



#### • 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  $H(X) = -\sum_{i=1}^{\infty} p_i log p_i$
- 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





'\	•					
1	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

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

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

$$= 0.971$$



·例·

$$H(D) = 0.971$$
  $H(D|A = 年龄) = 0.888$   $H(D|A = 有工作) = 0.647$ 

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

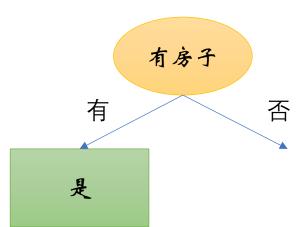
$$H(D|A = 有房子) = 0.551$$
 √

$$H(D|A = 信贷情况) = 0.608$$

取信息增益H(D)-H(D|A)最大的特征



4	•					
1	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
Ī	13	老年	是	否	好	是
	14	老年	是	否	非常好	是
_	15	老年	否	否	一般	否

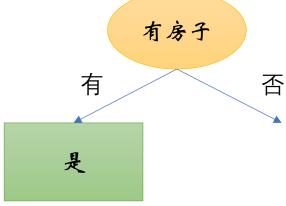


#### 类别不够纯

#### 房子特征用过了



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(年龄|D<sub>1</sub>)
$$= -\frac{4}{9} \left( \frac{3}{4} log_2 \frac{3}{4} + \frac{1}{4} log_2 \frac{1}{4} \right) - \frac{2}{9} \left( \frac{2}{2} log_2 \frac{2}{2} \right)$$

$$-\frac{3}{9} \left( \frac{1}{3} log_2 \frac{1}{3} + \frac{2}{3} log_2 \frac{2}{3} \right) = 0.667$$

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

H(信贷|D<sub>1</sub>)
$$= -\frac{4}{9} \left( \frac{4}{4} log_2 \frac{4}{4} \right) - \frac{4}{9} \left( \frac{2}{4} log_2 \frac{2}{4} + \frac{2}{4} log_2 \frac{2}{4} \right) - \frac{1}{9} \left( \frac{1}{1} log_2 \frac{1}{1} \right)$$

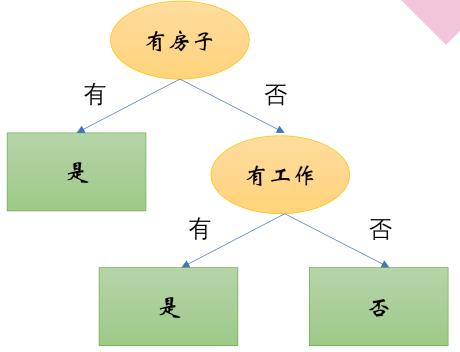
$$= 0.889$$

#### 类别不够纯

#### 房子特征用过了



1 •					
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	
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
  - $g_A(D,A) = \frac{g(D,A)}{H_A(D)}$

Ц(	D)	) =	Λ	Q'	71
П(	U,	) =	U.	7	/ <b>L</b>

$$H(D|A = 年龄) = 0.888$$

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

•	$H_A(D) = -$	$\sum_{i=1}^{n} rac{ D_i }{ D } log_2 rac{ D_i }{ D }$	其中,	n表示特征A的取值个数
---	--------------	--	-----	-------------

• 
$$H_{\text{++}}(D) = -\frac{5}{15} log_2 \frac{5}{15} - \frac{5}{15} log_2 \frac{5}{15} - \frac{5}{15} log_2 \frac{5}{15} = 1.58$$

