WDS

DataScience@Web

机器学习讨论班

2018年暑期

7. 决策树

郭雪纯

介绍内容

- 模型介绍
- 划分选择和决策树的生成
- 剪枝处理
- 连续与缺失值
- 多变量决策树
- 实验

假设空间与决策树

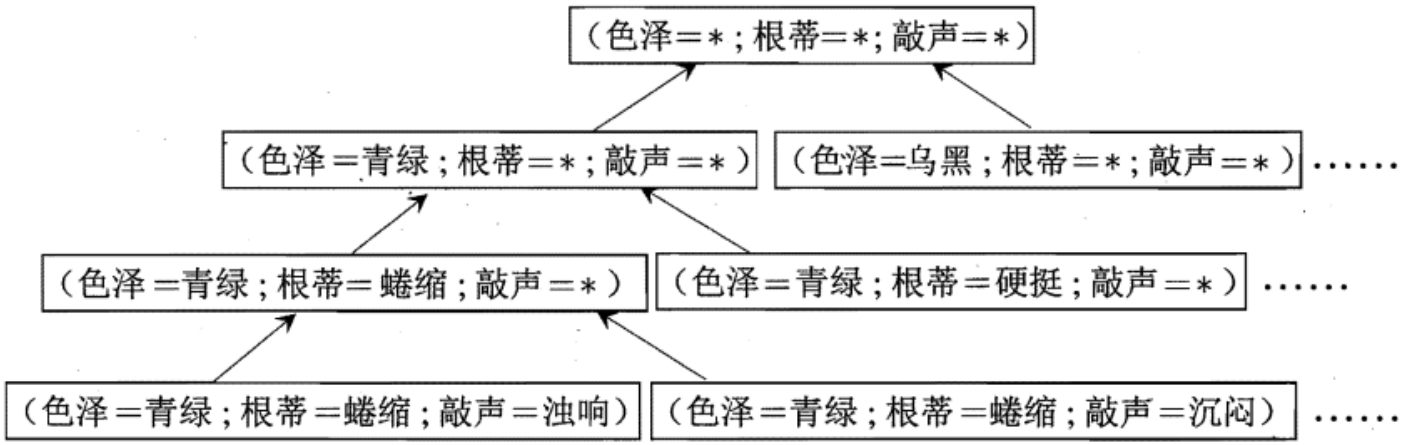


图1：西瓜问题的假设空间

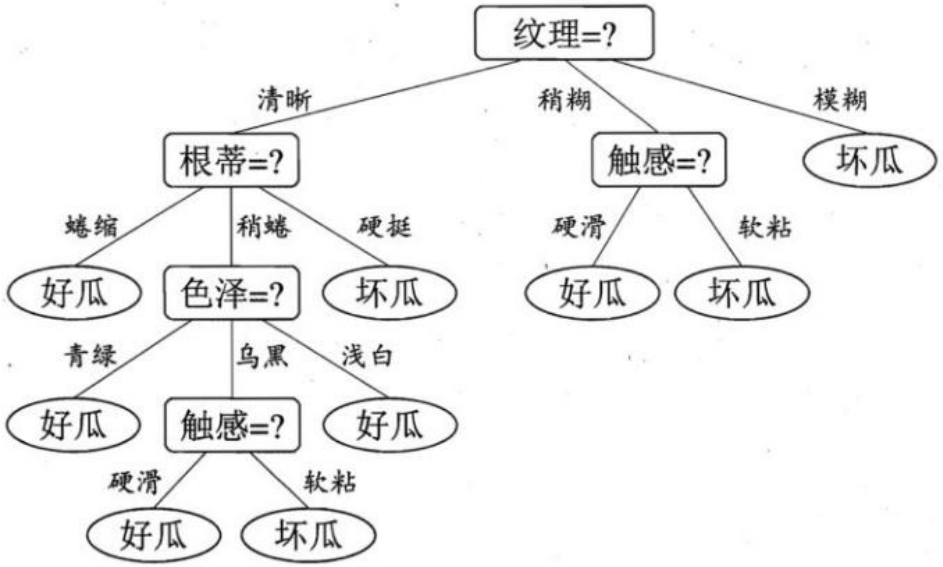
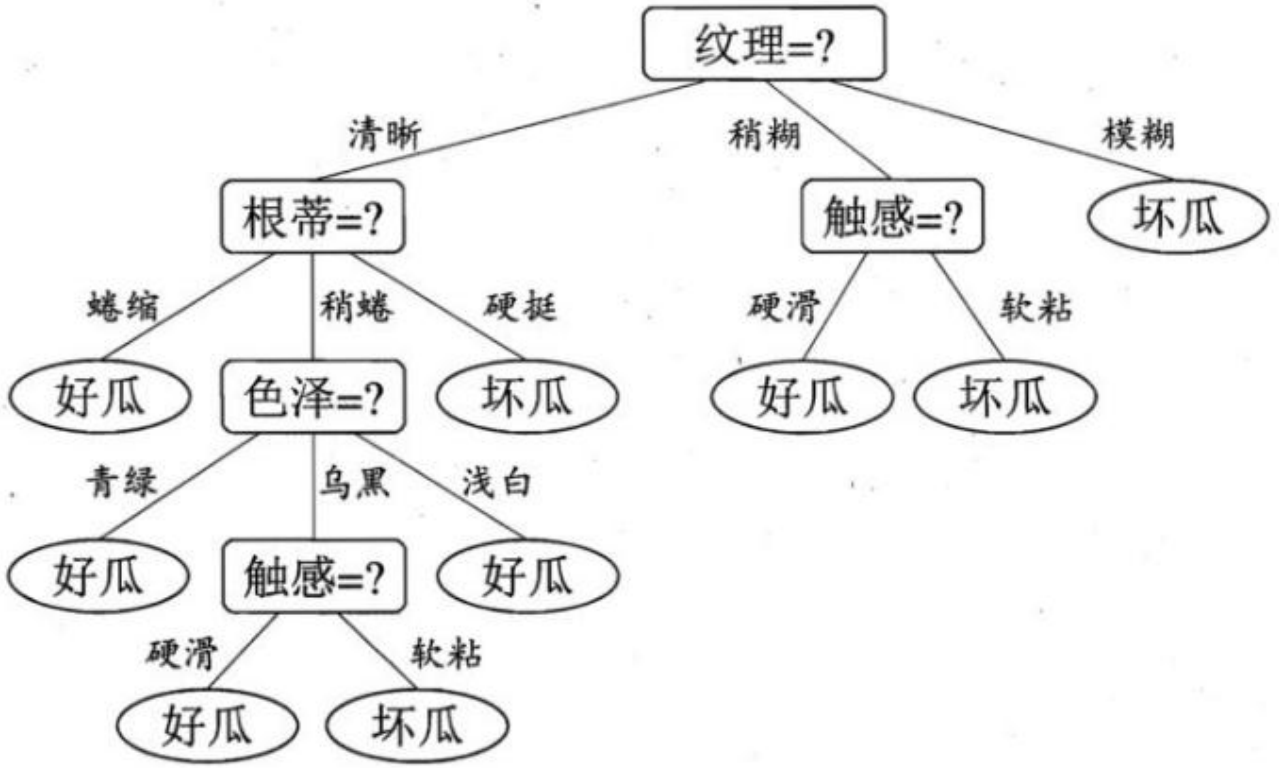


