第五章 非线性方法

决策树
（Decision Tree）

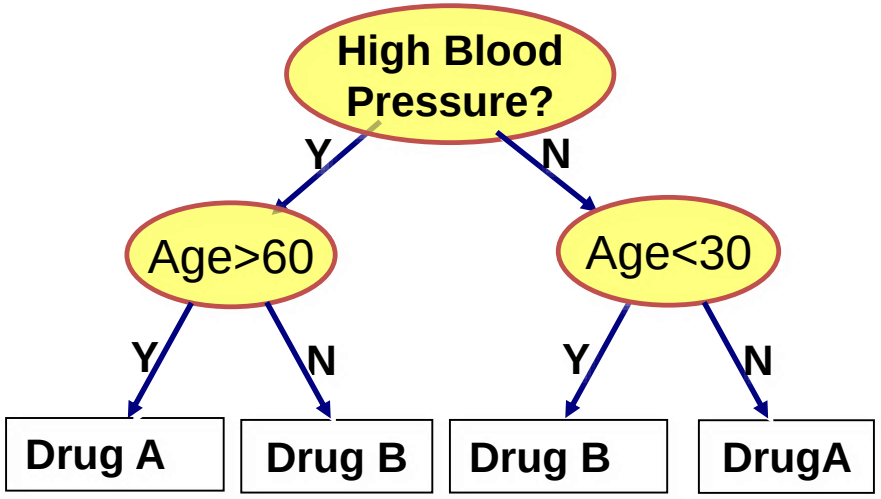
问问题，猜东西

颜色、形状、大小，…

?

决策树(Decision Trees)

- **游戏：二十个问题**
  - http://y.20q.net
- **树状的决策过程**

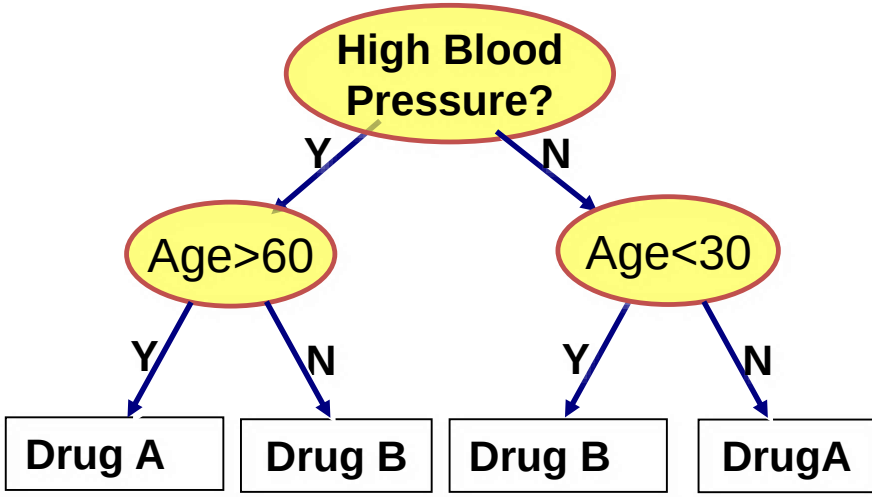


二分查找 （binary search）

- 样本有排序情况下的一种高效搜索方法
- 对于N个元素的列表，比较次数：

$$k \sim \log_2(N)$$

决策树



- 结构
  - 一个根节点
  - 若干内部节点
  - 若干叶节点

- 叶结点对应决策结果，其余节点各自对应一个属性测试（决策）
- 根节点包含所有样本，每个中间节点对样本进行一次划分
- 从根节点到叶结点的路径对应一个判定测试序列

决策树构建中的几个重要问题

- 每个节点应该用哪个（些）特征进行划分
- 什么情况下一个节点应该被认为是叶节点
- 树太大、太复杂时如何修剪
- 如果一个叶结点不是纯的，如何定义其类别标签

ID3 算法（Quinlan, 1979）
（交互式二分法Interactive Dichotomizer-3）

Claude E. Shannon (1916-2001)

- 信息量、熵（不纯度）(1949)

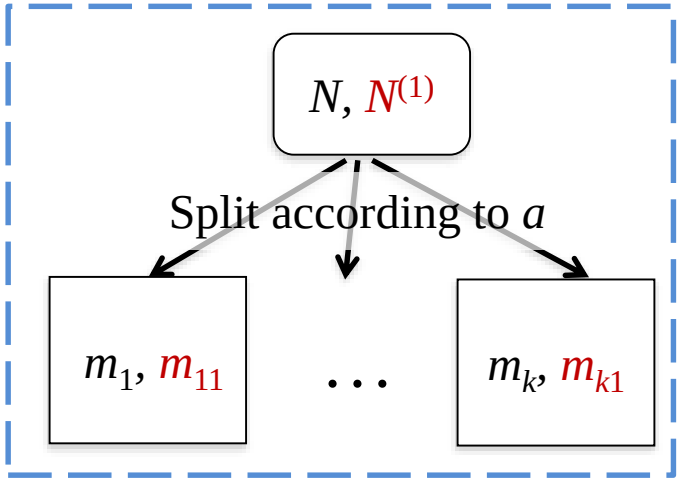
$$I = -(P_1 \log_2 P_1 + P_2 \log_2 P_2 + \cdots + P_k \log_2 P_k) = -\sum_{i=1}^{k} P_i \log_2 P_i$$

  - Example 1: 设 k =4, P_1=0.25, P_2=0.25, P_3=0.25, P_4=0.25
    I = - (0.25 × $\log_2$(0.25) × 4) = 2
  - Example 2: 设k =4, P_1=0, P_2=0.5, P_3=0, P_4=0.5
    I = - (0.5 × $\log_2$(0.5) × 2) = 1
  - Example 3: 设k =4, P_1=1, P_2=0, P_3=0, P_4=0
    I = - (1 × $\log_2$(1)) = 0

