Student: Kien D. Vu

Professor: Dr. Yao Xie

Computational Data Analytics

ISYE-6740-OAN

Homework 1

Problem a:

[a-b] Given N data points $\mathrm{x}^n (n = 1, \ldots, N)$, K-means clustering algorithm groups them into K clusters by minimizing the distortion function over $\{r^{nk}, \mu^k\}$

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r^{nk} \|\mathrm{x}^n - \mu^k\|^2,$$

where $r^{nk} = 1$ if x^n belongs to the k-th cluster and $r^{nk} = 0$ otherwise.

Since clusters are independent to each other, setting the derivative of the distortion fucntion w.r.t μ^k to find the minumum

Thus, $2 \sum_{n=1}^{N} r^{nk}(x^n - \mu^k) = 0$

$$\Leftrightarrow \quad \mu^{\boldsymbol{k}} = \frac{\sum_{\boldsymbol{n}} r^{nk} x^n}{\sum_{\boldsymbol{n}} r^{nk}}$$

Problem b: K-mean clustering algorithm is convergence in finite steps because

First, there are at max K^{N} ways to cluster N data points to K clusters, which is a finite number

Second, the distortion function is always decreases after each iteration of K-mean algorithm

Problem c: When we use the bottom-up hierarchical clustering to realize the partition of data, the complete linkage distance metrics would most likely result in cluster most similar to those given by K-mean with second-order Manhattan distance function. Because, in 2nd-order Manhattan distance function, the dominant term is the farthest distance between 2 data points, which is similar to complete linkage.

Problem d: In 2-moon dataset, the single linkage can successfully separate 2 clusters. (This was experimented in Python)