การบ้าน Decision Tree

Information Gain: An Attribute Selection Measure

- Select the attribute with the highest information gain (used in typical decision tree induction algorithm: ID3/C4.5)
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

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

- Information needed (after using A to split D into v partitions) to classify D:

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

- Information gained by branching on attribute A

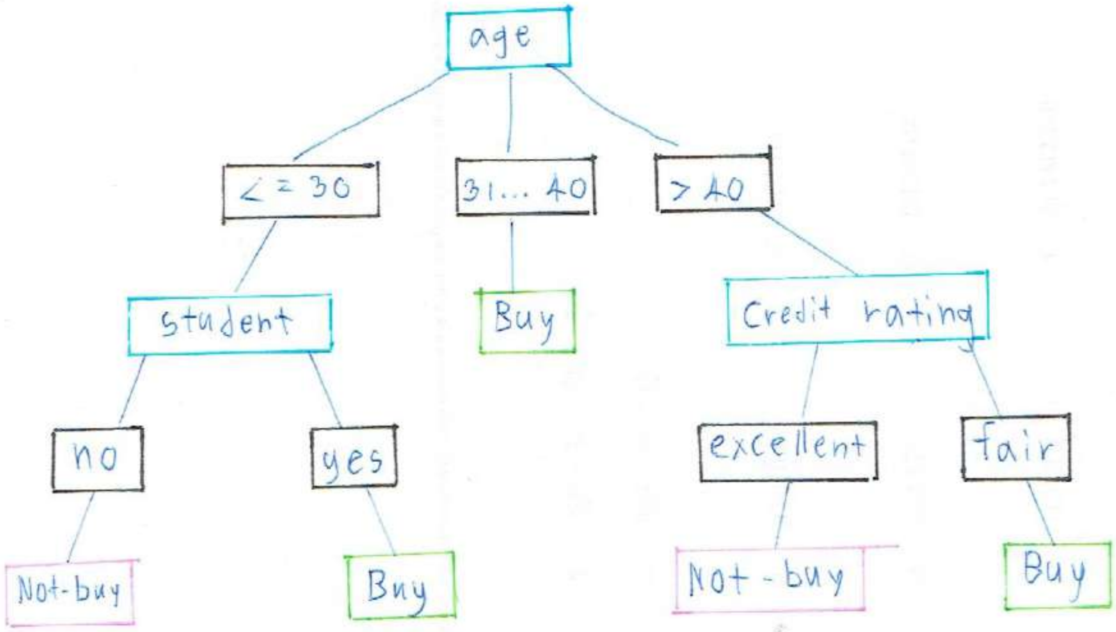
$$Gain(A) = Info(D) - Info_A(D)$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

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

Resulting tree :

นายเกียรติศักดิ์ หลอจินทร์ 623021041-9



หา Info (D)

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

(Y, N)

หา $Info_{age}(D)$

(<=30, 31-40, >40)

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

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

$$I(4,0) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - 0 = 0$$

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

แทนค่า $Info_{age}(D) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971)$

$= 0.694$

หา Gain (age)

$Gain(age) = Info(D) - Info_{age}(D)$

แทนค่า $Gain(age) = 0.94 - 0.694 = 0.246$

หา $Info_{income}(D)$

(high, medium, low)

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

(Y, N)

$I(2,2) = 1$

$I(4,2) = 0.918$

$I(3,1) = 0.811$

แทนค่า $Info_{income}(D) = \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811)$

$= 0.9108$

หา Gain (income)

$Gain_{income} = 0.94 - 0.9108$

$= 0.02918$

หา Info student (D)

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

(Yes: 6, 1 — No: 3, 4)

$$I(6,1) = 0.592$$

$$I(3,4) = 0.985$$

แทนค่า $Info_{student}(D) = \frac{7}{14}(0.592) + \frac{7}{14}(0.985) = 0.789$

หา Gain (student)

$$Gain(student) = 0.94 - 0.789 = 0.151$$

หา Info credit-rating (D)

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

(fair: 6, 2 — excellent: 3, 3)

แทนค่า $Info_{credit\text{-}rating}(D) = \frac{8}{14}(0.8113) + \frac{6}{14}(1) = 0.892$

หา Gain Credit-rating

$$Gain(creditrating = 0.99 - 0892 = 0.098$$

ได้ค่า Gain ดังนี้

Gain (age) = 0.246

Gain (income) = 0.029

Gain (student) = 0.151

Gain (credit-rating = 0.048

ที่เลือก Gain ที่มีค่ามากที่สุดมาเป็นคุณลักษณะแรก ได้ Gain (age)

age (<=30)

หา Info(D) ของ age (<=30)

$Info(D) = I(2,3) = 0.971$

หา Info income (D) ของ age

Hight Median low

Info income ของ age (<=30) $= \frac{2}{5} I(0,2) + \frac{2}{5} I(1,1) + \frac{1}{5} I(1,0)$

$I(0,2) = 0$

$I(1,1) = 1$

$I(1,0) = 0$

แทนค่า info income (D) ของ age (<=30)

$= \frac{2}{5}(0) + \frac{2}{5}(1) + \frac{1}{5}(0)$

$= 0.4$

หา Gain (income ของ age (<=30)

$= 0.971 - 0.4 = 0.571$

หา Info student (D) ของ age (<=30)

Info student (D) ของ age (<=30) $= \frac{2}{5} I(2,0) + \frac{3}{5} I(0,3)$

$I(2,0) = 0$ $I(0,3) = 0$

สิ้นสุด yes → yes (buy-computer), No → No (buy-computer)

แบ่งด้วย student เพราะสามารถแบ่งข้อมูลได้สมบูรณ์

age (>40)

หา Info(D) ของ age (>40)

Info (D) ของ age (>40) = $I(3,2)$ = 0.971

หา Info income (D) ของ age (>40)

Info income (D) ของ age (>40) = $\frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$ (Medium, low)

$I(2,1) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})$ = 0.918

$I(1,1)$ = 1

แทนค่า $Info_{income}$ (D) ของ age (>40)

$= \frac{3}{5}(0.918) + \frac{2}{5}(1)$ = 0.951

หา Gain income ของ age (>40)

Gain income ของ age (>40) = 0.971 − 0.951 = 0.02

หา $Info_{student}$ (D) ของ age (>40)

$Info_{student}$ (D) ของ age (>40) = $\frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$

$I(2,1) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})$ = 0.918

$I(1,1)$ = 1

แทนค่า $Info_{student}$ (D) ของ age (>40)

$= \frac{3}{5}(0.918) + \frac{2}{5}(1)$

= 0.951

หา Gain (student) ของ age (>40)

Gain (student) age (>40) = 0.971 − 0.951

= 0.02

ขั้น $Info_{credit\_rating}(D)$ ของ age (>40)

$Info_{credit\_rating}(D)$ ของ age (>40) = $\frac{3}{5} I(3,0) + \frac{2}{5} I(0,2)$ (fair, Excellent)

$I(3,0) = 0$ $I(0,2) = 0$

สิ้นสุด fair → yes(buy_computer), excellent → no (buy_computer)

เลือกแบ่งด้วย credit_rating ซึ่งจะได้แบ่งข้อมูลได้ถูกต้องทั้งหมด

สรุป