112-1 Decision Tree Take Home Exam

學號：412770116　　姓名：許嘉隆

題目：假設我們有 8 筆資料如下：

序號	有房	婚姻	年收入	拖欠
1	是	單身	125K	否
2	否	已婚	100K	否
3	否	單身	70K	否
4	是	已婚	120K	否
5	否	離婚	95K	是
6	否	已婚	60K	否
7	是	離婚	220K	否
8	否	單身	85K	是

目的是要找出**拖欠**的分類標準。若我們使用熵(Entropy)為指標，並使用訊息增益(Information Gain)為分枝的標準。請回答下列問題。

欄位：序號、有房、婚姻、年收入、拖欠

1. 計算拖欠的熵H(拖欠)。(10%)

$$Entropy = H(p_1, p_2, \ldots, p_K) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

拖欠	人數	*Ratio*
是	2	$\frac{2}{8}$
否	6	$\frac{6}{8}$

$$\begin{aligned} H(\text{拖欠}) &= -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right) \\ &\approx -(-0.5 + (-0.31)) \\ &\equiv 0.81 \end{aligned}$$ ■

2. (a) 計算條件熵(Conditional Entropy)，已知有房的訊息，拖欠的條件熵H(拖欠|有房)。(10%)
   (b) 計算有房的資訊增益(Gain有房(拖欠))。(10%)

(*a*) *Conditional Entropy* :

$$H(D|A) = \sum_{i,k} P(A_i)H(D_K \vee A_i) = -\sum_{i=1}^{n} P(A_i) \sum_{k=1}^{K} P(D_k \vee A_i) \log_2 P(D_k \vee A_i)$$

有房	拖欠	人數	*Ratio*
是 ($D_1 = 3$)	是 (D_{11})	0	$\frac{0}{3}$
$A_1 = \frac{3}{8}$	否 (D_{12})	3	$\frac{3}{3}$
否 ($D_2 = 5$)	是 (D_{21})	2	$\frac{2}{5}$
$A_2 = \frac{5}{8}$	否 (D_{22})	3	$\frac{3}{5}$

$$
\begin{aligned}
H(\text{拖欠}|\text{有房}) &= -\frac{3}{8}\left(\frac{0}{3}\log_2\left(\frac{0}{3}\right)+\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right)-\frac{5}{8}\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)\\
&\approx -\frac{3}{8}(0)-\frac{5}{8}(-1.32\ -\ 0.74)\\
&= 1.29
\end{aligned}
$$

■

(*b*) $Gain_A(D) = H(D) - H(D \vee A)$

$Gain_{\text{有房}}(\text{拖欠}) = H(\text{拖欠}) - H(\text{拖欠}|\text{有房}) = 0.81 - 1.29 = -0.48$ ■

3. (a) 計算條件熵(Conditional Entropy)，已知婚姻的訊息，拖欠的條件熵H(拖欠|婚姻)。(10%)
(b) 計算婚姻的資訊增益(Gain婚姻(拖欠))。(10%)

(*a*) *Conditional Entropy* :

$$H(D|A) = \sum_{i,k} P(A_i)H(D_K \vee A_i) = -\sum_{i=1}^{n} P(A_i)\sum_{k=1}^{K} P(D_k \vee A_i)\log_2 P(D_k \vee A_i)$$

婚姻	拖欠	人數	*Ratio*
單身($D_1 = 3$)	是 (D_{11})	1	$\frac{1}{3}$
$A_1 = \frac{3}{8}$	否 (D_{12})	2	$\frac{2}{3}$
已婚($D_2 = 3$)	是 (D_{21})	0	$\frac{0}{3}$
$A_2 = \frac{3}{8}$	否 (D_{22})	3	$\frac{3}{3}$
離婚($D_3 = 2$)	是 (D_{31})	1	$\frac{1}{2}$
$A_3 = \frac{2}{8}$	否 (D_{32})	1	$\frac{1}{2}$

$$
\begin{aligned}
H(\text{拖欠}|\text{婚姻}) &= -\frac{3}{8}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right)+\frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)-\frac{3}{8}\left(\frac{0}{3}\log_2\left(\frac{0}{3}\right)+\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right)-\frac{2}{8}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right)+\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\\
&\approx -\frac{3}{8}(-0.53-0.39)-\frac{3}{8}(0)-\frac{2}{8}(-0.5\ -0.5)\\
&= 0.345+0.25\\
&= 0.595
\end{aligned}
$$