图2：决策树举例

决策树

- 基于树结构来进行决策
- 决策过程最终对应了判定结果
- 遵循分而治之策略

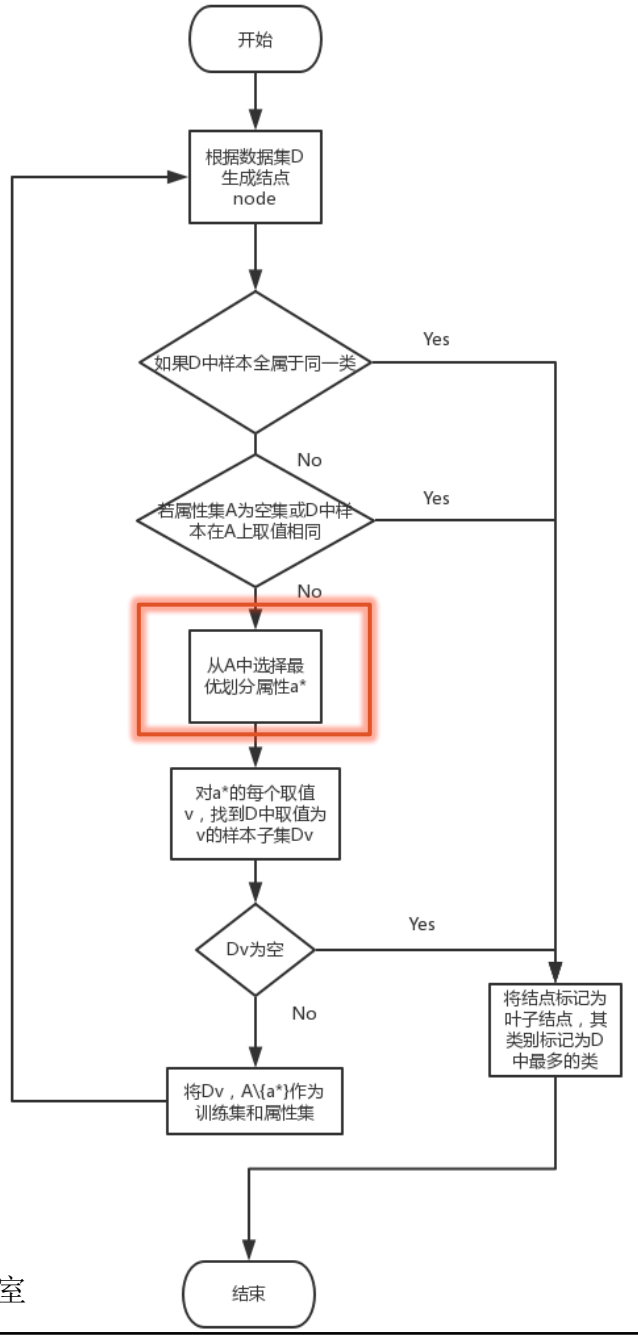


决策树基本学习算法

训练集 $D = \{(\boldsymbol{x}_1, y_1), \{\boldsymbol{x}_2, y_2\}, \ldots, \{\boldsymbol{x}_m, y_m\}\}$
属性集 $A = \{a_1, a_2, \ldots, a_d\}$

递归过程中返回情形

1. 当前结点包含的样本全属于同一类别
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



特征选择问题

特征选择在于选取对训练数据具有分类能力的特征
如果利用一个特征进行分类的结果与随机分类的结果没有很大差别则称这个特征是没有分类能力的

随着划分过程的进行，希望分支结点所包含的样本尽可能属于同一类别，即结点纯度（purity）越来越高

信息增益
增益率
基尼系数

信息熵(information entropy)

- 熵的概念源自热物理学，表示分子状态混乱程度的物理量
- 信息熵是由香农提出，把信息中排除冗余后的平均信息量成为“信息熵”
  - 不确定性越大，信息量越大，熵越大
  - 不确定性越小，信息量越大，熵越小
- 信息熵公式

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

$\mathrm{Ent}(D)$值越小
则D的纯度越高

https://blog.csdn.net/qq280929090/article/details/78135417

信息增益

假定离散属性a有V个可能取值$\{a^1, a^2, \ldots, a^V\}$
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

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

不同分支结点包含样本数不同，给分支结点赋予权重

信息增益越大，使用属性a来进行划分所获得的“纯度提升”越大

ID3决策树学习算法

用信息增益来进行决策树的划分属性选择

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

$$a_*=\arg\max_{a\in A}\mathrm{Gain}(D,a)$$

表1：西瓜数据集

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
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

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

对于每个属性，计算出信息增益，选择信息增益最大的属性作为当前划分的属性，以色泽举例，分别计算色泽为青绿、乌黑、浅白的分支结点的信息熵

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

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

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

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

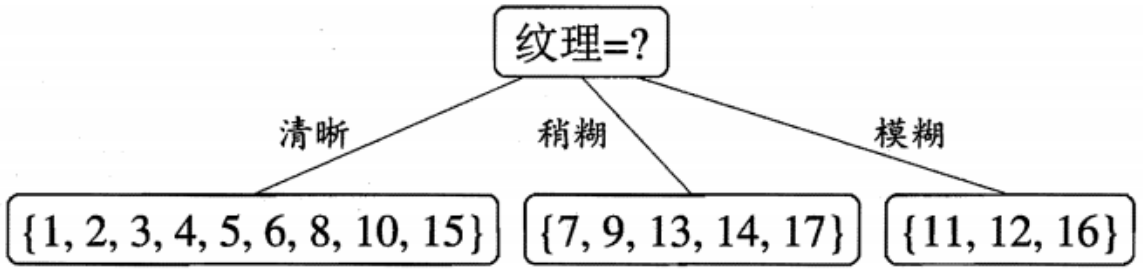
若以编号也作为属性，信息增益达到0.998，产生17个分支，每个分支的纯度达到最大，但是不具有泛化能力

类似的, 我们可计算出其他属性的信息增益:

