2.4.1.4 Information Gain

ID3 Algorithm as attribute selection measure minimizes information (entropy / uncertainty) achieve the least randomness or impurity

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

m = classes number

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

sample number

$p_i = \frac{|C_{i,D}|}{|D|}$ sample belongs to class C_i

Attribute A split D into v subset

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

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

original remaining

Q : calculate the Gain(A)

PID	Fever	Cough	Tiredness	COVID-19
1	no	mild	no	–
2	yes	no	no	–
3	yes	mild	yes	+
4	mild	yes	no	+
5	mild	mild	yes	+
6	no	mild	yes	–

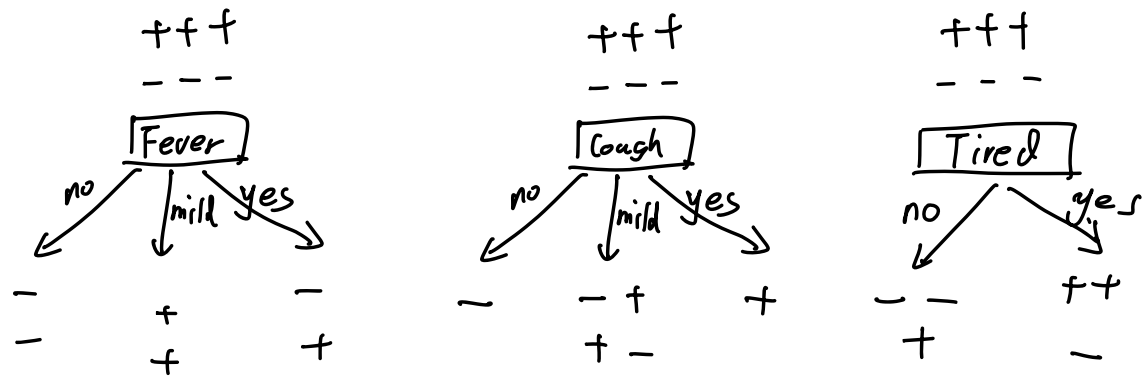
Solution ① $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

$Info(D) = -\left[p_1 \log_2(p_1) + p_2 \log_2(p_2)\right]$

$p_1 = \frac{3}{6} = 0.5 \qquad p_2 = \frac{3}{6} = 0.5$

$Info(D) = -\left[0.5 \log_2 0.5 + 0.5 \log_2 0.5\right]$

$= 1$

② 

③ Fever

$$Info_{Fever}(D) = \sum_{j=1}^{3} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$= \frac{2}{6} Info(D_{no}) + \frac{2}{6} Info(D_{mild}) + \frac{2}{6} Info(D_{yes})$$

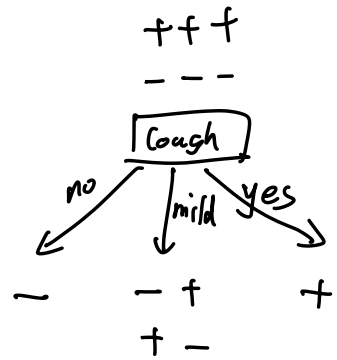
where $Info(D_{no}) = -1 \log_2(1) = 0$

$Info(D_{mild}) = 0$

$Info(D_{yes}) = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1$

So $Info_{Fever}(D) = \frac{1}{3}$

④ Cough



$$Info_{Cough}(D) = \sum_{j=1}^{3} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$= \frac{2}{6} Info(D_{no}) + \frac{2}{6} Info(D_{mild}) + \frac{2}{6} Info(D_{Yes})$$

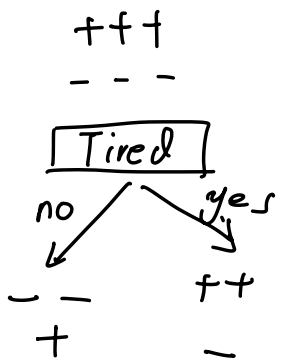
where $Info(D_{no}) = -1 \log_2(1) = 0$

$Info(D_{mild}) = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1$

$Info(D_{Yes}) = 0$

So $Info_{Cough}(D) = \frac{1}{3}$

⑤ Tired



$$Info_{Tired}(D) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$= \frac{3}{6} Info(D_{no}) + \frac{3}{6} Info(D_{Yes})$$

where $Info(D_{no}) = -\left[\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3}\right] = 0.5850$

$Info(D_{Yes}) = 0.5850$

So $Info_{Tired}(D) = 0.5850$

⑥ Gain(Fever) = $1 - \frac{1}{3} = 0.6667$

Gain(cough) = $1 - \frac{1}{3} = 0.6667$

Gain(Tired) = $1 - 0.5850 = 0.4150$

So, we can choose Fever or cough as the first attribute.

⑦ Assume we choose Fever as first attribute, we can choose cough or Tiredness as the second attribute due to they have same Gain. And the classification task is over.

PID	Fever	Cough	Tiredness	COVID-19
1	no	mild	no	–
2	yes	no	no	–
3	yes	mild	yes	+
4	mild	yes	no	+
5	mild	mild	yes	+
6	no	mild	yes	–