机器学习

苏州大学计算机科学与技术学院
自然语言处理实验室
主讲：周夏冰
邮箱：zhouxiabing@suda.edu.cn

前情回顾

- 决策树
  - 决策树学习本质：从训练数据集中归纳出一组分类规则
  - 学习目标：根据给定的训练数据集构建一个决策树模型，使它能够对实例进行正确的分类，泛化能力强
  - 能够度量节点“纯度”的指标：所包含的样本尽可能的属于同一类别
    - 熵：随机变量不确定性的度量 $H(X) = -\sum_{i=1}^{n} p_i log p_i$

- ID3: 信息增益表示得知特征X的信息而使得类Y的信息的不确定性减少的程度

前情回顾

- 例：

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

$$\mathbf{H(D)} = -\sum_{k=1}^{K} \frac{|D_k|}{|D|} \log_2 \frac{|D_k|}{|D|}$$

$$= -\frac{6}{15} log_2 \frac{6}{15} - \frac{9}{15} log_2 \frac{9}{15}$$

$$= \mathbf{0.971}$$

前情回顾

- 例：

$\mathbf{H(D) = 0.971}$　$H(D|A = \text{年龄}) = 0.888$　$H(D|A = \text{有工作}) = 0.647$

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

$H(D|A = \text{有房子}) = 0.551$ √

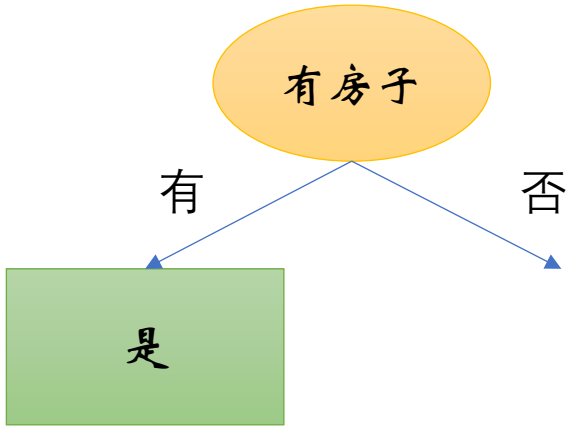
$H(D|A = \text{信贷情况}) = 0.608$

取信息增益H(D)-H(D|A)最大的特征

前情回顾

- 例：

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



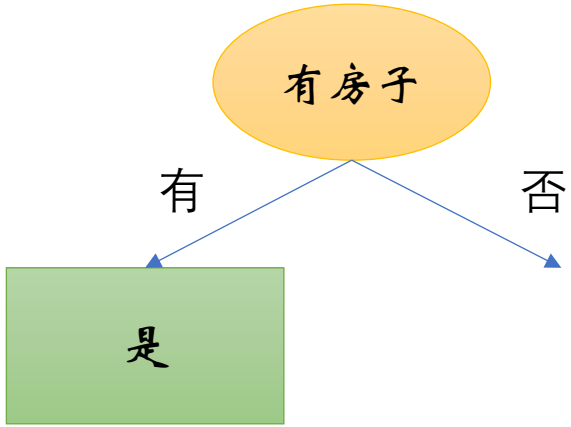
前情回顾

类别不够纯

房子特征用过了

- 例：

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



$$H(D_1) = -\frac{3}{9}log_2\frac{3}{9} - \frac{6}{9}log_2\frac{6}{9} = 0.918$$

$$H(\text{年龄}|D_1) = -\frac{4}{9}\left(\frac{3}{4}log_2\frac{3}{4} + \frac{1}{4}log_2\frac{1}{4}\right) - \frac{2}{9}\left(\frac{2}{2}log_2\frac{2}{2}\right) - \frac{3}{9}\left(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}\right) = 0.667$$

$$H(\text{有工作}|D_1) = -\frac{6}{9}\left(\frac{6}{6}log_2\frac{6}{6}\right) - \frac{3}{9}\left(\frac{3}{3}log_2\frac{3}{3}\right) = 0$$

$$H(\text{信贷}|D_1) = -\frac{4}{9}\left(\frac{4}{4}log_2\frac{4}{4}\right) - \frac{4}{9}\left(\frac{2}{4}log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}\right) - \frac{1}{9}\left(\frac{1}{1}log_2\frac{1}{1}\right) = 0.889$$