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

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

$$g_{\perp 4 =}(D, \perp 1) = 0.705$$

$$g_{ ext{信贷}}(D$$
,信贷 $)=0.38$ 

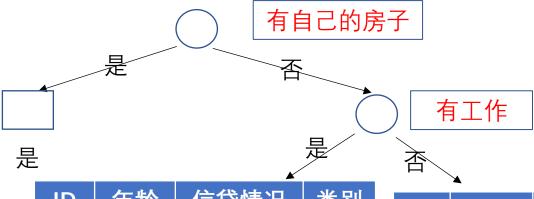
#### 前倾回顾



• 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

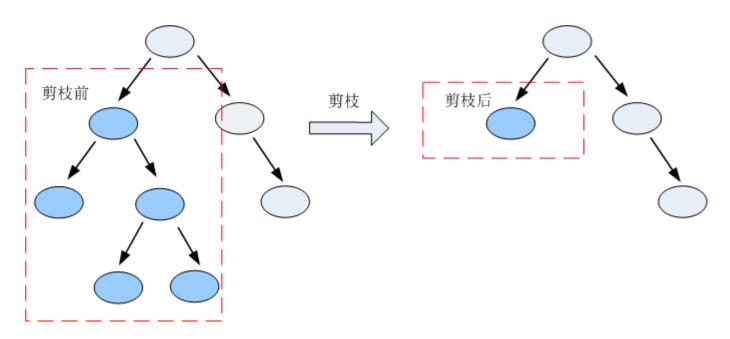


• 决策树的剪枝: 预测误差与树的复杂度之间的平衡, 以缓解决策树过拟

合问题

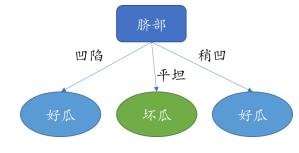
• 预剪枝: 边构建边剪枝

• 后剪枝: 先构建后剪枝



#### 好瓜

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



脐部
色泽 坏瓜 好瓜
青绿
好瓜好瓜坏瓜

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	<b>√</b>
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	X
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	X

色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	<b>✓</b>
浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	<b>✓</b>
青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
浅白	硬挺	清脆	模糊	平坦	硬滑	否	<b>✓</b>
浅白	蜷缩	浊响	模糊	平坦	软粘	否	<b>✓</b>
乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	<b>✓</b>
乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×
	青浅青浅浅乌	青绿蜷缩浅白稀蜷浅白硬挺浅白蜷缩乌黑稍蜷	青绿     蜷缩     沉闷       浅白     蜷缩     浊响       青绿     稍蜷     浊响       浅白     硬挺     清脆       浅白     蜷缩     浊响       乌黑     稍蜷     浊响	青绿     蜷缩     沉闷     清晰       浅白     蜷缩     浊响     清晰       青绿     稍蜷     浊响     稍糊       浅白     硬挺     清脆     模糊       浅白     蜷缩     浊响     模糊       乌黑     稍蜷     浊响     清晰	青绿     蜷缩     沉闷     清晰     凹陷       浅白     蜷缩     浊响     清晰     凹陷       青绿     稍蜷     浊响     稍糊     凹陷       浅白     硬挺     清脆     模糊     平坦       浅白     蜷缩     浊响     模糊     平坦       乌黑     稍蜷     浊响     清晰     稍凹	青緑     蜷缩     沉闷     清晰     凹陷     硬滑       浅白     蜷缩     浊响     清晰     凹陷     硬滑       青绿     稍蜷     浊响     稍糊     凹陷     硬滑       浅白     硬挺     清脆     模糊     平坦     软粘       乌黑     稍蜷     浊响     清晰     稍凹     硬滑	青緑     蜷縮     沉闷     清晰     凹陷     硬滑     是       浅白     蜷縮     浊响     清晰     凹陷     硬滑     是       青绿     稍蜷     浊响     稍糊     凹陷     硬滑     否       浅白     硬挺     清脆     模糊     平坦     软粘     否       乌黑     稍蜷     浊响     清晰     稍凹     硬滑     是

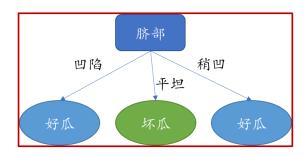
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	$\checkmark$
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	×
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	<b>  ✓</b>
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×

