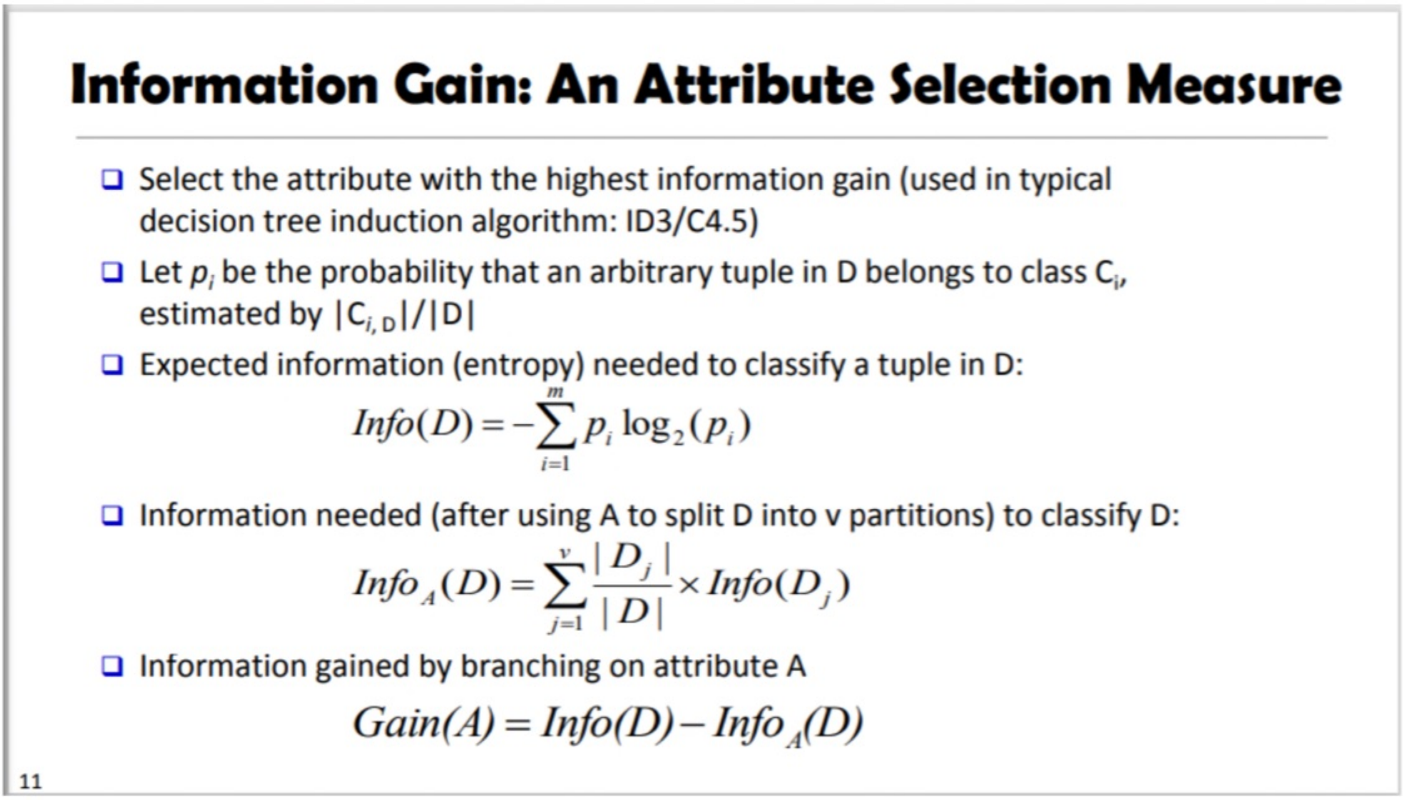


Information Gain: An Attribute Selection Measure

- ❑ 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:

