4.2划分选择

4.2.1信息增益，决策树ID3训练算法

信息熵(information entropy)

信息增益(information gain)

信息增益例子：西瓜数据集

表 4.1 西瓜数据集 2.0

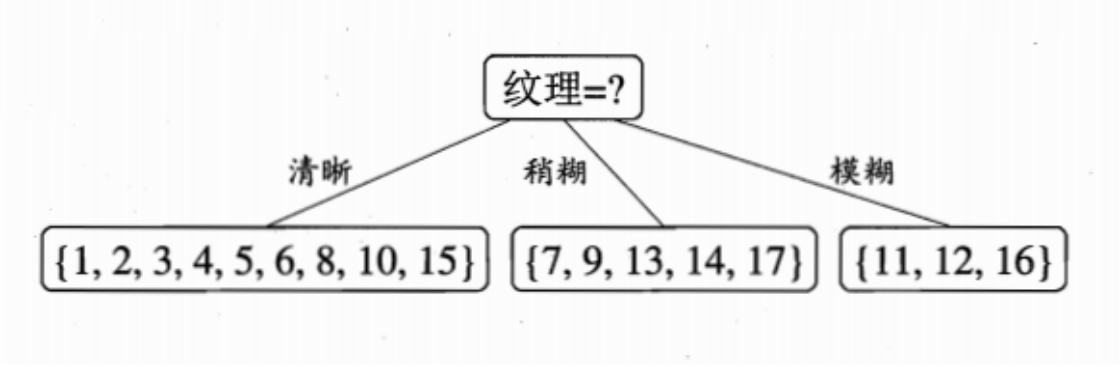
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	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

目标：分出是好瓜/坏瓜

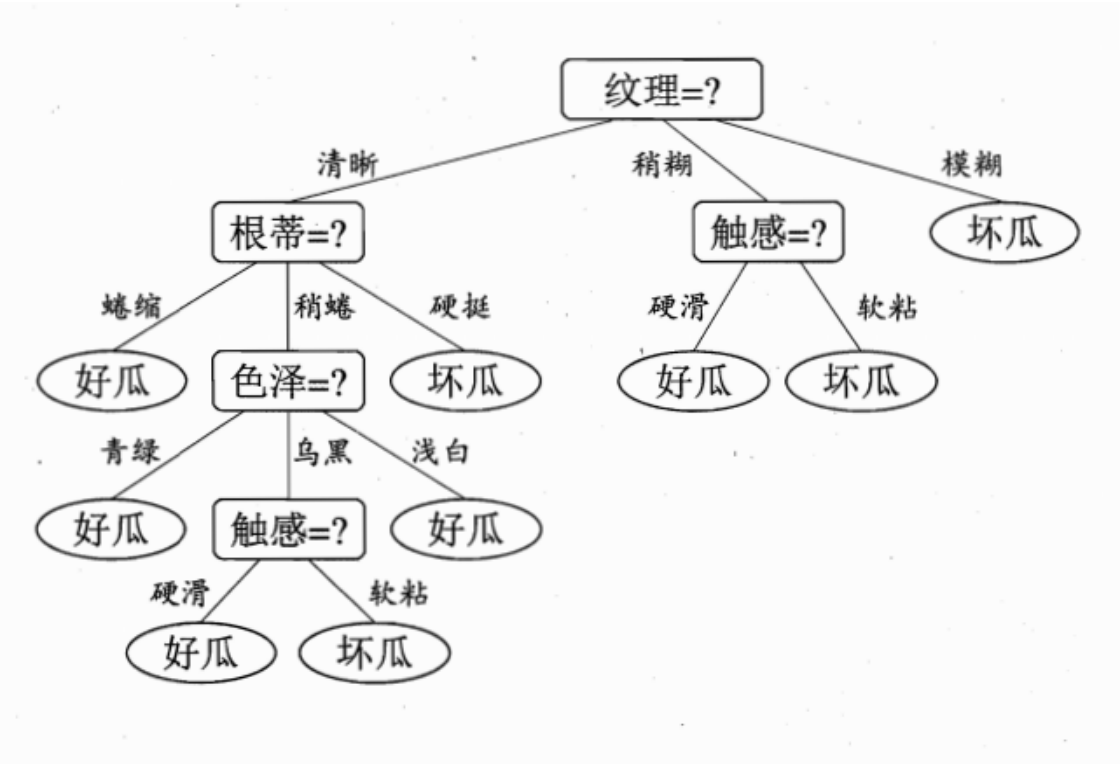
1. 根$Ent(0)=\frac{8}{17}log_2\frac{17}{8}+\frac{9}{17}log_2\frac{17}{9}=0.998$
2. 设ABCDEF分别代表色泽、根蒂、敲声、纹理、脐部、触感
3. A色泽分为：青绿、乌黑、浅白
4. 青绿中好瓜3个，坏瓜3个共6个；乌黑中好瓜4个，坏瓜2个共6个；浅白中好瓜1个，坏瓜4个共5个
5. $Ent(\text{青绿})=\frac{3}{6}log_2\frac{6}{3}+\frac{3}{6}log_2\frac{6}{3}=1.000,$
   $Ent(\text{乌黑})=\frac{4}{6}log_2\frac{6}{4}+\frac{2}{6}log_2\frac{6}{2}=0.918,$
   $Ent(\text{浅白})=\frac{1}{5}log_2\frac{5}{1}+\frac{4}{5}log_2\frac{5}{4}=0.722.$
6. $Ent(A)=\frac{6}{17}Ent(\text{青绿})+\frac{6}{17}Ent(\text{乌黑})+\frac{5}{17}Ent(\text{浅白})=0.889$，则信息增益Grain(D,色泽)=0.998-0.889=0.109
7. 同理，得到：
   Grain(D,根蒂)=0.413，Grain(D,敲声)=0.141，