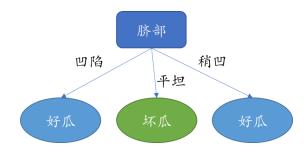




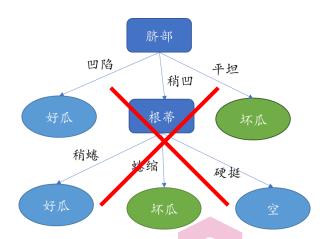
#### • 预剪枝

- 先构建单节点树
- 再构建深度为1
- 再构建深度为2





好瓜



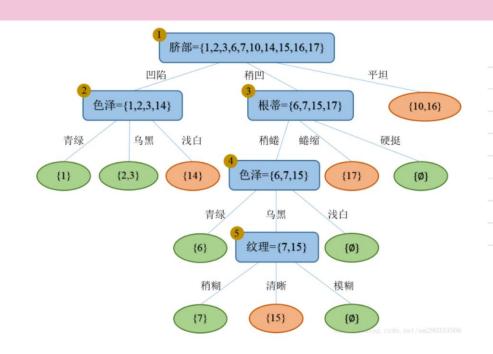
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	✓
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	✓
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	<b>√</b>
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	×
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	×
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×

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	<b>√</b>
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	<b>✓</b>
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	<b>✓</b>
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×

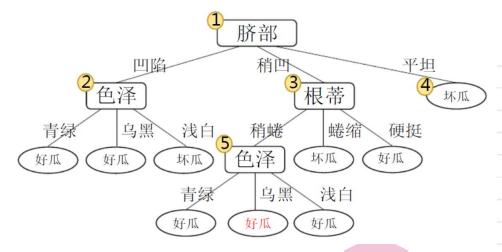
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	$\checkmark$
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	$\checkmark$
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	$\checkmark$
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	$\checkmark$
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	$\checkmark$
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×



#### • 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	<b>✓</b>
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	<b>✓</b>
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	<b>✓</b>
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	$\overline{}$
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	x
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	x

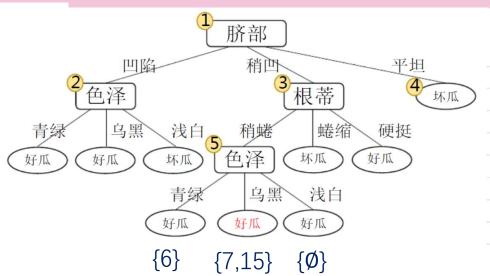
样本: {7,15}



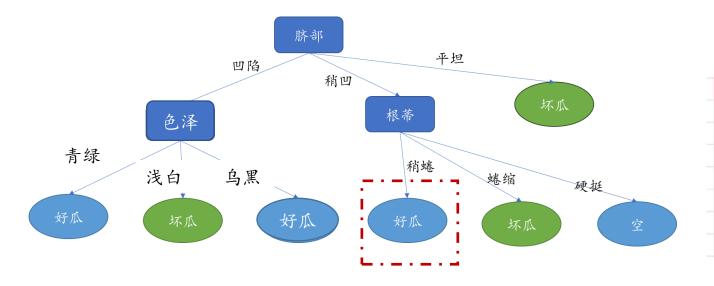


顺

• 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	x
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	] x

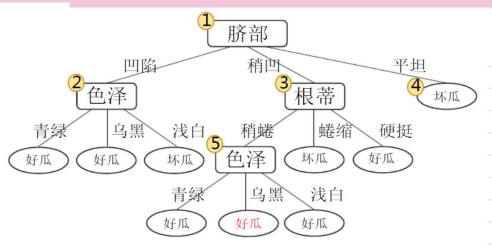


编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	$\checkmark$
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	x
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	x
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	] 🗸
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×

