《统计学方法》第五章
决策树

李航 著

演讲人：王宗威

决策树学习的三个步骤

- 特征选择：信息增益、信息增益比、基尼指数
- 决策树的生成：
- ID3（信息增益），C4.5（信息增益比），CART（基尼指数）
- 决策树的修剪

快速过一遍

- 决策树模型 P55定义5.1
- If-then规则（互斥并且完备）P56 第一段
- 条件概率分布 P（Y|X）P56第二段
- 问题：图5.2（b)
- NP完全问题：多项式复杂程度的非确定问题。（确定不了复杂度)
- 启发式方法：人在解决问题时以经验规则发现的一种方法，不是系统地、以确定的步骤去寻求答案。

特征选择

- 熵：单位（比特bit或纳特nat　1nat=10^-9bit）表示随机变量X不确定性的度量。取值范围为0≤H(p)≤logn

$$H(p) = -\sum_{i=1}^{n} p_i log p_i$$

- 条件熵：X给定条件Y的条件概率的熵对X的数学期望

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i)$$

特征选择

- 经验熵
- 条件经验熵
- 信息增益=经验熵-条件经验熵
- 表示特征A给定条件下对数据集D的分类的不确定性减少的程度

$$g(D,A) = H(D) - H(D|A)$$

- 信息增益比：可能存在偏向于取值较多的特征的问题

$$g_R(D,A) = \frac{g(D,A)}{H_A(D)}$$

特征选择

- 最小二乘回归树
- 基尼指数：表示经过分割后D的不确定性，基尼指数越大，不确定性越大。

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

- 对于给定的样本集合D，基尼指数为：

$$Gini(D) = 1 - \sum_{k=1}^{K} (\frac{|C_k|}{|D|})^2$$

- 如果样本被特征A是否取a被分割为D1和D2两部分，那么集合D的基尼指数定义为

$$Gini(D,A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

决策树的生成

- Id3（信息增益）见例5.2 P62 例5.3P64
- （1）判断两种条件
- （2）计算经验熵和经验条件熵，算出信息增益
- （3）判断信息增益是否小于阈值，若不超过则选出最优特征，化为节点及其子节点
- （4）重复（1）-（3）

- C4.5（信息增益比） 见习题5.1 P75
- 与id3基本相同，判断的值为信息增益比
- CART（基尼指数）见例5.4
- （1）求出所有特征的基尼指数和每个特征的最优切分点
- （2）选择最优特征，作为根节点生成两个子节点
- （3）重复（1）（2）

- 算法最后展示

三种算法的优缺点

- Id3算法：无法处理特征为连续值的情况，优先选择有较多属性值的特征,因为属性值多的特征会有相对较大的信息增益。
- C4.5算法：可以处理*连续型属性*，解决了id3算法倾向于选择取值较多特征的问题。
- 见http://blog.sina.com.cn/s/blog_60acd6780100djcf.html
- CART算法：生成分类和回归树，天然的解决了id3倾向于选择取值较多特征的问题。

决策树剪枝

- 比较损失函数大小，进行剪枝

- P65 5.4

损失函数为

$$C_\alpha(T) = C(T) + \alpha|T|$$

- 右边第一项代表的预测误差，即拟合程度
- 第二项|T|为节点数，即复杂度，α控制着二者的影响。
- α小促使树较复杂，α大促使树较简单
- 实质就是比较叶节点回缩到父节点之前和之后的损失函数大小
- 这个算法α是确定的
- 剪枝过程模拟见习题5.1
- 问题：剪枝之后如何确定类？

设树T的叶节点的个数为$|T|$，t是树T的叶节点，该叶节点上有N_t个样本点，其中k类样本点有N_{tk}个，$k=1,2,...,K$, $H_t(T)$为叶节点t上的经验熵，$\alpha \geq 0$为参数，则决策树的损失函数可以定义为：

$$C_\alpha(T) = \sum_{t=1}^{|T|} N_t H_t(T) + \alpha|T| (1)$$

其中，经验熵为：

$$H_t(T) = -\sum_k \frac{N_{tk}}{N_t} \log \frac{N_{tk}}{N_t} \ (2)$$

将（1）式的第一项记为：

$$C(T) = \sum_{t=1}^{|T|} N_t H_t(T) = -\sum_{t=1}^{|T|}\sum_{k=1}^{K} N_{tk} \log \frac{N_{tk}}{N_t}$$

则：

$$C_\alpha(T) = C(T) + \alpha|T| (3)$$

- CART剪枝 $C_\alpha(T) = C(T) + \alpha|T|$
- 第一项为基尼指数
- 此时变为比较不同α所对应的最优子树的基尼指数。
- 如何确定α见书P73
- 用独立的验证数据集进行交叉验证