Grain(D,纹理)=0.381，Grain(D,脐部)=0.289，
Grain(D,触感)=0.006

8. 纹理得到的信息增益最大，则优先基于纹理对根节点划分：



9. 再根据三个方向重复上述步骤，计算下一层的划分，最终得到决策树：



4.2.2增益率，决策树C4.5训练算法

增益率(gain ratio)

公式

Gain_ratio(D,a)=$\frac{Gain(D,a)}{IV(a)}$，其中IV(a)=$-\sum_{v=1}^{V}\frac{|D^v|}{|D|}log_2\frac{|D^v|}{|D|}$

对可取值数目较少的属性有所偏好

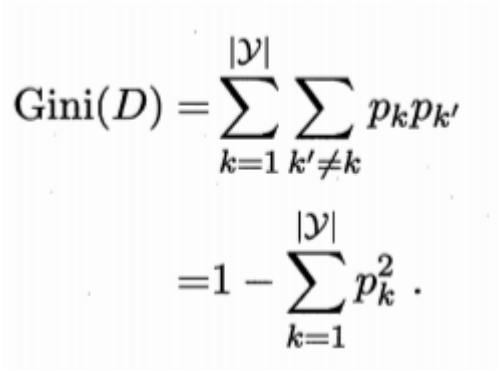
使用方法

先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的

4.2.3基尼指数，决策树CART训练算法(Classification and Regression)

分类树：基尼指数最小原则

基尼指数:



