HW 2 Decision Tree

Training data set: Who buys computer?

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

X : feature　　class

		yes	No
age :	≤ 30	2	3
	31…40	4	0
	> 40	3	2
income :	high	2	2
	med	4	2
	Low	3	1
Student :	yes	6	1
	No	3	5
Credit :	ex	3	3
	fair	6	2

1. หา Class

Yes No

(9, 5)

สูตร $\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

$= I(9,5)$

$= -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

$= 0.41 + 0.53$

$\therefore \text{Info}(D) = 0.94$ ※

2. หา Feature

$\text{Info}_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2)$

$= \frac{5}{14}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]$

$+ \frac{4}{14}\left[-\frac{4}{4}\log_2\left(\frac{4}{2}\right) - \frac{0}{4}\log\left(\frac{0}{4}\right)\right]$

$+ \frac{5}{14}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$

$= \frac{5}{14}\left[0.5287 + 0.44217\right] + \frac{5}{14}\left[0.44217 + 0.52877\right]$

$\therefore \text{Info}_{age}(D) = 0.694$

※ age

$\text{Info}_{incom}(D) = \frac{4}{14} I(2,2) + \frac{6}{14} I(4,2) \frac{4}{14} I(3,1)$

$= \frac{4}{14}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right] + \frac{4}{14}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$

$\therefore \text{Info}_{incom}(D) = 0.911$

※ income

$\text{Info}_{student}(D) = \frac{7}{14} I(6,1) + \frac{7}{14} I(3,4)$

$= \frac{7}{14}\left[-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right] + \frac{7}{14}\left[-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right]$

$\therefore \text{Info}_{student}(D) = 0.7883$

※ student

$$Info_{credit}(D) = \frac{6}{14} I(3,3) + \frac{8}{14} I(6,2)$$

$$= \frac{6}{14}\left[-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right)\right] + \frac{8}{14}\left[-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right]$$

$\therefore Info_{credit}(D) = 0.892$

✗ Credit

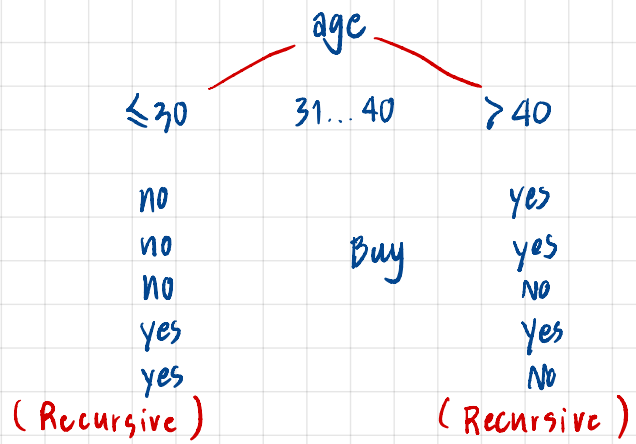
- จาก $Gain(A) = Info(D) - Info_A(D)$

$Gain(age) = 0.94 - 0.694 = 0.246$ (มากที่สุดจึงเลือกเป็น root node)

$Gain(income) = 0.94 - 0.911 = 0.029$

$Gain(student) = 0.94 - 0.788 = 0.151$

$Gain(credit\ rating) = 0.94 - 0.892 = 0.048$



age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Recursive ≤30 (2,3) Yes No

① ใน class * ≤30

$$Info(D) = I(2,3)$$

$$= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$\therefore Info(D) = 0.971$$

② ใน feature

$$Info_{income}(D) = \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0)$$

$$= \frac{2}{5}\left[-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] + \frac{1}{5}\left[-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right]$$

$$\therefore Info_{Income}(D) = 0.4$$

* income

$$Info_{student}(D) = \frac{2}{5}I(2,0) + \frac{3}{5}I(0,3)$$

$$= \frac{2}{5}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right] + \frac{3}{5}\left[-\frac{0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right)\right]$$

$$\therefore Info_{student}(D) = 0$$

* student

$$Info_{credit}(D) = \frac{2}{5} I(1,1) + \frac{3}{5} I(1,2)$$

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

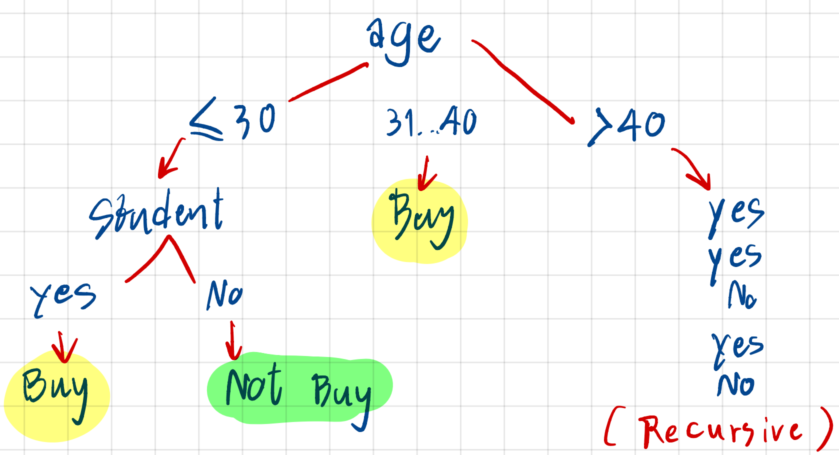
$\therefore$ $Info_{credit}(D) = 0.9509$

\# Credit

Gain (income) = 0.971 - 0.4 = 0.5710

Gain (studet) = 0.971 - 0 = 0.971 (Gain มากสุด) เลือก

Gain (Credit_rating) = 0.971 - 0.9509 = 0.0201



Recursive > 40

1. หา Class (3, 2) (Yes, No)

$$Info(D) = I(3,2)$$

$$= -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$\therefore$ $Info(D) = 0.971$

q1 feature

$$\text{Info}(D) = \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$$

$$= \frac{3}{5}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$

$\therefore \text{Info}_{\text{incom}}(D) = 0.9509$

$$\text{Info}(D) = \frac{3}{5} I(1,1) + \frac{2}{5} I(1,1)$$

$$= \frac{3}{5}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right)\right]$$

$\therefore \text{Info}_{\text{student}}(D) = 0.9509$

$$\text{Info}_{\text{credit}}(D) = \frac{2}{5} I(0,2) + \frac{3}{5} I(3,0)$$

$$= \frac{2}{5}\left[-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right)\right] + \frac{3}{5}\left[-\frac{3}{5}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right]$$

$\therefore \text{Info}_{\text{credit}}(D) = 0$

Gain (income)	= 0.971 − 0.9509	= 0.0201
Gain (student)	= 0.9710 − 0.9509	= 0.0201
Gain (Credit_rating)	= 0.971 − 0	= 0.971