(1) 根据《统计学习方法》中表5.1所给的训练集数据，利用信息增益比算法（C4.5算法）生成决策树。

第一次：

$$H(D) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$

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

1° 年龄：

$$H(D|A_1) = \frac{1}{3}\times\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right) + \frac{1}{3}\times\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right)$$

$$+ \frac{1}{3}\times\left(-\frac{4}{5}\log\frac{4}{5} - \frac{1}{5}\log\frac{1}{5}\right) = 0.888$$

$$H_{A_1}(D) = -\frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} = 1.585$$

$$g_R(D,A_1) = \frac{g(D,A_1)}{H_{A_1}(D)} = 0.052$$

2° 有工作：

$$H(D|A_2) = \frac{2}{3}\times\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right) + \frac{1}{3}\times 0$$

$$= 0.647$$

$$H_{A_2}(D) = -\frac{2}{3}\times\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

$$g_R(D,A) = \frac{g(D,A)}{H_{A_2}(D)} = 0.354$$

3° 有自己的房子

$$H(D|A_3) = \frac{3}{5}\times\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right) + \frac{2}{5}\times 0$$

$$= 0.551$$

$$H_{A_3}(D) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$

$$g_R(D,A_3) = \frac{g(D,A_3)}{H_{A_3}(D)} = 0.433$$

4° 信贷情况

$$H(D|A_4) = \frac{5}{15}\times\left(-\frac{4}{5}\log\frac{4}{5} - \frac{1}{5}\log\frac{1}{5}\right) + \frac{6}{15}\times\left(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right)$$

$$+ \frac{4}{15}\times 0$$

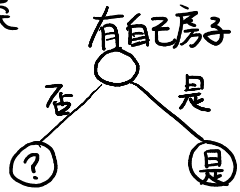
$$= 0.608$$

$$H_{A_4}(D) = -\frac{5}{15}\log\frac{5}{15} - \frac{4}{15}\log\frac{4}{15} - \frac{6}{15}\log\frac{6}{15}$$

$$= 1.566$$

$$g_R(D,A_4) = \frac{g(D,A_4)}{H_{A_4}(D)} = 0.232$$

按有自己房子分两类



第二次:

$$H(D)=-\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3}=0.918$$

1° 年龄

$$H(D|A)=\frac{4}{9}\times\left(-\frac{1}{4}\log\frac{1}{4}-\frac{3}{4}\log\frac{3}{4}\right)+\frac{2}{9}\times 0+\frac{3}{9}\times\left(-\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3}\right)$$

$$=0.667$$

$$H_{A_1}(D)=-\frac{4}{9}\log\frac{4}{9}-\frac{3}{9}\log\frac{3}{9}-\frac{2}{9}\log\frac{2}{9}=1.530$$

$$g_R(D|A_1)=\frac{g(D|A_1)}{H_{A_1}(D)}=0.164$$

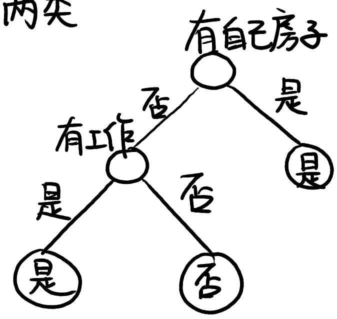
2° 有工作

$$H(D|A_2)=\frac{1}{3}\times 0+\frac{2}{3}\times 0$$

$$H_{A_2}(D)=0.918$$

$$g_R(D,A_2)=\frac{g(D,A_2)}{H_{A_2}(D)}=1$$ (显然增益比不会比1大了)

?可以再按有工作分两类



至此全部分完，结束

(2) 已知表 1所示的训练数据，试用平方损失准则生成一个二叉回归树。（提示：写出计算步骤）

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

表 1: 训练数据表

第一次划分：

通过观察，划分点出现在 $s=4,5,6$ 比较合理

需要求 $\min\limits_{s} \sum\limits_{x\le s}(y_i-c_1)^2+\sum\limits_{x>s}(y_i-c_2)^2$ 记为 $R(s)$