用熵来度量特征作为分类结点的效果
  - 熵大，不纯度大，分类差
  - 熵小，不纯度小，分类好

节点选择原则：信息增益（Information Gain）

- 上层结点，样本数N，第一类样本数$N^{(1)}$，信息熵 $I(N, N^{(1)})$
- 下层结点，在属性a上有k个取值，取值i下m_i个样本m_{i1}个第一类，熵$E(N,a)$
- 信息增益（不纯度减少量）



$$Gain\ (N,a) = I(N, N^{(1)}) - E(N,a),$$

其中，

$$I(N, N^{(1)}) = -((N^{(1)}/N)log_2(N^{(1)}/N) + (1-N^{(1)}/N)log_2(1-N^{(1)}/N))$$

$$E(N,a) = (m_1/N)I(m_1, m_{11}) + (m_2/N)I(m_2, m_{21}) + \ldots + (m_k/N)I(m_k,m_{k1})$$

举例

表6-1 顾客数据

编号	年龄	性别	收入	是否购买
1	21	男	4000	否
2	33	女	5000	否
3	30	女	3800	否
4	38	女	2000	否
5	25	男	7000	否
6	32	女	2500	否
7	20	女	2000	否
8	26	女	9000	是
9	32	男	5000	是
10	24	男	7000	否
11	40	女	4800	否
12	28	男	2800	否
13	35	女	4500	否
14	33	男	2800	是
15	37	男	4000	是
16	31	女	2500	否

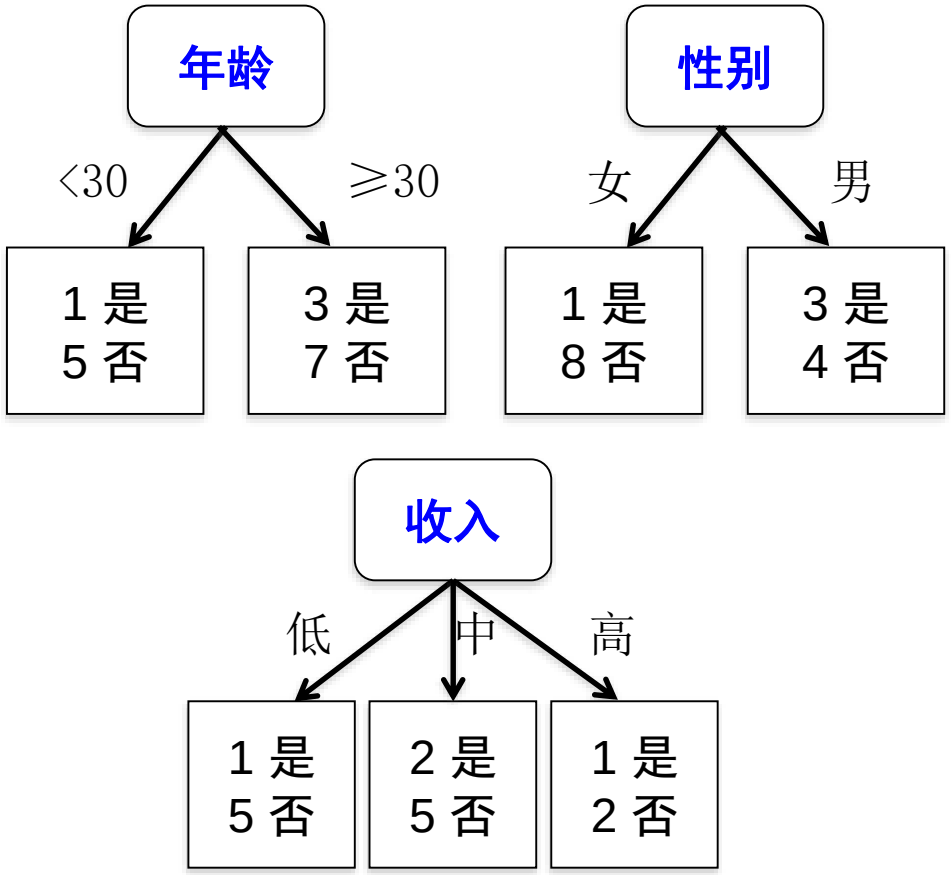
整理后的数据

表6-2 顾客数据

编号	年龄	性别	收入	是否购买
1	<30	男	中	否
2	≥30	女	中	否
3	≥30	女	中	否
4	≥30	女	低	否
5	<30	男	高	否
6	≥30	女	低	否
7	<30	女	低	否
8	<30	女	高	是
9	≥30	男	中	是
10	<30	男	高	否
11	≥30	女	中	否
12	<30	男	低	否
13	≥30	女	中	否
14	≥30	男	低	是
15	≥30	男	中	是
16	≥30	女	低	否

编号	年龄	性别	收入	是否购买
1	<30	男	中	否
2	≥30	女	中	否
3	≥30	女	中	否
4	≥30	女	低	否
5	<30	男	高	否
6	≥30	女	低	否
7	<30	女	低	否
8	<30	女	高	是
9	≥30	男	中	是
10	<30	男	高	否
11	≥30	女	中	否
12	<30	男	低	否
13	≥30	女	中	否
14	≥30	男	低	是
15	≥30	男	中	是
16	≥30	女	低	否

- 根节点
- 中间节点 ？
- 停止分支



举例

编号	年龄	性别	收入	是否购买
1	<30	男	中	否
2	≥30	女	中	否
3	≥30	女	中	否
4	≥30	女	低	否
5	<30	男	高	否
6	≥30	女	低	否
7	<30	女	低	否
8	<30	女	高	是
9	≥30	男	中	是
10	<30	男	高	否
11	≥30	女	中	否
12	<30	男	低	否
13	≥30	女	中	否
14	≥30	男	低	是
15	≥30	男	中	是
16	≥30	女	低	否

