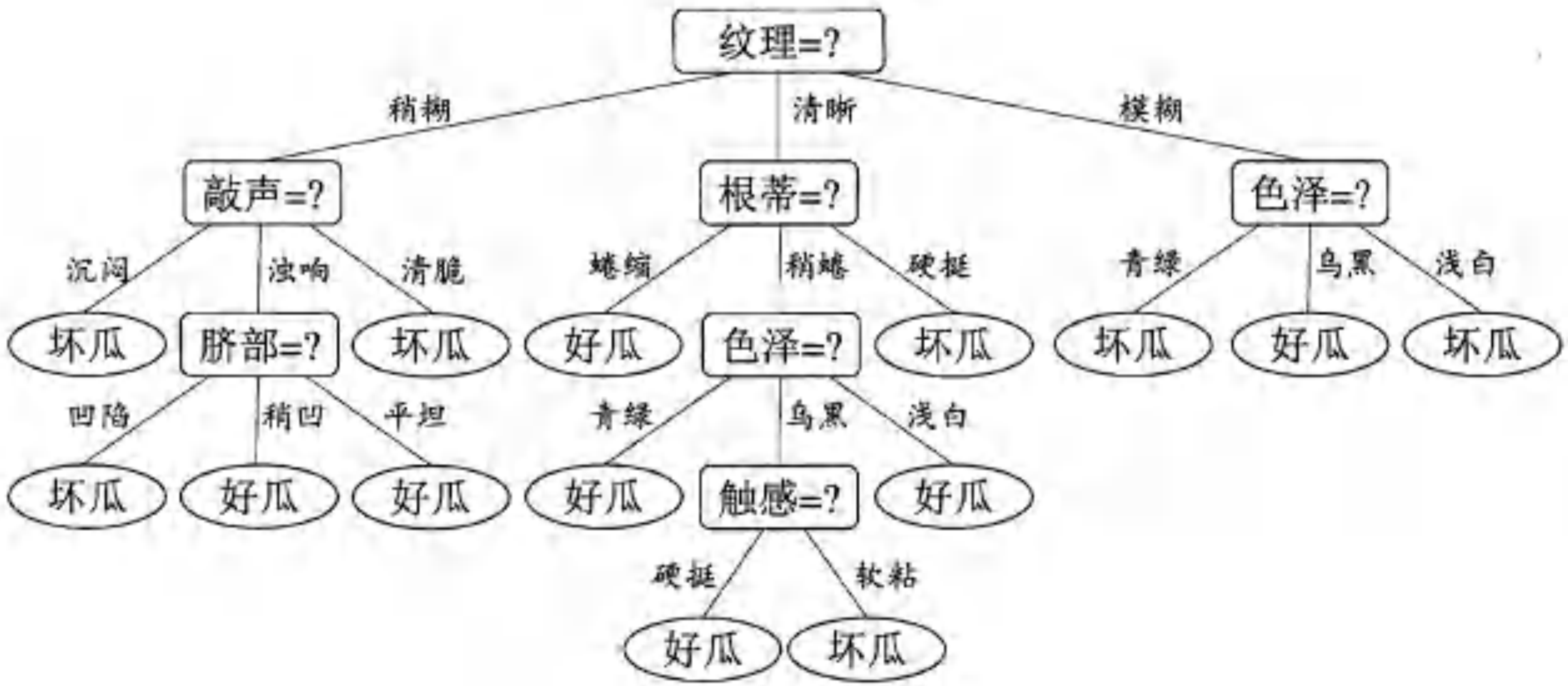
$$\mathrm{Gain}(D, 色泽) = 0.252 \qquad \mathrm{Gain}(D, 根蒂) = 0.171$$

