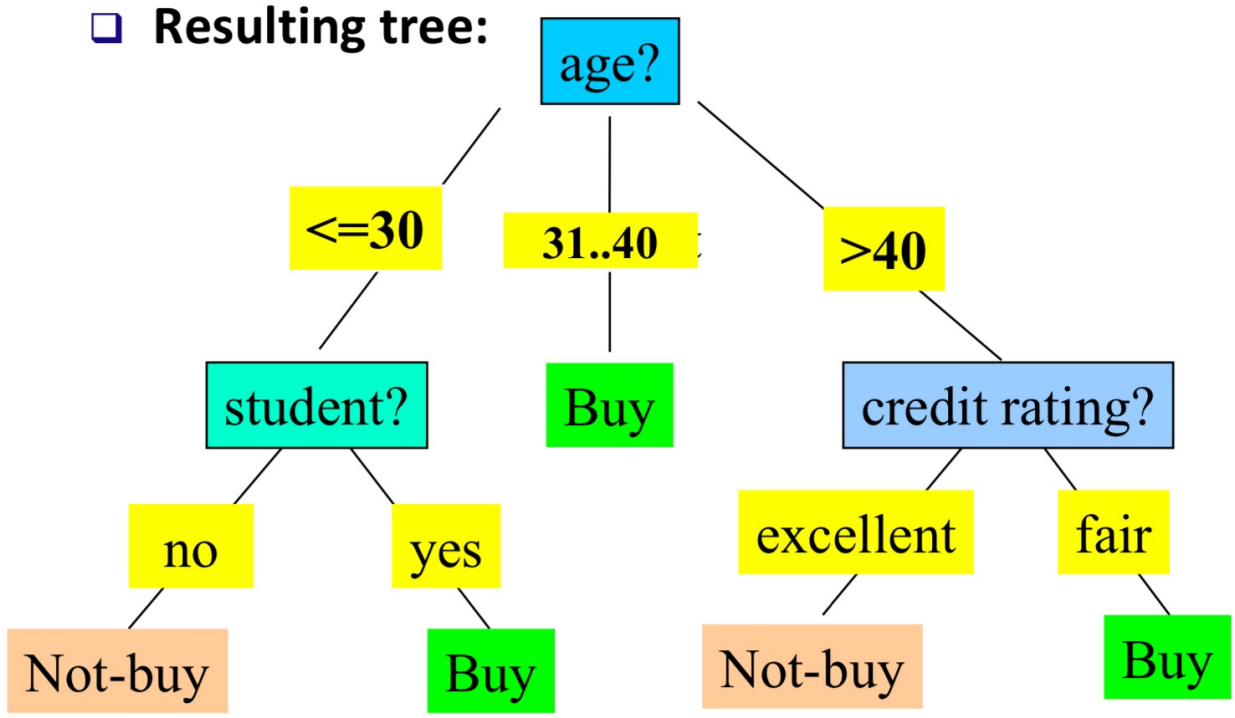
HW 5

- **Resulting tree:**




Training data set: Who buys computer?

สูตรที่ใช้ในการคำนวณ

| 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 |

- Expected information (entropy) needed to classify a tuple in D:

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

↳ prob ของการเป็น Y และ N — Class สูตรที่ใช้

- Information needed (after using A to split D into v partitions) to classify D:

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

↳ ดูของ feature — Feature สูตรที่ใช้

- Information gained by branching on attribute A

คำนวนว่ากิ่งไหนดีที่สุด

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

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

- Class P: buys_computer = “yes”
- Class N: buys_computer = “no”

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

Y N — Prob Y — Prob N — Class — n(จำนวนทั้งหมด)

2. $$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- age แบ่งออกเป็น 3 กลุ่ม คือ age ≤ 30, 31 ≥ age ≤ 40, age > 40

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

- income แบ่งออกเป็น 3 กลุ่ม คือ low, medium, high

$Info_{income}(D) = \frac{4}{14} I(3,1) + \frac{6}{14} I(4,2) + \frac{4}{14} I(2,2) = \frac{4}{14}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{6}{14}\left(-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) + \frac{4}{14}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)$

$= 0.911$

- student แบ่งออกเป็น 2 กลุ่ม คือ yes, no

$Info_{student}(D) = \frac{7}{14} I(6,1) + \frac{7}{14} I(3,4) = \frac{7}{14}\left(-\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right)\right) + \frac{7}{14}\left(-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)\right) = 0.788$

- credit_rating แบ่งออกเป็น 2 กลุ่ม fair, excellent

$Info_{credit\_rating}(D) = \frac{8}{14} I(6,2) + \frac{6}{14} I(3,3) = \frac{8}{14}\left(-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right) + \frac{6}{14}\left(-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right) = 0.892$

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

$Gain(age) = 0.940 - 0.694 = 0.246$

$Gain(income) = 0.940 - 0.911 = 0.029$

$Gain(student) = 0.940 - 0.788 = 0.152$

$Gain(Credit\_rating) = 0.940 - 0.892 = 0.048$

ดังนั้นเลือก Gain(age) เป็นรูท เพราะมีค่ามากที่สุด

4. แยกกลุ่มของ feature ตามค่าในราก (root node)

4.1 ≤ 30

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |

$Info(D) = I(\overset{Y}{2},\overset{N}{3}) = 0.971$

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

$= 0.4$

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

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

คำนวณ Information Grain

$Grain(income) = 0.971 - 0.4 = 0.571$

$Grain(student) = 0.971 - 0 = 0.971$

$Grain(Credit\_rating) = 0.971 - 0.951 = 0.02$

เลือก Grain(student) เป็น node ในอายุ ≤ 30

4.2 31...40

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| 31...40 | high | no | fair | yes |
| 31...40 | medium | no | excellent | yes |
| 31...40 | high | yes | fair | yes |
| 31...40 | low | yes | excellent | yes |

จาก buys_computer จะเห็นได้ว่ามีแต่ yes ไม่มี no

∴ เมื่ออายุ 31...40 สามารถตอบ yes ใน buys_computer ได้เลย

4.3 >40

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| >40 | medium | yes | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| >40 | medium | no | excellent | no |

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

$= 0.971$

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

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

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

คำนวณ Information Grain

$Grain(income) = 0.971 - 0.951 = 0.02$

$Grain(student) = 0.971 - 0.951 = 0.02$

$Grain(Credit\_rating) = 0.971 - 0 = 0.971$

เลือก Grain(Credit_rating) เป็น node ในอายุ > 40

5. สามารถสร้าง Decision Tree ได้ดังนี้