■

(*b*) $Gain_A(D) = H(D) - H(D \vee A)$

$Gain_{\text{婚姻}}(\text{拖欠}) = H(\text{拖欠}) - H(\text{拖欠}|\text{婚姻}) = 0.81 - 0.595 = 0.215$ ■

4. (a) 年收入為數值型特徵，若為二元(Binary)分類，試計算條件熵H(拖欠|年收入)最適合的分枝值為何。(20%)
(b) 計算年收入的資訊增益(Gain年收入(拖欠))。(10%)

(*a*)

	60 ~	70 ~	85 ~	95 ~	100 ~	120 ~	125 ~	220
	65	78	90	98	110	123	173	
	⩽ >	⩽ >	⩽ >	⩽ >	⩽ >	⩽ >	⩽ >	
是	0 2	0 2	1 1	2 0	2 0	2 0	2 0	
否	1 5	2 4	2 4	2 4	3 3	4 2	5 1	
Entropy	H_1 0.755	H_2 0.689	H_3 0.796	H_4 0.5	H_5 0.607	H_6 0.439	H_7 0.755	

$$
\begin{aligned}
H_1 &= -\frac{1}{8}\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right)+\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right)-\frac{7}{8}\left(\frac{2}{7}\log_2\left(\frac{2}{7}\right)+\frac{5}{7}\log_2\left(\frac{5}{7}\right)\right) &\approx 0.755\\
H_2 &= -\frac{2}{8}\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right)+\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)-\frac{6}{8}\left(\frac{2}{6}\log_2\left(\frac{2}{6}\right)+\frac{4}{6}\log_2\left(\frac{4}{6}\right)\right) &\approx 0.689\\
H_3 &= -\frac{3}{8}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right)+\frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)-\frac{5}{8}\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right)+\frac{4}{5}\log_2\left(\frac{4}{5}\right)\right) &\approx 0.796\\
H_4 &= -\frac{4}{8}\left(\frac{2}{4}\log_2\left(\frac{2}{4}\right)+\frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)-\frac{4}{8}\left(\frac{0}{4}\log_2\left(\frac{0}{4}\right)+\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right) &\approx 0.5\\
H_5 &= -\frac{5}{8}\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)-\frac{3}{8}\left(\frac{0}{3}\log_2\left(\frac{0}{3}\right)+\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right) &\approx 0.607\\
H_6 &= -\frac{6}{8}\left(\frac{2}{6}\log_2\left(\frac{2}{6}\right)+\frac{4}{6}\log_2\left(\frac{4}{6}\right)\right)-\frac{2}{8}\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right)+\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) &\approx 0.439\\
H_7 &= -\frac{7}{8}\left(\frac{2}{7}\log_2\left(\frac{2}{7}\right)+\frac{5}{7}\log_2\left(\frac{5}{7}\right)\right)-\frac{1}{8}\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right)+\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) &\approx 0.755
\end{aligned}
$$

因為H_6的值最小(0.439)，所以最適合的分枝值為 123K ■

(b) $Gain_{年收入}$(拖欠) = H(拖欠) − H(拖欠|年收入) = 0.81 − 0.439 = 0.371 ■

5. (a) 建構第一層的決策樹，以哪個特徵為主。(10%)
   (b) 建構第一層的決策樹。(10%)

(a) 選最大的$Imformation\ Gain$

$Gain_{有房}$(拖欠) = −0.48

$Gain_{婚姻}$(拖欠) = 0.215

$Gain_{年收入}$(拖欠) = 0.371

∵ $Gain_{年收入}$(拖欠)最大

∴ 建構第一層的決策樹，以年收入為主

(b)