课后习题

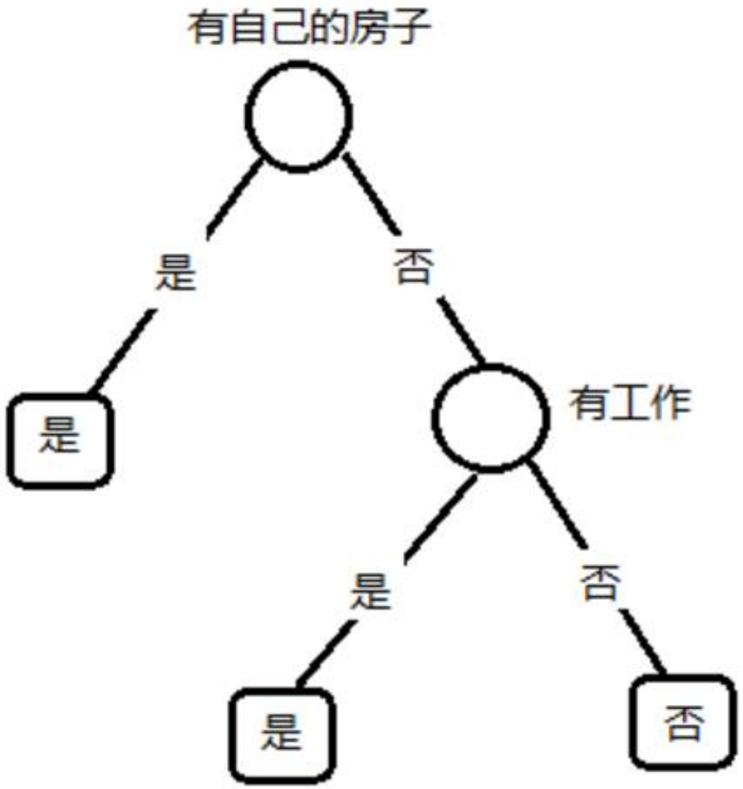
- 5.1
- （1）计算经验熵
- H(D） = 0.97095

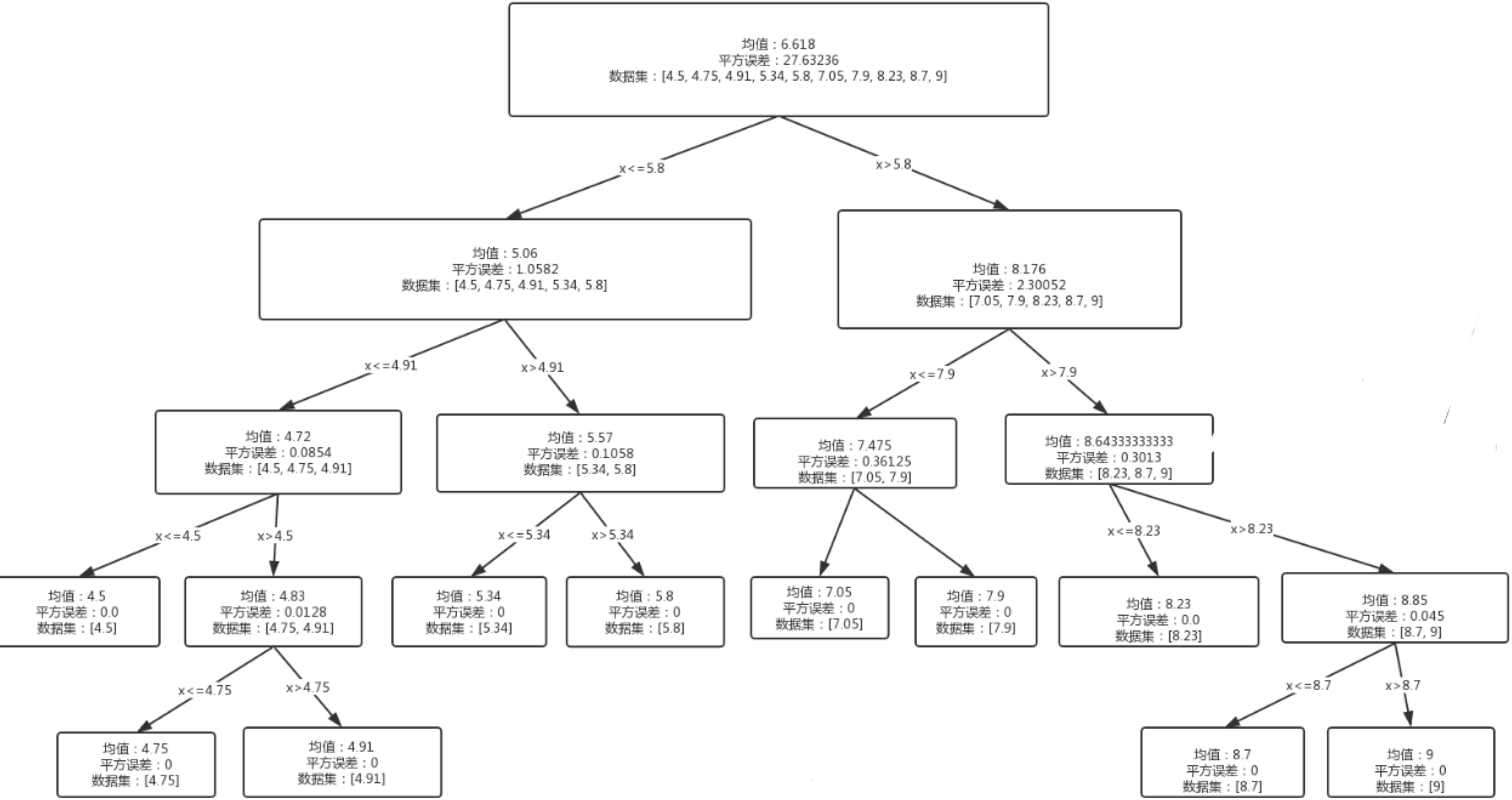
- （2）计算各特征的经验条件熵
- H(D|A1) = 0.88794
- H(D|A2) = 0.64730
- H(D|A3) = 0.55098
- H(D|A4) = 0.60796

