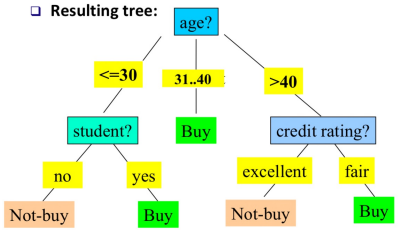
Attribute selection with Information gain

- Resulting tree:



age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

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

class P: buys_computer = "yes"

class N: buys_computer = "no"

Yes : 9

No : 5

Age

<= 30 yes: 2 / No: 3

31...40 yes: 4 / No: 0

>40 yes: 3 / No: 2

income

high yes: 2 / No: 2

medium yes: 4 / No: 2

low yes: 3 / No: 1

student

Yes yes: 6 / No: 1

No yes: 3 / No: 4

Credit

Fair yes: 6 / No: 2

Excellent yes: 3 / No: 3

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

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

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

$= 0.694$

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

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

$= 0.911$

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

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

$= 0.789$

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

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

$= 0.892$

$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246$

$Gain(income) = Info(D) - Info_{income}(D) = 0.940 - 0.911 = 0.029$

$Gain(student) = Info(D) - Info_{student}(D) = 0.940 - 0.789 = 0.151$

$Gain(credit\_rating) = Info(D) - Info_{credit\_rating}(D) = 0.940 - 0.892 = 0.048$

<= 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

Age
yes: 2 / No: 3

Income
high yes: 0 / No: 2
Medium yes: 1 / No: 1
low yes: 1 / No: 0

student
yes: 2 / No: 3

Credit
fair yes: 1 / No: 2
excellent yes: 1 / No: 1

Soln

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

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

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

$$= 0.4$$

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

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

$$= 0$$

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

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

$$= 0.951$$

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

Gain maxm

$$Gain_{student}(D) = Info(D) - Info_{student}(D) = 0.971 - 0 = 0.971$$

$$Gain_{credit\_rating}(D) = Info(D) - Info_{credit\_rating}(D) = 0.971 - 0.951 = 0.02$$

31 - 40

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

Age

yes: 4 / No: 0

Income

High yes: 2 / No: 0

Medium yes: 1 / No: 0

low yes: 1 / No: 0

Student

Yes

yes: 2 / No: 0

No

yes: 2 / No: 0

Credit

Fair

yes: 2 / No: 0

Excellent:

yes: 2 / No: 0

> 40

Training data set: Who buys computer?

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

Age
yes: 3 / No: 2

Income
High yes: 0 / No: 0
Medium yes: 2 / No: 1
low yes: 1 / No: 1

Student
Yes
yes: 2 / No: 1
No
yes: 1 / No: 1

Credit
fair
yes: 3 / No: 0
Excellent
yes: 0 / No: 2

Soln

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

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

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

$$= 0.951$$

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

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

$$= 0.951$$

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

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

$$= 0$$

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

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

Gain มากสุด

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

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

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

Gain student = 0.971

"เลือก Gain มากที่สุด"

student	buys_computer
no	no
no	no
no	no
yes	yes
yes	yes

No → Not buy

Yes → Buy

age	income	student	credit_rating	buys_computer
31…40	high	no	fair	yes
31…40	low	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes

Buy

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

Gain credit_rating = 0.971

"เลือก Gain ที่มากสุด"

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

Excellent → Not buy

fair → Buy