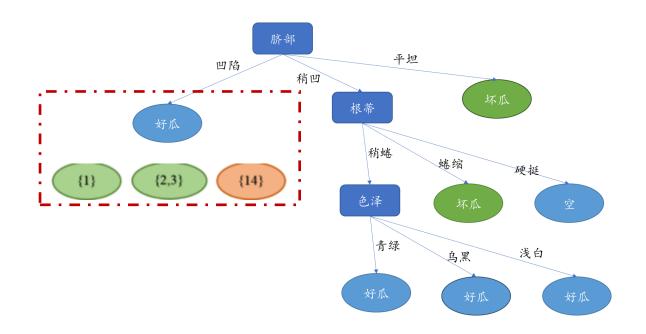




• 后剪枝



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	$\overline{}$
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	×
	4 5 8 9 11 12	4青绿5浅白8乌黑9乌黑11浅白12浅白	4     青绿     蜷缩       5     浅白     蜷缩       8     乌黑     稍蜷       9     乌黑     稍蜷       11     浅白     硬挺       12     浅白     蜷缩	4     青绿     蜷缩     沉闷       5     浅白     蜷缩     浊响       8     乌黑     稍蜷     浊响       9     乌黑     稍蜷     沉闷       11     浅白     硬挺     清脆       12     浅白     蜷缩     浊响	4     青绿     蜷缩     沉闷     清晰       5     浅白     蜷缩     浊响     清晰       8     乌黑     稍蜷     浊响     清晰       9     乌黑     稍蜷     沉闷     稍糊       11     浅白     硬挺     清脆     模糊       12     浅白     蜷缩     浊响     模糊	4     青绿     蜷缩     沉闷     清晰     凹陷       5     浅白     蜷缩     浊响     清晰     凹陷       8     乌黑     稍蜷     浊响     清晰     稍凹       9     乌黑     稍蜷     沉闷     稍糊     稍凹       11     浅白     硬挺     清脆     模糊     平坦       12     浅白     蜷缩     浊响     模糊     平坦	4     青绿     蜷缩     沉闷     清晰     凹陷     硬滑       5     浅白     蜷缩     浊响     清晰     凹陷     硬滑       8     乌黑     稍蜷     浊响     清晰     稍凹     硬滑       9     乌黑     稍蜷     沉闷     稍糊     稍凹     硬滑       11     浅白     硬挺     清脆     模糊     平坦     软粘       12     浅白     蜷缩     浊响     模糊     平坦     软粘	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     e     接縮     平坦     硬滑     否       12     浅白     蜷缩     浊响     模糊     平坦     软粘     否



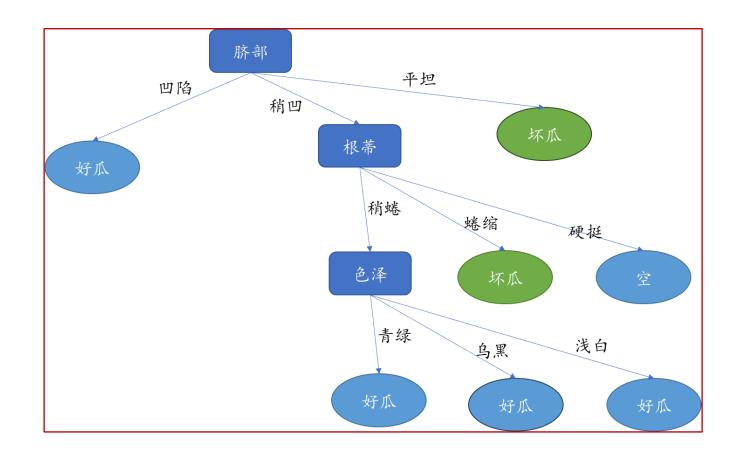
	编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	正确?
	4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	$\checkmark$
	5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是	✓
	8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是	✓
	9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否	×
	11	浅白	硬挺	清脆	模糊	平坦	硬滑	否	✓
	12	浅白	蜷缩	浊响	模糊	平坦	软粘	否	✓
	13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否	×
_									







