

## 2.4.1.4.1 Information Gain: Example

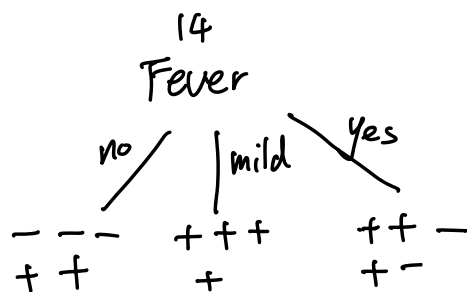
Q: compute gain information of 4 attributes.

PID	Fever	Cough	Sore Throat	Tiredness	Flu
1	no	yes	no	yes	-
2	no	yes	no	no	-
3	mild	yes	no	yes	+
4	yes	mild	no	yes	+
5	yes	no	yes	yes	+
6	yes	no	yes	no	-
7	mild	no	yes	no	+
8	no	mild	no	yes	-
9	no	no	yes	yes	+
10	yes	mild	yes	yes	+
11	no	mild	yes	no	+
12	mild	mild	no	no	+
13	mild	yes	yes	yes	+
14	yes	mild	no	no	-

Solution ①  $\text{Info}(D)$

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

② Fever



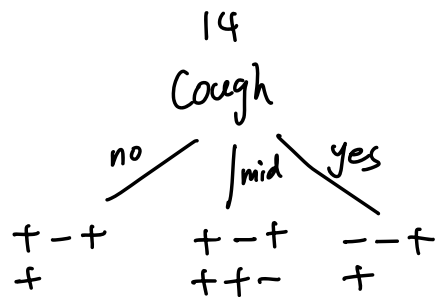
$$\begin{aligned} \text{Info}_{\text{Fever}}(D) &= \frac{5}{14} \text{Info}(D_{\text{no}}) + \frac{4}{14} \text{Info}(D_{\text{mild}}) + \frac{5}{14} \text{Info}(D_{\text{yes}}) \\ &= \frac{5}{14} \left[ -\frac{3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right) \right] + \frac{5}{14} \left[ -\frac{3}{5} \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \log_2\left(\frac{2}{5}\right) \right] \end{aligned}$$

$$= 2 \times \frac{5}{14} \times 0.9710$$

$$= 0.694 \text{ bits}$$

$$\begin{aligned} \text{Gain (Fever)} &= \text{Info}(D) - \text{Info}_{\text{Fever}}(D) \\ &= 0.940 - 0.694 = 0.246 \text{ bits} \end{aligned}$$

③ Similarly, cough.



$$\begin{aligned} \text{Gain(cough)} &= \text{Info}(D) - \text{Info}_{\text{cough}}(D) \\ &= 0.940 - \left[ \frac{4}{14} \text{Info}(D_{\text{no}}) + \frac{6}{14} \text{Info}(D_{\text{mid}}) + \frac{4}{14} \text{Info}(D_{\text{yes}}) \right] \\ &= 0.940 - \left[ \frac{4}{14} \left( -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) + \frac{6}{14} \left( -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) \right. \\ &\quad \left. + \frac{4}{14} \left( -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \right] \\ &= 0.029 \text{ bits} \end{aligned}$$

④ Sore Throat

$$\text{Gain (Sore Throat)} = 0.151 \text{ bits}$$

⑤ Tiredness

Gain (Tiredness) = 0.048 bits

⑥  $0.246 > 0.151 > 0.048 > 0.029$

Since the "fever" attribute has the highest information gain

it is selected as the splitting attribute