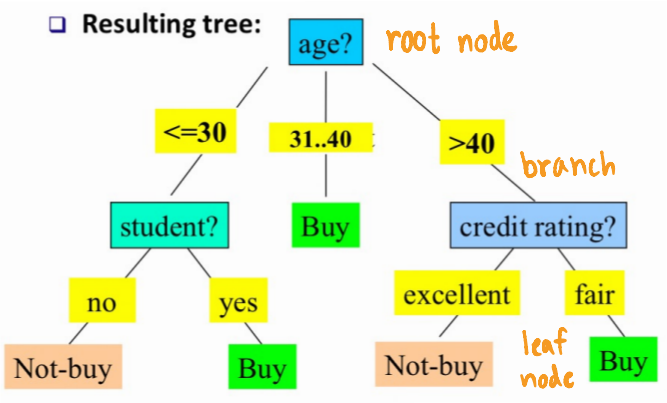
Decision Tree Induction – HW5



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

Class P: buys_computer = yes
Class N: buys_computer = no

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

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.904$$

(yes: 9, no: 5)

2. $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

(<=30, 31…40, >40)

age	Pi	Ni	I(Pi,Ni)
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

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

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

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

$$= 0.911$$

income	Pi	Ni	I(Pi,Ni)
high	2	2	1
median	4	2	0.918
low	3	1	0.811

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

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

$$= 0.789$$

student	Pi	Ni	I(Pi,Ni)
yes	6	1	0.592
no	3	4	0.985

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

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

$$= 0.892$$

credit_r	Pi	Ni	I(Pi,Ni)
fair	6	2	0.811
excellent	3	3	1

3. $Gain(A) = Info(D) - Info_A(D)$

$Gain(age) = 0.940 - 0.694 = 0.246$ ค่ามากสุด (root node)

$Gain(income) = 0.940 - 0.911 = 0.029$

$Gain(student) = 0.940 - 0.789 = 0.151$

$Gain(credit\_rating) = 0.940 - 0.892 = 0.048$

4. แยกกลุ่ม feature ตามค่าใน ราก feature age

<= 30

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

$Info_{income}(D) = \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0) = 0.4$

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

$Info_{credit\_r}(D) = \frac{3}{5}I(1,2) + \frac{2}{5}I(1,1) = 0.951$

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

31...40

yes, no

$Info(D) = I(4,0)$

ถ้าอายุ 31... 40 สามารถตอบ yes ใน buys_computer ได้เลย

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

> 40

$Info(D) = I(3,2) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$

$Info_{income}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = 0.951$ (median, low)

$Info_{student}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}(1,1) = 0.951$

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

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

Information Gain = ?

$Gain(Income) = 0.971 - 0.4 = 0.571$

$Gain(student) = 0.971 - 0 = 0.971$

$Gain(credit\_rating) = 0.971 - 0.951 = 0.02$

node ในอายุ <= 30

$Gain(Income) = 0.971 - 0.951 = 0.02$

$Gain(student) = 0.971 - 0.951 = 0.02$

$Gain(credit\_rating) = 0.971 - 0 = 0.971$

node ในอายุ >40

จะได้ Decision ดังนี้