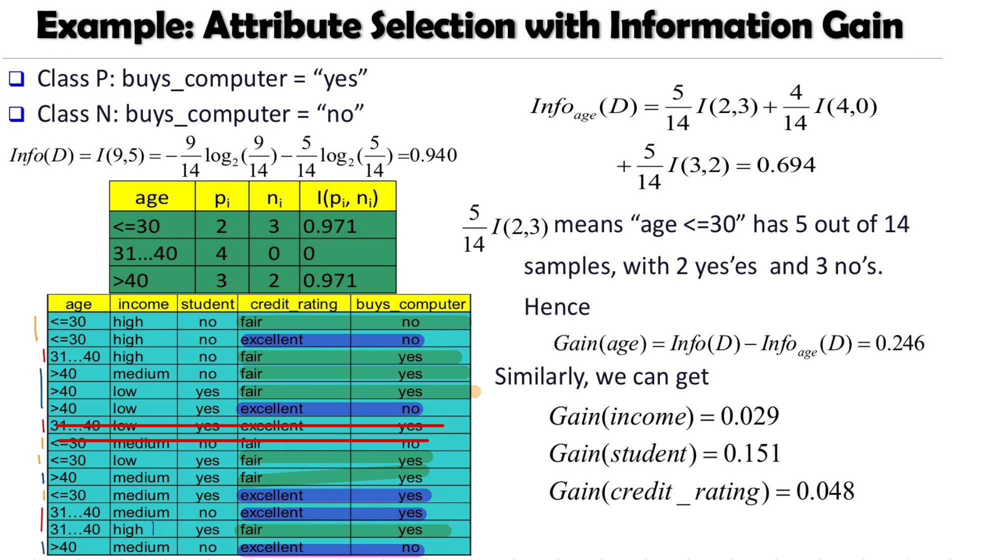

Example: Attribute Selection with Information Gain

- ❑ 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
~~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

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

$\frac{5}{14}I(2,3)$ 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

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

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

age	p_i	n_i	$I(p_i, n_i)$
<= 30	2	2	1
31… 40	3	0	0
> 40	3	2	0.971

$$Info(D) = I(8,4) = -\frac{8}{12}\log_2(\frac{8}{12}) - \frac{4}{12}\log_2(\frac{4}{12}) = 0.918$$

$$Info_{age}(D) = \frac{4}{12}i(2,2) + \frac{3}{12}i(3,0) + \frac{5}{12}i(3,2)$$

$$= \frac{4}{12}\left[-\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4})\right] + \frac{3}{12}\left[-\frac{3}{3}\log_2(\frac{3}{3}) - \frac{0}{3}\log_2(\frac{0}{3})\right] + \frac{5}{12}\left[-\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5})\right]$$

$$= \frac{4}{12}(1) + \frac{3}{12}(0) + \frac{5}{12}(0.971)$$

$$= 0.738$$

income	P_i	n_i	$I(p_i, n_i)$
h	2	2	1
me	4	1	0. 722
l	2	1	0. 918

$$Info_{income}(D) = \frac{4}{12}i(2,2) + \frac{5}{12}i(4,1) + \frac{3}{12}i(2,1)$$

$$= \frac{4}{12}(1) + \frac{5}{12}\left[-\frac{4}{5}\log_2(\frac{4}{5}) - \frac{1}{5}\log_2(\frac{1}{5})\right] + \frac{3}{12}\left[-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})\right]$$

$$= \frac{4}{12} + \frac{5}{12}(0.722) + \frac{3}{12}(0.918)$$

$$= 0.864$$

Student	P_i	n_i	$i(p_i\ n_i)$
Y	3	3	1
n	5	1	

$$Info_{student}(D) = \frac{6}{12}i(3,3) + \frac{6}{12}i(5,1)$$

$$= \frac{6}{12}(1) + \frac{6}{12}\left[-\frac{5}{6}\log_2(\frac{5}{6}) - \frac{1}{6}\log_2(\frac{1}{6})\right]$$

$$= \frac{6}{12} + \frac{6}{12}(0.650)$$

$$= 0.825$$

Credit	P_i	n_i	$i(p_i\ n_i)$
f	6	1	0.592
e	2	3	0.971

$$Info_{cadit}(D) = \frac{7}{12}i(6,1) + \frac{5}{12}i(2,3)$$

$$= \frac{7}{12}\left[-\frac{6}{7}\log_2(\frac{6}{7}) - \frac{1}{7}\log_2(\frac{1}{7})\right] + \frac{5}{12}\left[-\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5})\right]$$

$$= \frac{7}{12}(0.592) + \frac{5}{12}(0.971)$$

$$= 0.750$$

Gain(age) = 0.918 - 0.738 = 0.18

Gain(income) = 0.918 - 0.864 = 0.054

Gain(student) = 0.918 - 0.825 = 0.093

Gain (credit) = 0.918 - 0.750 ≈ 0.168

* Gain (Age) มีค่ามากที่สุด จึงได้ Age เป็น root node