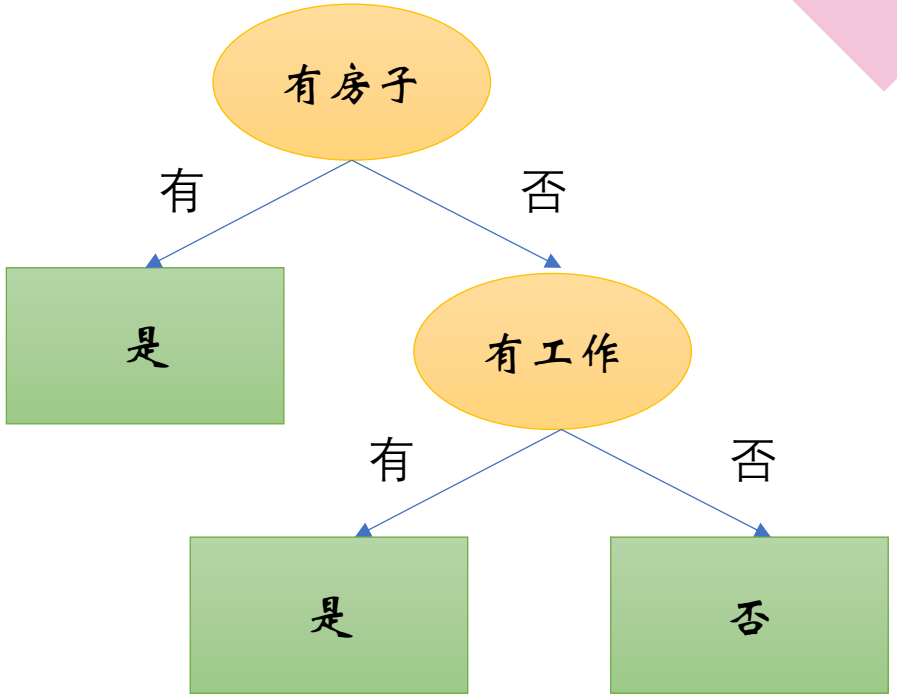
前情回顾

类别不够纯

房子特征用过了

- 例：

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



前倾回顾

C4.5算法

- 以信息增益作为划分训练数据集的特征，特征的问题
- 信息增益比

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

$H(D) = 0.971$

$H(D|A = \text{年龄}) = 0.888$

g(D, 年龄）=0.971-0.888=0.083

- $g_A(D,A) = \frac{g(D,A)}{H_A(D)}$
- $H_A(D) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} log_2 \frac{|D_i|}{|D|}$ 其中，n表示特征A的取值个数
- $H_{\text{年龄}}(D) = -\frac{5}{15} log_2 \frac{5}{15} - \frac{5}{15} log_2 \frac{5}{15} - \frac{5}{15} log_2 \frac{5}{15} = 1.58$

$g_{\text{年龄}}(D, \text{年龄}) = \frac{0.083}{1.58} = 0.053$

- $H_{\text{有工作}} = 0.918$ $H_{\text{有房子}} = 0.971$ $H_{\text{信贷}} = 1.566$ $H_{\text{编号}} = 3.907$

$g_{\text{工作}}(D, \text{工作}) = 0.705$

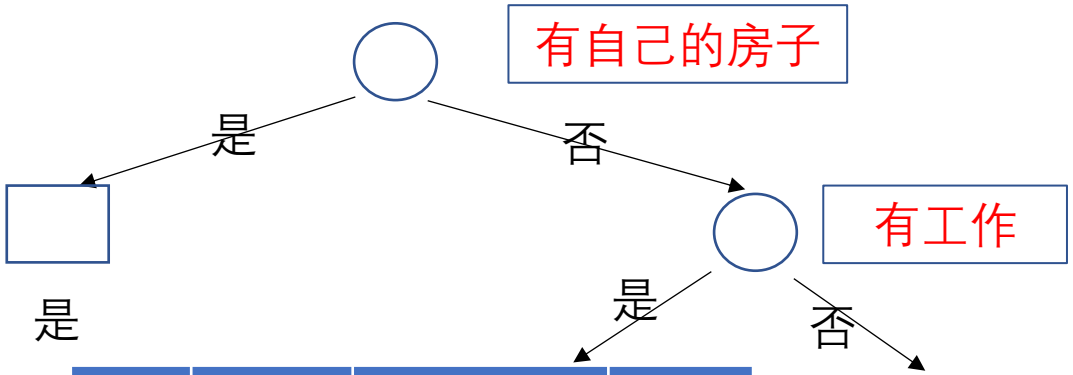
$g_{\text{有房子}}(D, \text{有房子}) = 0.567$

$g_{\text{信贷}}(D, \text{信贷}) = 0.388$

$g_{\text{编号}}(D, \text{编号}) = 0.235$

前倾回顾

- 例
  - 这里生成的决策树只用到两个特征（两个内节点），ID3算法容易存在过拟合问题。



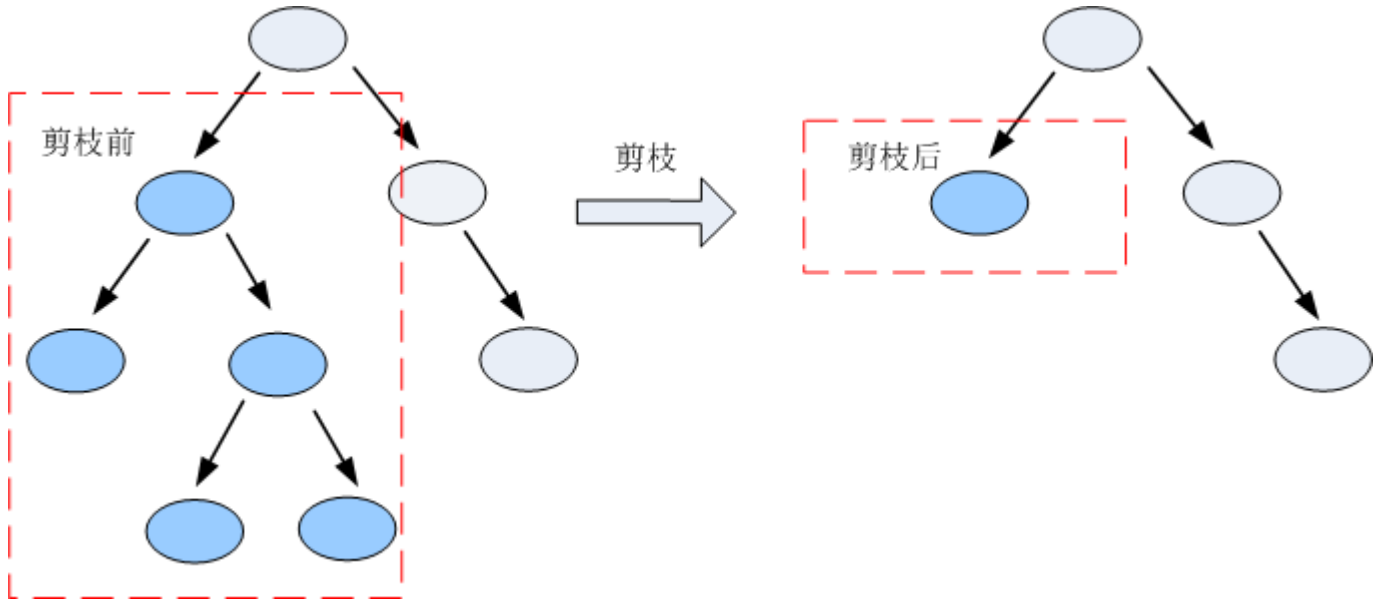
ID	年龄	信贷情况	类别
3	青年	好	是
13	老年	好	是
14	老年	非常好	是