#### • 后剪枝





· CART: 分类回归树

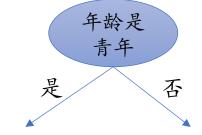
• 二叉树

• 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



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





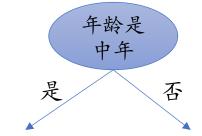


- · CART: 分类回归树
  - 二叉树
  - 分类指标: 基尼指数

• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

•  $Gini(D, A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$ 

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



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





· CART: 分类回归树

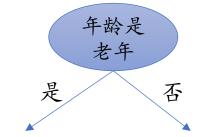
• 二叉树

• 分类指标: 基尼指数

• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

•  $Gini(D, A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$ 

					1/ 1/8
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 '1 = 1 + 7 + Th	
1	青年	0	0	一般	(1: 是, 0: 否)
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



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





- · 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) =	$= \sum_{i=1}^{m} \frac{ I }{ I }$	$\frac{ D^i }{ D }Gini(D^i)$
---	-------------	------------------------------------	------------------------------

					//_/
编号	X:年龄	X:是否有工作 (1: 是, 0: 否)	X:是否买房 (1: 是, 0: 否)		
1	青年	0	0	一般	0 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

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

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

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







- · CART: 分类回归树
  - 二叉树
  - 分类指标: 基尼指数

• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

• $Gini(D,A) = \sum_{i=1}^{m} \frac{ D^i }{ D } Gini(D^i)$	Gini(D, A2 = 0) = $\frac{10}{15}$	$\left(1-\left(\frac{6}{10}\right)^{\frac{1}{2}}\right)$	$\left(-\left(\frac{4}{10}\right)^2\right)$	$+\frac{5}{15}$	(1-(
--	-----------------------------------	--	---	-----------------	------

Gini(D, A1 = 青年) = 0.44

					// /
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) 0 0
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



- · CART: 分类回归树
  - 二叉树
  - 分类指标: 基尼指数

• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

• 
$$Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$$

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

	Gini(	D, A1	= 青年)	= 0.44
--	-------	-------	-------	--------

$$Gini(D, A2 = 0) = 0.32$$





- · CART: 分类回归树
  - 二叉树
  - 分类指标: 基尼指数

• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

• 
$$Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$$

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

X:是否有工作

(1: 是, 0: 否)

X:年龄

青年青年

青年

青年 青年 中年

中年中年

中年

中年

老年

老年

老年

老年

老年

编号

10

11 12

13

14 15 X:是否买房

0

0

(1: 是, 0: 否)

Y:是否放贷

0

0

(1: 是, 0: 否)

X:信贷表现

好

好

好

非常好

非常好

非常好

好

好

非常好

一般

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

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

Gini(D, A2 = 0) = 0.32
Gini(D, A3 = 0) = 0.27
Gini(D, A4 = 0) = 0.32





· CART: 分类回归树

• 二叉树

• 分类指标: 基尼指数

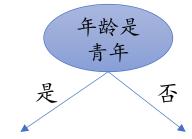
• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

• 
$$Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$$

#### • 回归指标:

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

							0
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		是否放贷 (, 0: 否)	Can Nic
1	青年	0	0	一般	0.1	0	
2	青年	0	0	好	0.3	0	7
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	8.0	1	
10	中年	0	1	非常好	8.0	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	



$$s(D,A1 = ff)$$

$$= \min_{c1} ((0.1 - c1)^2 + (0.3 - c1)^2 + (0.7 - c1)^2 + (0.6 - c1)^2 + (0.1 - c1)^2) + \min_{c2} ((0.4 - c2)^2 + \dots + (0.4 - c2)^2)$$



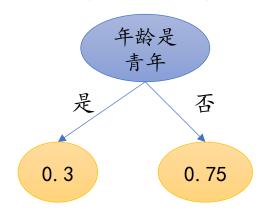
- · CART: 分类回归树
  - 二叉树
  - 分类指标: 基尼指数

• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

- $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$
- 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 = 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		否放贷 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	8.0	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 = 0.3

