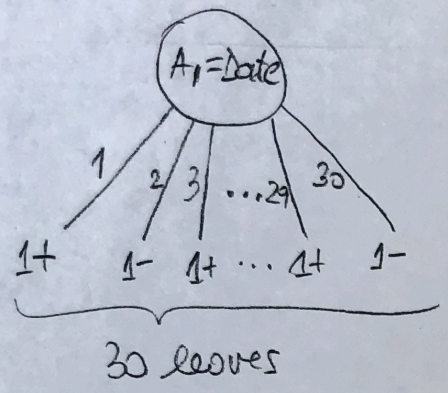
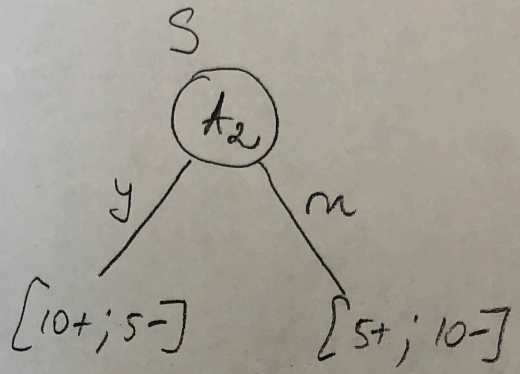
Example: Gain Ratio (4) vs. Info. gain (2)

$S = 30$ ex. $= [15+, 15-]$





⇓ perfectly predicts training data but poorly the test data

---

(1) $E(S) = -\frac{15}{30}\log_2\frac{15}{30} - \frac{15}{30}\log_2\frac{15}{30} = 1$

---

(2) $Gain(S, Date) = 1 - \underbrace{\frac{1}{30}(0) - \frac{1}{30}\cdot 0 - \ldots - \frac{1}{30}\cdot 0}_{30 \text{ times}} = 1$

$Gain(S, A_2) = 1 - \frac{15}{30}\left(-\frac{10}{15}\log_2\frac{10}{15} - \frac{5}{15}\log_2\frac{5}{15}\right) - \frac{15}{30}\left(-\frac{5}{15}\log_2\frac{5}{15} - \frac{10}{15}\log_2\frac{10}{15}\right)$

$= 0.8617$

Because $1 > 0.8617 \xrightarrow{\text{Info Gain}}$ $A_1 = Date$ will be selected → NOT GOOD!

---

(3) $SplitInfo(S, Date) = \underbrace{-\frac{1}{30}\log_2\frac{1}{30} - \ldots - \frac{1}{30}\log_2\frac{1}{30}}_{30 \text{ times}} = -\log_2\frac{1}{30} = 4.9069$

$SplitInfo(S, A_2) = -\frac{15}{30}\log_2\frac{15}{30} - \frac{15}{30}\log_2\frac{15}{30} = -\log_2\frac{1}{2} = 1$

---

(4) $GainRatio(S; Date) = \frac{1}{4.9069} = 0.2038 < GainRatio(S; A_2) = \frac{0.8617}{1} = 0.8617$