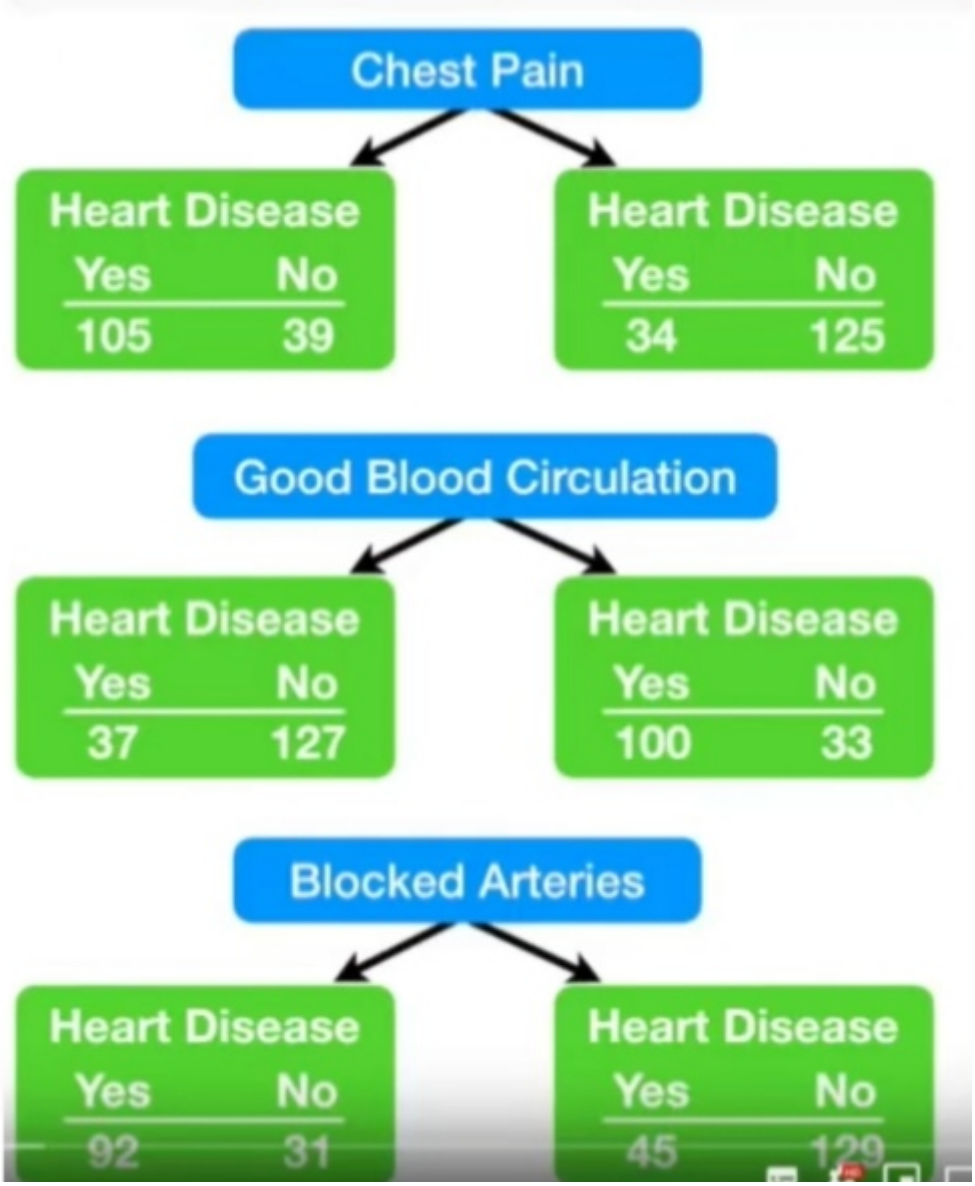
$$\begin{aligned}\mathrm{Gini}(D) &= \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} \\ &= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2 \ .\end{aligned}$$

基尼指数越小，数据集D的纯度越高

e.g.心脏病患者三指标：

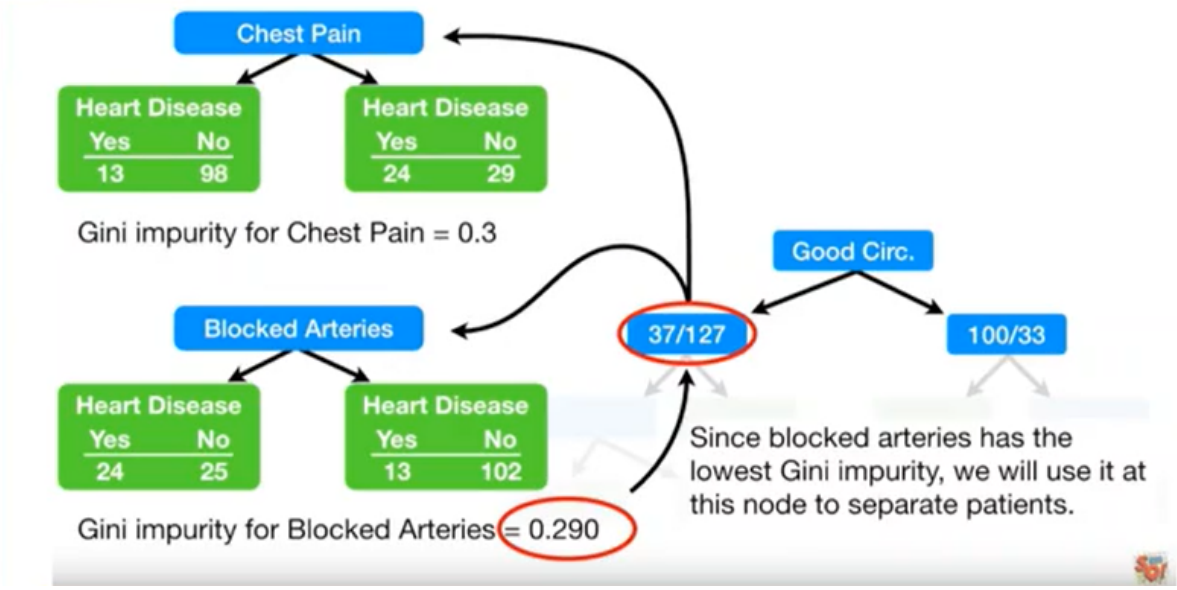
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