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

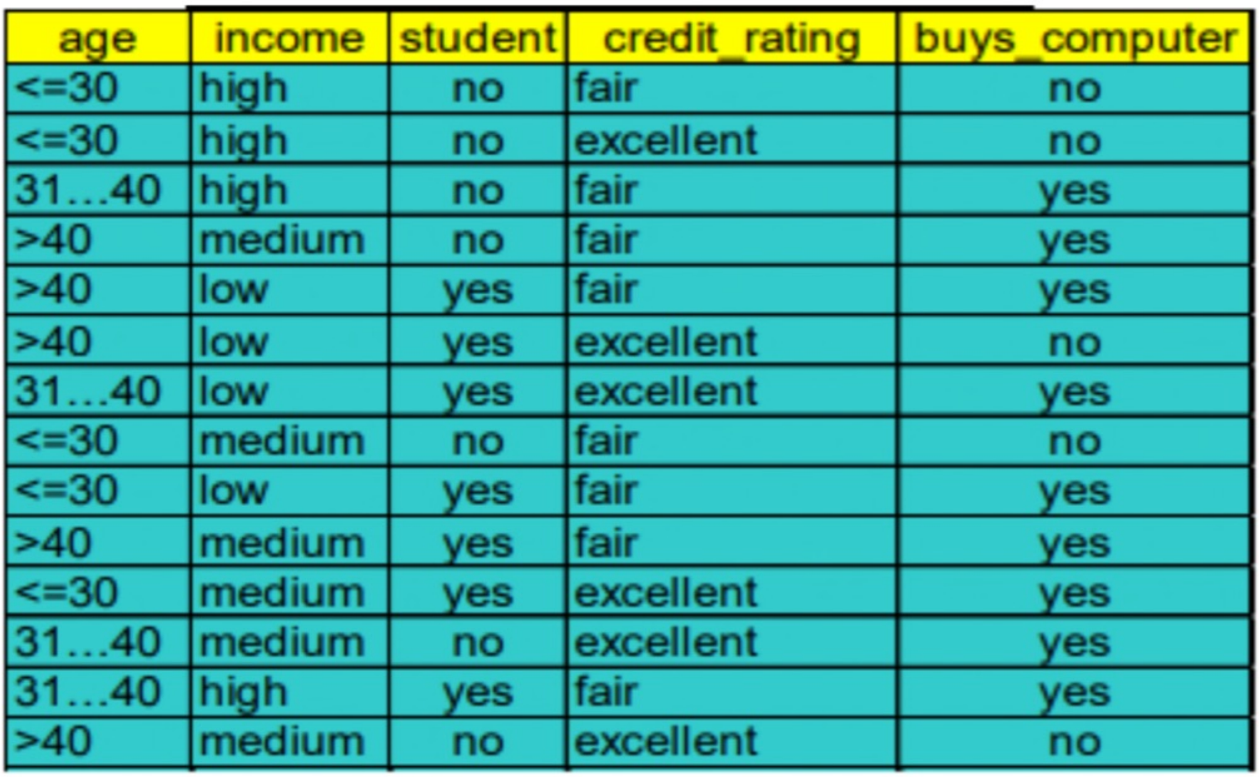
- ❑ Information needed (after using A to split D into v partitions) to classify D:

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

- ❑ Information gained by branching on attribute A

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

Example:



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

- หา Info (D)

$$\text{Info}(D) = I(\overset{y}{9}.\overset{n}{5}) = \frac{9}{4}\overset{y}{\log_{(2)}}\left(\frac{9}{14}\right) - \frac{5}{14}\overset{n}{\log_{(2)}}\left(\frac{5}{14}\right)$$

$$= 0.94$$

- หา $\text{Info}_{age}(D)$

$$\text{Info}_{age}(D) = \frac{5}{14}\overset{<=30}{I(2,3)} + \frac{4}{14}\overset{31-40}{I(4,0)} + \frac{5}{14}\overset{>40}{I(3,2)}$$

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

$$I(4,0) = -\frac{4}{4}\log_{(2)}\left(\frac{4}{4}\right) - \frac{0}{4}\log_{(2)}\left(\frac{0}{4}\right) = 0$$

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

แทนค่า $\text{Info}_{age}(D) = \frac{5}{4}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.644$

- หา Gain (age)

$$\text{Gain}(age) = 0.94 - 0.694$$

$$= 0.246$$

- หา $\text{Info}_{income}(D)$

$$\text{Info}_{income}(D) = \frac{4}{14}\overset{high}{I(\overset{y}{2},\overset{n}{2})} + \frac{6}{14}\overset{medium}{I(\overset{y}{4},\overset{n}{2})} + \frac{4}{14}\overset{low}{I(\overset{y}{3},\overset{n}{1})}$$

$$I(\overset{y}{2},\overset{n}{2}) = -\frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) - \frac{2}{4}\log_{(2)}\left(\frac{2}{4}\right) = 1$$

$$I(\overset{y}{4},\overset{n}{2}) = -\frac{4}{6}\log_{(2)}\left(\frac{4}{6}\right) - \frac{2}{6}\log_{(2)}\left(\frac{2}{6}\right) = 0.918$$

$$I(\overset{y}{3},\overset{n}{1}) = -\frac{3}{4}\log_{(2)}\left(\frac{3}{4}\right) - \frac{1}{4}\log_{(2)}\left(\frac{1}{4}\right) = 0.811$$

