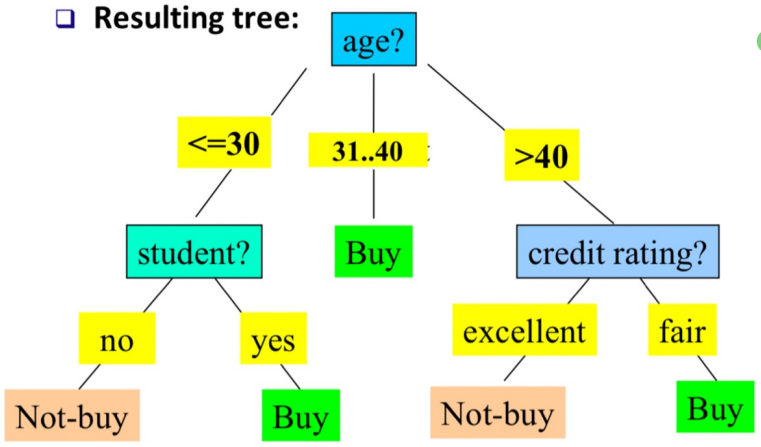
HW 5

Decision Tree Induction: An Example

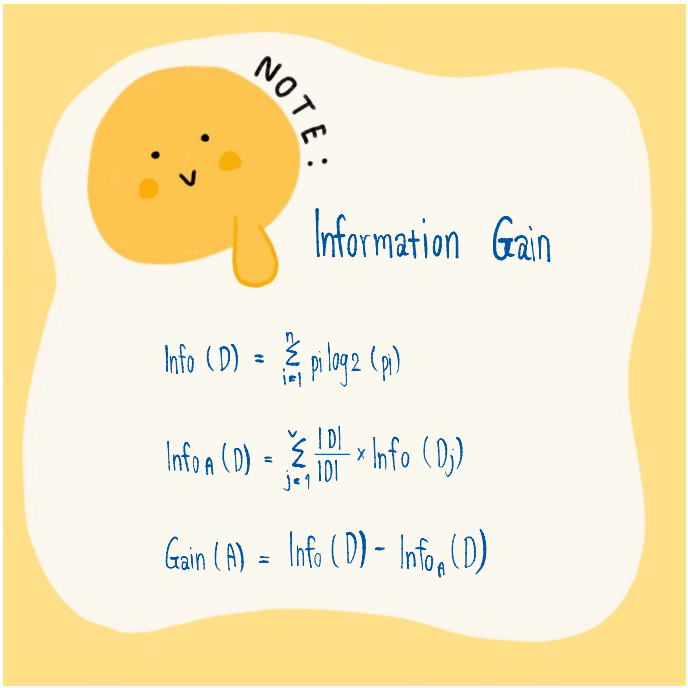
- **Decision tree construction**:
  - A top-down, recursive, divide-and-conquer process
- **Resulting tree:**



Training data set: Who buys computer?

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

Note: The data set is adapted from "Playing Tennis" example of R. Quinlan



Class P : buys_computer = "yes" → 9

Class N : buys_computer = "No" → 5

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

$= 0.940$

age	p_i	n_i
<= 30	2	3
31 … 40	4	0
>40	3	2

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

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

$= 0.940 - 0.694 = 0.246$

Gain (age) มีค่ามากที่สุด

income	p_i	n_i
high	2	2
medium	4	2
low	3	1

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

$$= \frac{4}{14}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right]$$

$$+ \frac{4}{14}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] = 0.911$$

$$Gain_{(income)} = Info(D) - Info_{income}(D)$$

$$= 0.940 - 0.911 = 0.029$$

Student	p_i	n_i
yes	6	1
no	3	4

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

$$= \frac{7}{14}\left[-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right] + \frac{7}{14}\left[-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right] = 0.789$$

$$Gain_{(student)} = Info(D) - Info_{student}(D)$$

$$= 0.940 - 0.789 = 0.151$$

---

credit_rating	p_i	n_i
fair	6	2
excellent	3	3

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

$$= \frac{8}{14}\left[-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right] + \frac{6}{14}\left[-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right] = 0.892$$

$$Gain_{(credit\_rating)} = Info(D) - Info_{credit\_rating}(D)$$

$$= 0.940 - 0.892 = 0.048$$

เรียงตามลำดับมากไปหาน้อย

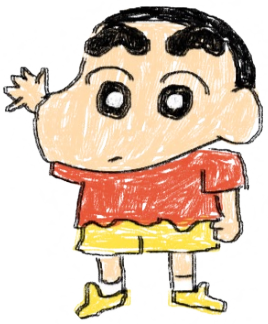
Grain (age) = 0.246 มีค่าสูงสุดจึงเป็น root node

Grain (student) = 0.151

Grain (credit_rating) = 0.048

Grain (income) = 0.029

Feature age



<= 30

$Info(D) = I(2,3) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}) = 0.971$

yes : 2

No : 3

$Info_{income}(D) = \frac{2}{5}I(0,2) + \frac{2}{5}I(1,1) + \frac{1}{5}I(1,0) = 0.4$

high : 2

medium : 1

low : 2

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

yes : 2

No : 3

$Info_{credit\_rating}(D) = \frac{3}{5}I(1,2) + \frac{2}{5}I(1,1) = 0.951$

fair : 3

excellent : 2

Information Grain

Grain (Income) = 0.971 - 0.4 = 0.571

Grain (Student) = 0.971 - 0 = 0.971 → node ไหลย, <= 30

Grain (Credit_rating) = 0.971 - 0.951 = 0.02

31 ... 40

yes : 4 ถ้าลย, 31...40 จะเป็น yes หมดเลยใน buys_Computer

No : 0

> 40

$Info(D) = I(3,2) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$

yes : 3

No : 2

$Info_{income}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = 0.951$

medium : 3

low : 2

$Info_{student}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = 0.951$

yes : 3

No : 2

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

fair : 3

excellent : 2

Information Grain

Grain (Income) = 0.971 - 0.951 = 0.02

Grain (Student) = 0.971 - 0.951 = 0.02

Grain (Credit_rating) = 0.971 - 0 = 0.971 → node ในสาย > 40

จะได้ Decision Tree ดังนี้