ISYE 6740 Homework 1

Yao Xie

August 19, 2019

1 Clustering. [100 points total. Each part is 25 points.]

[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.

(a) Prove that using the squared Euclidean distance $\|\mathbf{x}^n - \mu^k\|^2$ as the dissimilarity function and minimizing the distortion function, we will have

$$\mu^k = \frac{\sum_n r^{nk}\mathbf{x}^n}{\sum_n r^{nk}}.$$

That is, μ^k is the center of k-th cluster. [5 pts]

(b) Prove that K-means algorithm converges to a local optimum in finite steps. [5 pts]

[c-d] In class, we discussed bottom-up hierarchical clustering. For each iteration, we need to find two clusters $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_m\}$ and $\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_p\}$ with the minimum distance to merge. Some of the most commonly used distance metrics between two clusters are:

- Single linkage: the minimum distance between any pairs of points from the two clusters, i.e.

$$\min_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

- Complete linkage: the maximum distance between any parts of points from the two clusters, i.e.

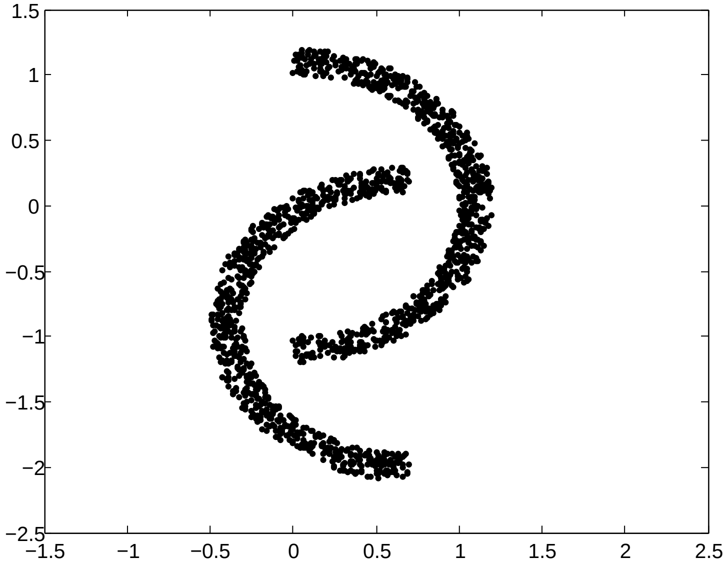
$$\max_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

- Average linkage: the average distance between all pair of points from the two clusters, i.e.

$$\frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} \|\mathrm{x}_i - \mathrm{y}_j\|$$

(c) When we use the bottom up hierarchical clustering to realize the partition of data, which of the three cluster distance metrics described above would most likely result in clusters most similar to those given by K-means? (Suppose K is a power of 2 in this case). [5 pts]

(d) For the following data (two moons), which of these three distance metrics (if any) would successfully separate the two moons? [5 pts]


1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5