Chap 5 决策树

图灵2201 张祎迪

5.1 根据表 5.1 所给的训练数据集，利用信息增益比（C4.5 算法）生成决策树

表 5.1　贷款申请样本数据表

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

取 A_1, A_2, A_3, A_4 分别表示特征 [年龄、有工作、有自己的房子和信贷] D表示数据集

Step1: 计算加入每个特征带来的信息增益比

> 注 $g_R(D, A_i)$ 的计算值已经在书中给出

$$g_R(D, A_1) = \frac{g(D,A_1)}{H_{A_1(D)}} = \frac{0.083}{-3*\frac{5}{15}log_2\frac{5}{15}} = 0.052$$
$$g_R(D, A_2) = \frac{g(D,A_1)}{H_{A_1(D)}} = \frac{0.324}{-\frac{5}{15}log_2\frac{5}{15}-\frac{10}{15}log_2\frac{10}{15}} = 0.353$$
$$g_R(D, A_3) = \frac{g(D,A_1)}{H_{A_1(D)}} = \frac{0.420}{-\frac{6}{15}log_2\frac{6}{15}-\frac{9}{15}log_2\frac{9}{15}} = 0.433$$
$$g_R(D, A_4) = \frac{g(D,A_1)}{H_{A_1(D)}} = \frac{0.363}{-\frac{5}{15}log_2\frac{5}{15}-\frac{6}{15}log_2\frac{6}{15}-\frac{4}{15}log_2\frac{4}{15}} = 0.232$$

$\because\ max(g_R(D, A_i)) = g_R(D, A_3) = 0.433$
$\therefore$ 取特征A_3 作为根结点的特征，将训练集分成两个子集D_1:是(有自己的房子) D_2: 否(没有自己的房子)

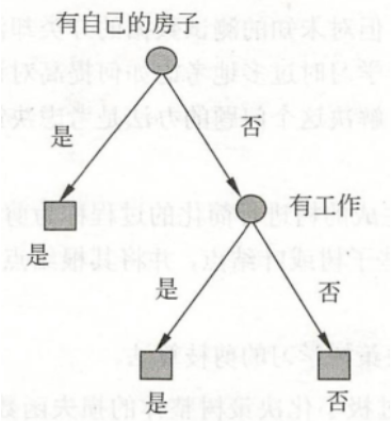
Step2 : 对划分后的子集依次递归地重复进行step1 知道满足终止条件，具体过程为

- D_1 中只包含“是”的样本点，满足终止条件，所以可以直接作为叶子结点
- 对D_2进行进一步的划分

  $$g_R(D_2, A_1) = \frac{g(D_2,A_1)}{H_{A_1(D_2)}} = \frac{0.251}{-\frac{4}{9}log_2\frac{4}{9}-\frac{2}{9}log_2\frac{2}{9}-\frac{3}{9}log_2\frac{3}{9}} = 0.164$$
  $$g_R(D_2, A_1) = \frac{g(D_2,A_1)}{H_{A_1(D_2)}} = \frac{0.918}{-\frac{3}{9}log_2\frac{3}{9}-\frac{6}{9}log_2\frac{6}{9}} = 1.000$$
  $$g_R(D_2, A_1) = \frac{g(D_2,A_1)}{H_{A_1(D_2)}} = \frac{0.363}{-\frac{4}{9}log_2\frac{4}{9}-\frac{4}{9}log_2\frac{4}{9}-\frac{1}{9}log_2\frac{1}{9}} = 0.340$$

  同理，选择A_2作为特征，并将D_2分割成两个子集D_3:是(有工作) D_2: 否(没有工作) 而D_3,D_4都只含有一类样本，所以终止条件。

最终决策树为：



5.2 已知如表 **5.2** 所示的训练数据，使用平方误差损失准则生成一个二叉回归树

表 5.2　训练数据表

x_i	1	2	3	4	5	6	7	8	9	10
y_i	4.50	4.75	4.91	5.34	5.80	7.05	7.90	8.23	8.70	9.00

- x_i为一维的，所以只需选择最优切分点

> $s=1$

$\hat{c_1}=4.5, \hat{c_2}=6.8633$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=22.648$

- 接下来的计算同理

> $s=2$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=17.702$

> $s=3$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=12.193$

> $s=4$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=7.379$

> $s=5$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=3.359$

> $s=6$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=5.074$

> $s=7$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=10.052$

> $s=8$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=15.178$

> $s=9$

$\sum_{x_i \in R_1(l,s)}(y_i-\hat{c_1})^2+\sum_{x_i \in R_2(l,s)}(y_i-\hat{c_2})^2=21.328$

> $s=10$

- 输入空间不改变

易得，当$s=5$时，取得最小值，可以将输入空间分为两个区域 $x \leq 5$ 和 $x>5$

对得到的自区域迭代地进行计算，设定阈值为0.2，最终可以得到二叉回归树为

```
      5
     / \
    3   7
   / \  / \
4.72 5.57 6  8
      /\ /\
     7.05 7.9 8.23 8.85

```

- 后面的计算由于原理类似，在计算时直接采用编程方法解决,结果如下