$n=16$

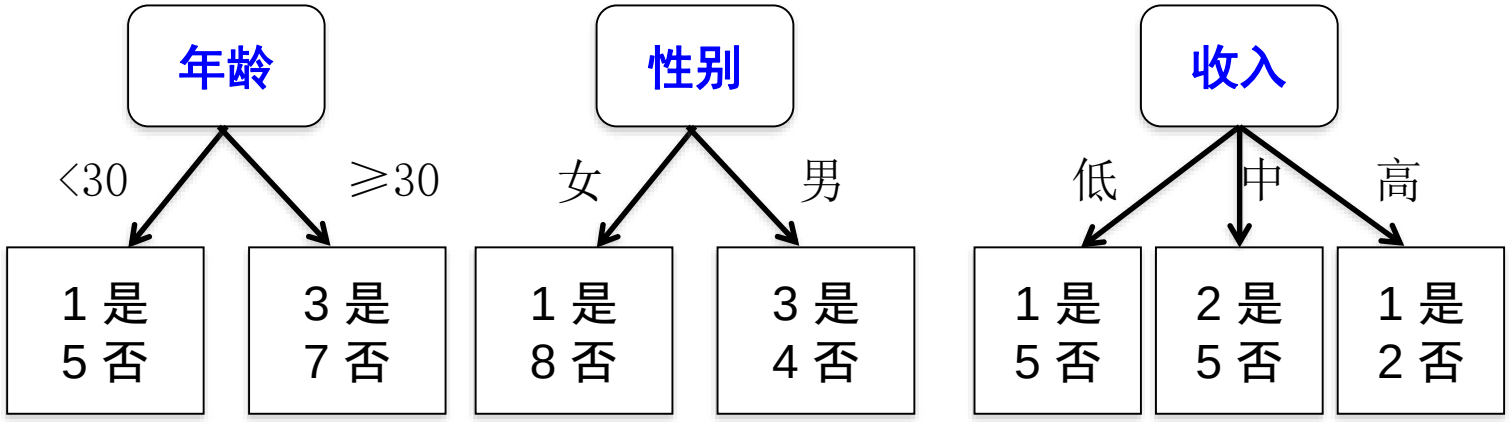
$n_1=4$

$I(16,4)= -((4/16)*\log_2(4/16)+(12/16)*\log_2(12/16))$

$= 0.8113$

E(年龄) = (6/16)*I(6,1)+(10/16)*I(10,3) = 0.7946

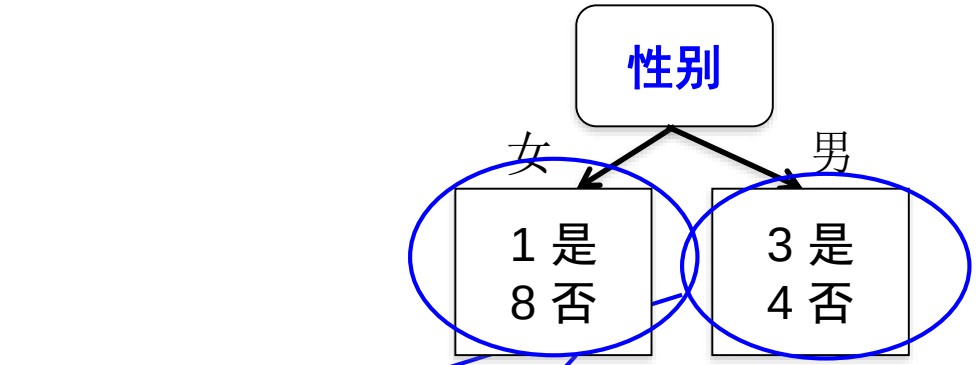
Gain(年龄)= $I(16,4) - E$(年龄) = 0.0167



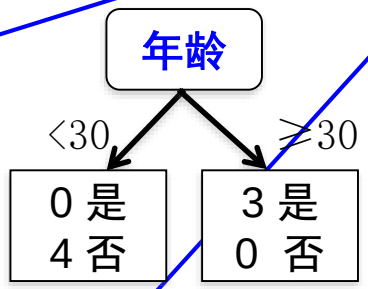
- Gain(年龄)=0.0167
- Gain(性别)=0.0972
- Gain(收入)=0.0177

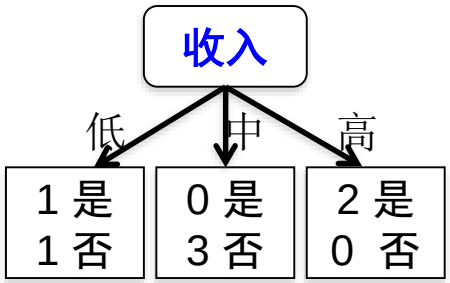
Max：作为第一个分类依据

编号	年龄	性别	收入	是否购买
1	<30	男	中	否
5	<30	男	高	否
9	≥30	男	中	是
10	<30	男	高	否
12	<30	男	低	否
14	≥30	男	低	是
15	≥30	男	中	是



