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

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

2 Clustering

Number of possible clusters:

$$B(n) = \sum_{k=0}^{\infty} \frac{k^n}{k!}$$

Sum of squared error (what we want to minimize):

$$\text{SSE} = \sum_{i-1}^{K} \sum_{x \in C_i} \mathsf{dist}(m_i, x)^2.$$

So we are trying to minimize the loss function, for the centroids of K clusters $\mathbf{c} = (c_1, \dots, c_K)$:

$$\text{L}(\mathbf{c}) = \sum_{i=1}^{n} \min_{j=1,\dots,K} \|x_i - c_j\|_2^2.$$

We alternate between

- updating $z_i = \arg\min_{j=1,\dots,k} \|x_i - c_j\|_2^2$ (maps point x_i to a cluster j);
- updating $c_j = \frac{1}{|\{i|z_i = j\}|} \sum_{i|z_i=j} x_i$ (recomputes the cluster centroids)

Unsupervised measures of cluster validity

- **Cohesion**: within-cluster sum of squares (SSW)

$$\text{SSW} = \sum_{i=1}^{K} \sum_{x \in C_i} (x - m_i)^2.$$

- **Separation**: between-cluster sum of squares (SSB)

$$\text{SSB} = \sum_{i} |C_i|(m - m_i)^2$$

  where $|C_i|$ is the size of cluster i and m is the global centroid.

- **Silhouette coefficient**: for a point P_i calculate the avg distance a to the points of the cluster and the minimum avg distance b to the points of another cluster. The silhouette coefficient is

$$s = \frac{b-a}{\max\{a,b\}}.$$

Supervised measures of cluster validity

- **Label probability per cluster**:

$$p_{ij} = \frac{m_{ij}}{m_j}$$

  where m_j: size of cluster j and m_{ij}: number of elements of cluster j that are labelled i.

- **Entropy of cluster** j:

$$h_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}.$$

  Total entropy:

$$h = \sum_{j=1}^{K} \frac{m_j}{m} h_j.$$

- **Purity**:

$$\text{purity}_j = \max\{p_{ij}\}.$$

  Total purity:

$$\text{purity} = \sum_{j=1}^{K} \frac{m_j}{m} \text{purity}_j.$$

- **Precision**:

$$\frac{TP}{TP+FP} = \frac{m_{ij}}{m_j} = p_{ij}.$$

- **Recall**:

$$\frac{TP}{TP+FN} = \frac{m_{ij}}{m_i}.$$

- **F-measure**:

$$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Remember that we have

		cluster	
		same	different
class	same	f_{11}	f_{10}
	different	f_{01}	f_{00}

More supervised measures

- **Rand statistic**:

$$R = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}.$$

- **Jaccard coefficient**:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}.$$

- **Adjusted Rand Index**:

$$\text{ARI} = \frac{R(L,C) - \mathbb{E}\left[R(L,C)\right]}{\max\left\{R(L,C), \mathbb{E}\left[R(L,C)\right\}\right]}$$

with

$$\mathbb{E}\left[R(L,C)\right] = \frac{\pi(L)\pi(C)}{\frac{n(n-1)}{2}}$$

$$\max\left\{R(L,C)\right\} = \frac{1}{2}(\pi(L) - \pi(C))$$

with $\pi(C)$: number of objects pairs that belong to the same group C.

3 Fuzzy clustering

We generalize k-mean objective function

$$\text{SSE} = \sum_{j=1}^{k}\sum_{i=1}^{n} w_{ij}^{p} \text{dist}\left(\mathbf{x}_i, \mathbf{c}_j\right)^2.$$

So the procedure is

1. choose random weights w_{ij};
2. until centroids do not change:
   (a) $\mathbf{c}_j = \frac{\sum_{i=1}^{n} w_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} w_{ij}}$ (updates centroids);
   (b) $w_{ij} = \frac{\left(\frac{1}{\text{dist}(\mathbf{x_i},\mathbf{c_j})^2}\right)^{\frac{1}{p-1}}}{\sum_{q=1}^{k}\left(\frac{1}{\text{dist}(\mathbf{x_i},\mathbf{c_q})^2}\right)^{\frac{1}{p-1}}}$ (updates weights).