Example 2: Gain Ratio (C4.5)

• **Example 2**. A test on *income* splits the given data into **three** partitions, namely *low*, *medium*, and *high*, containing **four**, **six**, and **four** tuples, respectively.

• To compute the **gain ratio** of *income*, we first use Eq. (8.5) to obtain

Example 2: Gain Ratio (C4.5)

TID	*age*	*income*	*student*	*credit_rating*	*Class: buy_computer*
9	youth	low	yes	fair	yes
5	senior	low	yes	fair	yes
7	middle_aged	low	yes	excellent	yes
6	senior	low	yes	excellent	no
11	youth	medium	yes	excellent	yes
10	senior	medium	yes	fair	yes
4	senior	medium	no	fair	yes
12	middle_aged	medium	no	excellent	yes
8	youth	medium	no	fair	no
14	senior	medium	no	excellent	no
13	middle_aged	high	yes	fair	yes
3	middle_aged	high	no	fair	yes
1	youth	high	no	fair	no
2	youth	high	no	excellent	no

Example 2: Gain Ratio (C4.5)

$$SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right)$$

$$SplitInfo_{income}(D) = 1.557.$$

- From Example 8.1, we have *Gain*(*income*) = 0.029. Therefore, *GainRatio*(*income*) = 0.029 / 1.557 = 0.019.
- Similarly, we can compute *GainRatio*(*age*) = 0.156 bits, *GainRatio*(*student*) = 0.152 bits, and *GainRatio*(*credit_rating*) = 0.049 bits [Exercise]. Because *age* has **highest gain ratio** among attributes, **it is selected as the splitting attribute**.