表3

ID	年龄	信贷情况	类别
1	青年	一般	否
2	青年	好	否
5	青年	一般	否
6	中年	一般	否
7	中年	好	否
15	老年	一般	否

前情回顾

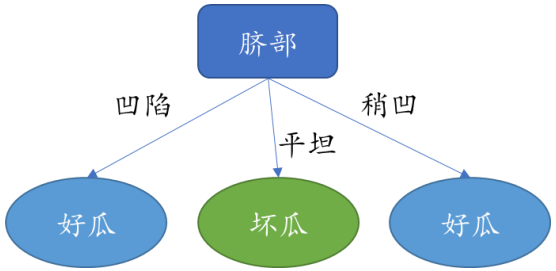
- 决策树的剪枝：预测误差与树的复杂度之间的平衡，以缓解决策树过拟合问题
  - 预剪枝：边构建边剪枝
  - 后剪枝：先构建后剪枝



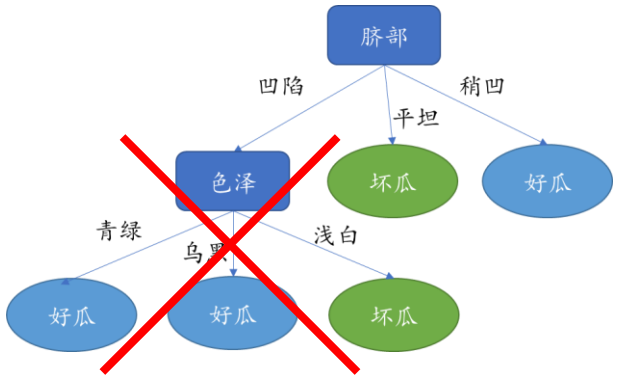
前情回顾

- 预剪枝
  - 先构建单节点树
  - 再构建深度为1
  - 再构建深度为2

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
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	×



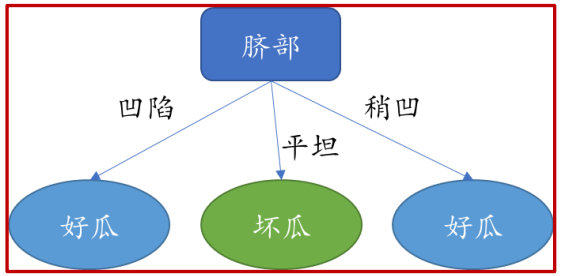
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	✓
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×



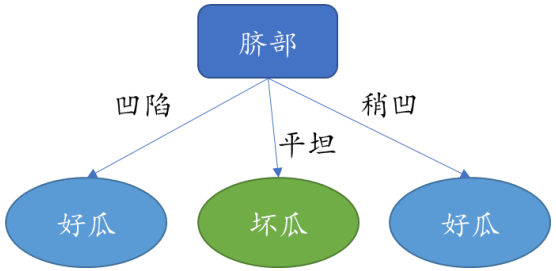
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	×
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×

前情回顾

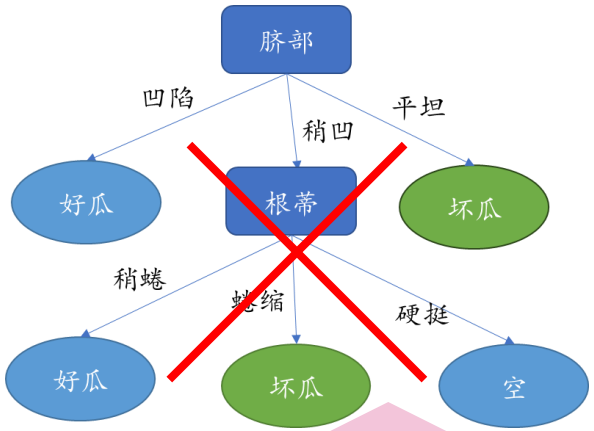
- 预剪枝
  - 先构建单节点树
  - 再构建深度为1
  - 再构建深度为2



编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
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	×



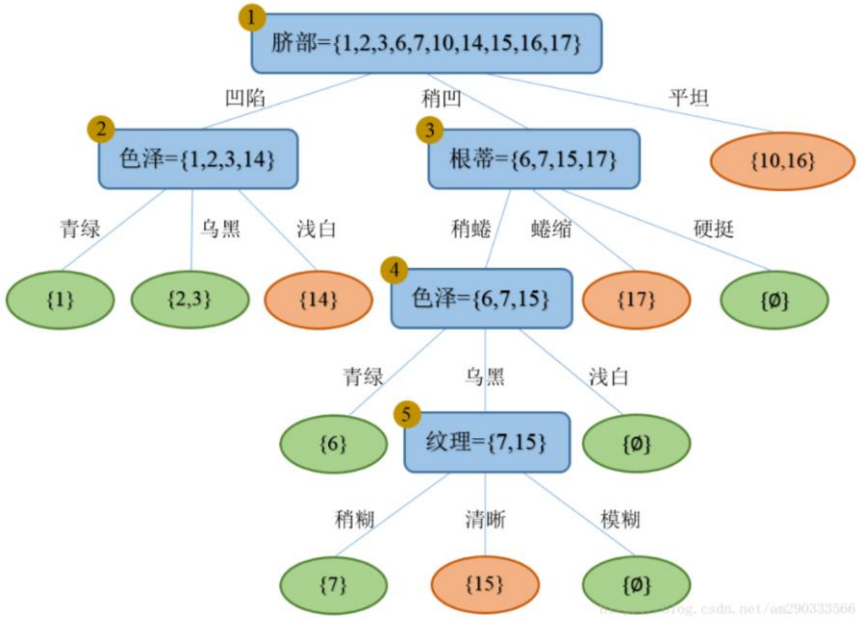
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	✓
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×