$I(7,3)=-((3/7)*\log_2(3/7)+(4/7)*\log_2(4/7))=0.9852$

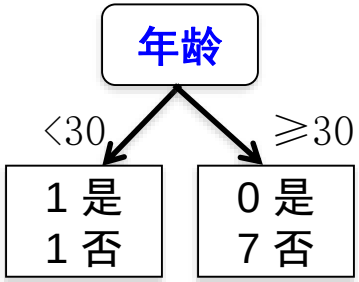


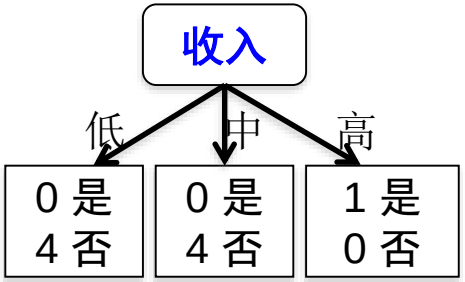


- Gain(年龄)=0.9852
- Gain(收入)=0.688

编号	年龄	性别	收入	是否购买
2	≥30	女	中	否
3	≥30	女	中	否
4	≥30	女	低	否
6	≥30	女	低	否
7	<30	女	低	否
8	<30	女	高	是
11	≥30	女	中	否
13	≥30	女	中	否
16	≥30	女	低	否

$I(9,1)=-((1/9)*\log_2(1/9)+(8/9)*\log_2(8/9))=0.5032$



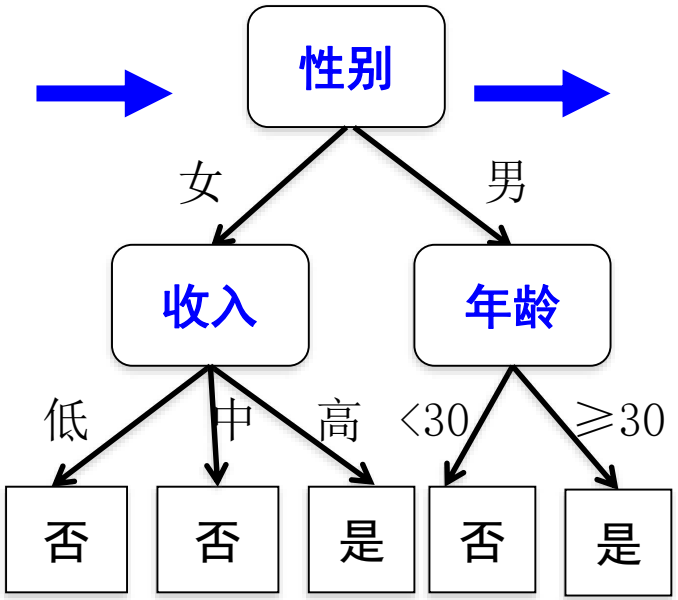


- Gain(年龄)=0.2222
- Gain(家庭所得)=0.5032

资料

编号	年龄	性别	收入	是否购买
1	<30	男	中	否
2	≥30	女	中	否
3	≥30	女	中	否
4	≥30	女	低	否
5	<30	男	高	否
6	≥30	女	低	否
7	<30	女	低	否
8	<30	女	高	是
9	≥30	男	中	是
10	<30	男	高	否
11	≥30	女	中	否
12	<30	男	低	否
13	≥30	女	中	否
14	≥30	男	低	是
15	≥30	男	中	是
16	≥30	女	低	否

决策树



分类规则

IF 性别=女 AND 收入= 低 THEN 购买=否

IF 性别=女 AND收入= 中 THEN 购买=否

IF 性别=女 AND收入= 高 THEN 购买=是

IF 性别=男 AND 年龄<30 THEN 购买=否

IF 性别=男 AND 年龄≥30 THEN 购买=是

Top-Down Induction of DTs (ID3)

```
proc growtree(data)
  if (data not perfectly classified)
    find `best' splitting attribute A
    for each (a in A)
      create child a
      data_a = data restricted to A=a
      growtree(data_a)
    endfor
  endif
endproc
```

递归算法

不纯度度量准则

基尼不纯度（Gini impurity）

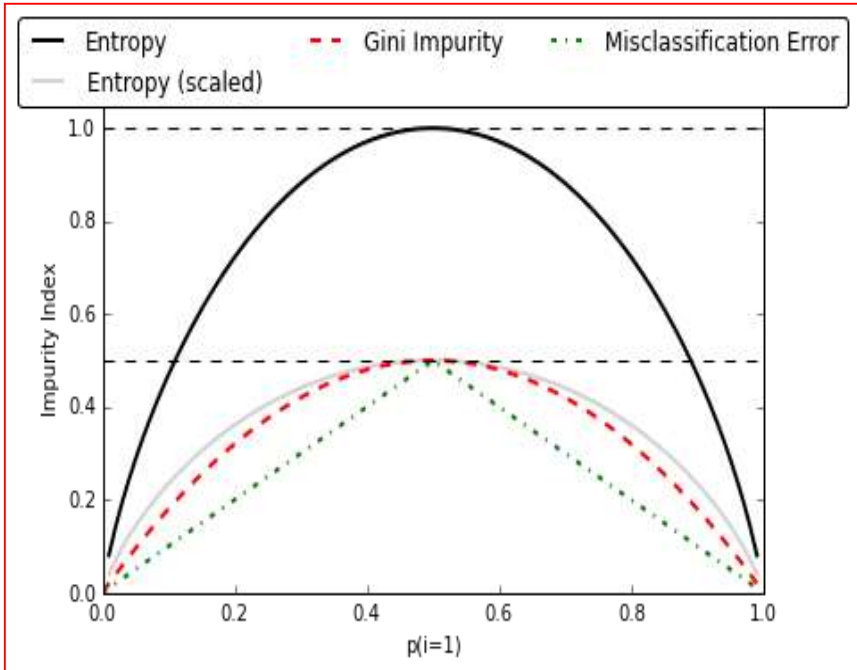
$$Gini(N) = \sum_{m \neq n} P(w_m)P(w_n) = 1 - \sum_{j=1}^{k} P(w_k)$$

属性a上的纯度增益：

$$Gain(N,a) = Gini(N) - \sum_{v=1}^{m} \frac{|N^{(m)}|}{|N|} Gini(N^{(m)})$$

误差不纯度 (Misclassification impurity)

$$I(N) = 1 - \max_j P(w_j)$$



多特征划分的归一化 (C4.5)

