## 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

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

11

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)

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

$$= 0.94$$

หา $Info_{age}(D)$

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

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

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

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

แทนค่า $Info_{age}(D) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$

หา Gain (age)

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

## หา Info income (D)

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

(y, n)

$$I(2,2) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

$$I(4,2) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.918$$

$$I(3,1) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$$

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

## หา Gain (income)

$$Gain_{income} = 0.94 - 0.911 = 0.029$$

## หา Info student (D)

Yes NO

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

(y, n)

$$I(6,1) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) = 0.592$$

$$I(3,4) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.985$$

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

## หา Gain (student)

$$Gain_{student} = 0.94 - 0.789 = 0.151$$

หา Info credit_rating (D)

$$Info_{credit\_rating}(D) = \frac{8}{14} I(6,2) + \frac{1}{14} I(3,3)$$

(fair: 6 yes, 2 no; excellent: 3 yes, 3 no)

$$I(6,2) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0.811$$

$$I(3,3) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1$$

แทนค่า $Info_{credit\_rating}(D) = \frac{8}{14}(0.811) + \frac{6}{14}(1) = 0.892$

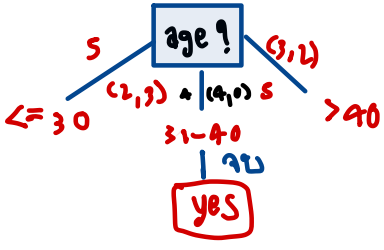
หา Gain (Credit_rating)

$$Gain(credit\_rating) = 0.94 - 0.892 = 0.048$$

จาก Gain

Gain (age) = 0.246
Gain (income) = 0.029
Gain (student) = 0.151
Gain (credit_rating) = 0.048

∴ เลือก Gain ที่มีค่ามากที่สุดมาเป็นตัวแบ่งแรก คือ Gain (age)



หา Info (D) ของ age (<=30)

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

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

$I(0,2) = -\frac{0}{2}\log_2(\frac{0}{2}) - \frac{2}{2}\log_2(\frac{2}{2}) = 0$

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

$I(1,0) = -\frac{1}{1}\log_2(\frac{1}{1}) - \frac{0}{1}\log_2(\frac{0}{1}) = 0$

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) $= \frac{2}{5} I(\overset{y}{2},\overset{n}{0}) + \frac{3}{5} I(\overset{y}{0},\overset{n}{3})$ (Yes, No)

สรุป yes → yes (buy_computer), No → No (buy_computer)

เลือกแบ่งด้วย student เพราะสามารถแบ่งข้อมูลได้สมบูรณ์

หา $Info(D)$ ของ age (>40) $= I(\overset{y}{3},\overset{n}{2}) = 0.971$

หา $Info_{income}(D)$ ของ age (>40) $= \frac{3}{5} I(\overset{y}{2},\overset{n}{1})$ (medium) $+ \frac{2}{5} I(\overset{y}{1},\overset{n}{1})$ (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 $Gain(Student)$ ของ age (>40) $= 0.971 - 0.951 = 0.02$

หา $Info_{credit\_rating}(D)$ ของ age (>40) $= \frac{5}{5} I(\overset{fair}{3},0) - \frac{5}{5} I(0,\overset{excellent}{2})$

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

เลือกแบ่งด้วย credit_rating

"สรุป"