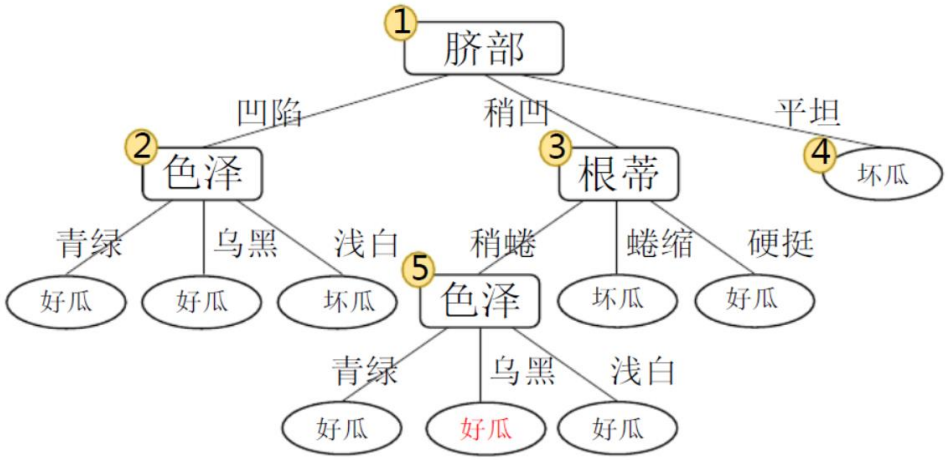
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	✓
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×

前情回顾

- 后剪枝



编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
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	×

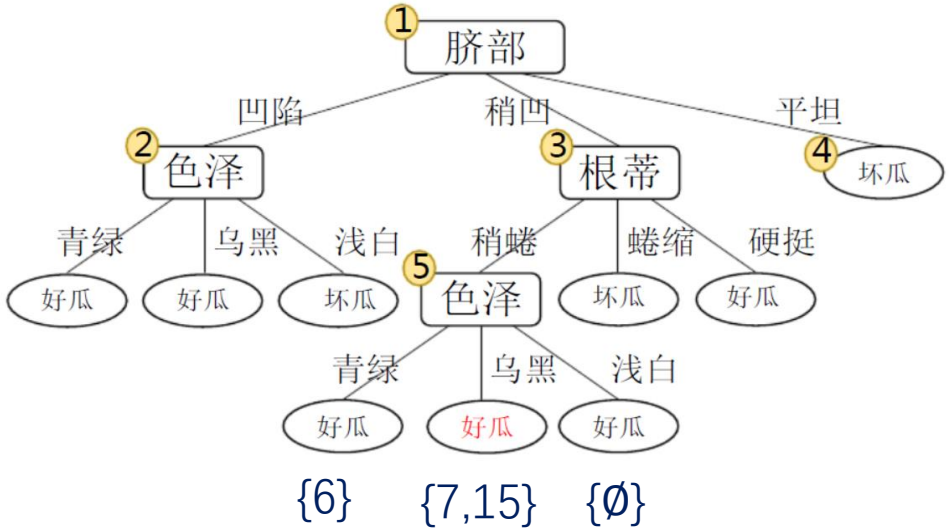


编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	×
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×