特征取值多(k大) 的划分带来的信息增益比特征取值少 (k小) 的划分要大（**例如按照样本编号划分**）

ID3： $Gain\ (X,a) = I(N, N^{(1)}) - E(N,a)$

C4.5： $\text{Gain\_ratio}(N,a)=\text{Gain}(N,a)/I(N,a)$

$$Gain\_ratio(N,a)= \frac{Gain(N,a)}{-\sum_{v=1}^{k} P_v \log P_v}$$

其中 $P_v = \frac{N_v}{N}$; $v=1,\ldots,k;$ 为特征a上的v种不同取值

用连续特征构造决策树（C4.5）

- 若特征x包含k个取值，则按大小排序，用二分法划分可有k-1种划分方案，根据其中信息增益率最大的划分选择特征节点

- 同理，也可以把特征离散化为多值

CART 算法

- 由Friedman等人提出，1980年代以来就开始发展，是基于树结构产生分类和回归模型的过程，是一种产生二元树的技术
- CART与C4.5算法的最大相异之处是其在每一个节点上都是采用二分法，也就是一次只能够有两个子节点，C4.5则在每一个节点上可以产生不同数量的分枝
- CART模型适用於目标变量为连续型和类别型的变量，如果目标变数是类别型变量，则可以使用分类树（classification trees），目标变数是连续型的，则可以采用回归树（regression trees）

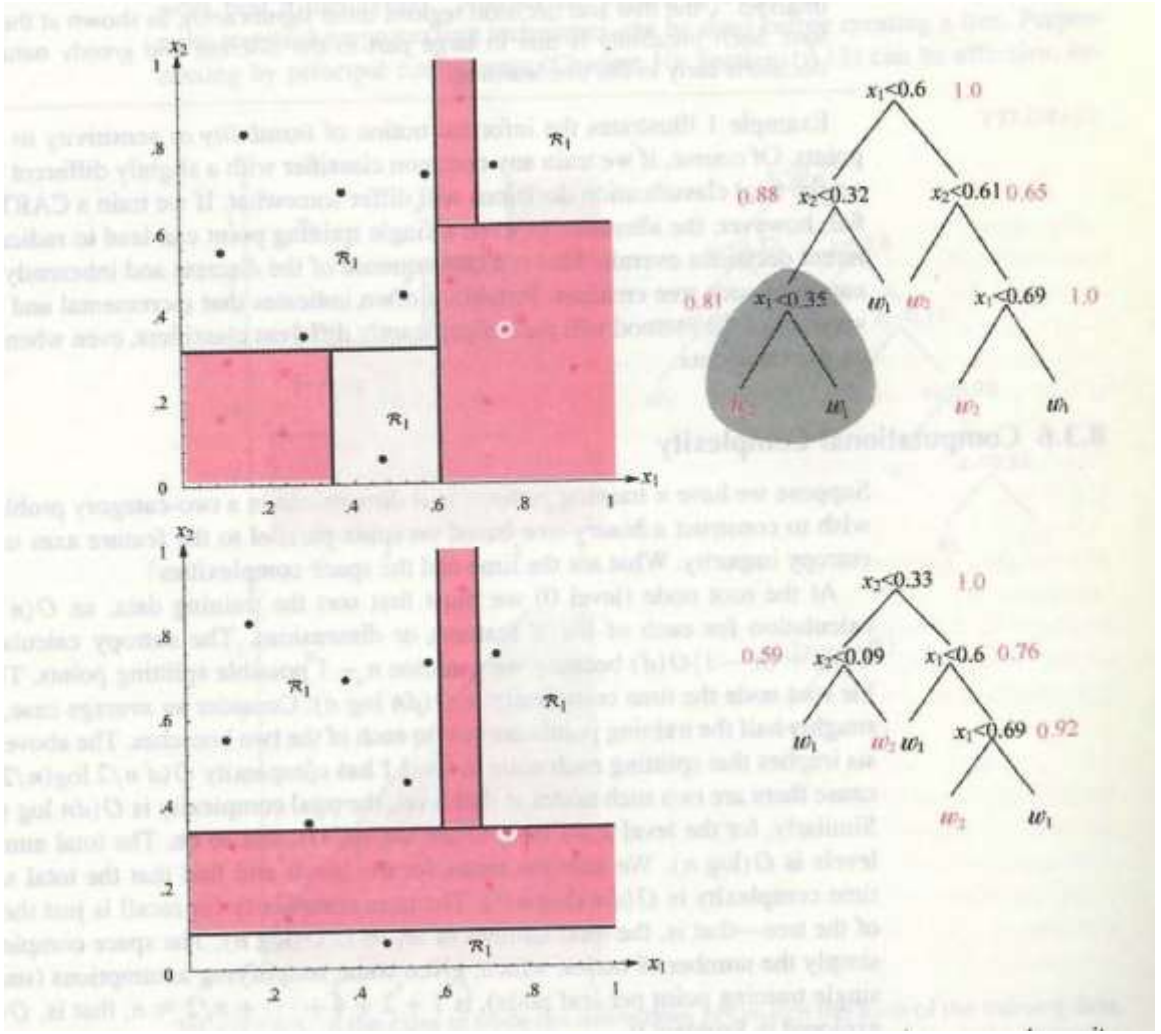
过学习（overfitting）问题

- 完全成长的树对未来新数据的预测一定是最好的吗？

- 过度学习问题（over fitting）
  - 过度学习是指模型过度训练，导致模型记住的不是训练集的一般性，反而是训练集的局部特性
  - **机器学习的关键是推广能力**