1° $s=4$. $c_1=\overline{y[1:4]}$ $c_2=\overline{y[5:10]}$

$R(4)=0.3737+7.005=7.3787$

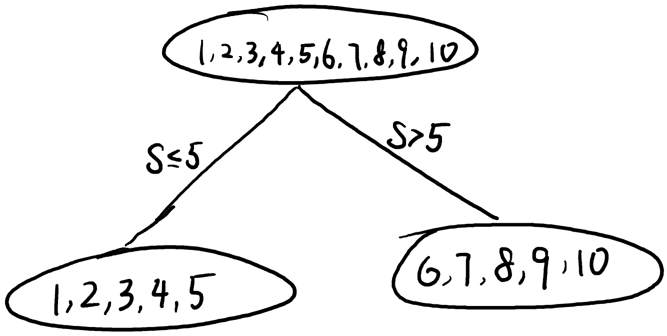
2° $s=5$ $c_1=\overline{y[1:5]}$ $c_2=\overline{y[6:10]}$

$R(5)=1.0582+2.3005=3.3587$

3° $s=6$ $c_1=\overline{y[1:6]}$ $c_2=\overline{y[7:10]}$

$R(6)=4.3582+0.7157=5.0739$

在 $s=5$ 处划分



第二次划分（$x=1,2,3,4,5$）

1° $s=1$ $c_1=y_1$, $c_2=\overline{y[2:5]}$

$R(1)=0+0.6662=0.6662$

2° $s=2$ $c_1=\overline{y[1:2]}$ $c_2=\overline{y[3:5]}$

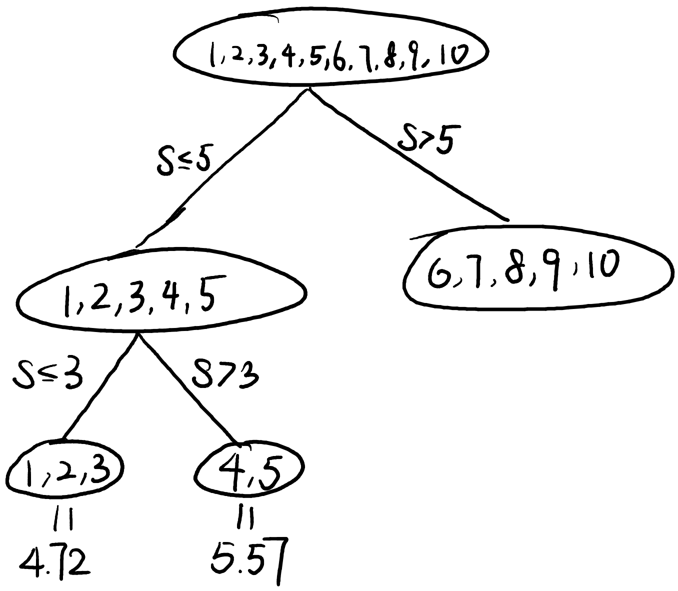
$R(2)=0.0313+0.3962=0.4273$

3° $s=3$ $c_1=\overline{y[1:3]}$ $c_2=\overline{y[4:5]}$

$R(3)=0.0854+0.1058=0.1912$

4° $s=4$ $c_1=\overline{y[1:4]}$ $c_2=y_5$

$R(4)=0.3737+0=0.3737$



在 $s=3$ 处划分，$c_1=4.72$，$c_2=5.57$

第二次划分（$x=6,7,8,9,10$）

1° $s=6$ $\quad c_1=y_6$, $c_2=\overline{y[7:10]}$

$R(6)=0+0.7157=0.7157$

2° $s=7$ $\quad c_1=\overline{y[6:7]}$, $c_2=\overline{y[8:10]}$

$R(7)=0.3613+0.3013=0.6626$

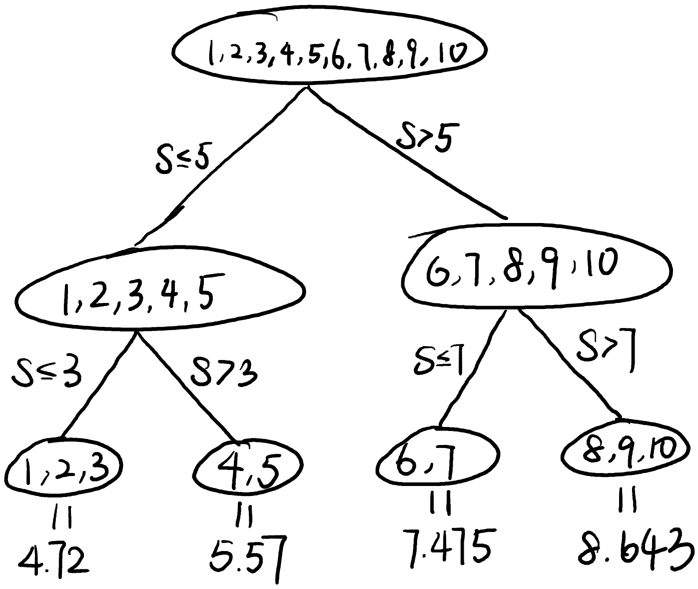
3° $s=8$ $\quad c_1=\overline{y[6:8]}$, $c_2=\overline{y[9:10]}$

$R(8)=0.7413+0.0450=0.7863$

4° $s=9$ $\quad c_1=\overline{y[6:9]}$ $\quad c_2=y_{10}$

$R(9)=1.4518+0=1.4518$

在$s=7$处划分, $c_1=7.475$, $c_2=8.643$



(3) 在CART剪枝过程中，假设第k步，对每个内部节点t计算$C(T_t)$、$|T_t|$以及

$$g_k(t)=\frac{C(t)-C\left(T_t\right)}{|T_t|-1}$$