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

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

$$\text{Gain}(D, \text{触感}) = 0.006.$$



C4.5决策树学习算法

用增益率来进行决策树的划分属性选择

启发式规则：先从候选划分属性中找出信息增益高于平均水平的属性，再从其中选择增益率最高的

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

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

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

属性a的固有值，即为数据集D关于特征a的熵。属性a取值越多，熵值就会越大

基尼指数(Gini index)

Gini(D)反映了从数据集中随机抽取两个两本，其类别不一致的概率。

Gini(D)越小，纯度越高

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

CART决策树

- 用基尼指数来进行决策树的划分属性选择，同时决定改特征的最优二值切分点

- CART假设决策树是二叉树，内部节点取值为“是”和“否”

不同分支结点包含样本数不同，给分支结点赋予权重

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

$$a_*=\arg\min_{a\in A}\ \text{Gini\_index}(D,a).$$

表2：贷款情况数据集

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

$$Gini(D, A_1 = 1) = \frac{5}{15}\left(2 \times \frac{2}{5} \times \left(1 - \frac{2}{5}\right)\right) + \frac{10}{15}\left(2 \times \frac{7}{10} \times \left(1 - \frac{7}{10}\right)\right) = 0.44$$

$$Gini(D, A_1 = 2) = 0.48$$

$$Gini(D, A_1 = 3) = 0.44$$

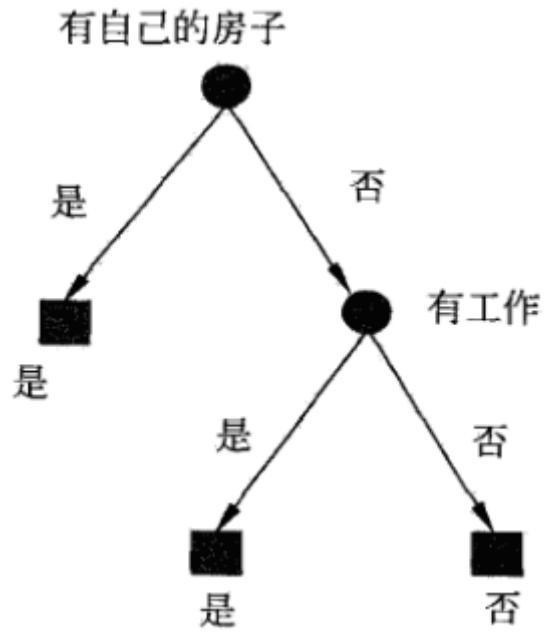
$$Gini(D, A_2 = 1) = 0.32$$

$$Gini(D, A_3 = 1) = 0.27$$

$$Gini(D, A_4 = 1) = 0.36$$

$$Gini(D, A_4 = 2) = 0.47$$

$$Gini(D, A_4 = 3) = 0.32$$



CART回归树

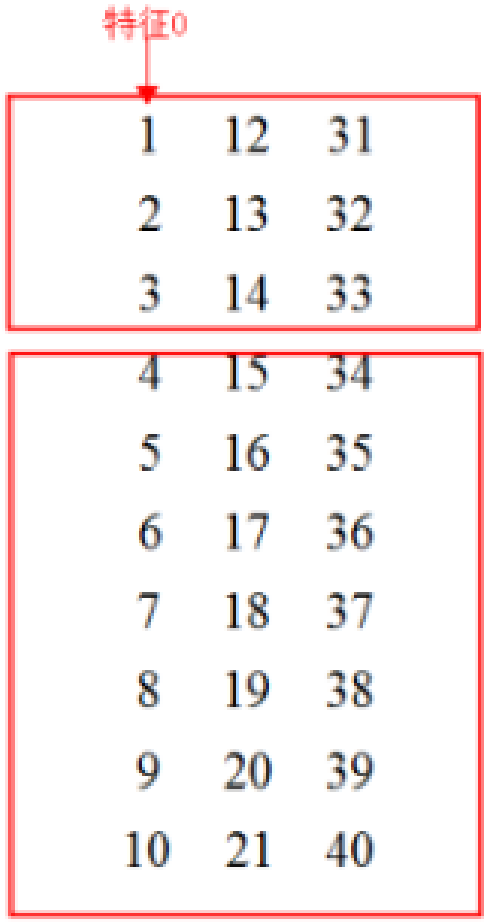
训练集 $D = \{(\boldsymbol{x}_1, y_1), \{\boldsymbol{x}_2, y_2\}, \ldots, \{\boldsymbol{x}_N, y_N\}\}$
回归树对应着输入空间的一个划分，以及划分在单元上的输出值
假设将输入空间划分为M个单元$R_1, R_2, \ldots, R_M$

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

$$\sum_{x_i \in R_m} (y_i - f(x_i))^2$$

$$\hat{c}_m = \text{ave}(y_i \mid x_i \in R_m)$$

1	12	31
2	13	32
3	14	33
4	15	34
5	16	35
6	17	36
7	18	37
8	19	38
9	20	39
10	21	40



通过特征0，特征切分点为3，把数据集分为两个部分

CART回归树

CART回归最重要的就是寻找最优切分特征和最优切分点

$$R_1(j,s)=\{x \mid x^{(j)} \leqslant s\} \quad 和 \quad R_2(j,s)=\{x \mid x^{(j)} > s\}$$

