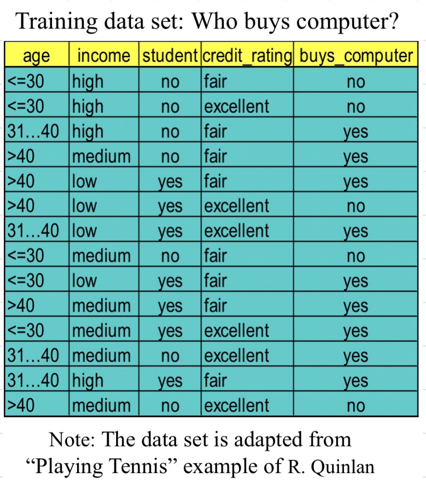
HW 5



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

- หา node ของ root

จากสูตร

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

จะได้

$$\mathrm{Info}(D) = I(9,5) \quad ; \quad 9 = yes \;,\; 5 = no$$

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

หาค่า Info ในแต่ละ feature

- $\mathrm{Info}_{age}(D) = \frac{5}{14}\underbrace{I(2,3)}_{<=30} + \frac{4}{14}\underbrace{I(4,0)}_{31..40} + \frac{5}{14}\underbrace{I(3,2)}_{>40}$

$$= \frac{5}{14}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{14}\left[-\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$= 0.694$$

- $\mathrm{Info}_{income}(D) = \frac{4}{14}\underbrace{I(2,2)}_{high} + \frac{6}{14}\underbrace{I(4,2)}_{medium} + \frac{4}{14}\underbrace{I(3,1)}_{low}$

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

$$= 0.911$$

- $\mathrm{Info}_{student}(D) = \frac{7}{14}\underbrace{I(6,1)}_{yes} + \frac{7}{14}\underbrace{I(3,4)}_{no}$

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

$$= 0.789$$

- $\mathrm{Info}_{credit\_rating}(D) = \frac{8}{14}\underbrace{I(6,2)}_{fair} + \frac{6}{14}\underbrace{I(3,3)}_{excellent}$

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

$$= 0.892$$

$\mathrm{Gain}(age) = \mathrm{Info}(D) - \mathrm{Info}_{age}(D) = 0.940 - 0.694 = 0.246$ ※

$\mathrm{Gain}(income) = \mathrm{Info}(D) - \mathrm{Info}_{income}(D) = 0.940 - 0.911 = 0.029$

$\mathrm{Gain}(student) = \mathrm{Info}(D) - \mathrm{Info}_{student}(D) = 0.940 - 0.789 = 0.151$

$\mathrm{Gain}(credit\_) = \mathrm{Info}(D) - \mathrm{Info}_{credit\_}(D) = 0.940 - 0.892 = 0.048$

◦ หาค่า Info ในแต่ละกิ่ง

- <= 30

age	income	student	credit_rating	buy
<= 30	high	no	fair	no
	high	no	excellent	no
	medium	no	fair	no
	low	yes	fair	yes
	medium	yes	excellent	yes

$$Info(D) = I(2,3) \quad ; \quad 2 = yes \; , \; 3 = no$$

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

$$Info_{income}(D) = \overbrace{\frac{2}{5}I(0,2)}^{high} + \overbrace{\frac{2}{5}I(1,1)}^{medium} + \overbrace{\frac{1}{5}I(1,0)}^{low}$$

$$= \frac{2}{5}\left[-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] + \frac{1}{5}\left[-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right]$$

$$= 0.4$$

$$Info_{student}(D) = \overbrace{\frac{2}{5}I(2,0)}^{yes} + \overbrace{\frac{3}{5}I(0,3)}^{no}$$

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

$$= 0$$

$$Info_{credit}(D) = \overbrace{\frac{3}{5}I(1,2)}^{fair} + \overbrace{\frac{2}{5}I(1,1)}^{excellent}$$

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

$$= 0.951$$

$$Gain(income) = Info(D) - Info_{income}(D) = 0.971 - 0.4 = 0.571$$

$$Gain(student) = Info(D) - Info_{student}(D) = 0.971 - 0 = 0.971 \quad ※$$

$$Gain(credit\_) = Info(D) - Info_{credit\_}(D) = 0.971 - 0.951 = 0.02$$

- 31...40

age	income	student	credit	buys
31...40	high	no	fair	yes
	low	yes	excellent	yes
	medium	no	excellent	yes
	high	yes	fair	yes

※※ จากตารางสามารถสรุปได้ว่าช่วงอายุ 31...40 มีข้อมูล buys_computer เป็น yes ทั้งหมด

- >40

age	income	student	credit_rating	buy
>40	medium	no	fair	yes
	low	yes	fair	yes
	low	yes	excellent	no
	medium	yes	fair	yes
	medium	no	excellent	no

$\text{Info}(D) = I(3,2)$; $3 = \text{yes}$, $2 = \text{no}$

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

$$\text{Info}_{\text{income}}(D) = \overbrace{\frac{3}{5}I(2,1)}^{\text{medium}} + \overbrace{\frac{2}{5}I(1,1)}^{\text{low}}$$

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

$$= 0.951$$

$$\text{Info}_{\text{student}}(D) = \overbrace{\frac{3}{5}I(2,1)}^{\text{yes}} + \overbrace{\frac{2}{5}I(1,1)}^{\text{no}}$$

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

$$= 0.951$$

$$\text{Info}_{\text{credit}}(D) = \overbrace{\frac{3}{5}I(3,0)}^{\text{fair}} + \overbrace{\frac{2}{5}I(2,0)}^{\text{excellent}}$$

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

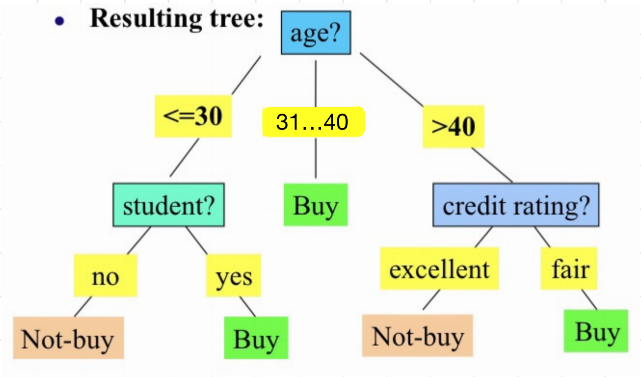
$$= 0$$

$\text{Gain}(\text{income}) = \text{Info}(D) - \text{Info}_{\text{income}}(D) = 0.971 - 0.951 = 0.02$

$\text{Gain}(\text{student}) = \text{Info}(D) - \text{Info}_{\text{student}}(D) = 0.971 - 0.951 = 0.02$

$\text{Gain}(\text{credit\_}) = \text{Info}(D) - \text{Info}_{\text{credit\_}}(D) = 0.971 - 0 = 0.971$ #

Descision Tree Induction


• Resulting tree:
age?
<=30
31…40
>40
student?
Buy
credit rating?
no
yes
excellent
fair
Not-buy
Buy
Not-buy
Buy