Homework 8 Data Analysis

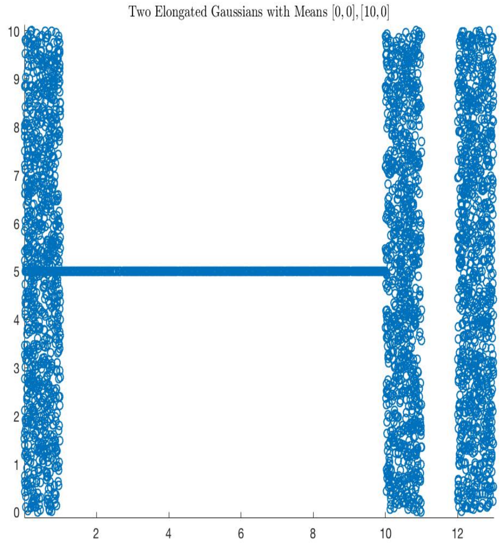
Hanyuan Zhu

November 8, 2018

Question 1

a.

1. Given N data points, $\{x_i | i = 1, 2, ...N\}$ and the distance frunction between Clusters , $\Delta(C_i, C_j)$.
2. $C_{\Delta_t} = \{C_i = \{x_i\} | i = 1, ..., N\}$ Initially each data point is a cluster, so we have a collection of clusters,
3. Iteratively, we merges two clusters with smallest distance, that is for $arg \min_{i,j} \Delta(C_i, C_j)$ then $C_i \cup C_j = C_i$.



b. Complete linkage clustering : it computes all pairwise distance between the elements in cluster 1 and the elements in cluster 2, and considers the largest value (i.e., maximum value) of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters.

Single linkage clustering : It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the smallest of these dissimilarities as a linkage criterion. It tends to produce long, loose clusters.

Question 2

a.

$$L_{sym} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

D is degree matrix, and W is adjacent matrix of graph .

b.

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

Method 1: L has a eigenvalue $\lambda = 0$ and the eigenvector is $Lv = 0$.

$$L_{sym} D^{\frac{1}{2}} = D^{-\frac{1}{2}} L$$

$$L_{sym} D^{\frac{1}{2}} v = D^{-\frac{1}{2}} L v = 0$$

Therefore L_{sym} has eigenvalue 0 with eigenvector $D^{\frac{1}{2}} v$.

Question 3

a. Because of complete unweighted graph, the we have laplacian below,

$$L_{rw} = I - D^{-1} W$$

where $D = \{d_i = n - 1 | i = 1, 2, ...n\}$.

$$(L_{rw})ij = \begin{cases} 1, & i = j \\ \frac{1}{n-1}, & i \neq j \end{cases} \tag{1}$$

b. It has eigenvalue $\frac{n}{n-1}$ with multiplicity n-1, and the last eigenvalue is 0.

c. By part b. we know such K =1 , which means 1 cluster. Since a unweighted complete graph has no real cluster structure inside it, every point is identical.

Question 4

a. For traning part , to find the best parameter n for Knn.

- 1. Compute the distance matrix for testting point respect to all training points.

- 2. For each point in test set, we choose the n nearest point and then to label it by the majority of labels among these n points.
- 3. Compupter the loss function by mislabelling. To minimize the loss function , we ge the optimized n.

To label incoming data

1. Compute the distance matrix for incoming point respect to all labelled points.
2. We choose the n nearest point and then to label it by the majority of labels among these n points.

b. It wont be a useful result, because all test points will have same label, which is the majority from train set.

Question 5

For unsupervise learning part, we can just use DBSCAN, because data intuitively are separated densely in 3 clusters. We just need to choose minPts and δ properly. Let δ be sufficiently large but should be smaller than the gap between those intuitive clusters.

$y_i^{(s)}$ is the lebel of each point $x_i \in C^{(s)}$ from , and l_i is the true Label

We have loss function $loss = \frac{1}{n} \sum_{i=1}^{n} |y_i^{(s)} - l_i|$