แทนค่า $\text{Info}_{income}(D) = \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.811) = 0.911$

- หา Gain (income)

$$\text{Gain}(income) = 0.94 - 0.911$$

$$= 0.029$$

- หา $\text{Info}_{student}(D)$

$$\text{Info}_{student}(D) = \frac{7}{14}\overset{yes}{I(\overset{y}{6},\overset{n}{1})} + \frac{7}{4}\overset{no}{I(\overset{y}{3},\overset{n}{4})}$$

$$I(6,1) = -\frac{6}{7}\log_{(2)}\left(\frac{6}{7}\right) - \frac{1}{7}\log_{(2)}\left(\frac{1}{7}\right) = 0.592$$

$$I(3,4) = -\frac{3}{7}\log_{(2)}\left(\frac{3}{7}\right) - \frac{4}{7}\log_{(2)}\left(\frac{4}{7}\right) = 0.985$$

แทนค่า $\text{Info}_{student}(D) = \frac{7}{14}(0.592) + \frac{7}{14}(0.985)$

- หา Gain (Student)

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

- หา $Info_{(redit\_rating)}(D)$

$$Info_{(redit\_rating)}(D) = \overset{\text{fair}}{\frac{8}{14}} I(\overset{y}{6},\overset{n}{2}) + \overset{\text{excellent}}{\frac{6}{14}} I(\overset{y}{3},\overset{n}{3})$$

$$I(6,2) = -\frac{6}{8}\log_{(2)}\left(\frac{6}{8}\right) - \frac{2}{8}\log_{(2)}\left(\frac{2}{8}\right) = 0.8111$$

$$I(3,3) = -\frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) - \frac{3}{6}\log_{(2)}\left(\frac{3}{6}\right) = 1$$

แทนค่า $Info_{credit\_rating}(D) = \frac{8}{14}(0.8111) + \frac{6}{14}(1) = 0.892$

- หา Gain(credit_rating)

$$Gain(credit\_rating) = 0.99 - 0.892$$
$$= 0.048$$

จาก Gain

$$Gain(age) = 0.246$$
$$Gain(income) = 0.029$$
$$Gain(Student) = 0.151$$
$$Gain(Credit\_rating) = 0.048$$

เลือก Gain ที่มีค่ามากที่สุดมาพิจารณาเป็นส่วนแรก ซึ่งในที่นี้คือ Gain (age)

age (<=30)

- หา Info (D) ของ age (<=30)

$$Info(D) = I(\overset{y}{2},\overset{n}{3}) = 0.971$$

- หา $Info_{income}(D)$ ของ age (<=30)

$$Info_{income}(D) \text{ ของ } age(<=30) = \overset{\text{high}}{\frac{2}{3}} I(\overset{y}{0},\overset{n}{2}) + \overset{\text{medium}}{\frac{2}{5}} I(\overset{y}{1},\overset{n}{1}) + \overset{\text{low}}{\frac{1}{5}} I(\overset{y}{1},\overset{n}{0})$$

$$I(0,2) = -\frac{0}{2}\log_{(2)}\left(\frac{0}{2}\right) - \frac{2}{2}\log_{(2)}\left(\frac{2}{2}\right) = 0$$

$$I(1,1) = -\frac{1}{2}\log_{(2)}\left(\frac{1}{2}\right) - \frac{1}{2}\log_{(2)}\left(\frac{1}{2}\right) = 1$$

$$I(1,0) = -\frac{1}{1}\log_{(2)}\left(\frac{1}{1}\right) - \frac{0}{1}\log_{(2)}\left(\frac{0}{1}\right) = 0$$

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)

Gain (income) ของ age (<=30) $= 0.971 - 0.4 = 0.571$

- หา $\text{Info}_{\text{student}}(D)$ ของ age $(<=30)$

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

สิ้นสุด yes → yes(buy_computer)
no → no (buy_ computer)

