

6.1

原表：

序号	年龄	是否有车	收入情况	信用情况	是否同意贷款
1	19	否	一般	非常好	否
2	32	否	一般	一般	否
3	75	否	良好	一般	否
4	21	否	一般	好	否
5	36	否	一般	一般	否
6	40	否	一般	好	否
7	69	是	一般	好	是
8	45	是	良好	好	是
9	61	是	一般	非常好	是
10	66	是	一般	非常好	是
11	25	否	良好	好	是
12	42	是	一般	非常好	是
13	62	否	良好	好	是
14	63	否	良好	非常好	是
15	29	是	良好	一般	是

对训练集 D，首先计算划分前样本的信息熵：

$$I(D) = -\frac{6}{15}\log_2\frac{6}{15} - \frac{9}{15}\log_2\frac{9}{15} = 0.9710$$

关于年龄：

$$\begin{aligned}
 \Delta I(D, A_1) &= I(D) - \left[\frac{4}{15}I(D_1) + \frac{5}{15}I(D_2) + \frac{6}{15}I(D_3) \right] \\
 &= 0.9710 - \left[\frac{4}{15} \left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} \right) + \frac{5}{15} \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \right) + \frac{6}{15} \left(-\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} \right) \right] \\
 &= 0.9710 - \left[\frac{4}{15} \times 1 + \frac{5}{15} \times 0.9710 + \frac{6}{15} \times 0.6500 \right] = 0.1207
 \end{aligned}$$

关于收入情况：

$$\begin{aligned}
 \Delta I(D, A_2) &= I(D) - \left[\frac{6}{15}I(D_1) + \frac{9}{15}I(D_2) \right] \\
 &= 0.9710 - \left[\frac{6}{15} \left(-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} \right) + \frac{9}{15} \left(-\frac{5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9} \right) \right] \\
 &= 0.9710 - \left[\frac{6}{15} \times 0.6500 + \frac{9}{15} \times 0.9911 \right] = 0.1163
 \end{aligned}$$

关于是否有车：

$$\begin{aligned}
 \Delta I(D, A_3) &= I(D) - \left[\frac{6}{15}I(D_1) + \frac{9}{15}I(D_2) \right] \\
 &= 0.9710 - \left[\frac{6}{15} \times 0 + \frac{9}{15} \left(-\frac{3}{9}\log_2\frac{3}{9} - \frac{6}{9}\log_2\frac{6}{9} \right) \right] = 0.4200
 \end{aligned}$$

关于信贷情况：

$$\Delta I(D, A_4) = I(D) - \left[\frac{5}{15}I(D_1) + \frac{6}{15}I(D_2) + \frac{4}{15}I(D_3) \right]$$

$$= 0.9710 - \left[\frac{5}{15} \left(-\frac{4}{5} \log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5} \right) + \frac{6}{15} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) + \frac{4}{15} \left(-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) \right]$$

$$= 0.9710 - \left[\frac{5}{15} \times 0.7219 + \frac{6}{15} \times 0.9183 + \frac{4}{15} \times 0.8113 \right] = 0.1467$$

综上所述，第一层应当按照 A_3 为分类特征，此时按照是否有车划分为“是”和“否”两类，此时有车的一类必然能得到贷款，因此仅对于无车的一类继续划分：

序号	年龄	收入情况	信用情况	是否同意贷款
1	19	一般	非常好	否
2	32	一般	一般	否
3	75	良好	一般	否
4	21	一般	好	否
5	36	一般	一般	否
6	40	一般	好	否
11	25	良好	好	是
13	62	良好	好	是
14	63	良好	非常好	是

$$I(D_2) = -\frac{3}{9} \log_2 \frac{3}{9} - \frac{6}{9} \log_2 \frac{6}{9} = 0.9183$$

关于年龄：

$$\Delta I(D_2, A_1) = 0.9183 - \left[\frac{3}{9} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) + 0 + \frac{3}{9} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) \right] = 0.3061$$

关于收入情况：

$$\Delta I(D_2, A_2) = 0.9183 - \left[0 + \frac{4}{9} \left(-\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} \right) \right] = 0.5577$$

关于信贷情况：

$$\Delta I(D_2, A_4) = 0.9183 - \left[\frac{2}{9} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) + \frac{4}{9} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) + 0 \right] = 0.2516$$

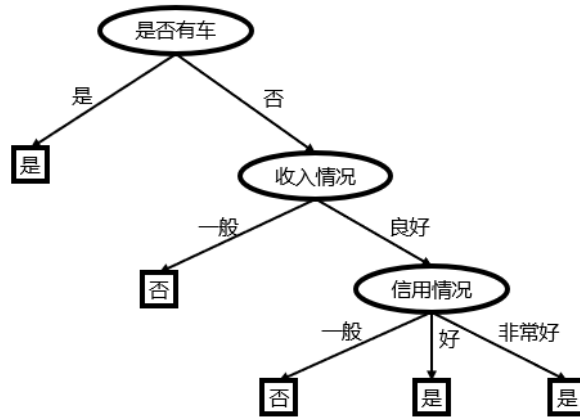
因此按照收入情况继续进行划分，收入情况一般的都不予贷款，收入情况良好的继续生长：

序号	年龄	信用情况	是否同意贷款
3	75	一般	否
11	25	好	是
13	62	好	是
14	63	非常好	是

$$I(D_3) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8113$$

此时按照年龄无法划分，按照信用情况可以完成划分

综上所述得到三层树为：



6.2

(1) 依题意：

$$\begin{aligned}
 E_{COM} &= \mathbb{E}_{\mathbf{x}} \left\{ \left[\frac{1}{M} \sum_{m=1}^M \epsilon_m(\mathbf{x}) \right]^2 \right\} = \frac{1}{M^2} \left\{ \sum_{m \neq l} \mathbb{E}_{\mathbf{x}} [\epsilon_m(\mathbf{x}) \epsilon_l(\mathbf{x})] + \sum_{m=1}^M \mathbb{E}_{\mathbf{x}} \{ [\epsilon_m(\mathbf{x})]^2 \} \right\} \\
 &= \frac{1}{M^2} \left\{ 0 + \sum_{m=1}^M \mathbb{E}_{\mathbf{x}} \{ [\epsilon_m(\mathbf{x})]^2 \} \right\} = \frac{1}{M} E_{AV}
 \end{aligned}$$

(2) 函数 $f(y) = y^2$ 是凸函数，由琴生不等式可知 $\mathbb{E}[f(y)] \geq f(\mathbb{E}[y])$ ，于是：

$$\begin{aligned}
 E_{COM} &= \mathbb{E}_{\mathbf{x}} \left\{ \left[\frac{1}{M} \sum_{m=1}^M \epsilon_m(\mathbf{x}) \right]^2 \right\} \leq \mathbb{E}_{\mathbf{x}} \left\{ \frac{1}{M} \sum_{m=1}^M [\epsilon_m(\mathbf{x})]^2 \right\} = \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{\mathbf{x}} \{ [\epsilon_m(\mathbf{x})]^2 \} \\
 &= E_{AV}
 \end{aligned}$$