二叉树分类：



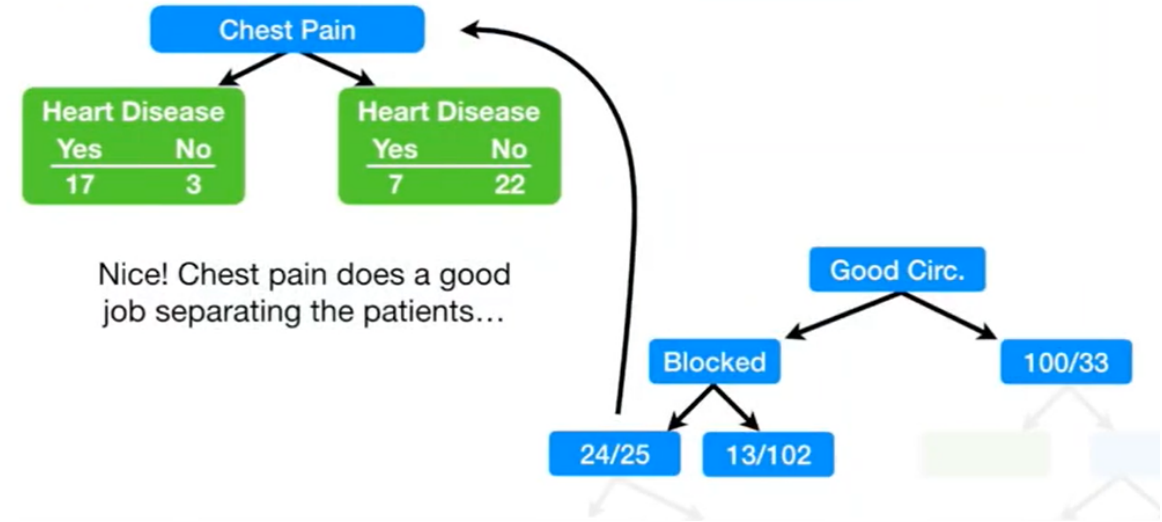
在Chest Pain为Yes的数据集中取两个样例，他们在相同集合的概率为$(\frac{105}{105+39})^2+(\frac{39}{105+39})^2$，则这个概率越大纯度越高；在不同集合的概率为$1-(\frac{105}{105+39})^2+(\frac{39}{105+39})^2$，即基尼指数，指数越小纯度越高。

上图算得三个基尼指数：
Gini(CHest Pain)=0.364
Gini(Good Blood Circulation)=0.360
Gini(Blocked Arteries)=0.381
从中选一个最小的0.360作为第一次分叉的依据