เลือกแบ่งด้วย Student เพราะสามารถแบ่งข้อมูลได้แบบสมบูรณ์

age (> 40)

$$\text{Info}(D) \text{ ของ age } (>40) = I(3,2) = 0.971$$

- หา $\text{Info}_{\text{income}}(D)$ ของ age $(>40) = \frac{3}{5} \overset{\text{medium}}{I(\overset{y}{2},\overset{n}{1})} + \frac{2}{5} \overset{\text{low}}{I(\overset{y}{1},\overset{n}{1})}$

$$I(2,1) = -\frac{2}{3}\log_{(2)}\left(\frac{2}{3}\right) + \frac{1}{3}\log_{(2)}\left(\frac{1}{3}\right) = 0.918$$

$$I(1,1) = 1$$

แทนค่า $\text{Info}_{\text{income}}(D)$ ของ age $(>40) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$

- หา Gain (incom) ของ age (> 40)

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

- หา $\text{Info}_{\text{student}}$ ของ age (> 40)

$$\text{Info}_{\text{student}} \text{ ของ age } (>40) = \frac{3}{5} \overset{\text{yes}}{I(\overset{y}{2},\overset{n}{1})} + \frac{2}{5} \overset{\text{no}}{I(\overset{y}{1},\overset{n}{1})}$$

$$I(2,1) = -\frac{2}{3}\log_{(2)}\left(\frac{2}{3}\right) - \frac{1}{3}\log_{(2)}\left(\frac{1}{3}\right) = 0.918$$

$$I(1,1) = 1$$

แทนค่า $\text{Info}_{\text{student}}$ ของ age $(>40) = \frac{3}{5}(0.918) + \frac{2}{5}(0.918) + \frac{2}{5}(1) = 0.951$

- หา Gain (Student) ของ age (>40)

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

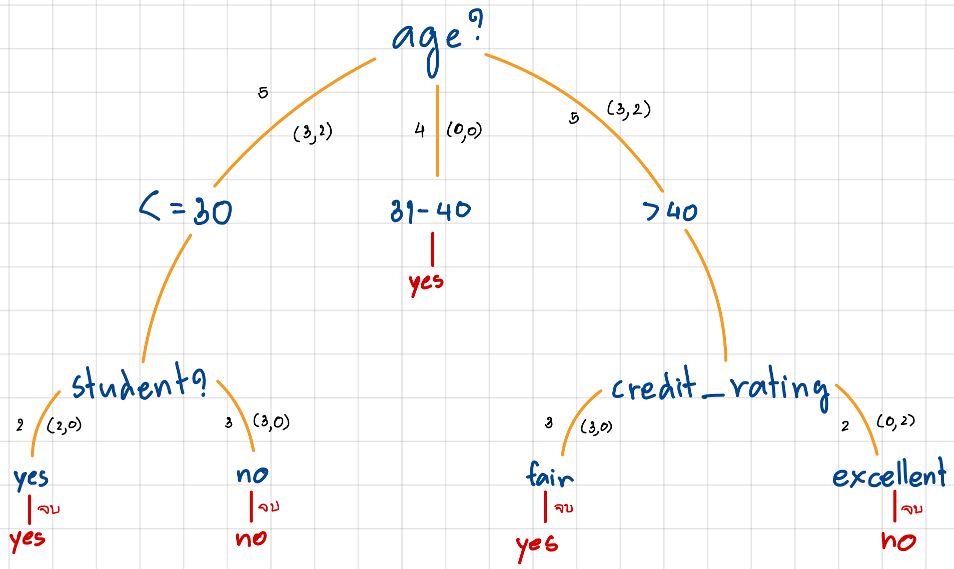
- หา $\text{Info}_{\text{credit\_rating}}(D)$ ของ age (>40)

$$\text{Info}_{\text{credit\_rating}}(D) \text{ ของ age } (>40) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2)$$

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

เลือกแบ่งด้วย Credit_rating เพราะสามารถแบ่งข้อมูลได้สมบูรณ์

สรุป


age?
5
(3,2)
4
(0,0)
5
(3,2)
<=30
31-40
>40
yes
student?
credit_rating
2
(2,0)
3
(3,0)
3
(3,0)
2
(0,2)
yes
no
fair
excellent
จบ
จบ
จบ
จบ
yes
no
yes
no