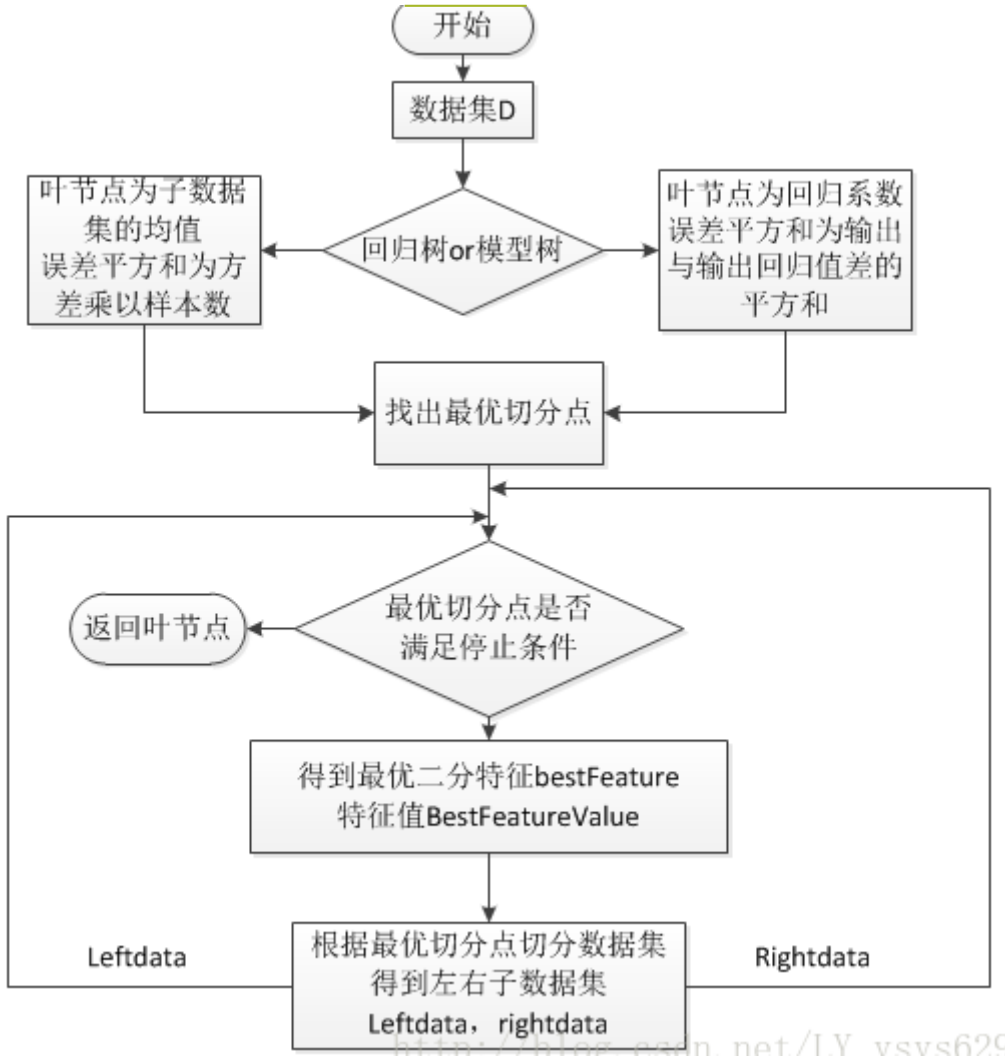
$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i-c_2)^2\right]$$

$$\hat{c}_1=\text{ave}(y_i \mid x_i \in R_1(j,s)) \quad 和 \quad \hat{c}_2=\text{ave}(y_i \mid x_i \in R_2(j,s))$$

启发式规则：选择第j个变量$x^{(j)}$和它的取值s，作为切分变量和切分点

- 对复杂的关系建模，一种可行的方式是使用树来对预测值分段，包括分段常数或分段直线。

- 若叶节点使用的模型是分段常数则称为回归树，若叶节点使用的模型是线性回归方程，则称为模型树



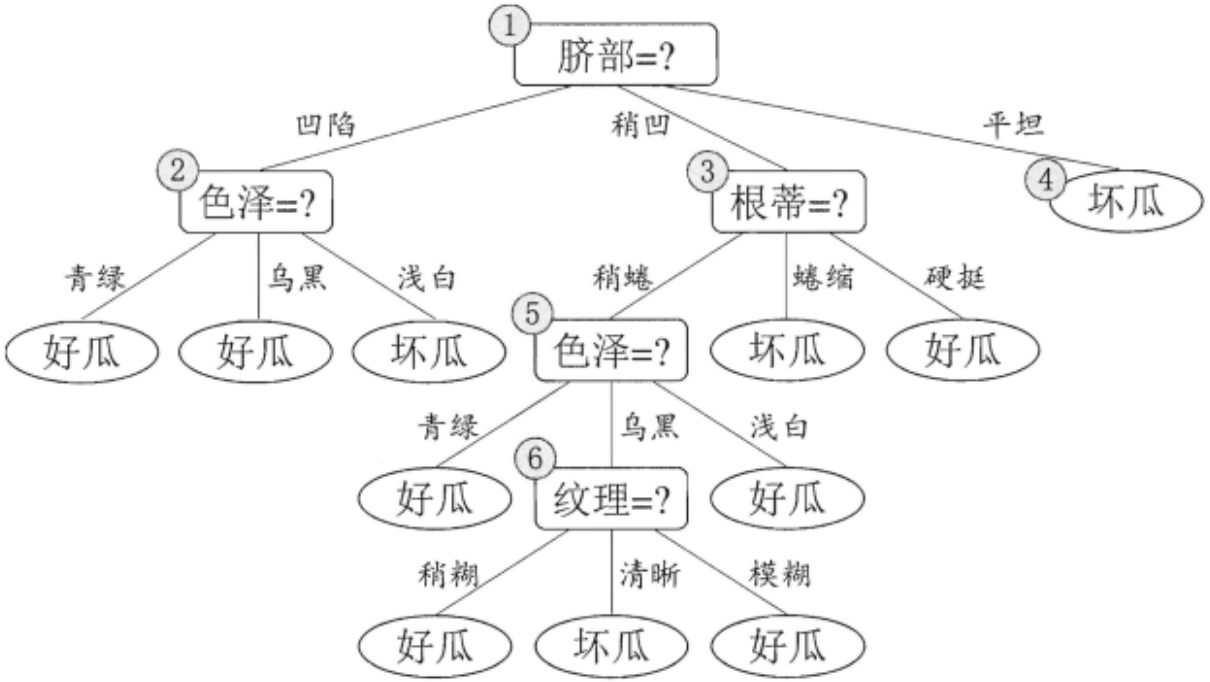
剪枝处理（pruning）

- 决策树学习算法对付“过拟合”重要手段
  - 预剪枝（prepruning）在决策树生成过程中，对每个结点划分前后进行估计
  - 后剪枝（postpruning）在生成一颗完整的决策树之后，自底向上对非叶子结点进行考察