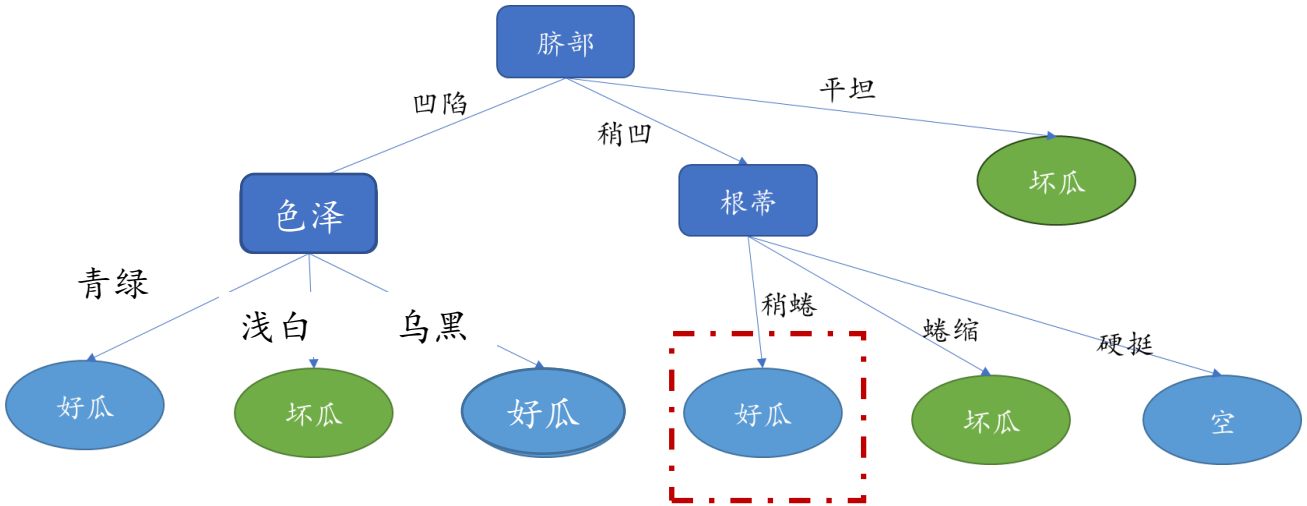
样本：{7,15}

前情回顾

- 后剪枝



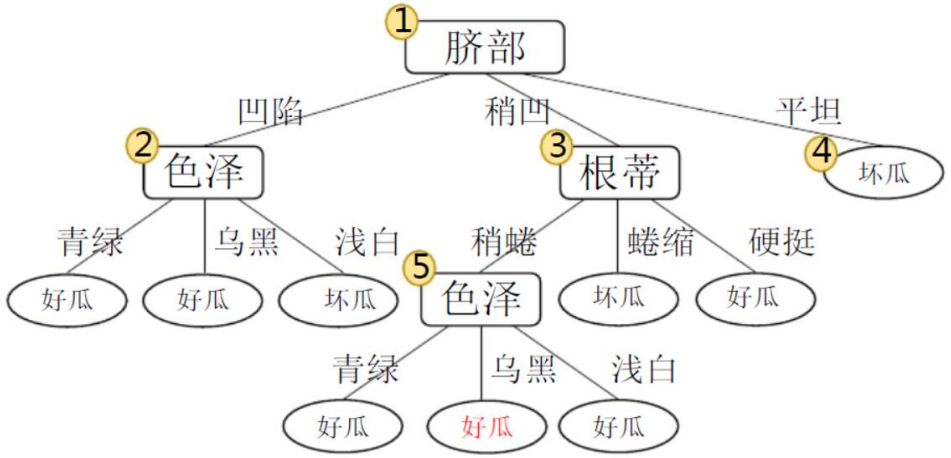
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	×
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×



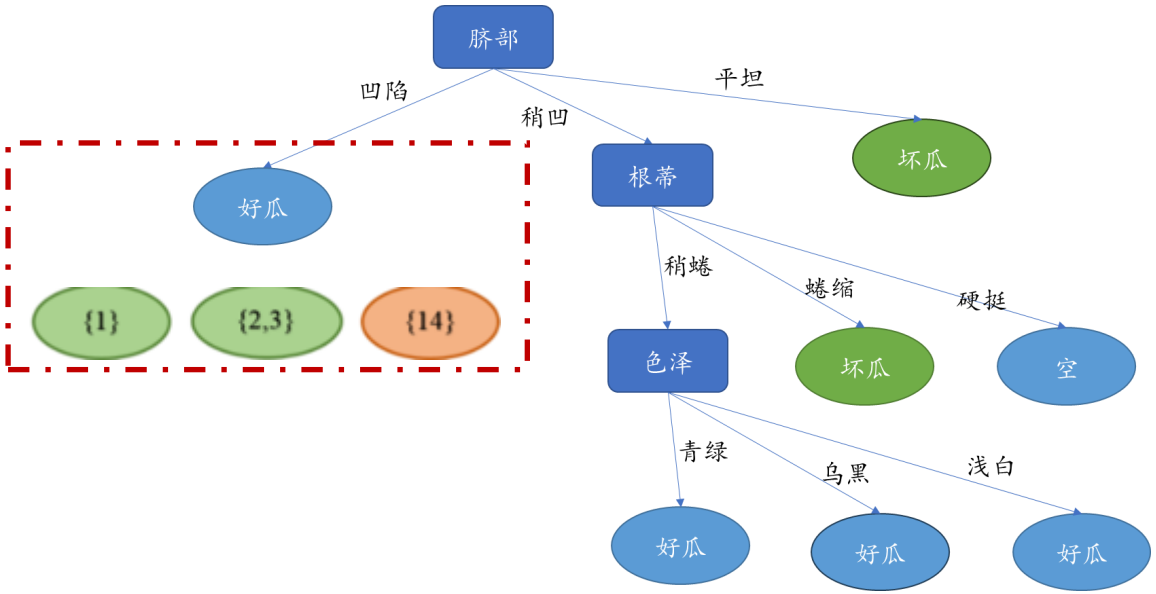
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	×
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×

前情回顾

- 后剪枝



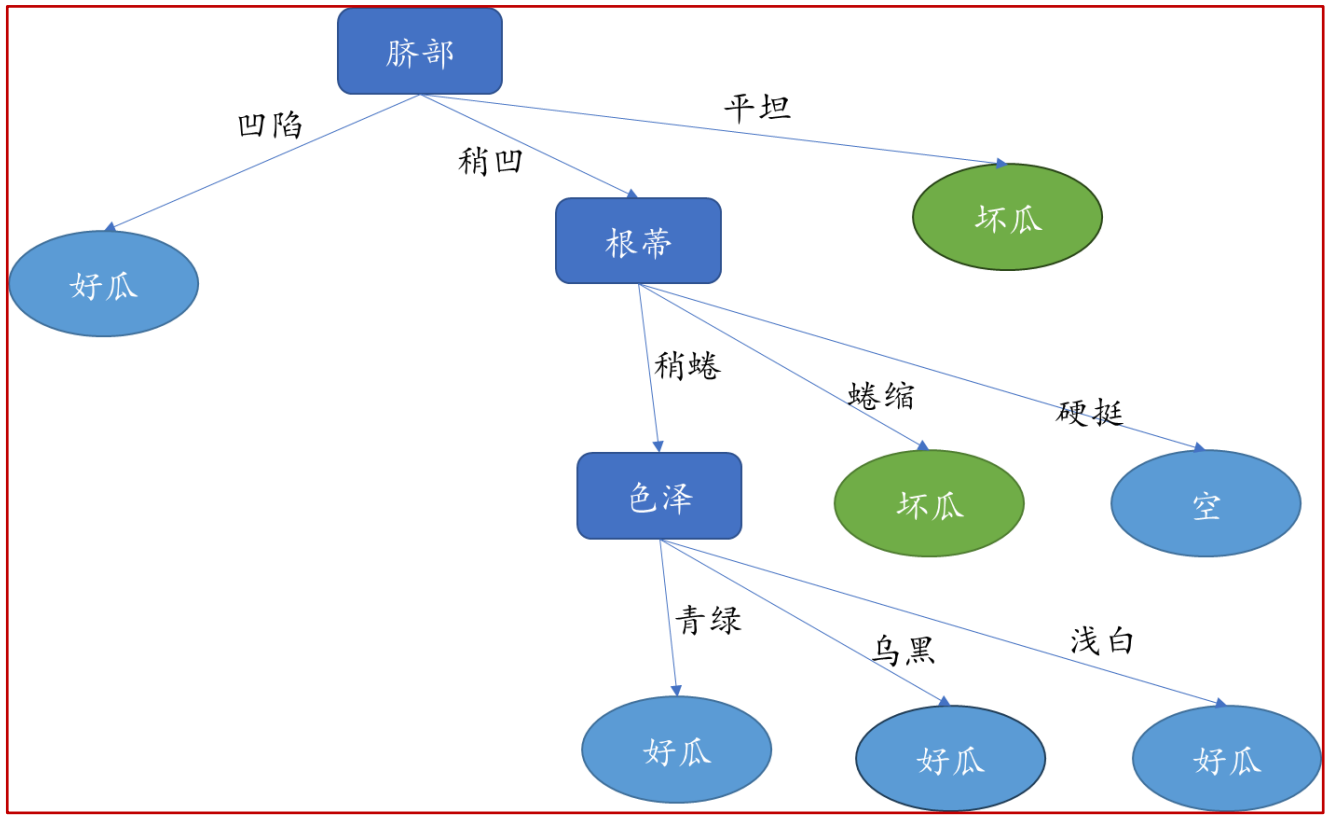
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	×
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×



