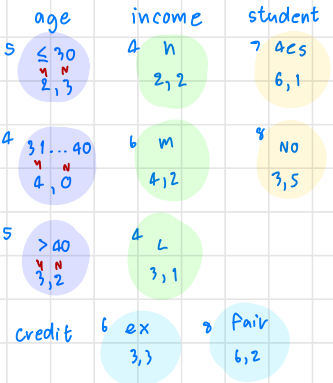
HW3 : Decision Tree



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

1 คำนวณ class

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

$= I(9, 5)$

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

$= 0.41 + 0.53$

$Info(D) = 0.940$

2 คำนวณ Feature $Info_A(D)$

สูตร $Info_A(D) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$

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

$= \frac{5}{14}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{14}\left[-\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$

$= \frac{5}{14}\left[0.5287 + 0.44217\right] + \left[\frac{4}{14} \times \text{ค่าไม่ได้}\right] + \frac{5}{14}\left[0.44217 + 0.52877\right]$

$= \left[\frac{5}{14} \times 0.9709\right] + \left[\frac{5}{14} \times 0.9709\right]$

$= 0.34676 + 0.34676 = 0.69352$

$Info_{age}(D) = 0.694$

นางสาวสิรภัทร ไชยมาตย์ SIDS 6430205527-6

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

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

$$= \frac{4}{14}[0.5 + 0.5] + \frac{6}{14}[0.3899 + 0.5283] + \frac{4}{14}[0.3112 + 0.5]$$

$$= \left[\frac{4}{14} \times 1\right] + \left[\frac{6}{14} \times 0.9182\right] + \left[\frac{4}{14} \times 0.8112\right]$$

$$= 0.2857 + 0.3935 + 0.2317 = 0.9109$$

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

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

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

$$= \frac{7}{14}[0.1906 + 0.4010] + \frac{7}{14}[0.5238 + 0.4613]$$

$$= \left[\frac{7}{14} \times 0.5916\right] + \left[\frac{7}{14} \times 0.9851\right]$$

$$= 0.2958 + 0.4925 = 0.78835$$

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

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

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

$$= \frac{6}{14}[0.5 + 0.5] + \frac{8}{14}[0.3113 + 0.5]$$

$$= \left[\frac{6}{14} \times 1\right] + \left[\frac{8}{14} \times 0.8113\right]$$

$$= 0.4285 + 0.4636 = 0.8921$$

$$\text{Info}_{\text{credit}}(D) = 0.892$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

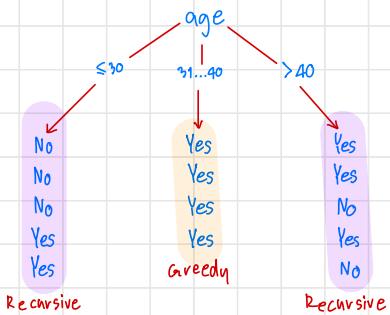
$\text{Gain}(\text{age}) = \text{Info}(D) - \text{Info}_{\text{age}}(D) = 0.940 - 0.694 = 0.246$ → Gain มากที่สุด เลือกเป็น root node

$\text{Gain}(\text{income}) = \text{Info}(D) - \text{Info}_{\text{income}}(D) = 0.940 - 0.911 = 0.029$

$\text{Gain}(\text{student}) = \text{Info}(D) - \text{Info}_{\text{student}}(D) = 0.940 - 0.788 = 0.152$

$\text{Gain}(\text{credit\_rating}) = \text{Info}(D) - \text{Info}_{\text{credit\_rating}}(D) = 0.940 - 0.892 = 0.048$

นางสาวสิรภัทร ไชยมาตย์ SIDS 6430205527-6



Recursive age ≤ 30

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~~

1 คำนวน Class

$$Info(D) = I(\overset{Y}{2}, \overset{N}{3})$$

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

$$= 0.5288 + 0.4422$$

$$Info(D) = 0.9710$$

2 คำนวน Feature Info(D)

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

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

$$= \left[\frac{2}{5} \times \text{หาค่าไม่ได้}\right] + \frac{2}{5}[0.5 + 0.5] + \left[\frac{1}{5} \times \text{หาค่าไม่ได้}\right]$$

$$= \frac{2}{5} \times 1$$

$$Info_{income}(D) = 0.4$$

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

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

$$= \left[\frac{2}{5} \times \text{หาค่าไม่ได้}\right] + \left[\frac{3}{5} \times \text{หาค่าไม่ได้}\right]$$

