Decision Tree

การสร้างต้นไม้ตัดสินใจ (Decision Tree) การที่เลือก node ที่ดีที่สุดจาก Feature ในการแยก class จะหาจาก "Information Gain" ที่สูงที่สุดจากสูตรของ "Entropy" ซึ่งมีสูตรคำนวณ

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$ (Gain ของ Class) ; P_i คือ ความน่าจะเป็น

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$ (Gain ของ Feature)

$$Gain(A) = Info(D) - Info_A(D)$$ (Information Gain จากการแยก Feature แต่ละตัว)

Example

Data

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

Feature | Class

1. คำนวณหา Gain ของ Class จากคอลัมน์ buys_computer ประกอบไปด้วย Class "Yes" กับ "No"

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

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

(9 = Yes, 5 = No)

$$= 0.940$$

Prob of Class (from values in feature)

Prob of Feature (in values)

2. คำนวณหา Gain ของ Feature แต่ละตัวซึ่งประกอบไปด้วย $\left(Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)\right)$

ซึ่งจะแบ่งตามความน่าจะเป็นของแต่ละ Feature และ Class

2.1) Gain ของ age

age ≤ 30 (Yes No) | age 31...40 (Yes No) | age > 40 (Yes No)

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

$$= \frac{5}{14}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{14}\left[-\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

Yes (2/5) | No (3/5)

$$= 0.694$$

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

Feature | Class

2.2) Gain ของ income

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

$$= \frac{4}{14}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] + \frac{6}{14}\left[-\frac{4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right] + \frac{4}{14}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$$

$$= 0.911$$

2.3 Gain ของ student

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

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

$$= 0.789$$

2.4 Gain ของ credit_rating

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

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

$$= 0.892$$

Gain of Class ↓ Gain of Feature ↓

3. คำนวณ Information Gain ของแต่ละ Feature ($Gain(A) = Info(D) - Info_A(D)$)

โดยจะเลือก Gain ที่สูงที่สุดให้เป็น Root Node

3.1 $Gain(age) = 0.940 - 0.911 = 0.246$ Gain สูงที่สุดให้เป็น Root Node

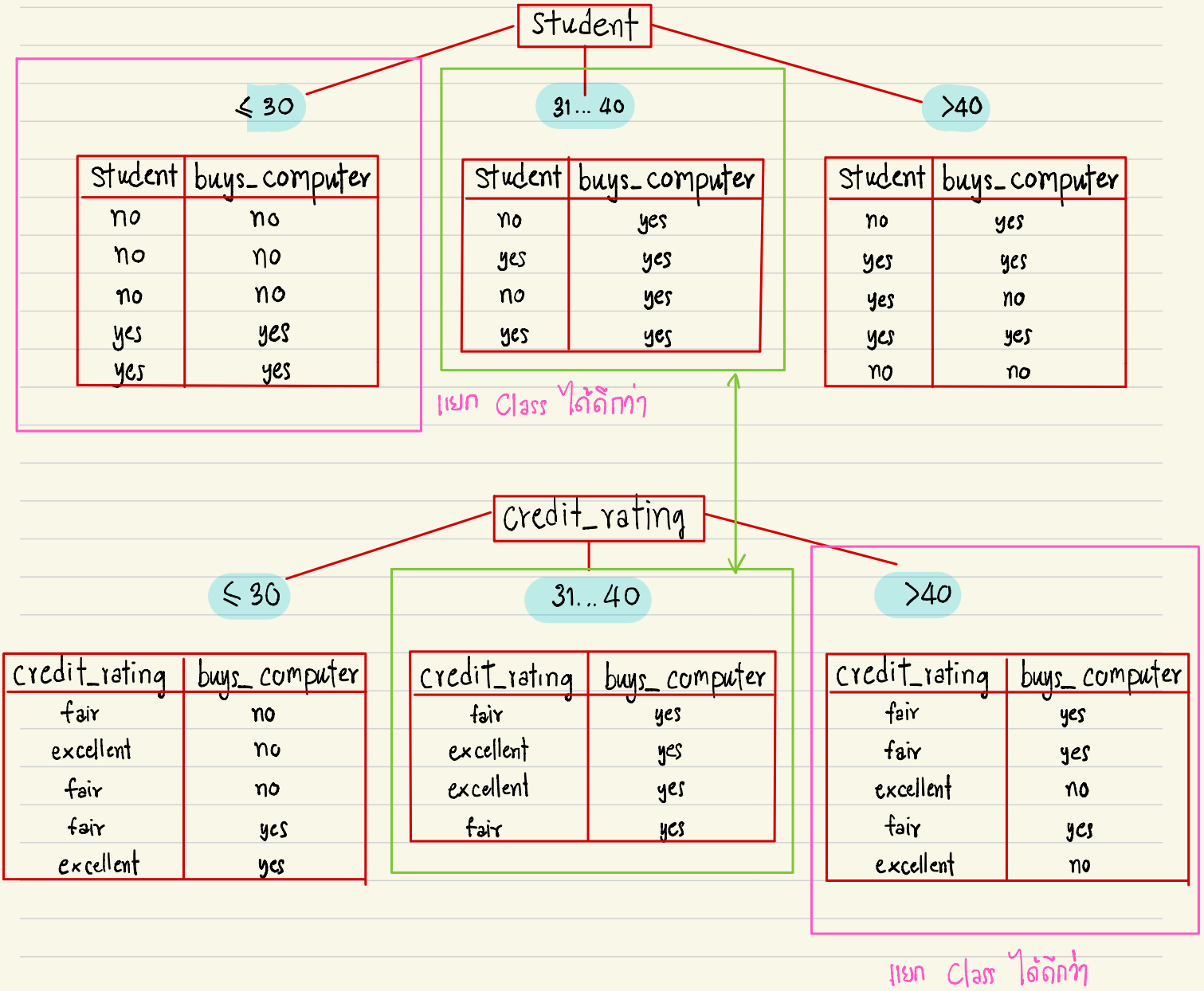
3.2 $Gain(income) = 0.940 - 0.911 = 0.029$ ต่ำสุดตัดทิ้ง เพราะ: Decision Tree อาจไม่มีประสิทธิภาพ

3.3 $Gain(student) = 0.940 - 0.789 = 0.151$

3.4 $Gain(credit\_rating) = 0.940 - 0.892 = 0.048$

Decision Tree

1) แยกกลุ่มของ Feature ตามค่าใน Root Node



Student	buys_computer
no	no
no	no
no	no
yes	yes
yes	yes

Student	buys_computer
no	yes
yes	yes
no	yes
yes	yes

Student	buys_computer
no	yes
yes	yes
yes	no
yes	yes
no	no

credit_rating	buys_computer
fair	no
excellent	no
fair	no
fair	yes
excellent	yes

credit_rating	buys_computer
fair	yes
excellent	yes
excellent	yes
fair	yes

credit_rating	buys_computer
fair	yes
fair	yes
excellent	no
fair	yes
excellent	no

2. แยก Values ออกมาแบ่ง Class ของ Feature และเขียนเป็น Decision Tree ได้ดังนี้