一、计算题

对下表分别用 ID3 和 CART 算法生成决策树（要求写出详细的计算步骤）。

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	好	是
5	青年	是	是	非常好	是
6	青年	是	是	一般	是
7	青年	否	否	一般	否
8	中年	否	否	一般	否
9	中年	否	否	好	否
10	中年	是	是	好	是
11	中年	否	是	非常好	是
12	中年	否	是	非常好	是
13	中年	是	是	好	是
14	老年	否	是	非常好	是
15	老年	否	是	好	是
16	老年	是	否	好	是
17	老年	是	否	非常好	是
18	老年	否	否	一般	否

答案：

分别以 A_1、A_2、A_3、A_4 表示特征“年龄”、“有工作”、“有房子”、“信贷情况”，

ID3 决策树生成过程如下：

计算各特征对数据集 D 的信息增益：

经验熵为：$H(D)=-(\frac{12}{18}\log_2\frac{12}{18}+\frac{6}{18}\log_2\frac{6}{18})=0.9183$

$$G(D,A_1)=H(D)-H(D,A_1)$$

$$=0.9183-[\frac{7}{18}\times(-\frac{4}{7}\log_2\frac{4}{7}-\frac{3}{7}\log_2\frac{3}{7})+\frac{6}{18}\times(-\frac{4}{6}\log_2\frac{4}{6}-\frac{2}{6}\log_2\frac{2}{6})+\frac{5}{18}\times(-\frac{4}{5}\log_2\frac{4}{5}-\frac{1}{5}\log_2\frac{1}{5})]$$

$$=0.028$$

$$G(D,A_2)=H(D)-H(D,A_2)$$

$$=0.9183-[\frac{8}{18}\times0+\frac{10}{18}\times(-\frac{4}{10}\log_2\frac{4}{10}-\frac{6}{10}\log_2\frac{6}{10})]$$

$$=0.379$$

$$G(D,A_3)=H(D)-H(D,A_3)$$

$$=0.9183-[\frac{9}{18}\times0+\frac{9}{18}\times(-\frac{3}{9}\log_2\frac{3}{9}-\frac{6}{9}\log_2\frac{6}{9})]$$

$$=0.459$$ （信息增益最大）

$$G(D,A_4)=H(D)-H(D,A_4)$$
$$=0.9183-[\frac{5}{18}\times(-\frac{1}{5}\log_2\frac{1}{5}-\frac{4}{5}\log_2\frac{4}{5})+\frac{8}{18}\times(-\frac{6}{8}\log_2\frac{6}{8}-\frac{2}{8}\log_2\frac{2}{8})+\frac{5}{18}\times 0]$$
$$=0.357$$

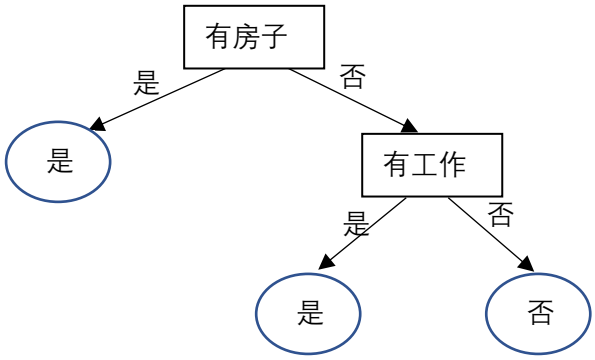
$G(D,A_3)$最大，**因此选择特征 A_3=“有房子”作为根节点**

（1）分支 1：“有房子”=是时，类别也全部为“是”，因此可以作为叶子，

（2）分支 2：“有房子”=否时，继续计算后续的信息增益：

$$H(D_2)=-(\frac{3}{9}\log_2\frac{3}{9}+\frac{6}{9}\log_2\frac{6}{9})=0.918$$
$$H(D_2,A_1)=-\frac{4}{9}\times(\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4})-\frac{2}{9}\times 0-\frac{3}{9}\times(\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3})=0.667$$
$$H(D_2,A_2)=-\frac{3}{9}\times 0-\frac{6}{9}\times 0=0$$
$$H(D_2,A_4)=-\frac{4}{9}\times 0-\frac{4}{9}\times(\frac{2}{4}\log_2\frac{2}{4}+\frac{2}{4}\log_2\frac{2}{4})-\frac{1}{9}\times 0=0.444$$
$$G(D_2,A_1)=H(D_2)-H(D_2,A_1)=0.918-0.667=0.251$$
$$G(D_2,A_2)=H(D_2)-H(D_2,A_2)=0.918-0=0.918$$ （信息增益最大）
$$G(D_2,A_4)=H(D_2)-H(D_2,A_4)=0.918-0.444=0.474$$