树结构的生成对数据敏感

ω_1 (black)		ω_2 (red)	
x_1	x_2	x_1	x_2
.15	.83	.10	.29
.09	.55	.08	.15
.29	.35	.23	.16
.38	.70	.70	.19
.52	.48	.62	.47
.57	.73	.91	.27
.73	.75	.65	.90
.47	.06	.75	.36* (.32†)



Richard Duda, Pattern Classification, 2nd Edition, Chapter 8

Overfitting vs. Size of the Tree

Overfitting

- Hypothesis h overfits iff ∃ h' with

$$error_{train}(h) < error_{train}(h')$$
$$error_{true}(h) > error_{true}(h')$$

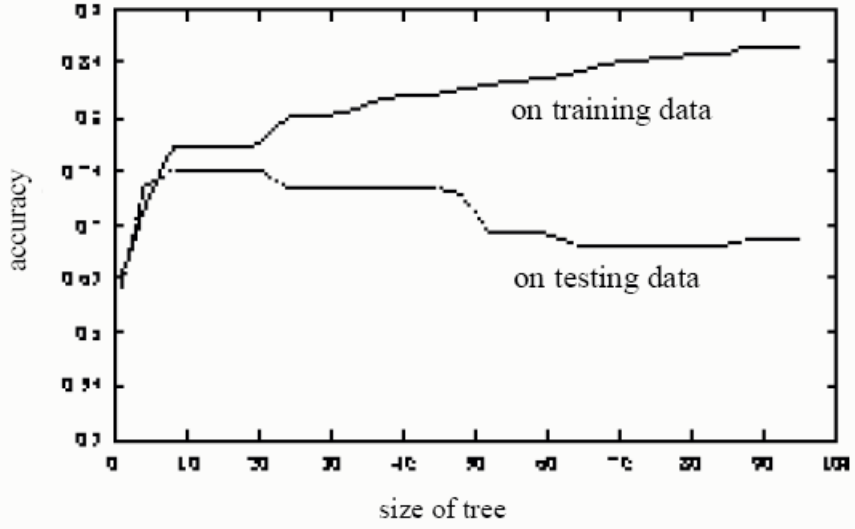
Bias Variance Trade-Off

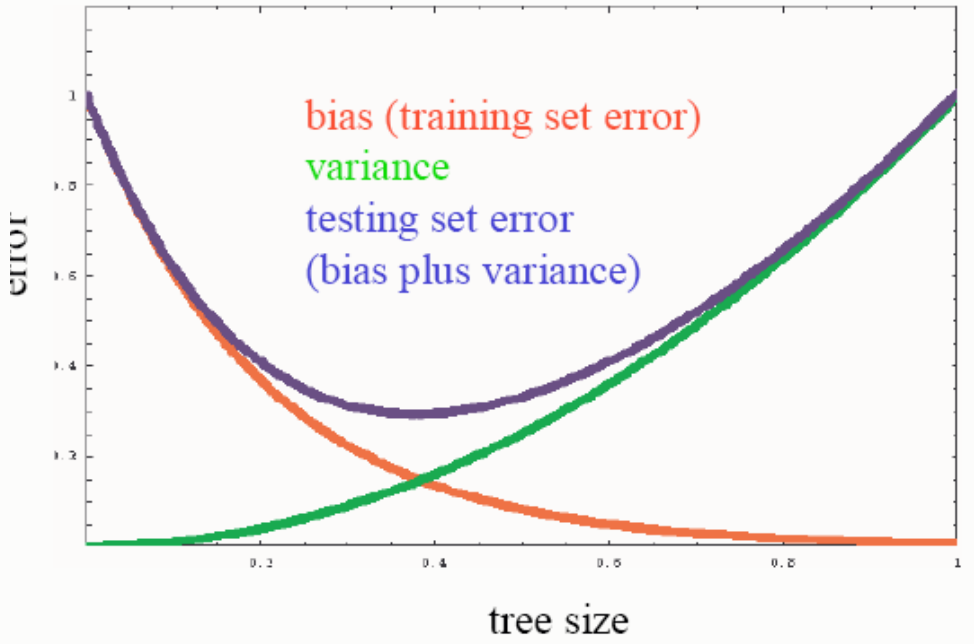
High Bias ⟷ **High Variance**

Small trees can't fit many data sets

Large trees sensitive to randomness in data selection

Overfitting in ID3





Ockham's Razor (Occam's Razor)

- William of Ockham (1285-1347):
  - "Non sunt multiplicanda entia praeter necessitatem."
  - Entities are not to be multiplied beyond necessity.
    - "law of parsimony"
- In machine learning (and science):
  - To prefer simpler hypotheses over more complex ones.

- Albert Einstein (1879-1955):
  - "Everything should be made as simple as possible, but not simpler."