$$\mathrm{Gain}(D, 敲声) = 0.145 \qquad \boxed{\mathrm{Gain}(D, 纹理) = 0.424}$$

$$\mathrm{Gain}(D, 脐部) = 0.289 \qquad \mathrm{Gain}(D, 触感) = 0.006$$

连续与缺失值 – 缺失值处理

- 最终决策树

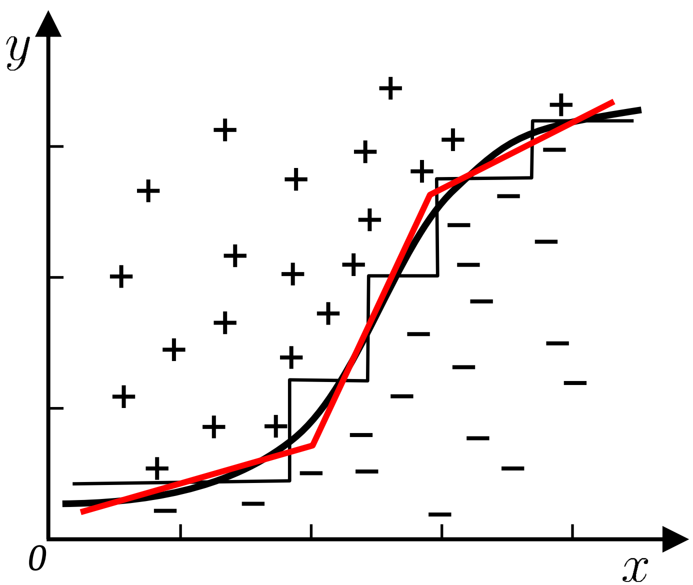


大纲

- 决策树简介（基本流程）
- 决策树算法的关键：划分选择
- 克服过拟合的问题：剪枝处理
- 处理多种类型数据：连续与缺失值
- 决策树的变体：多变量决策树

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

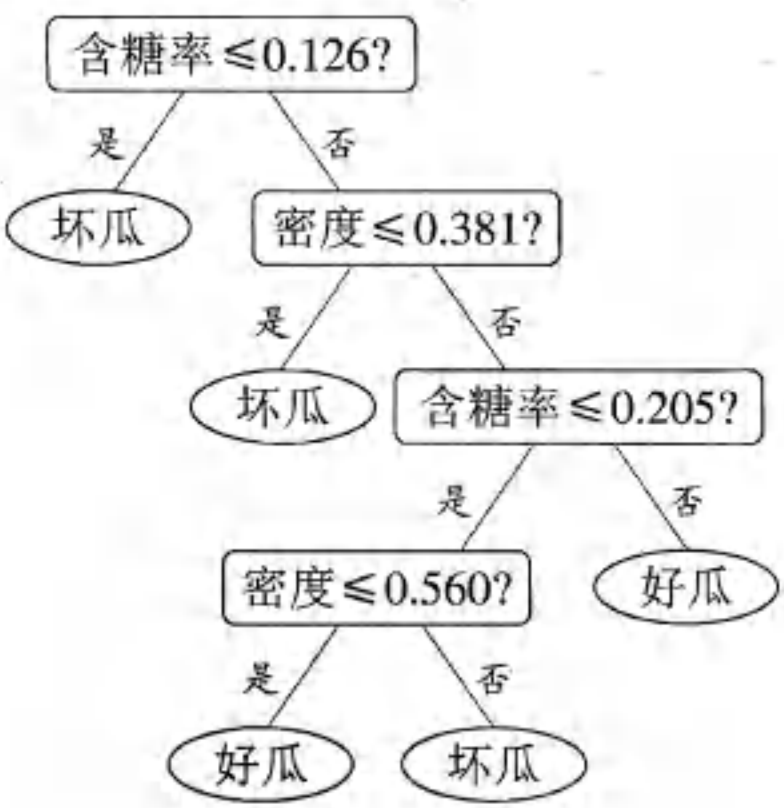