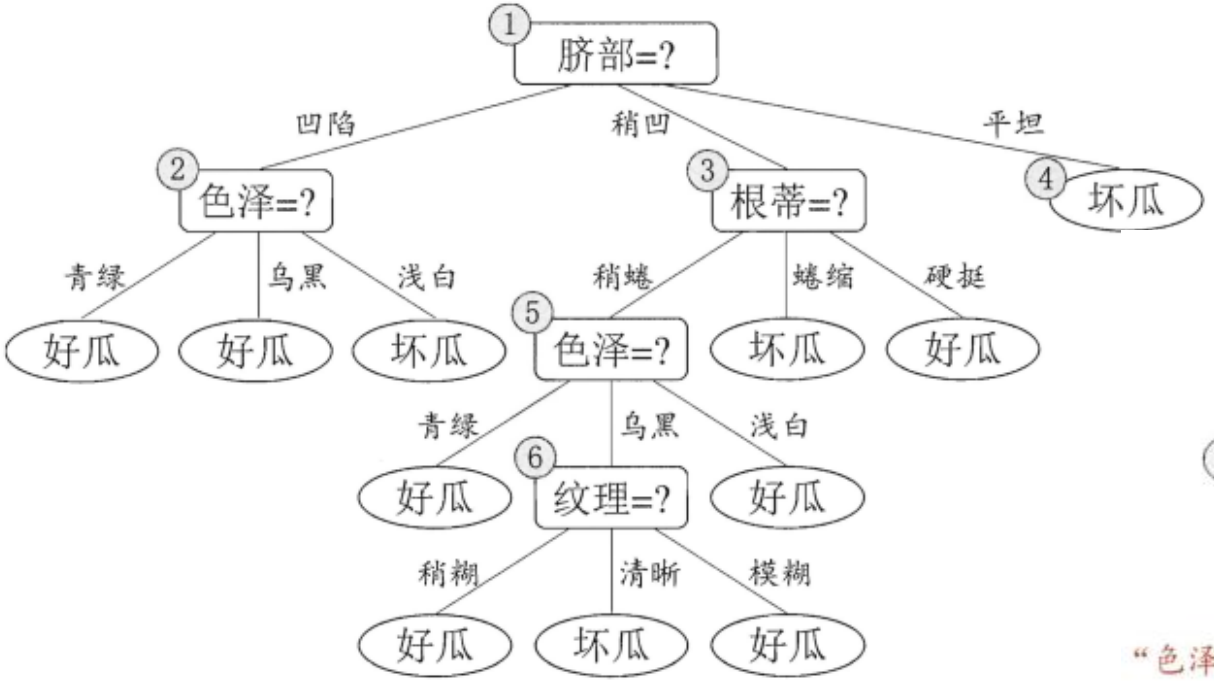
表3：西瓜数据集

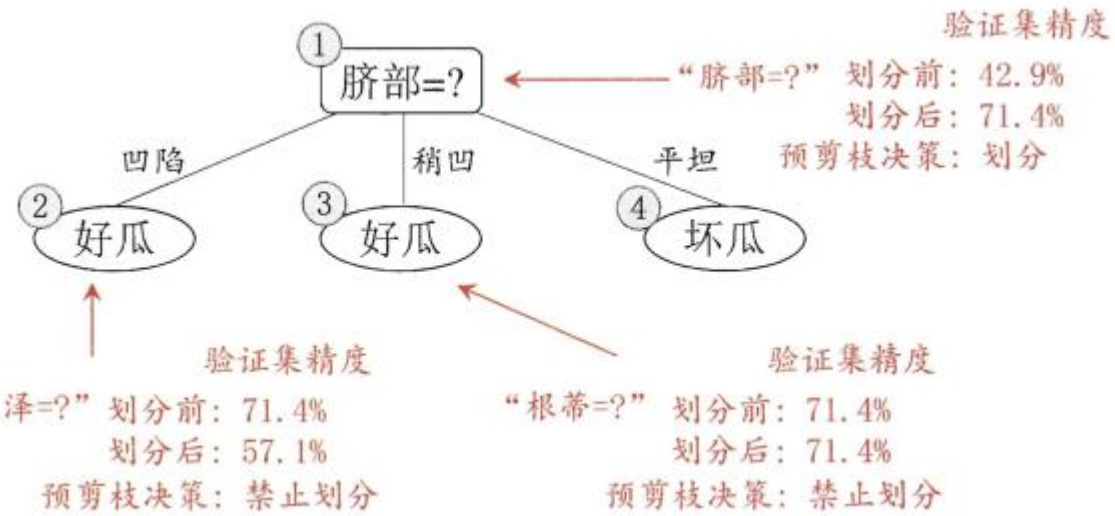
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

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

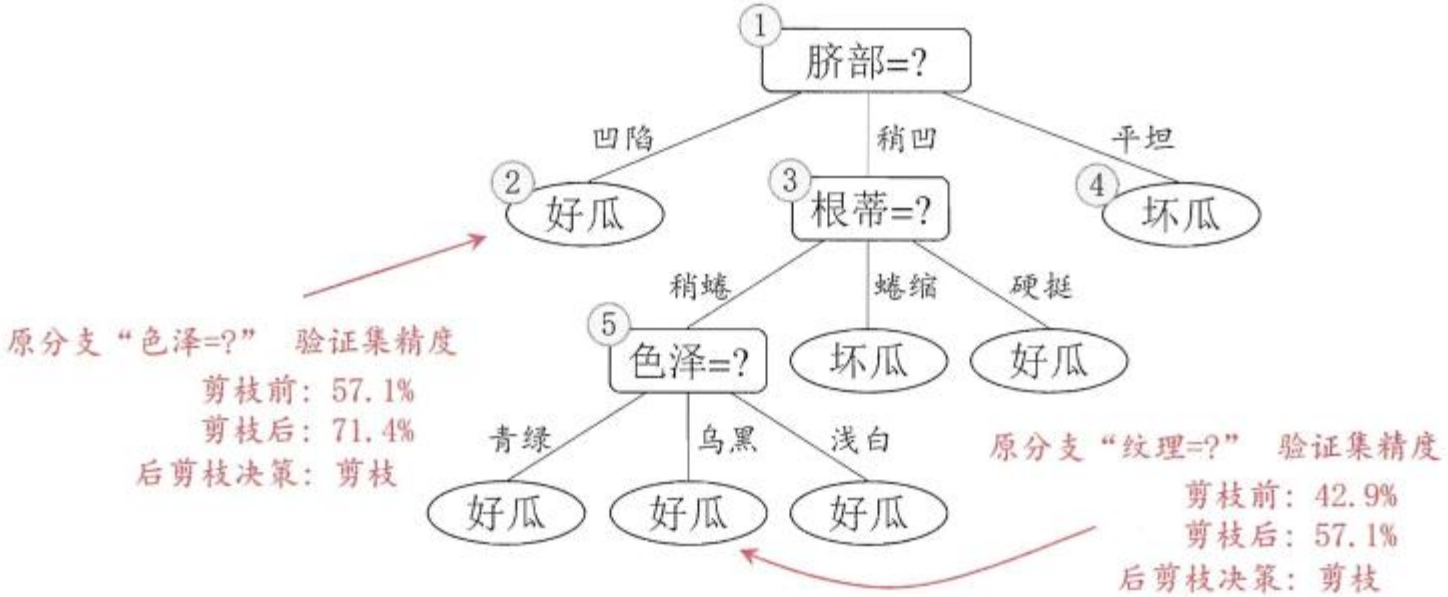


预剪枝（pruning）


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


①
脐部=?
验证集精度
“脐部=?” 划分前：42.9%
划分后：71.4%
预剪枝决策：划分
凹陷
稍凹
平坦
②
好瓜
③
好瓜
④
坏瓜
验证集精度
“色泽=?” 划分前：71.4%
划分后：57.1%
预剪枝决策：禁止划分
验证集精度
“根蒂=?” 划分前：71.4%
划分后：71.4%
预剪枝决策：禁止划分

