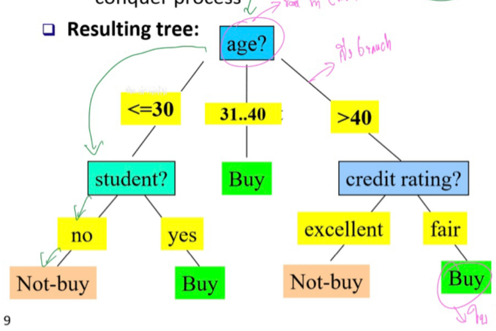
น.ส. กาญจนสุดา พุ่มมูลศรี 633020438-6

- **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

9

มี yes = 9 , no = 5

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

- $Info(D) = I(9,5) = -\frac{9}{14}\log_2 - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$
- $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$

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

$= \frac{5}{14}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{14}\left[-\frac{4}{4}\log_2\left(\frac{4}{4}\right)\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right] = 0.694$

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

$= \frac{4}{14}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right] + \frac{4}{14}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] = 0.911$

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

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

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

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

- หา Information Gain

สูตร $Gain(A) = Info(D) - Info_A(D)$

1. $Gain(age) = 0.940 - 0.694 = 0.246$ → ได้ Gain ที่ถูกเลือก เพราะมีค่าสูงที่สุด
2. $Gain(income) = 0.940 - 0.911 = 0.029$
3. $Gain(student) = 0.940 - 0.789 = 0.151$
4. $Gain(credit\_rating) = 0.940 - 0.892 = 0.048$

- แยก feature หาค่าของ root node

1. 1 $<= 30$

ใช้ค่าเดิม $Info(D) = I(2,3) = 0.971$

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

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

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

$Gain(Income) = 0.971 - 0.4 = 0.571$

$Gain(Student) = 0.971 - 0 = 0.971$

$Gain(credit\_rating) = 0.971 - 0.951 = 0.02$

∴ เลือก Gain (student) เป็น node $<= 30$ ∆

2. 31..40

yes = 4, no = 0

∴ 31..40 สามารถตอบ yes ได้เลย ∆

3. > 40

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

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

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

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

$Gain(Income) = 0.971 - 0.951 = 0.02$

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

$Gain(credit\_rating) = 0.971 - 0 = 0.971$

∴ เลือก Gain (credit_rating) node > 40 ∆

- Decision Tree ที่ได้ออกมา