

Example: Attribute Selection with Information Gain

□ Class P: buys_computer = "yes"

□ Class N: buys_computer = "no"

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

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

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	no	excellent	yes
<=30	medium	no	excellent	yes
<=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

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

yān N mǎn hū

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

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

Similarly, we can get

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

$$Info(D) = I(8,4) = -\frac{8}{12} \log_2\left(\frac{8}{12}\right) - \frac{4}{12} \log_2\left(\frac{4}{12}\right) = 0.9113$$

$$Info_{age}(D) = \frac{4}{12} I(2,2) + \frac{3}{12} I(3,0) + \frac{5}{12} I(3,2)$$

$$= \frac{4}{12} (1) + \frac{3}{12} (0) + \frac{5}{12} (0.9710)$$

$$= 0.9761$$

$$I(2,2) = -\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right) = 1$$

$$I(3,0) = -\frac{3}{3} \log_2\left(\frac{3}{3}\right) = 0$$

$$I(3,2) = -\frac{3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right) = 0.9710$$

$$Gain(age) = 0.9113 - 0.9761$$

$$= 0.3422$$

Example: Attribute Selection with Information Gain

□ Class P: buys_computer = "yes"

□ Class N: buys_computer = "no"

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

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

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	low	no	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

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

yau N mau lu

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

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

Similarly, we can get

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

$$\begin{aligned} Info_{income}(D) &= \frac{4}{12} I(2,2) + \frac{5}{12} I(4,1) + \frac{3}{12} I(2,1) \\ &= \frac{4}{12} (1) + \frac{5}{12} (0.7219) + \frac{3}{12} (0.9183) \\ &= 0.8637 \end{aligned}$$

$$I(4,1) = -\frac{4}{5} \log_2\left(\frac{4}{5}\right) - \frac{1}{5} \log_2\left(\frac{1}{5}\right) = 0.7219$$

$$I(2,1) = -\frac{2}{3} \log_2\left(\frac{2}{3}\right) - \frac{1}{3} \log_2\left(\frac{1}{3}\right) = 0.9183$$

$$\begin{aligned} Gain(income) &= 0.9183 - 0.8637 \\ &= 0.0546 \end{aligned}$$

Example: Attribute Selection with Information Gain

□ Class P: buys_computer = "yes"

□ Class N: buys_computer = "no"

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

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

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	low	no	fair	yes
≤ 30	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

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

yay N mnduku

$\frac{5}{14} I(2,3)$ means "age ≤ 30 " has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

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

Similarly, we can get

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

$$\begin{aligned} Info_{student}(D) &= \frac{6}{12} I(5,1) + \frac{6}{12} I(3,3) \\ &= \frac{6}{12} (0.6500) + \frac{6}{12} (1) \\ &= 0.825 \end{aligned}$$

$$I(5,1) = -\frac{5}{6} \log_2\left(\frac{5}{6}\right) - \frac{1}{6} \log_2\left(\frac{1}{6}\right) = 0.6500$$

$$I(3,3) = -\frac{3}{6} \log_2\left(\frac{3}{6}\right) - \frac{3}{6} \log_2\left(\frac{3}{6}\right) = 1$$

$$\begin{aligned} Gain(student) &= 0.9183 - 0.8250 \\ &= 0.0933 \end{aligned}$$

Example: Attribute Selection with Information Gain

Class P: buys_computer = "yes"

Class N: buys_computer = "no"

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

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

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	no	excellent	yes
<=30	low	no	fair	yes
<=30	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

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

yān N nān fū

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

Hence

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

Similarly, we can get

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

$$\begin{aligned} Info_{credit_rating}(D) &= \frac{7}{12} I(6,1) + \frac{5}{12} I(2,3) \\ &= \frac{7}{12} (0.6906) + \frac{5}{12} (0.9710) \\ &= 0.8074 \end{aligned}$$

$$I(6,1) = -\frac{6}{7} \log_2\left(\frac{6}{7}\right) - \frac{1}{7} \log_2\left(\frac{1}{7}\right) = 0.6906$$

$$I(2,3) = -\frac{2}{5} \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right) = 0.9710$$

$$\begin{aligned} Gain(credit_rating) &= 0.9183 - 0.8074 \\ &= 0.1109 \end{aligned}$$