后剪枝（pruning）


1
脐部=?
凹陷
稍凹
平坦
2
好瓜
3
根蒂=?
4
坏瓜
稍蜷
蜷缩
硬挺
5
色泽=?
坏瓜
好瓜
青绿
乌黑
浅白
好瓜
好瓜
好瓜
原分支“色泽=?” 验证集精度
剪枝前：57.1%
剪枝后：71.4%
后剪枝决策：剪枝
原分支“纹理=?” 验证集精度
剪枝前：42.9%
剪枝后：57.1%
后剪枝决策：剪枝

连续值处理

https://wenku.baidu.com/view/0ef3a961bcd126fff7050bba.html

连续属性的可取值不再有限，不能直接根据连续属性的可取值来对结点进行划分，用到连续属性离散化技术——二分法（bi-partition）

给定样本集D和连续属性a,假定a在D上出现了n个不同的取值，将这些值进行排序得到$\{a^1, a^2, \ldots, a^n\}$

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

在候选集上找划分点t

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

表4：包含连续属性的西瓜数据集

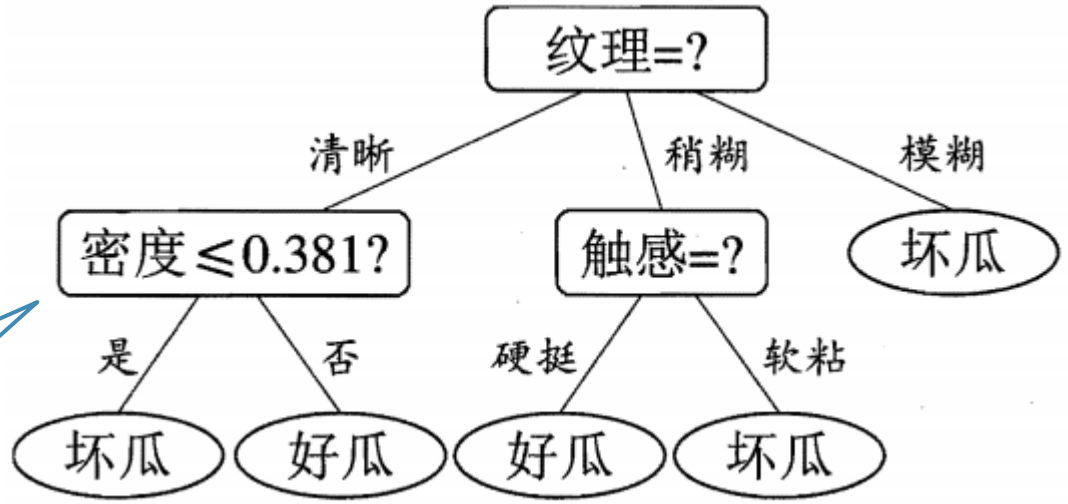
编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	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

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

属性“密度”最大信息增益为0.262，所对应的划分点为0.381

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

属性“纹理”被选作根结点划分属性，此后结点划分过程递归进行



若当前结点划分的属性，
该属性还可作为后代结点
的划分属性

缺失值处理

https://wenku.baidu.com/view/0ef3a961bcd126fff7050bba.html

- 如何在属性值缺失的情况下进行划分属性选择

表示D中属性a上没有缺失值的样本

表示$\tilde{D}$中属于第k类的样本子集

表示$\tilde{D}$中在属性a上取值为a^v的样本子集

$$\rho = \frac{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in D} w_{\boldsymbol{x}}},$$

$$\tilde{p}_k = \frac{\sum_{\boldsymbol{x} \in \tilde{D}_k} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}} \quad (1 \leqslant k \leqslant |\mathcal{Y}|),$$

$$\tilde{r}_v = \frac{\sum_{\boldsymbol{x} \in \tilde{D}^v} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}} \quad (1 \leqslant v \leqslant V).$$

- 对于属性a，ρ表示无缺失值样本所占比例
- $\tilde{p}_k$表示无缺失值样本中第k类所占的比例
- $\tilde{r}_v$表示无缺失值样本中在属性a上取值为a^v的样本子集

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

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

缺失值处理

- 给定划分属性，若样本在该属性值上的值缺失该如何对样本进行划分

根结点中各样本的权重初始化为1

$$\rho = \frac{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in D} w_{\boldsymbol{x}}},$$

$$\tilde{p}_k = \frac{\sum_{\boldsymbol{x} \in \tilde{D}_k} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}} \quad (1 \leqslant k \leqslant |\mathcal{Y}|),$$

$$\tilde{r}_v = \frac{\sum_{\boldsymbol{x} \in \tilde{D}^v} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}} \quad (1 \leqslant v \leqslant V).$$

若样本$\boldsymbol{x}$在划分属性a上取值已知，则可直接划分到其对应的子结点，权值保持不变。

若样本$\boldsymbol{x}$在划分属性a上取值未知，则将$\boldsymbol{x}$划分入所有子结点，在对应子结点中调整样本权值为$\tilde{r}_v * \omega_x$

表5：包含缺失值的西瓜数据集

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

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

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

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

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

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

$$\mathrm{Gain}(\tilde{D}, \text{色泽}) = \mathrm{Ent}(\tilde{D}) - \sum_{v=1}^{3} \tilde{r}_v \mathrm{Ent}(\tilde{D}^v)$$

$$= 0.985 - \left(\frac{4}{14}\times 1.000 + \frac{6}{14}\times 0.918 + \frac{4}{14}\times 0.000\right)$$

