Example: Attribute Selection with Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

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

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
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31…40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31…40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31…40	medium	no	excellent	yes
13	31…40	high	yes	fair	yes
14	>40	medium	no	excellent	no

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

$\frac{5}{14}I(2,3)$ 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

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

น.ส. ภาวิณี หน่อสีดา 663020288-2

A11

income

high		Med		Low	
1	N	4	Y	5	Y
2	N	8	N	6	N
3	Y	10	Y	7	Y
13	Y	11	Y	9	Y
		12	Y		
		14	N		

Income	P_i	n_i	$I(P_i, n_i)$
high	2	2	1
medium	4	2	0.9183
Low	3	1	0.8113

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

$I(4,2) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.9183$

$I(3,1) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.8113$

$Info_{income}(D) = \frac{4}{14}(1) + \frac{6}{14}(0.9183) + \frac{4}{14}(0.8113)$

$= 0.9111$

จาก $Info(D) = I(9,5) = 0.940$

ดังนั้น $Gain(Income) = 0.940 - 0.9111$

$= 0.0289$ #

student

No		Yes	
1	N	5	Y
2	N	6	N
3	Y	7	Y
4	Y	9	Y
8	N	10	Y
12	Y	11	Y
14	N	13	Y

Student	P_i	n_i	$I(P_i, n_i)$
No	3	4	0.9852
Yes	6	1	0.5917

$I(3,4) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.9852$

$I(6,1) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) = 0.5917$

$Info_{student}(D) = \frac{7}{14}(0.9852) + \frac{7}{14}(0.5917)$

$= 0.7885$

ดังนั้น $Gain_{student} = 0.940 - 0.7885$

$= 0.1515$ #

Credit_rating

fair	excellent
N	N
Y	N
Y	Y
Y	Y
N	N
Y	Y
Y	
Y	

Credit	P_i	n_i	$I(P_i, n_i)$
fair	6	2	0.8113
excellent	3	3	1

$I(6,2) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0.8113$

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

$Info_{credit}(D) = \frac{8}{14}(0.8113) + \frac{6}{14}(1)$

$= 0.8922$

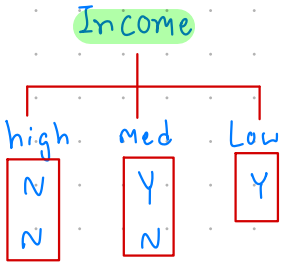
$Gain_{credit} = 0.94 - 0.8922$

$= 0.0478$ #

น.ส. ภาวิณี หน่อสีดา 6630202288-2

$Info(D) = I(2,3) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.9701$

Age ≤ 30



Income	Pi	ni	I(Pi,ni)
high	0	2	0
medium	1	1	1
Low	1	0	0

$I(0,2) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$

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

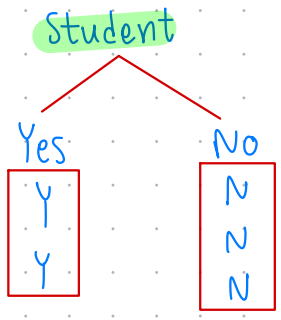
$I(1,0) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{1}{1}\right) = 0$

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

$= 0.8$

$Gain_{income} = 0.9701 - 0.8$

$= 0.1701$ #



Student	Pi	ni	I(Pi,ni)
Yes	2	0	0
No	0	3	0

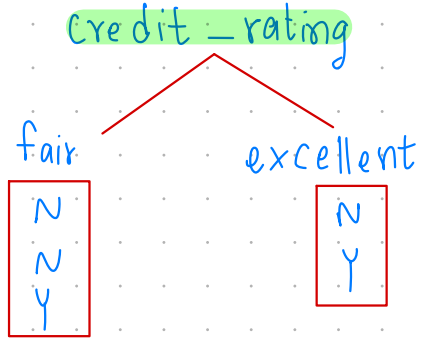
$I(2,0) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) = 0$

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

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

$Gain_{student} = 0.9701 - 0$

$= 0.9701$ #



Credit_rating	Pi	ni	I(Pi,ni)
fair	1	2	0.9183
excellent	1	1	1

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

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

$Info_{credit}(D) = \frac{3}{5}(0.9183) + \frac{2}{5}(1)$

$= 0.951$

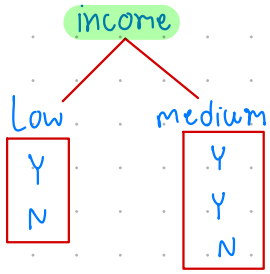
$Gain_{(credit)} = 0.9701 - 0.951$

$= 0.0199$ #

น.ส. ภาวิณี หน่อสีดา 6630202288-2

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

Age > 40



income	P_i	n_i	$I(P_i, n_i)$
Low	1	1	1
medium	2	1	0.918

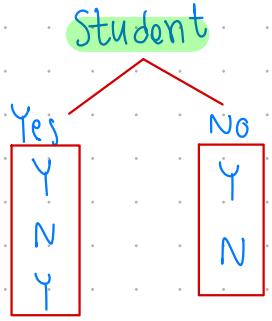
$I(1,1) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$

$I(2,1) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{2}{3}\log_2(\frac{2}{3}) = 0.918$

$Info_{income}(D) = \frac{2}{5}(1) + \frac{3}{5}(0.918) = 0.551$

$Gain_{income} = 0.971 - 0.551$

$= 0.42$ #



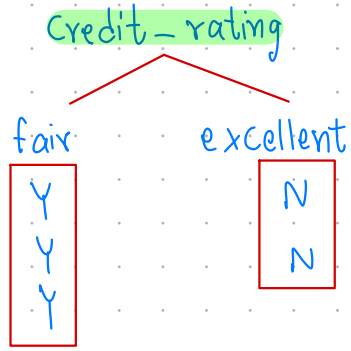
Student	P_i	n_i	$I(P_i, n_i)$
Yes	2	1	0.918
No	1	1	1

$I(2,1) = -\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3}) = 0.918$

$I(1,1) = -\frac{1}{1}\log_2(\frac{1}{1}) - \frac{1}{1}\log_2(\frac{1}{1}) = 1$

$Info_{student}(D) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.551$

$Gain_{student} = 0.971 - 0.551 = 0.42$ #



credit	P_i	n_i	$I(P_i, n_i)$
fair	3	0	0
excellent	0	2	0

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

$I(0,2) = -\frac{0}{2}\log_2(\frac{0}{2}) - \frac{2}{2}\log_2(\frac{2}{2}) = 0$

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

$= 0$

$Gain_{credit} = 0.971 - 0$

$= 0.971$ #