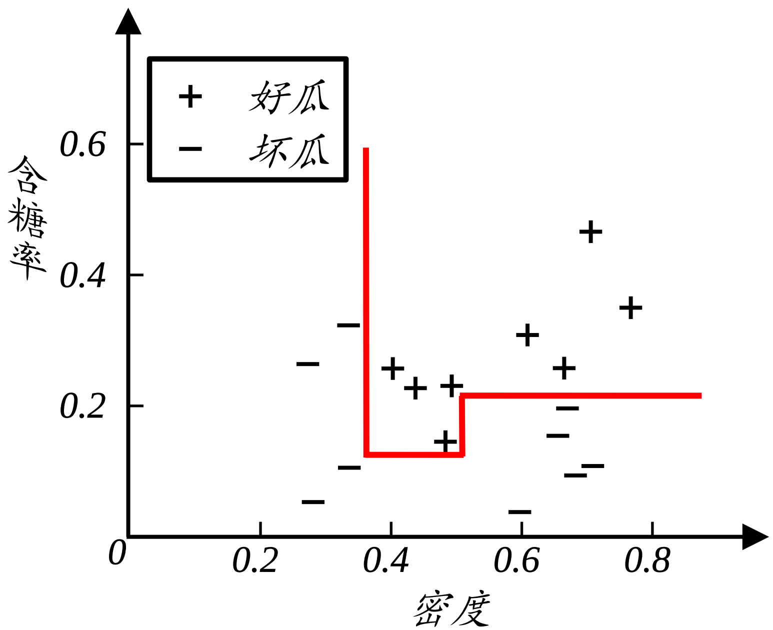
- 非叶节点不再是仅对某个属性，而是对属性的线性组合

- 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

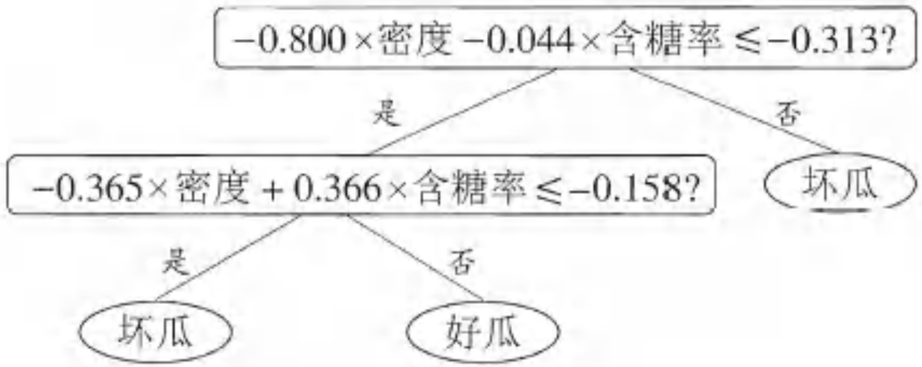
- 单变量决策树

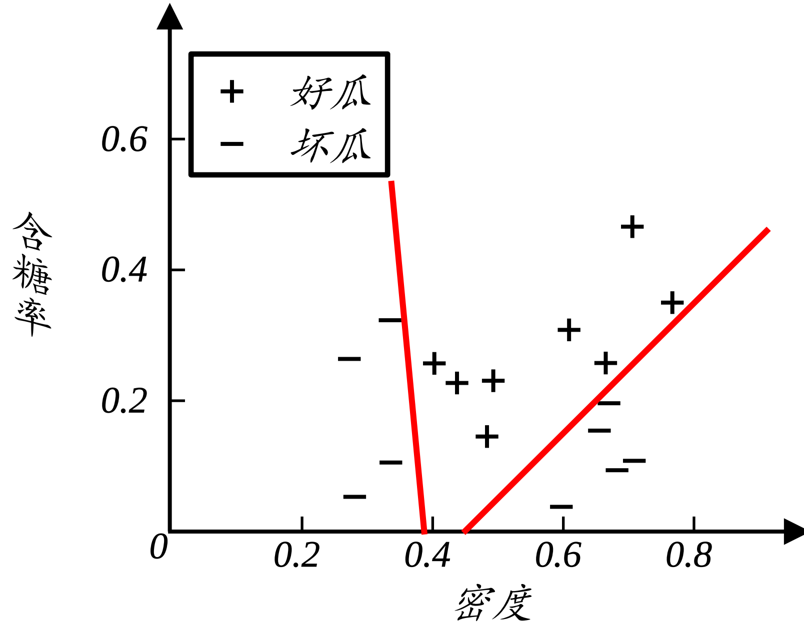




多变量决策树

- 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
  - 熟悉三种划分准则
- 克服过拟合的问题：剪枝处理
  - 预剪枝 vs 后剪枝
- 处理多种类型数据：连续与缺失值
  - 了解基本原理
- 决策树的变体：多变量决策树
  - 了解基本原理