Exercise: Week 1 : Lecture 1

We consider the following data set

$D = \{(x_i, y_i)\}_{i=1}^{5} = \{(1,0), (2,1), (3,0), (4,1), (5,1)\}$

and compute the information gain associated with a split at $\xi = 3.5$.

The entropy of the whole dataset is

$$H(D) = -(P_0 \log_2 P_0 + P_1 \log_2 P_1), \quad P_0 = \frac{2}{5}, P_1 = \frac{3}{5}$$

$$H(D) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right)$$

$$= -\frac{1}{5}\left(2\log_2\frac{2}{5} + 3\log_2\frac{3}{5}\right)$$

$$= -\frac{1}{5}\left(2\log_2 2 - 2\log_2 5 + 3\log_2 3 - 3\log_2 5\right)$$

$$H(D) = -\frac{1}{5}\left(2 + 3\log_2 3 - 5\log_2 5\right).$$

$\log_2(2) = 1$

$\log_2(1) = 0$

$\log_2(x)$

x must be +ve

The first set after the split is

$D_1 = \{(1,0), (2,1), (3,0)\}$

and its entropy is

$H(D_1) = -(P_0 \log_2 P_0 + P_1 \log_2 P_1), \quad P_0 = \frac{2}{3}, P_1 = \frac{1}{3}$

$$H(D_1) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$= -\frac{1}{3}\left(2\log_2\frac{2}{3} + \log_2\frac{1}{3}\right) = -\frac{1}{3}\left(2\log_2 2 - 2\log_2 3 + \log 1 - \log_2 3\right)$$

$$= -\frac{1}{3}\left(2 - 3\log_2 3\right)$$

The second set after the split is $D_2 = \{(4,1), (5,1)\}$

and its entropy is $H(D_2) = -(P_0 \log_2 P_0 + P_1 \log_2 P_1), P_0 = 0, P_1 = 1$

$$H(D_2) = -(0\log_2 0 + 1\log_2 1) = 0$$

The information gain is

$$\Delta H = 5H(\mathcal{D}) - (3H(\mathcal{D}_1) + 2H(\mathcal{D}_2))$$

$$\Delta H = 5\left(-\frac{1}{5}(2+3\log_2 3 - 5\log_2 5)\right) - \left(3\left(-\frac{1}{3}(2-3\log_2 3)\right) + 2(0)\right)$$

$$= -(2+3\log_2 3 - 5\log_2 5) + (2 - 3\log_2 3)$$

$$= -(2+3\log_2 3 - 5\log_2 5 - 2 + 3\log_2 3$$

$$= -(6\log_2 3 - 5\log_2 5) \simeq -\left(6\frac{16}{10} - 5\frac{23}{10}\right) = \frac{19}{10} = \frac{1.9}{2}$$