Data-X - Homework 7

Louis Tilloy

Question 1

The information gained by one split of the data T on the feature a is given by the K-L divergence :

$$\begin{aligned} IG(T,a) &= \mathrm{H}(T) - \mathrm{H}(T|a) \\ &= \mathrm{H}(T) - \sum_{v \in \{0,1\}} \frac{|\{\mathbf{x} \in T | x_a = v\}|}{|T|} \cdot \mathrm{H}(\{\mathbf{x} \in T | x_a = v\}) \end{aligned} \tag{1}$$

We can compute the information gain for each feature with the formula (1), let us first compute the different entropies :

$$\begin{aligned} H(T) &= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) \\ &= 1 \end{aligned}$$

$$\begin{aligned} H(T|HasJob) &= -\frac{5}{8}(\frac{2}{5}\log_2(\frac{2}{5}) + \frac{3}{5}\log_2(\frac{3}{5})) - \frac{3}{8}(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) \\ &\approx 0.95 \end{aligned}$$

$$\begin{aligned} H(T|HasFamily) &= -\frac{1}{2}(\frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4})) - \frac{1}{2}(\frac{1}{4}\log_2(\frac{1}{4}) + \frac{3}{4}\log_2(\frac{3}{4})) \\ &\approx 0.81 \end{aligned}$$

$$\begin{aligned} H(T|IsAbove30years) &= -\frac{3}{4}(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) - \frac{1}{4}(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) \\ &= 1 \end{aligned}$$

Thus :

$$\begin{aligned} IG(T, HasJob) &\approx 1 - 0.95 &&= 0.05 \\ IG(T, HasFamily) &\approx 1 - 0.81 &&= 0.19 \\ IG(T, IsAbove30years) &= 1 - 1 &&= 0 \end{aligned}$$

Thus the best feature to do the first split is the feature $HasFamily$.

Question 2

$$\begin{aligned} \mathrm{H}(S) &= -\mathrm{P}(A)\log_2 \mathrm{P}(A) - \mathrm{P}(B)\log_2 \mathrm{P}(B) - \mathrm{P}(C)\log_2 \mathrm{P}(C) \\ &= -0.7\log_2(0.7) - 0.2\log_2(0.2) - 0.1\log_2(0.1) \\ &\approx 1.16 \end{aligned}$$

Intuitively it means that each information given by the signal S could be on average coded on roughly 1.16 bits.

The real Shanon source coding theorem says that for an optimal coding function f from the set of possible values of S to {0, 1} :

$$\begin{array}{rcccl} & H(S) & \leq & \mathbb{E}[length(f(S))] & < & H(S)+1 \\ \Rightarrow & 1.16 & \leq & \mathbb{E}[length(f(S))] & < & 2.16 \end{array}$$