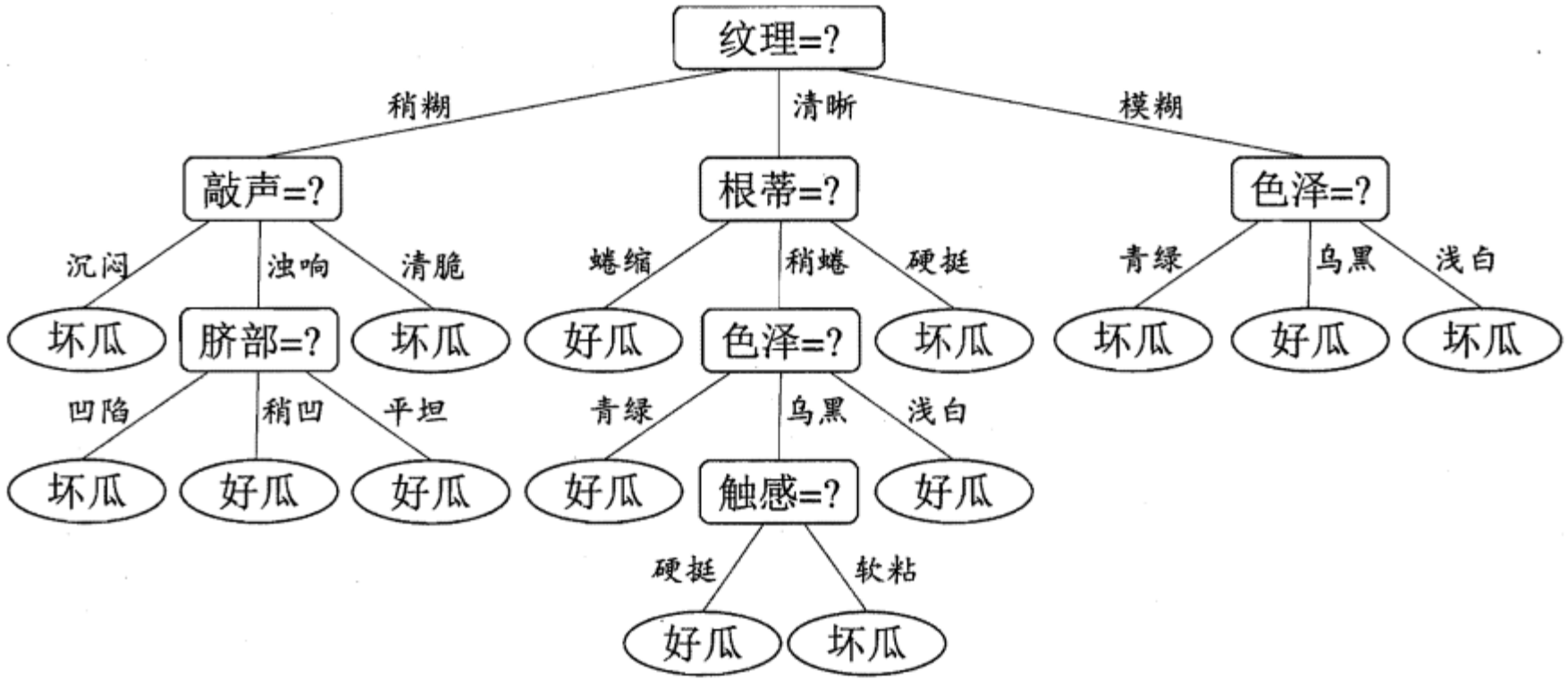
$$= 0.306\ .$$

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

$\mathrm{Gain}(D, \text{色泽}) = 0.252;\quad \mathrm{Gain}(D, \text{根蒂}) = 0.171;$

$\mathrm{Gain}(D, \text{敲声}) = 0.145;\quad \mathrm{Gain}(D, \text{纹理}) = 0.424;$

$\mathrm{Gain}(D, \text{脐部}) = 0.289;\quad \mathrm{Gain}(D, \text{触感}) = 0.006.$

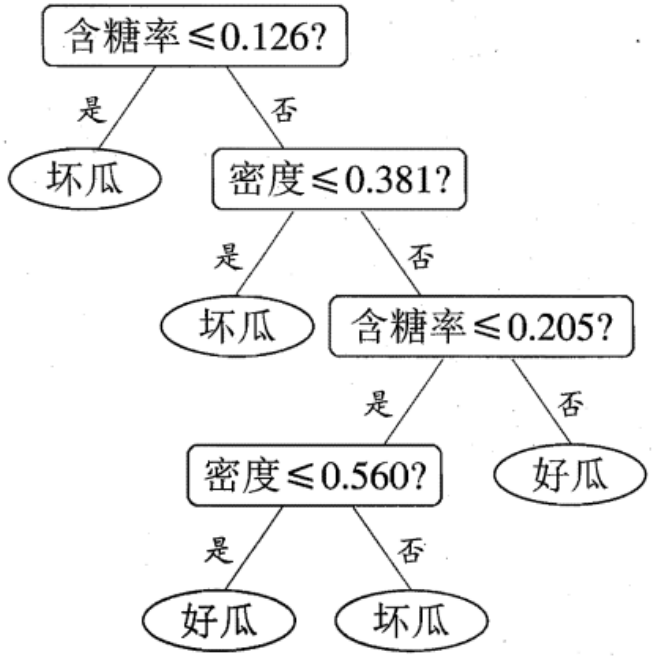

纹理=?
稍糊
清晰
模糊
敲声=?
根蒂=?
色泽=?
沉闷
浊响
清脆
蜷缩
稍蜷
硬挺
青绿
乌黑
浅白
坏瓜
脐部=?
坏瓜
好瓜
色泽=?
坏瓜
坏瓜
好瓜
坏瓜
凹陷
稍凹
平坦
青绿
乌黑
浅白
坏瓜
好瓜
好瓜
好瓜
触感=?
好瓜
硬挺
软粘
好瓜
坏瓜