编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
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	×

前情回顾

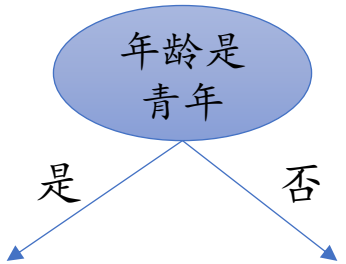
- 后剪枝



前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0

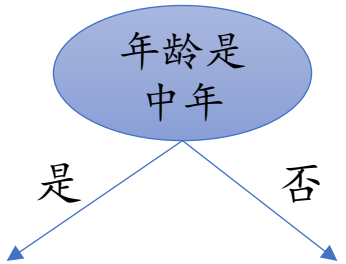


$$\text{Gini}(D, A1 = \text{青年}) = \frac{5}{15}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{10}{15}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) = 0.44$$

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0

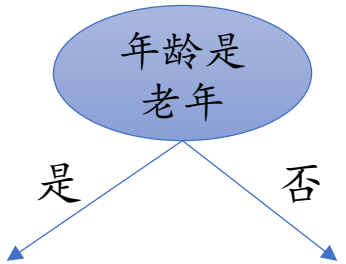


$$\text{Gini}(D, A1 = \text{中年}) = \frac{5}{15}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{10}{15}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) = 0.48$$

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0



$$\text{Gini}(D, A1 = \text{老年}) = \frac{5}{15}\left(1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2\right) + \frac{10}{15}\left(1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2\right) = 0.44$$

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0

$$\text{Gini}(D, A1 = \text{青年}) = \frac{5}{15}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{10}{15}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) = 0.44$$

$$\text{Gini}(D, A1 = \text{中年}) = \frac{5}{15}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{10}{15}\left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right) = 0.48$$

$$\text{Gini}(D, A1 = \text{老年}) = \frac{5}{15}\left(1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2\right) + \frac{10}{15}\left(1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2\right) = 0.44$$

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0

$$\text{Gini}(D, A2=0) = \frac{10}{15}\left(1-\left(\frac{6}{10}\right)^2-\left(\frac{4}{10}\right)^2\right)+\frac{5}{15}\left(1-\left(\frac{5}{5}\right)^2\right) = 0.32$$

$\text{Gini}(D, A1 = \text{青年}) = 0.44$

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D, A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0

$$\text{Gini}(D, A3 = 0) = \frac{9}{15}\left(1 - \left(\frac{3}{9}\right)^2 - \left(\frac{6}{9}\right)^2\right) + \frac{6}{15}\left(1 - \left(\frac{6}{6}\right)^2\right) = 0.27$$

$\text{Gini}(D, A1 = \text{青年}) = 0.44$

$\text{Gini}(D, A2 = 0) = 0.32$

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0

$$\text{Gini}(D, A4 = \text{一般}) = \frac{5}{15}\left(1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2\right) + \frac{10}{15}\left(1 - \left(\frac{8}{10}\right)^2 - \left(\frac{2}{10}\right)^2\right) = 0.32$$

$$\text{Gini}(D, A4 = \text{好}) = \frac{6}{15}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right) + \frac{9}{15}\left(1 - \left(\frac{5}{9}\right)^2 - \left(\frac{4}{9}\right)^2\right) = 0.47$$

$$\text{Gini}(D, A4 = \text{非常好}) = \frac{4}{15}\left(1 - \left(\frac{4}{4}\right)^2\right) + \frac{11}{15}\left(1 - \left(\frac{5}{11}\right)^2 - \left(\frac{6}{11}\right)^2\right) = 0.36$$