避免过学习：剪枝（pruning）

- 预剪枝（prepruning）：
  - 利用训练集决定节点划分
  - 根据测试集或者不纯度减少的阈值决定是否停止

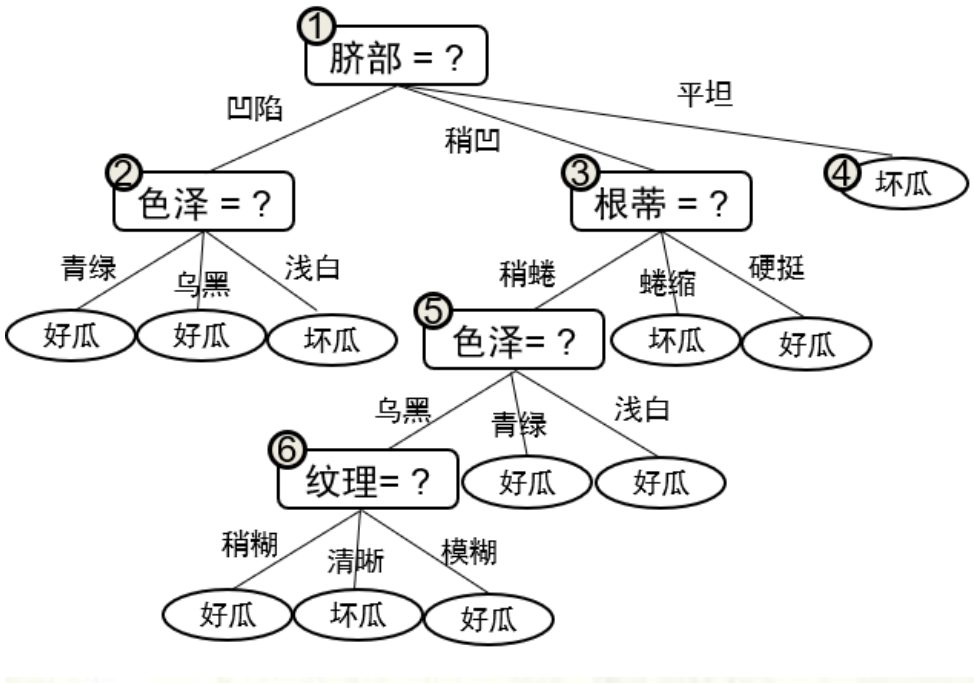
- 后剪枝（postpruning）：
  - 用训练集生成完整的树
  - 在独立剪枝集上减少分类错误的修剪法