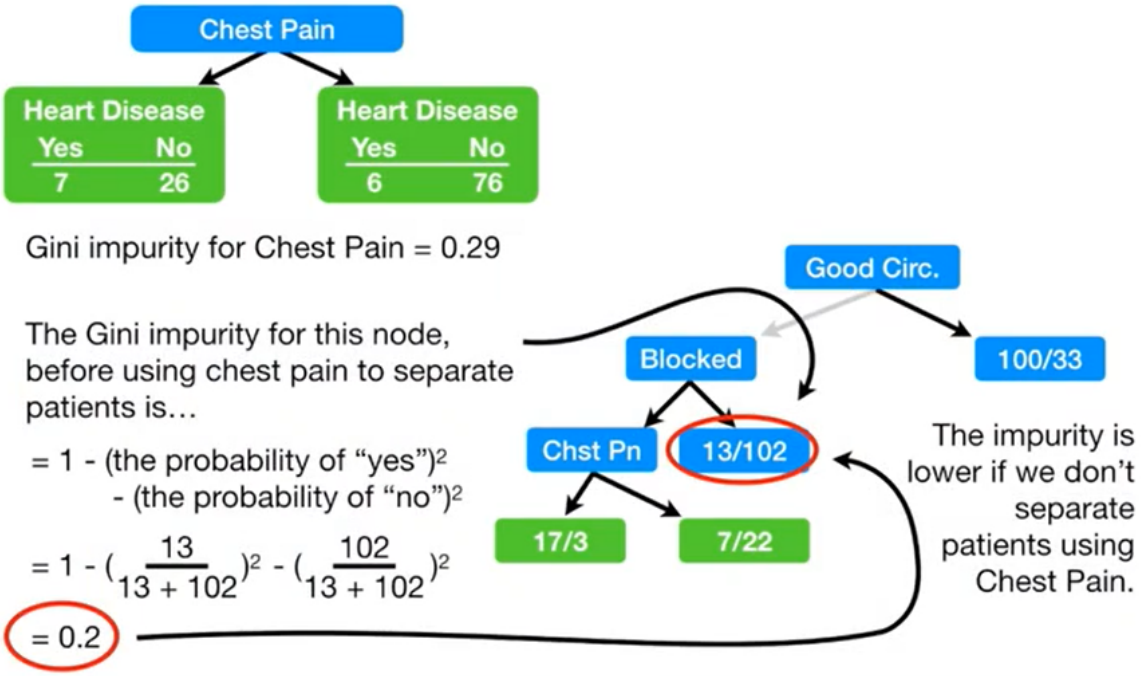


Gini(Chest Pain)=0.30
Gini(Blocked Arteries)=0.29

选0.29的Blocked Arteries作为第二次分叉



左侧指数为0.5，用Chest Pain第三次划分后指数为0.33，所以左侧应该再分一次



右侧指数为0.2，用Chest Pain第三次划分后指数为0.29，指数上升，纯度变低，所以右侧不用在分

代码例子：
https://www.bilibili.com/video/av79015715?p=59
https://www.bilibili.com/video/av79015715?p=60
https://www.bilibili.com/video/av79015715?p=61

回归树：平方误差最小原则（Sum of Squared Residuals残差平方和）

回归树是决策树的一种

Equation 6-4. CART cost function for regression

$$J(k, t_k) = \frac{m_{\text{left}}}{m}\text{MSE}_{\text{left}} + \frac{m_{\text{right}}}{m}\text{MSE}_{\text{right}} \quad \text{where} \begin{cases} \text{MSE}_{\text{node}} = \sum_{i \in \text{node}} \left(\hat{y}_{\text{node}} - y^{(i)}\right)^2 \\ \hat{y}_{\text{node}} = \frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)} \end{cases}$$

参考：hands on ml

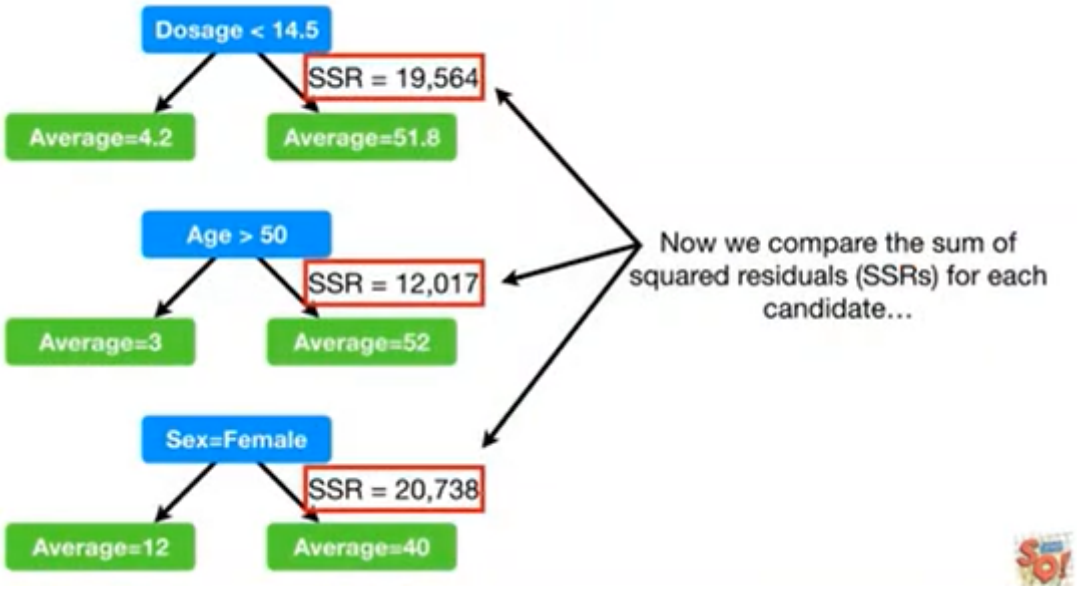
例子：https://www.youtube.com/watch?v=g9c66TUylZ4

一个特征

移动阈值，找到均方误差(sum of squared residuals简称SSR)最小值

多个特征

算出三个特征的均方误差最小值，选最小的



防止over fit

如：设定小于20个不再分割