Decision Tree Induction

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

- Info (D)

$$\text{Info}(D) = I(\overset{Y}{9},\overset{N}{5}) = \frac{9}{14}\log_{(2)}\left(\frac{9}{14}\right) - \frac{5}{14}\log_{(2)}\left(\frac{5}{14}\right)$$

$$= 0.94$$

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

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

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

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

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

$$\text{minm } \text{Info}_{age}(D) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$$

- Gain (age)

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

$$= 0.246$$

- หา $Info_{income}(D)$

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

(high, medium, low)

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

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

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

เหมือน $Info_{income}(D) = \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.311) = 0.911$

- หา Gain (income)

$$Gain(income) = 0.94 - 0.911$$
$$= 0.029$$

- หา $Info_{student}(D)$

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

(Yes, NO)

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

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

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

$$= 0.789$$

- หา Gain (student)

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

- หา $Info_{credit\_rating}(D)$

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

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

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

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 ที่มีค่ามากที่สุดมาเป็นจากเป็นลำดับแรก ซึ่งในที่นี้คือ Gain (age)

age (<=30)

- หา Info (D) ของ age (<=30)

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

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

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

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

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

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

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 $Info_{student}(D)$ ของ age (<=30)

  $Info_{student}(D)$ ของ age (<=30) $= \frac{2}{5} I(\overset{yes}{2},\overset{NO}{0}) + \frac{3}{5} I(0,3)$

  สิ้นสุด Yes -> Yes (buy_Computer)
  No -> No (buy_Computer)

เลือกแบ่งด้วย Student เพราะสามารถแบ่งข้อมูลได้แบบสมบูรณ์

age (>40)

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

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

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

  $I(1,1) = 1$

แทนค่า $Info_{income}(D)$ 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) $= \frac{3}{5} I(\overset{yes}{2},1) + \frac{2}{5} I(\overset{NO}{1},1)$

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

  $I(1,1) = 1$

  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) $= 0.971 - 0.951 = 0.02$

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

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

  สิ้นสุด fair → Yes (buy_computer)
  excellent → No (buy_computer)

เลือกแบ่งด้วย Credit_rating เพราะสามารถแบ่งข้อมูลได้สมบูรณ์

สรุป