因此 A_2=“有工作”为最优特征，作为下一个划分节点：此时，“有工作”=是时，所有样本类别均为“是”；“有工作”=否时，所有样本类别均为“否”。后续全部为叶子。因此构建的 ID3 决策树如下：



CART 决策树生成过程如下：

计算各特征对数据集 D 的基尼指数:

$$Gini(D,A_1=\text{“青年”})=\frac{7}{18}\times 2\times\frac{4}{7}\times(1-\frac{4}{7})+\frac{11}{18}\times 2\times\frac{8}{11}\times(1-\frac{8}{11})=0.423$$
$$Gini(D,A_1=\text{“中年”})=\frac{6}{18}\times 2\times\frac{4}{6}\times(1-\frac{4}{6})+\frac{12}{18}\times 2\times\frac{8}{12}\times(1-\frac{8}{12})=0.444$$
$$Gini(D,A_1=\text{“老年”})=\frac{5}{18}\times 2\times\frac{4}{5}\times(1-\frac{4}{5})+\frac{13}{18}\times 2\times\frac{8}{13}\times(1-\frac{8}{13})=0.431$$

$$Gini(D,A_2)=\frac{8}{18}\times 2\times\frac{8}{8}\times(1-\frac{8}{8})+\frac{10}{18}\times 2\times\frac{4}{10}\times(1-\frac{4}{10})=0.267$$

$$Gini(D,A_3)=\frac{9}{18}\times 2\times\frac{9}{9}\times(1-\frac{9}{9})+\frac{9}{18}\times 2\times\frac{3}{9}\times(1-\frac{3}{9})=0.222$$ （基尼指数最小）

$$Gini(D,A_4=\text{“一般”})=\frac{5}{18}\times 2\times\frac{1}{5}\times(1-\frac{1}{5})+\frac{13}{18}\times 2\times\frac{11}{13}\times(1-\frac{11}{13})=0.277$$

$$Gini(D,A_4=\text{“好”})=\frac{8}{18}\times 2\times\frac{6}{8}\times(1-\frac{6}{8})+\frac{10}{18}\times 2\times\frac{6}{10}\times(1-\frac{6}{10})=0.433$$

$$Gini(D,A_4=\text{“非常好”})=\frac{5}{18}\times 2\times\frac{5}{5}\times(1-\frac{5}{5})+\frac{13}{18}\times 2\times\frac{7}{13}\times(1-\frac{7}{13})=0.359$$

$Gini(D,A_3)$ 最小，**因此选择特征 A_3=“有房子”作为根节点。**

（1）分支 1：“有房子”=是时，类别也全部为“是”，因此可以作为叶子；

（2）分支 2：“有房子”=否时，继续计算后续的基尼指数：

$$Gini(D_2,A_1=\text{“青年”})=\frac{4}{9}\times 2\times\frac{1}{4}\times(1-\frac{1}{4})+\frac{5}{9}\times 2\times\frac{2}{5}\times(1-\frac{2}{5})=0.433$$

$$Gini(D_2,A_1=\text{“中年”})=\frac{2}{9}\times 2\times\frac{2}{2}\times(1-\frac{2}{2})+\frac{7}{9}\times 2\times\frac{3}{7}\times(1-\frac{3}{7})=0.381$$

$$Gini(D_2,A_1=\text{“老年”})=\frac{3}{9}\times 2\times\frac{2}{3}\times(1-\frac{2}{3})+\frac{6}{9}\times 2\times\frac{1}{6}\times(1-\frac{1}{6})=0.333$$

$$Gini(D_2,A_2)=\frac{3}{9}\times 2\times\frac{3}{3}\times(1-\frac{3}{3})+\frac{6}{9}\times 2\times\frac{6}{6}\times(1-\frac{6}{6})=0$$ （基尼指数最小）

此时，因为 $Gini(D_2,A_2)$ 最小，达到可能取到的最小值，因此直接选择 A_2=“有工作”作为下一个最优特征，并且由 A_2=“有工作”作为中间节点，后续分支全部为叶子，从而构建的 CART 决策树与上述 ID3 决策树完全一致。