Hw 5

นางสาว อภิสรา พงษ์เชี่ยงชา 633020027-2

- Class

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

= I(9,5)

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

= $-\frac{9}{14}\log_2\frac{9}{14}-\frac{5}{14}\log_2\frac{5}{14}$

= $-\frac{9}{14}(-0.637)-\frac{5}{14}(-1.485)$

= 0.940

- Feature

- $\text{Info}_{age}(D)$ = $\sum_{j=1}^{v}\left|\frac{D_j}{D}\right| \times \text{Info}(D_j)$

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

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

= $\frac{5}{14}(0.529+0.442)+\frac{4}{14}(0)+\frac{5}{14}(0.442+0.529)$

= $\frac{5}{14}(0.971)+$ผลลัพธ์ไม่ได้$+\frac{5}{14}(0.971)$

= 0.347 + 0.347 + 0.694

Gain(age) = Info(D) - $\text{Info}_{age}(D)$

= 0.940 - 0.694 = 0.291

- $\text{Info}_{income}(D)$ = $\sum_{j=1}^{v}\left|\frac{D_j}{D}\right| \times \text{Info}(D_j)$

= $\frac{4}{14}I(2,2)+\frac{6}{14}I(4,2)+\frac{4}{14}I(3,1)$

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

= $\frac{4}{14}+\frac{6}{7}(0.918)+\frac{4}{14}(0.811)$

= 0.286 + 0.394 + 0.232 = 0.912

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

= 0.940 - 0.912 = 0.028

- $\text{Info}_{student}(D)$ = $\sum_{j=1}^{v}\left|\frac{D_j}{D}\right| \times \text{Info}(D_j)$

= $\frac{7}{14}I(3,4)+\frac{7}{14}(6,1)$

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

= $\frac{7}{14}(0.524+0.461)+\frac{7}{14}(0.191+0.401)$

= $\frac{7}{14}(0.985)+\frac{7}{14}(0.592)$

= 0.493 + 0.296 = 0.789

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

= 0.940 - 0.893 = 0.047

- $\text{Info}_{credit}(D)$ = $\sum_{j=1}^{v}\left|\frac{D_j}{D}\right| \times \text{Info}(D_j)$

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

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

= $\frac{8}{14}(0.311+0.5)+\frac{6}{14}(0.5+0.5)$

= $\frac{8}{14}(0.811)+\frac{6}{14}$

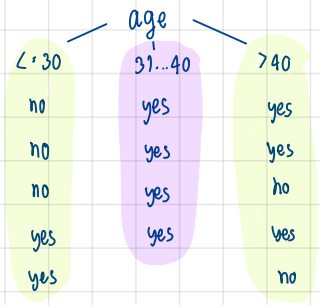
= 0.464 + 0.429 = 0.893

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

= 0.940 - 0.893 = 0.047

∴ เลือก Gain(age) เพราะมีค่าเยอะที่สุด

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 Greedy Recursive

F_1 age < = 30

age	income	student	credit	buys
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

$Info(D) = \sum_{i=1}^{n} p_i \log_2(p_i)$

$= I(2,3)$

$= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$

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

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

$= 0.4$

$Gain_{income} = Info(D) - Info_{income}(D)$

$= 0.971 - 0.4 = 0.571$

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

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

$= 0$

$Gain_{student} = Info(D) - Info_{student}(D)$

$= 0.971 - 0 = 0.971$

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

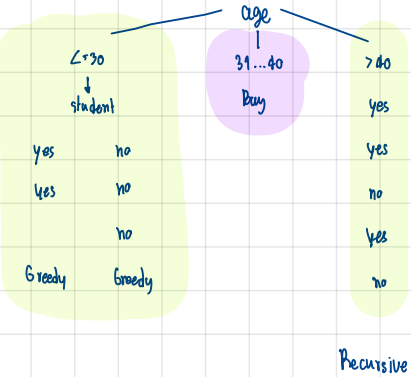
$= \frac{3}{5}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$

$= 0.551 + 0.4 = 0.951$

$Gain_{credit} = Info(D) - Info_{credit}(D)$

$= 0.971 - 0.951 = 0.020$

∴ Gain income > Gain student



Recursive

F_2 age >40

age	income	student	credit	buys
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

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

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

$= 0.971$

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

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

$= 0.551 + 0.4 = 0.951$

$Gain_{income} = Info(D) - Info_{income}(D)$

$= 0.971 - 0.951 = 0.02$ *

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

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

$= 0.4 + 0.551 = 0.951$

$Gain_{student} = Info(D) - Info_{student}(D)$

$= 0.971 - 0.951 = 0.02$ *

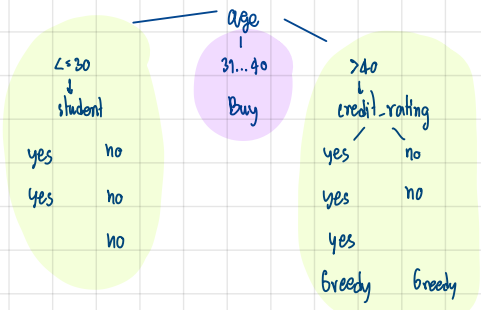
$Info_{credit}(D) = \frac{3}{5} I(3,0) + \frac{2}{5} I(1,1)$ (unaffected)

$= \frac{3}{5}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$

$= 0.4$

$Gain_{credit} = Info(D) - Info_{credit}(D)$

$= 0.971 - 0.4 = 0.571$ *



Decision Tree Induction