6434476923 พัทธ์ อาคาปิยธรรม

Created a decision tree using information gain as a measure

1. Expected information

มี class 2 class — C0, C1

class P : class = "C0" → 10 แถว
class N : class = "C1" → 10 แถว
} มีทั้งหมด 20 แถว

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad p_i = \frac{|C_{i,D}|}{|D|}$$

ดังนั้น $Info(D) = I(10,10)$; $m = 2$

$$= -p_{C0}\log_2(p_{C0}) - p_{C1}\log_2(p_{C1})$$

$$= -\frac{10}{20}\log_2\left(\frac{10}{20}\right) - \frac{10}{20}\log_2\left(\frac{10}{20}\right) = 1$$

2. Information

มี 3 แบบ → Gender, Car Type และ Shirt Size

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D_i|} \times Info(D_j)$$

Gender :

gd	p_i	n_i
F	4	6
M	6	4

$$Info_{Gender}(D) = \frac{10}{20} I(4,6) + \frac{10}{20} I(6,4)$$

$$= \frac{10}{20}\left[-\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}\right] \rightarrow \text{Female}$$

$$+\frac{10}{20}\left[-\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}\right] \rightarrow \text{Male} \qquad \approx 0.97095$$

Car Type :

ct	p_i	n_i
Fam	1	3
Sport	8	0
Lux	1	7

$$Info_{Car\ Type}(D) = \frac{4}{20} I(1,3) + \frac{8}{20} I(8,0) + \frac{8}{20} I(1,7)$$

$$= \frac{4}{20}\left[-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right] \rightarrow \text{Family}$$

$$+\frac{8}{20}\left[-\frac{8}{8}\log_2\frac{8}{8}\right] \rightarrow \text{Sports}$$

$$+\frac{8}{20}\left[-\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8}\right] \rightarrow \text{Luxury} \qquad \approx 0.37968$$

Shirt Size :

ss	p_i	n_i
S	3	2
M	3	4
L	2	2
XL	2	2

$$Info_{Shirt\ Size}(D) = \frac{5}{20} I(3,2) + \frac{7}{20} I(3,4) + \frac{4}{20} I(2,2) + \frac{4}{20} I(2,2)$$

$$= \frac{5}{20}\left[-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right] \rightarrow \text{small}$$

$$+\frac{7}{20}\left[-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right] \rightarrow \text{Medium}$$

$$+\frac{4}{20}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right] \rightarrow \text{Large}$$

$$+\frac{4}{20}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right] \rightarrow \text{Extra Large} \qquad \approx 0.98757$$

3. Information gained

$$Grain(A) = Info(D) - Info_A(D)$$

Gain (Gender) = 1 − 0.97095 = 0.02905

Gain (Car Type) = 1 − 0.37968 = 0.62032 → splitting attribute

Gain (Shirt Size) = 1 − 0.98757 = 0.01243

6434476923 พัทธ์ อาภาปิยธรรม

Car Type?

Family

Gender	Shirt Size	class
M	Small	C0
M	Large	C1
M	Extra Large	C1
M	Medium	C1

Sport

Gender	Shirt Size	class
M	Medium	C0
M	Medium	C0
M	Large	C0
M	Extra Large	C0
M	Extra Large	C0
F	Small	C0
F	Small	C0
F	Medium	C0

Luxury

Gender	Shirt Size	class
F	Large	C0
M	Extra Large	C1
F	Small	C1
F	Small	C1
F	Medium	C1
F	Medium	C1
F	Medium	C1
F	Large	C1

pure

Recursive

Gender	Shirt Size	class
M	Small	C0
M	Large	C1
M	Extra Large	C1
M	Medium	C1

1. Expected information

$$\text{Info}(D) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.81128$$

2. Information

มี 2 แบบ → Gender และ Shirt size

Gender :

gd	p_i	n_i
F	0	0
M	1	3

$$\text{Info}_{\text{Gender}}(D) = \frac{4}{4}\left[-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right] = 0.81128$$

class P : class = "C0" → 1 แบบ

class N : class = "C1" → 3 แบบ

Shirt Size :

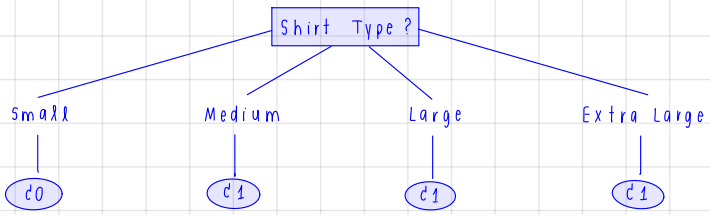
ss	p_i	n_i
S	1	0
M	0	1
L	0	1
XL	0	1

$$\text{Info}_{\text{Shirt size}}(D) = \frac{1}{4}(-1\log_2 1) \rightarrow \text{small}$$
$$+ \frac{1}{4}(-1\log_2 1) \rightarrow \text{Medium}$$
$$+ \frac{1}{4}(-1\log_2 1) \rightarrow \text{Large}$$
$$+ \frac{1}{4}(-1\log_2 1) \rightarrow \text{Extra Large}$$
$$= 0$$

3. Information gained

Gain (Gender) = 0.81128 − 0.81128 = 0

Gain (Shirt Size) = 0.81128 − 0 = 0.81128 → splitting attribute



Recursive

6434476923 พัทธ์ อาคาปิยธรรม

Gender	Shirt Size	Class
F	Large	C0
M	Extra Large	C1
F	Small	C1
F	Small	C1
F	Medium	C1
F	Medium	C1
F	Medium	C1
F	Large	C1

1. Expected information

$$Info(D) = -\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8} \approx 0.54356$$

2. Information

ที่ 2 แบบ → Gender และ Shirt Size

Gender :

gd	p_i	n_i
F	1	6
M	0	1

$$Info_{Gender}(D) = \frac{7}{8}\left[-\frac{1}{7}\log_2\frac{1}{7} - \frac{6}{7}\log_2\frac{6}{7}\right] \rightarrow Female$$

$$+ \frac{1}{8}\left[-1\log_2 1\right] \rightarrow male \approx 0.51771$$

Class P : Class = "C0" → 1 แถว

Class N : Class = "C1" → 7 แถว

Shirt Size :

ss	p_i	n_i
S	0	2
M	0	3
L	1	1
XL	0	1

$$Info_{Shirt\ size}(D) = \frac{2}{8}(-1\log_2 1) \rightarrow small$$

$$+ \frac{3}{8}(-1\log_2 1) \rightarrow Medium$$

$$+ \frac{2}{8}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] \rightarrow Large$$

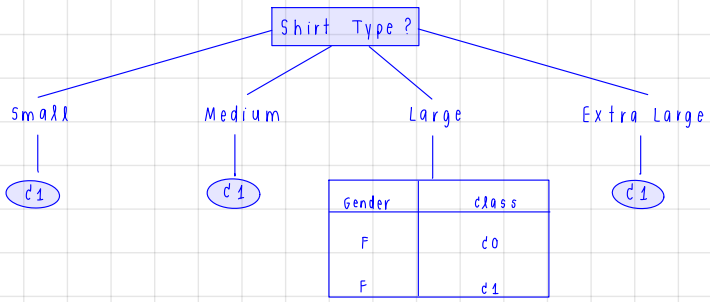
$$+ \frac{1}{8}(-1\log_2 1) \rightarrow Extra\ Large$$

$$\approx 0.25$$

3. Information gained

Gain (Gender) = 0.54356 − 0.51771 = 0.02585

Gain (Shirt Size) = 0.54356 − 0.25 = 0.29356 → splitting attribute



คำตอบ