Homework 8
MATH 123 - Fall 2018
Tufts University, Department of Mathematics
Due: November 8, 2018

Question 1

(a) Write down the algorithm for an agglomerative hierarchical clustering algorithm with distance function between sets Δ.

(b) Give an example of data where single linkage clustering and complete linkage clustering produce very different dendrograms. Explain.

Question 2

Let $\mathcal{G}$ be a graph with weight matrix $W \in \mathbb{R}^{n \times n}$.

(a) What is the formula for the normalized graph Laplacian L_{sym}?

(b) Prove that L_{sym} is not invertible.

Question 3

Suppose we build a *fully connected* graph in which each node is connected to every other node i.e. $W_{ij} = 1, i \neq j$ and $W_{ii} = 0, i = 1, \ldots, n$.

(a) What is the formula for the random walk graph Laplacian L_{rw} in this case?

(b) What can be said about the eigenvalues $\{\lambda_i\}_{i=1}^n$ of L_{rw}?

(c) What would happen if the number of clusters was estimated with the eigengap statistic $\hat{K} = \text{argmax}_k \lambda_{k+1} - \lambda_k$? Explain if this estimate makes sense or not.

Question 4

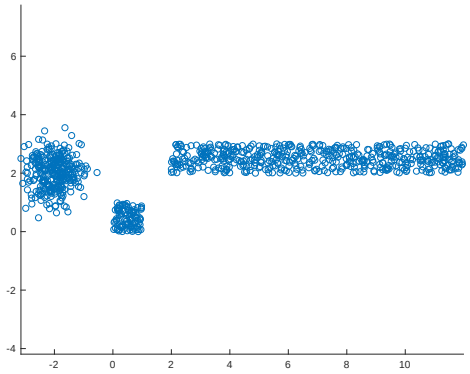
Suppose you are given testing data $\{x_i^*\}_{i=1}^m$ and labeled training data $\{(x_i, y_i)\}_{i=1}^n$.

(a) Write down the algorithm for k_{NN}-classification.

(b) What happens if $k_{\text{NN}} = n$? Is this likely to yield a good or bad result? Explain.

Question 5

Consider the dataset shown below.



How would you cluster this data with the purpose of learning the important structure in the data? Discuss how to set relevant parameters, and how you would validate performance if, after you cluster, you are provided with labels.