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

นางสาวสิรภัทร ไชยมาตย์ SIDS 6470205527-6

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

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

$$= \frac{2}{5}[0.5 + 0.5] + \frac{3}{5}[0.5283 + 0.38899]$$

$$= \left[\frac{2}{5} \times 1\right] + \left[\frac{3}{5} \times 0.9182\right]$$

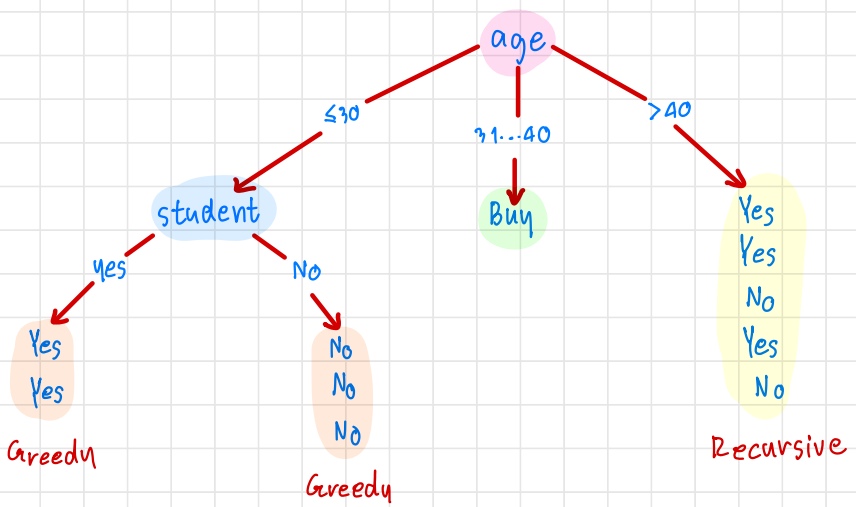
$$= 0.4 + 0.5509$$

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

Gain (income) = $Info(D) - Info_{income}(D) = 0.9710 - 0.4 = 0.5710$

Gain (student) = $Info(D) - Info_{student}(D) = 0.9710 - 0 = 0.9710$ → Gain มากที่สุด

Gain (credit_rating) = $Info(D) - Info_{credit\_rating}(D) = 0.9710 - 0.9509 = 0.0201$



นางสาวสิรลักษ์ ไชยมาตย์ SIDS 6470205527-6

Recursive age > 40

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

1 คำนวน class

$$Info(D) = I(\overset{Y}{3}, \overset{N}{2})$$

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

$$= 0.4422 + 0.5288$$

$$Info(D) = 0.971$$

2 คำนวน Feature $Info_A(D)$

$$Info_{income}(D) = \frac{3}{5}\overset{M}{I}(2,1) + \frac{2}{5}\overset{L}{I}(1,1)$$

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

$$= \frac{3}{5}[0.3899 + 0.5283] + \frac{2}{5}[0.5 + 0.5]$$

$$= \left[\frac{3}{5} \times 0.9182\right] + \left[\frac{2}{5} \times 1\right]$$

$$= 0.5509 + 0.4$$

$$Info_{income}(D) = 0.9509$$

$$Info_{student}(D) = \underbrace{\frac{3}{5} I(2,1)}_{Y} + \underbrace{\frac{2}{5} I(1,1)}_{N}$$

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

$$= \frac{3}{5}[0.3899 + 0.5283] + \frac{2}{5}[0.5 + 0.5]$$

$$= \left[\frac{3}{5} \times 0.9182\right] + \left[\frac{2}{5} \times 1\right]$$

$$= 0.5509 + 0.4$$

$$Info_{student}(D) = 0.9509$$

$$Info_{credit}(D) = \underbrace{\frac{2}{5} I(0,2)}_{ex} + \underbrace{\frac{3}{5} I(3,0)}_{fair}$$

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

$$= \left[\frac{2}{5} \times \text{หาค่าไม่ได้}\right] + \left[\frac{3}{5} \times \text{หาค่าไม่ได้}\right]$$

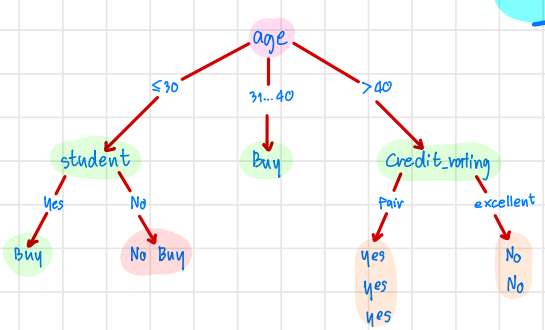
$$Info_{credit}(D) = 0$$

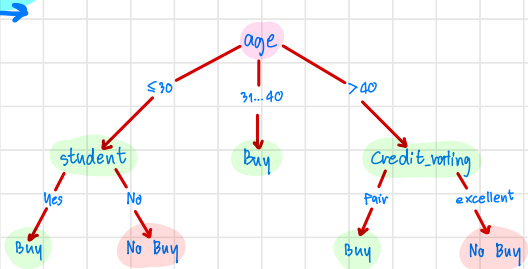
$$Gain(income) = Info(D) - Info_{income}(D) = 0.9710 - 0.9509 = 0.0201$$

$$Gain(student) = Info(D) - Info_{student}(D) = 0.9710 - 0.9509 = 0.0201$$

$$Gain(credit\_rating) = Info(D) - Info_{credit\_rating}(D) = 0.9710 - 0 = 0.9710$$ → Gain ที่มากที่สุด

Final →





นางสาวสิรภัทร ไชยมาตย์ SIDS 6430205527-6