C2=0.75



· CART: 分类回归树

二叉树

• 分类指标: 基尼指数

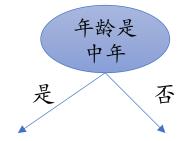
• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

• 
$$Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$$

#### • 回归指标:

• 
$$\min_{j,s} \left[ \min_{c_1} \sum_{x_i \in R_{1(j,s)}} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_{2(j,s)}} (y_i - c_2)^2 \right]$$
 +  $\left( 0.8 - c1 \right)^2 \right) + \min_{c_2} \left( (0.1 - c2)^2 + \dots + (0.4 - c2)^2 \right)$ 

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 (1: 是	:否放贷
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	8.0	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



$$s(D,A1 = \mp \pm)$$

$$= \min_{c1} ((0.4 - c1)^2 + (0.5 - c1)^2 + (0.9 - c1)^2 + (0.8 - c1)^2 + (0.8 - c1)^2 + \min_{c2} ((0.1 - c2)^2 + \dots + (0.4 - c2)^2)$$

C1=0.68

C2 = 0.55





- · CART: 分类回归树
  - 二叉树
  - 分类指标: 基尼指数

• 
$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

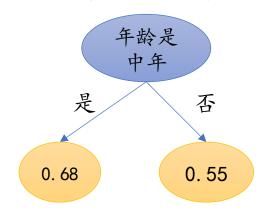
- $Gini(D,A) = \sum_{i=1}^{m} \frac{|D^i|}{|D|} Gini(D^i)$
- 回归指标:

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

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 (1: 是	是否放贷 , 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	8.0	1
10	中年	0	1	非常好	8.0	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





#### • 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

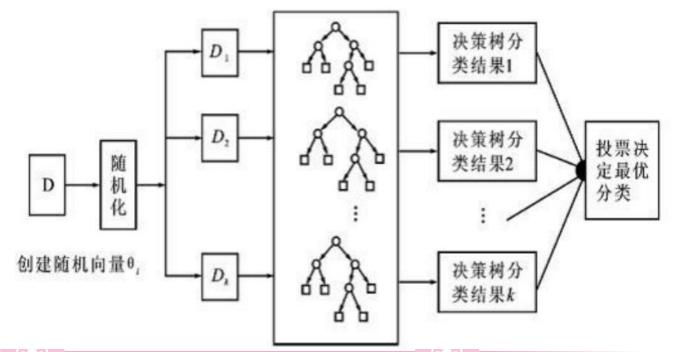




- ❖ 集成学习通过构建并结合多个学习器来完成学习任务
  - > 好而不同
- Boosting & Bagging



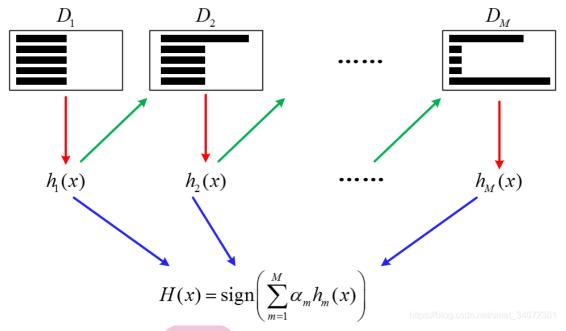
- Bagging
  - ❖ 随机森林





#### Boosting

· AdaBoost ——基于"基学习器的线性组合"



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

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

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



#### Boosting

❖ 提升树——基于CART决策树

 $* min \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m))$ 

 $L(y, f(x)) = (y - f(x))^2 = [y - f_{m-1}(x) - T(x, \theta_m)]^2 = [r - T(x, \theta_m)]^2$ 

❖ 梯度提升:  $r_{mi} = \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}$