Il Diocane: Introduction to Data Mining Cheatsheet

Kotatsu, the Bringer of Jewishness

1 Similarity and dissimilarity

Entropy:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Sample entropy:

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

Mutual information:

$$I(X,Y) = H(X) + H(X) - H(X,Y)$$

where $H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \log_2 p_{ij}$. For discrete variables the maximum mutual information is

$$\log_2(\min\{n_x, n_y\})$$

where n_x is th number of values that X can take.

We can combine similarities with

$$\text{similarity}(\mathbf{x},\mathbf{y}) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k((\mathbf{x},\mathbf{y}))}{\sum_{k=1}^{n} w_k \delta_k}$$

with

$$\delta_k = \begin{cases} & \text{if both attribute are asymmetric AND they are both zero or if one of them is missing} \\ 0 & \\ 1 & \text{otherwise} \end{cases}$$