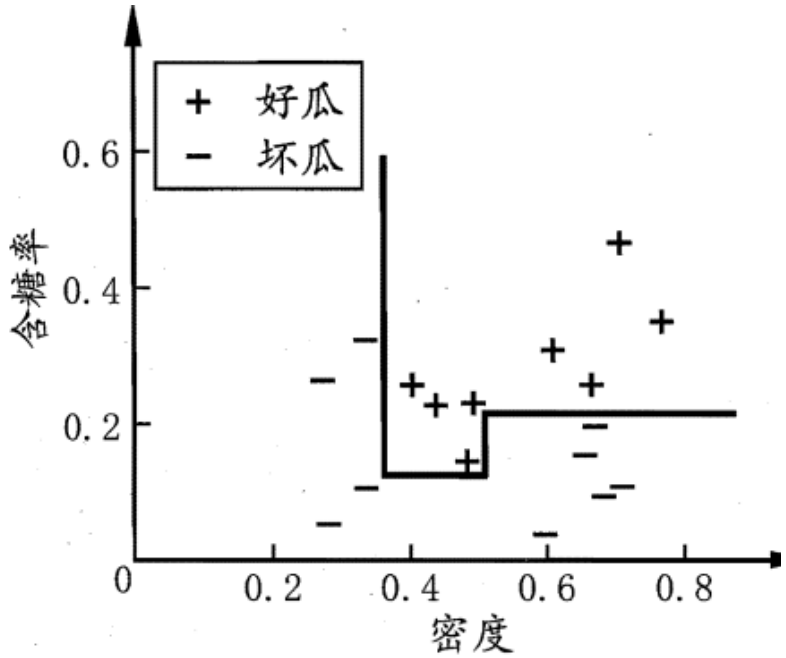
多变量决策树

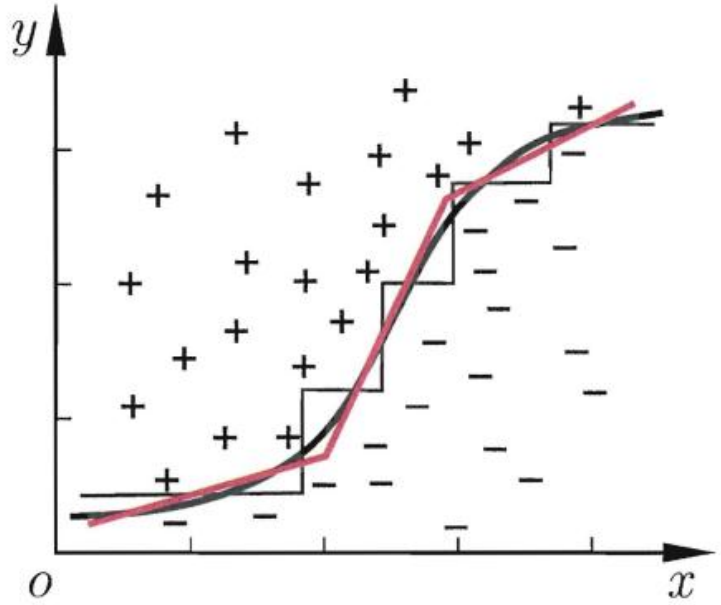
- 决策树所形成的分类边界：轴平行（axis-parallel）
- 每一段划分都直接对应了属性取值，较好的可解释性

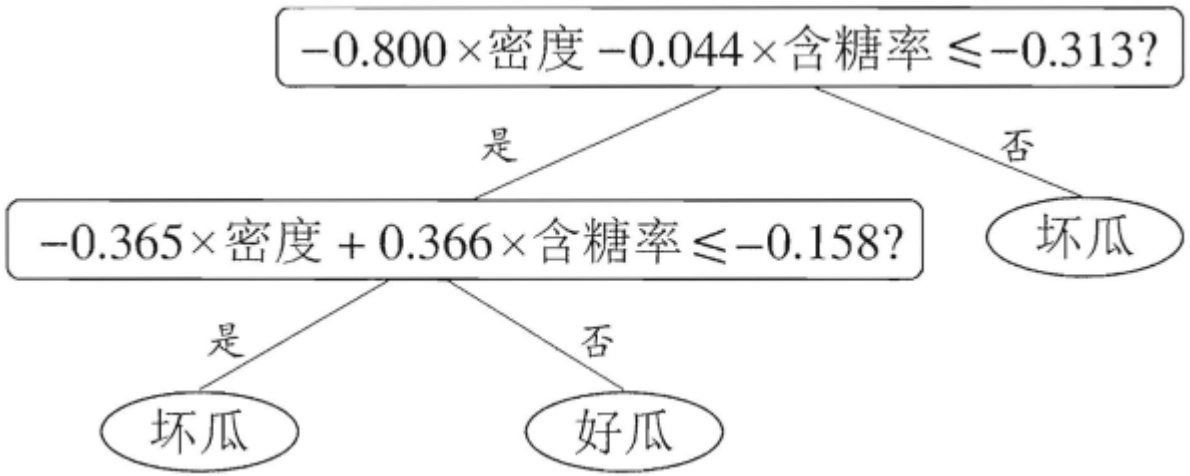
表5：只包含连续属性的西瓜数据集

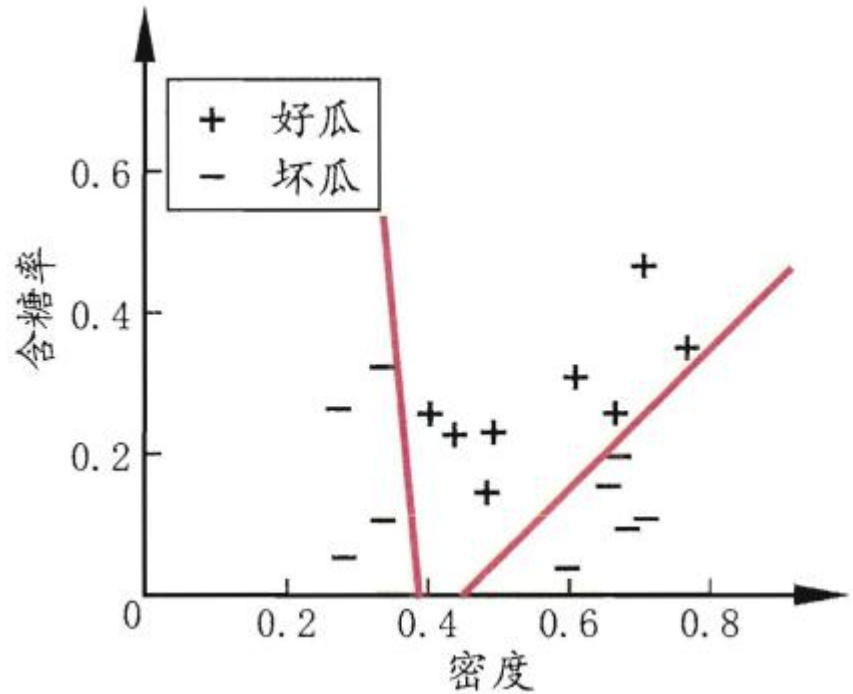
编号	密度	含糖率	好瓜
1	0.697	0.460	是
2	0.774	0.376	是
3	0.634	0.264	是
4	0.608	0.318	是
5	0.556	0.215	是
6	0.403	0.237	是
7	0.481	0.149	是
8	0.437	0.211	是
9	0.666	0.091	否
10	0.243	0.267	否
11	0.245	0.057	否
12	0.343	0.099	否
13	0.639	0.161	否
14	0.657	0.198	否
15	0.360	0.370	否
16	0.593	0.042	否
17	0.719	0.103	否


含糖率≤0.126?
是
否
坏瓜
密度≤0.381?
是
否
坏瓜
含糖率≤0.205?
是
否
密度≤0.560?
好瓜
是
否
好瓜
坏瓜


+ 好瓜
− 坏瓜
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
含糖率
密度


y
O
x


−0.800×密度 −0.044×含糖率 ≤−0.313?
是
否
−0.365×密度 + 0.366×含糖率 ≤−0.158?
坏瓜
是
否
坏瓜
好瓜


好瓜
坏瓜
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
密度
含糖率

实验

- 使用scikit-learn提供的决策树实验解决分类问题
- http://scikit-learn.org/stable/modules/tree.html

谢 谢