$\text{Gini}(D, A1 = \text{青年}) = 0.44$

$\text{Gini}(D, A2 = 0) = 0.32$

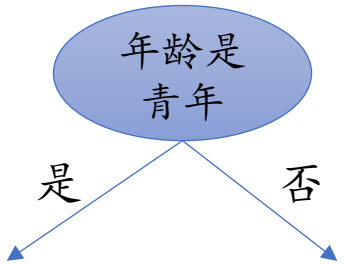
$\text{Gini}(D, A3 = 0) = 0.27$ ✓

$\text{Gini}(D, A4 = 0) = 0.32$

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$
  - 回归指标：
    - $$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right]$$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0.1 0
2	青年	0	0	好	0.3 0
3	青年	1	0	好	0.7 1
4	青年	1	1	一般	0.6 1
5	青年	0	0	一般	0.1 0
6	中年	0	0	一般	0.4 0
7	中年	0	0	好	0.5 0
8	中年	1	1	好	0.9 1
9	中年	0	1	非常好	0.8 1
10	中年	0	1	非常好	0.8 1
11	老年	0	1	非常好	0.9 1
12	老年	0	1	好	0.8 1
13	老年	1	0	好	0.7 1
14	老年	1	0	非常好	0.9 1
15	老年	0	0	一般	0.4 0



$$s(D, A1 = \text{青年}) = \min_{c1}\left((0.1-c1)^2 + (0.3-c1)^2 + (0.7-c1)^2 + (0.6-c1)^2\right) + \min_{c2}\left((0.4-c2)^2 + \cdots + (0.4-c2)^2\right)$$

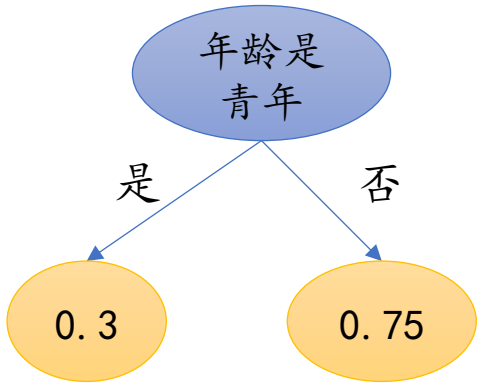
C1=0.3　　C2=0.75

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$
  - 回归指标：
    - $$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right]$$

$s(D, A1 = \text{青年}) = 0.715$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0

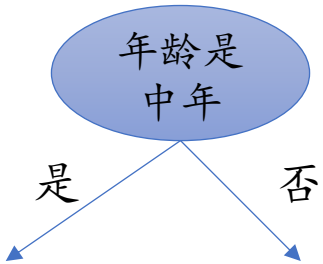


C1=0.3　　C2=0.75

前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$
  - 回归指标：
    - $$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right]$$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0
2	青年	0	0	好	0
3	青年	1	0	好	1
4	青年	1	1	一般	1
5	青年	0	0	一般	0
6	中年	0	0	一般	0
7	中年	0	0	好	0
8	中年	1	1	好	1
9	中年	0	1	非常好	1
10	中年	0	1	非常好	1
11	老年	0	1	非常好	1
12	老年	0	1	好	1
13	老年	1	0	好	1
14	老年	1	0	非常好	1
15	老年	0	0	一般	0



$$s(D, A1 = \text{中年}) = \min_{c1}\left((0.4-c1)^2 + (0.5-c1)^2 + (0.9-c1)^2 + (0.8-c1)^2 + (0.8-c1)^2\right) + \min_{c2}\left((0.1-c2)^2 + \cdots + (0.4-c2)^2\right)$$

C1=0.68　　C2=0.55

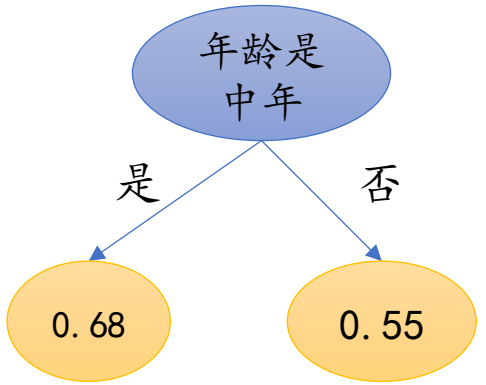
前情回顾

- **CART: 分类回归树**
  - 二叉树
  - 分类指标：基尼指数
    - $Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$
    - $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$
  - 回归指标：
    - $$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right]$$

$s(D, A1 = 中年) = 1.16$

$s(D, A1 = 青年) = 0.715$

编号	X:年龄	X:是否有工作 (1：是，0：否)	X:是否买房 (1：是，0：否)	X:信贷表现	Y:是否放贷 (1：是，0：否)
1	青年	0	0	一般	0.1 0
2	青年	0	0	好	0.3 0
3	青年	1	0	好	0.7 1
4	青年	1	1	一般	0.6 1
5	青年	0	0	一般	0.1 0
6	中年	0	0	一般	0.4 0
7	中年	0	0	好	0.5 0
8	中年	1	1	好	0.9 1
9	中年	0	1	非常好	0.8 1
10	中年	0	1	非常好	0.8 1
11	老年	0	1	非常好	0.9 1
12	老年	0	1	好	0.8 1
13	老年	1	0	好	0.7 1
14	老年	1	0	非常好	0.9 1
15	老年	0	0	一般	0.4 0



