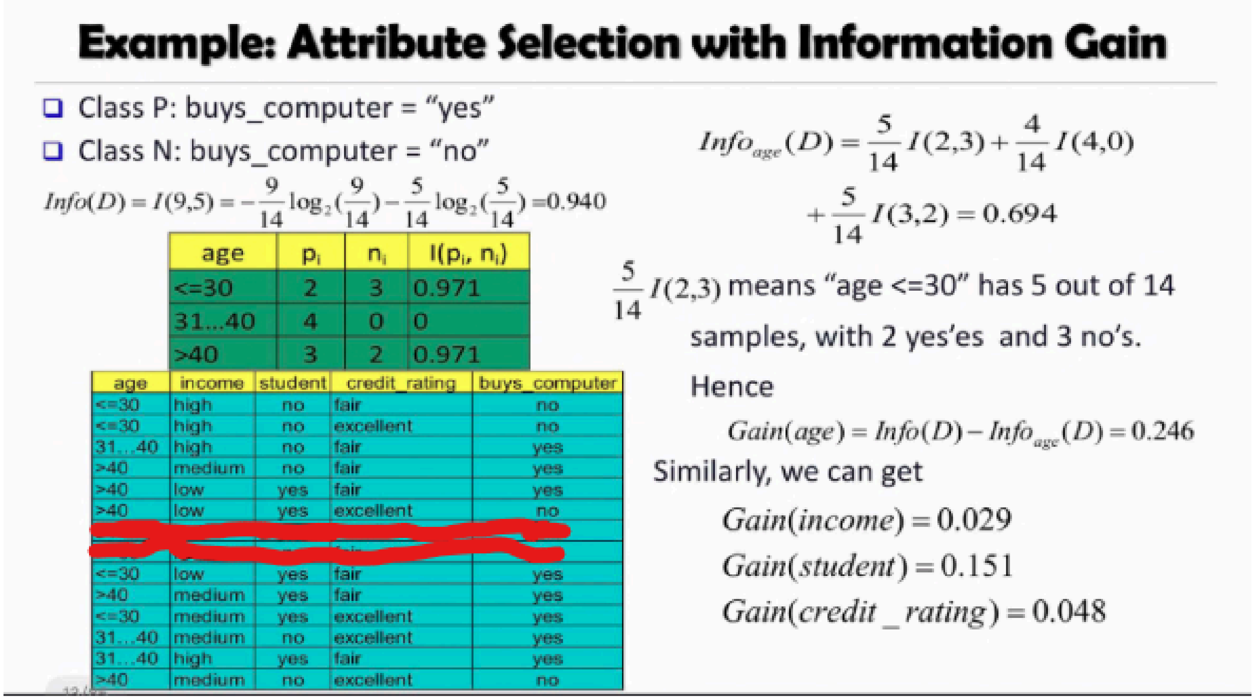

Example: Attribute Selection with Information Gain

- ❑ Class P: buys_computer = "yes"
- ❑ Class N: buys_computer = "no"

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

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
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
<=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

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

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly, we can get

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

1. คำนวณ Info(D)

Info(D) = I (8,4)

= - 8/12 $\log_2$ (8/12) - 4/12 $\log_2$ (4/12)

= 0.9183

2. คำนวณ Info$_{age\ ,income\ ,student,\ credit}$ (D)

2.1. Info$_{age}$ (D) = 4/12 I (2,2) + 3/12 I (3,0) + 5/12 I (3,2)

= 4/12 (1) + 3/12 (0) + 5/12 (0.9710)

Info$_{age}$ (D) = 0.5761

2.2. Info$_{income}$ (D) = 4/12 I (2,2) + 5/12 I (4,1) + 3/12 I (2,1)

= 4/12 I (1) + 5/12 I (0.7219) + 3/12 I (0.9183)

= 0.8637

2.3. Info$_{student}$ (D) = 6/12 I (5,1) + 6/12 I (3,3)

= 6/12 I (0.6500) + 6/12 I (1)

= 0.8250

2.4. Info$_{credit}$ (D) = 7/12 I (6,1) + 5/12 I (2,3)

= 7/12 I (0.6906) + 5/12 I (0.9710)

= 0.8074

3. คำนวณหา Gain

3.1 Gain (*age*) = 0.9183 - 0.5761
= 0.3422

3.2 Gain (*income*) = 0.9183 - 0.8637
= 0.0546

3.3 Gain (*student*) = 0.9183 - 0.8250
= 0.0933

3.4 Gain (*credit*) = 0.9183 - 0.8074
= 0.1109

สรุป ได้ค่า Gain (*age*) = 0.3422 เป็นค่าที่มากที่สุดจึงให้ age เป็น root note