Property Demonstration on the Clustering Algorithm with Various Examples

陈子蔚 Chen Ziwei

Department of Computer Science and Engineering

Recent Research Topic: Generative Model

Supervisor: 郑锋

Task 7-1

Find a function ϕ which would project the space of $\mathcal{S}$ into a rougher feature space.

$$\phi: \mathbb{R}^{\dim(\mathcal{S})} \to \mathbb{R}^{\dim(\mathcal{H})}$$

Then, $\mathcal{H} = \{h_1, \cdots, h_n\}$ would be obtained. $O(n)$

Then find $|C|$ plane marginals of regions in $\mathbb{R}^{\dim(\mathcal{H})}$ that contain almost the same number of features, which means $\mathcal{H}$ could be separated into $|C|$ parts, $\mathcal{H} = \bigcup_{1 \leq i \leq |C|} \mathcal{H}_i$. $O(n \log n)$

Find the corresponding sites stemmed from $\mathcal{S}$, initial clustering, $\mathcal{S} = \bigcup_{1 \leq i \leq |C|} \mathcal{S}_i$, would be obtained and the center could be calculated by $c_i = \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} s, 1 \leq i \leq |C|$. $O(n)$

Overall, the come complexity of the designed initialization is $O(n \log n)$.

For example, $n = 8$ and $|C| = 2$ with the following figure.

$$\phi: s \mapsto \|s\|_1$$

Trace back the obtained marginal, initial centers $(2, 2.5), (8, 2.5)$ are obtained. The initialization is good enough to avoid the worst case.
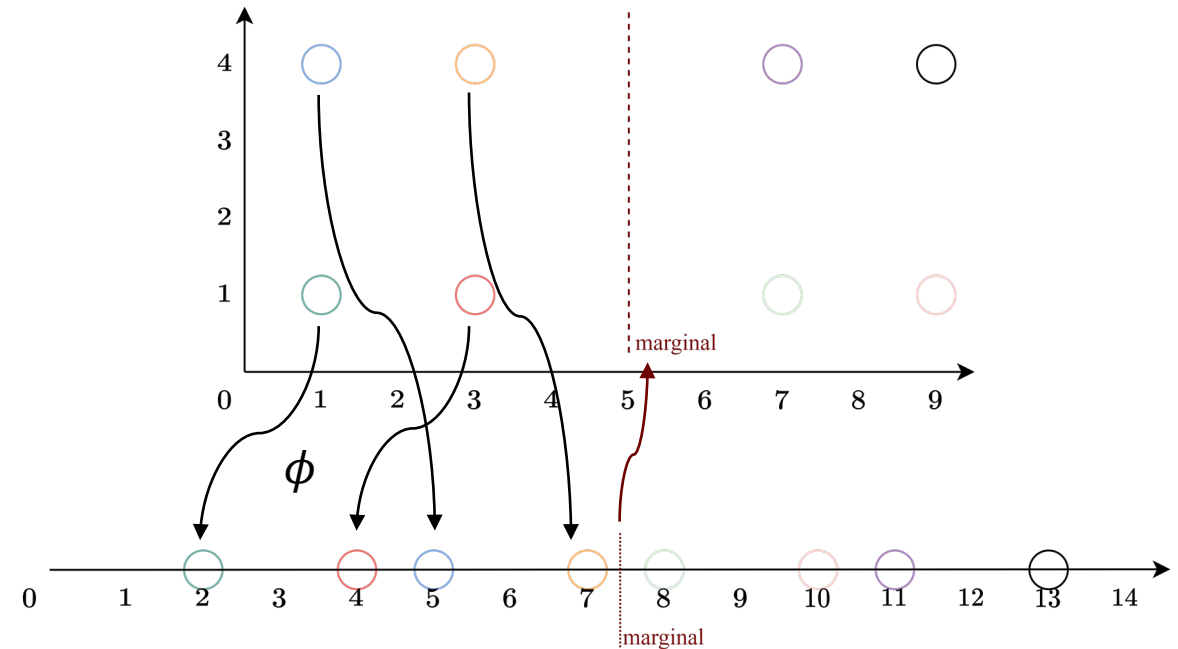
Task 7-2

Consider a simple example with $k = 1$, then

$$\text{dist}(s, C) = \text{dist}(s, c) = \|c - s\|$$

The first row is the case with k-means and solution is 16.
The second row is the case with k-medoids and the solution is 19.

Obviously, k-medoids paid less attention to outlier samples since the solution of k-medoids would remain the third site even if the first site is extremely far from the residual.

Objective Funtion

$$\min_c \sum_{s \in S} \|c - s\|^2$$

$$\min_c \sum_{s \in S} \|c - s\|$$

0   17 19 21 23