C1=0.68　　C2=0.55

前情回顾

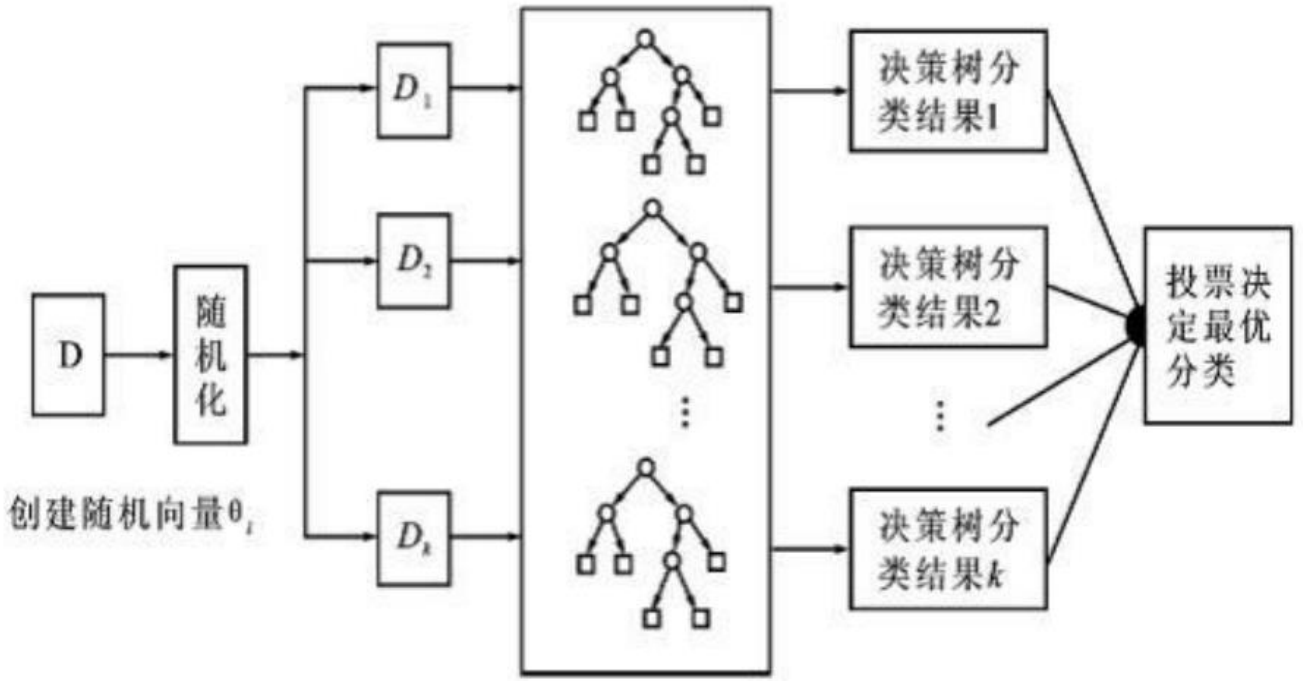
- 决策树
  - ❖ 属性连续数值
    - ❖ 切分点，二叉树
  - ❖ 缺失值
    - ❖ 特征选择：每个属性按照未缺失计算
    - ❖ 缺失样例子集划分：每个子集均有，赋予权重

前情回顾

- 集成学习通过构建并结合多个学习器来完成学习任务
  - 好而不同
- Boosting & Bagging

前情回顾

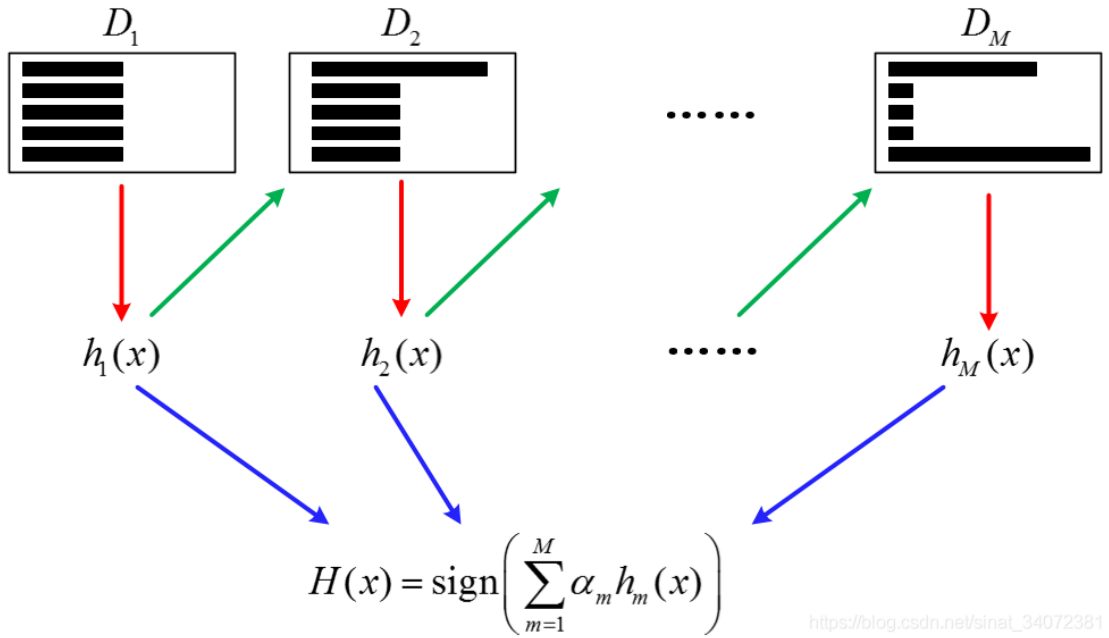
- Bagging
  - 随机森林



前情回顾

❖ Boosting

- AdaBoost——基于“基学习器的线性组合”



$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right);$$

$$\mathcal{D}_{t+1}(\boldsymbol{x}) = \frac{\mathcal{D}_t(\boldsymbol{x})}{Z_t} \times \begin{cases} \exp(-\alpha_t), & \text{if } h_t(\boldsymbol{x}) = f(\boldsymbol{x}) \\ \exp(\alpha_t), & \text{if } h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}) \end{cases}$$

$$= \frac{\mathcal{D}_t(\boldsymbol{x})\exp(-\alpha_t f(\boldsymbol{x}) h_t(\boldsymbol{x}))}{Z_t}$$

前情回顾

- Boosting
  - 提升树——基于CART决策树
  - $\min_{\theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m))$
  - $L\big(y, f(x)\big) = \big(y - f(x)\big)^2 = [y - f_{m-1}(x) - T(x, \theta_m)]^2 = [r - T(x, \theta_m)]^2$
  - 梯度提升： $r_{mi} = \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}$