记第k步所有内部节点的集合为$\mathcal{M}_k$, 记 $\alpha_k=g_k(a)=\min_{t\in\mathcal{M}_k} g_k(t)$，即节点$a$是使函数$g_k(t)$取值最小的内部节点（假设此内部节点唯一），则将$a$剪枝。记剪枝后内部节点的集合是$\mathcal{M}_{k+1}$，定义$\alpha_{k+1}=g_{k+1}(b)=\min_{t\in\mathcal{M}_{k+1}} g_{k+1}(t)$。请证明$\alpha_{k+1}>\alpha_k$.

证： $\because g_k(a)=\min\limits_{t\in M_k} g(t)=\alpha_k$

$C_k(T_a)$表示第k轮损失，$|T_a|_k$ 表示第k轮节点数

显然 $\forall t\in M_{k+1}$，若t的子节点在k轮未被剪枝，结构完全一样

$C_{k+1}(T_t)=C_k(T_t)$, $|T_t|_{k+1}=|T_t|_k$

若t的子节点被剪枝了 $|T_t|_{k+1}=|T_t|_k-|T_a|_k+1$（有$|T_a|_k$节点被1个节点代替）

$C_{k+1}(T_t)=C_k(T_t)-C_k(T_a)+C_k(a)$（$T_t$的损失中$T_a$子树的损失被$a$单节点代替）

$C_k(a)-C_k(T_a)=\alpha_k(|T_a|_k-1)$

$\forall t\in M_k\setminus\{a\}$, $C_k(t)-C_k(T_t)>\alpha_k(|T_t|_k-1)$

$\forall s\in M_{k+1}$，则显然$s\neq a$，

若a不是s曾经的子节点

则 $C_{k+1}(s)-C_{k+1}(T_s)=C_k(s)-C_k(T_s)>\alpha_k(|T_s|_k-1)=\alpha_k(|T_s|_{k+1}-1)$

若a曾经是s的子节点

则 $C_{k+1}(s)-C_{k+1}(T_s)=C_k(s)-C_{k+1}(T_s)$

$C_{k+1}(T_s)=C_k(T_s\setminus T_a)+C_k(a)=C_k(T_s)-C_k(T_a)+C_k(a)$

$\qquad =C_k(T_s)-\alpha_k(|T_a|_k-1)$

$C_{k+1}(s)-C_{k+1}(T_s)=C_k(s)-C_k(T_s)+\alpha_k(|T_a|_k-1)$

$\qquad >\alpha_k(|T_s|_k-1-|T_a|_k+1)=\alpha_k(|T_s|_{k+1}-1)$

$\therefore \forall s\in M_{k+1}$, $g(s)=\dfrac{C_{k+1}(s)-C_{k+1}(T_s)}{|T_s|_{k+1}-1}>\alpha_k$

则 $\alpha_{k+1}=\min\limits_{s\in M_{k+1}} g(s)>\alpha_k$