例子：西瓜分类

- 数据划分法：训练样本和测试样本

表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

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

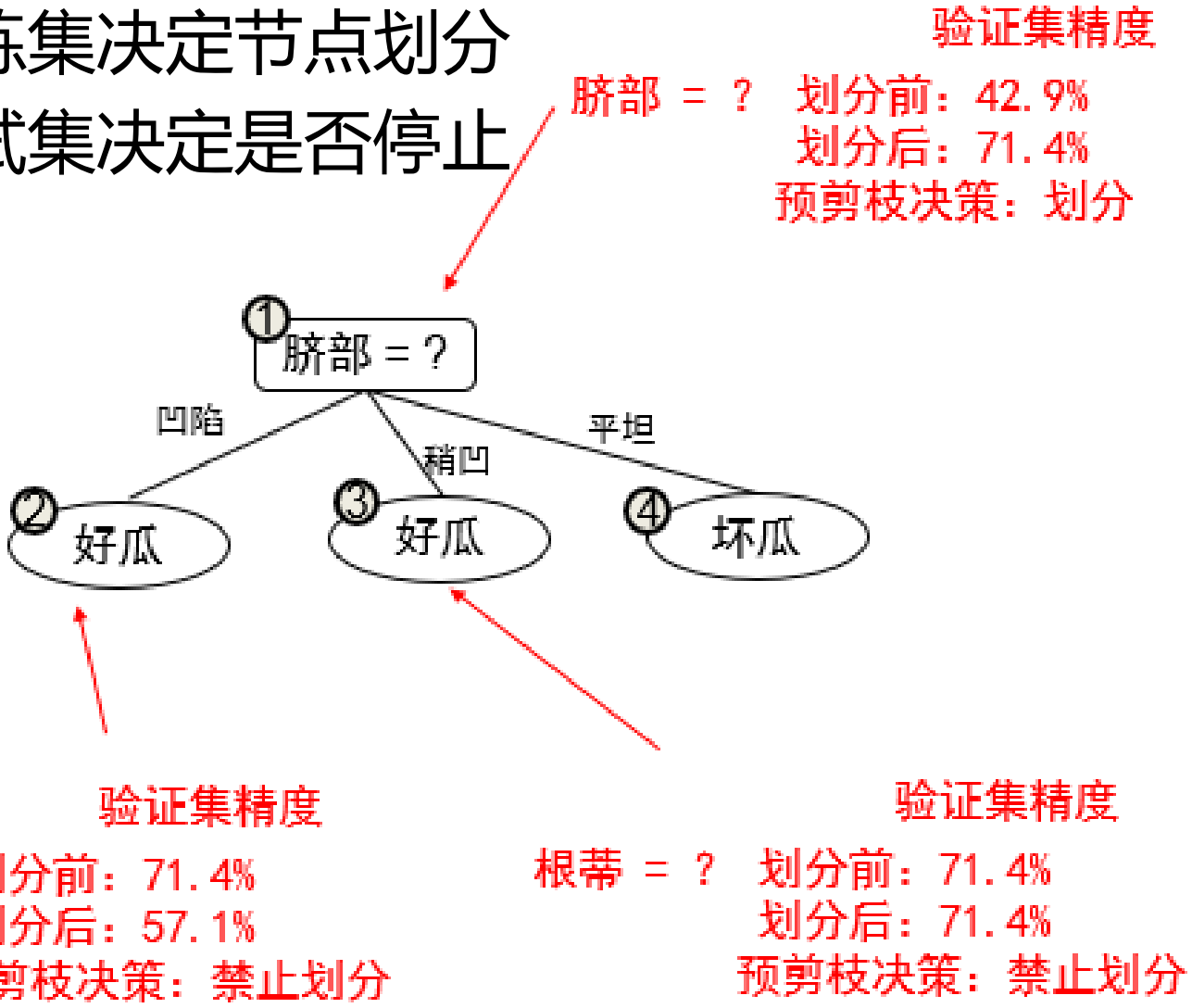


图：基于表4. 2生成的未剪枝决策树

周志华，《机器学习》，清华大学出版社

预剪枝

- 利用训练集决定节点划分
- 根据测试集决定是否停止

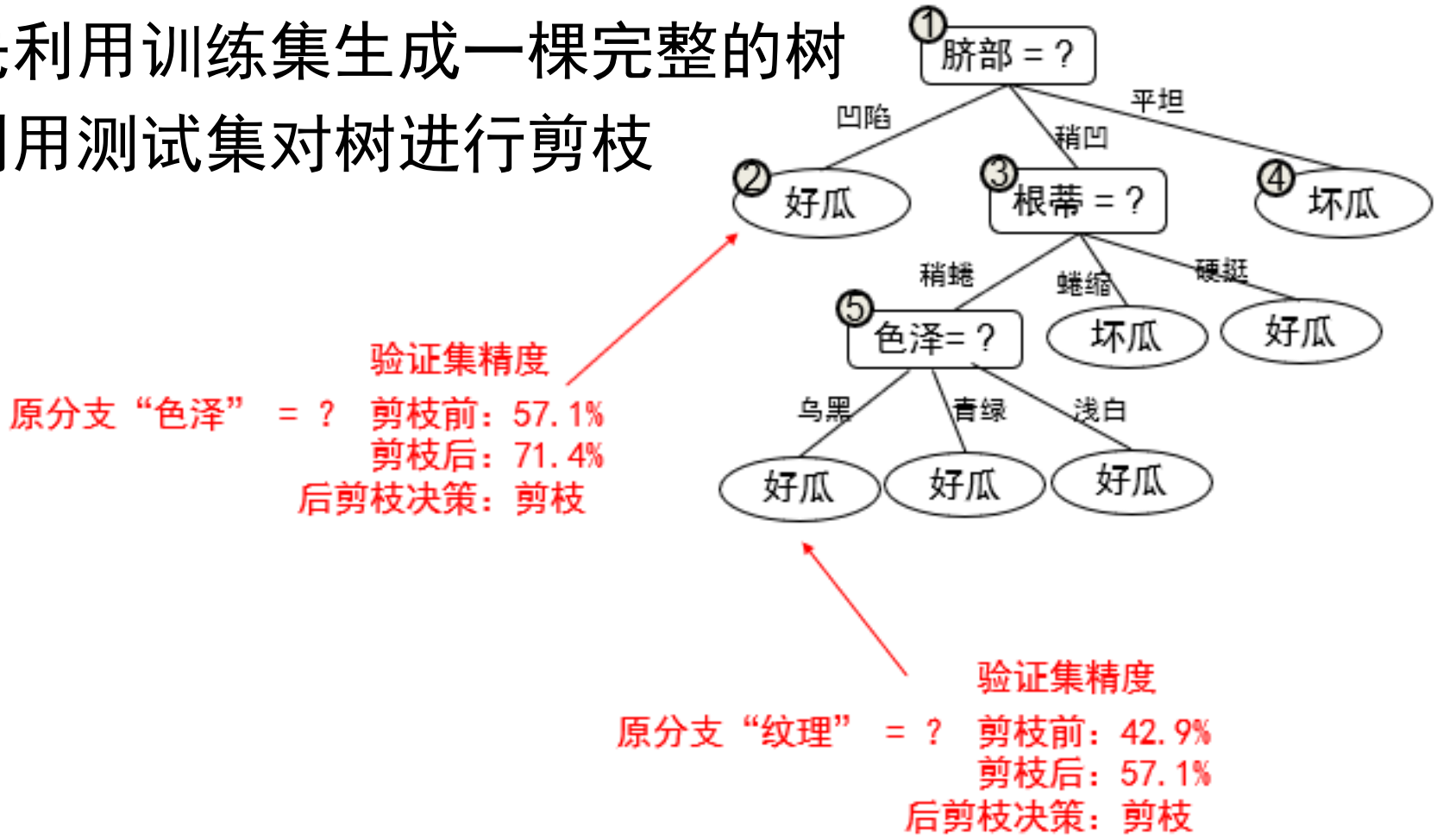


图：基于表4.2生成的预剪枝决策树

周志华，《机器学习》，清华大学出版社

后剪枝

- 先利用训练集生成一棵完整的树
- 利用测试集对树进行剪枝



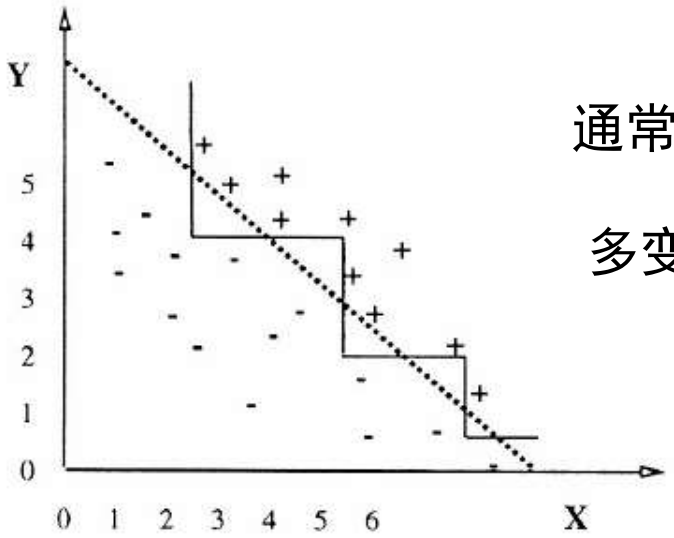
图：基于表4.2生成的后剪枝决策树

后剪枝树通常比预剪枝树保留更多的分支

多变量决策树
(multivariate decision tree)

- 每一个节点是一个线性分类器

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$



通常决策树的分类边界与轴平行（axis-parallel）

多变量决策树实现“斜划分”

参考文献：
CE Brodley, PE Utgoff, Multivariate decision tree, *Machine Learning*, 1995, 19(1):45-77

- 随机森林 Random Forests

(Leo Breiman, *Machine Learning, 45: 5-32, 2001*)
(http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm)

- Many decision trees

➔ Random Forest

Leo Breiman (1928-2005)