- （3）计算各特征A的值的熵
- HA1(D)=1.58496
- HA2(D)=0.91830
- HA3(D)=0.97095
- HA4(D)=1.56605

- （4）计算信息增益比
- gR(D,A1）= 0.05237
- gR(D,A2）= 0.35244
- gR(D,A3）= 0.43253
- gR(D,A4）= 0.23179

- 选择最大的，A3特征，房子，将数据分成两部分，一部分有房，发现一定是，建立单节点，类别是“是”。另一边同样方法继续计算，发现是A2特征，工作，同样道理，最后建出树。


有自己的房子
是
否
是
有工作
是
否
是
否



均值：6.618
平方误差：27.63236
数据集：[4.5, 4.75, 4.91, 5.34, 5.8, 7.05, 7.9, 8.23, 8.7, 9]
x<=5.8
x>5.8
均值：5.06
平方误差：1.0582
数据集：[4.5, 4.75, 4.91, 5.34, 5.8]
均值：8.176
平方误差：2.30052
数据集：[7.05, 7.9, 8.23, 8.7, 9]
x<=4.91
x>4.91
x<=7.9
x>7.9
均值：4.72
平方误差：0.0854
数据集：[4.5, 4.75, 4.91]
均值：5.57
平方误差：0.1058
数据集：[5.34, 5.8]
均值：7.475
平方误差：0.36125
数据集：[7.05, 7.9]
均值：8.64333333333
平方误差：0.3013
数据集：[8.23, 8.7, 9]
x<=4.5
x>4.5
x<=5.34
x>5.34
x<=8.23
x>8.23
均值：4.5
平方误差：0.0
数据集：[4.5]
均值：4.83
平方误差：0.0128
数据集：[4.75, 4.91]
均值：5.34
平方误差：0
数据集：[5.34]
均值：5.8
平方误差：0
数据集：[5.8]
均值：7.05
平方误差：0
数据集：[7.05]
均值：7.9
平方误差：0
数据集：[7.9]
均值：8.23
平方误差：0.0
数据集：[8.23]
均值：8.85
平方误差：0.045
数据集：[8.7, 9]
x<=4.75
x>4.75
x<=8.7
x>8.7
均值：4.75
平方误差：0
数据集：[4.75]
均值：4.91
平方误差：0
数据集：[4.91]
均值：8.7
平方误差：0
数据集：[8.7]
均值：9
平方误差：0
数据集：[9]

- 5.3
- 假设存在两个或两个以上子树使损失函数Cα最小

- 那么他们的损失函数相同
- 但是每个子树结构一定不同，那么复杂度一定不同，所以必定存在简单的子树
- 矛盾，原理得证

- 5.4
- 假设最优树TA在[αi, αi+1)区间上有一个更优树TB

- 分两种情况讨论
- （1）TA的α>TB的α
- （2）TA的α<TB的α