Clustering: Linkage Clustering

Given points $x_1, \ldots, x_n$, a clustering mapping c is one that maps each x_t to one of K clusters. For instance, a clustering of the n points given by c might have for instance $c(x_1) = 1, c(x_2) = 3, c(x_3) = 1 \ldots$ meaning that x_1 and x_3 belong to cluster one and x_2 belongs to cluster 3 etc. Also, a given cluster mapping C provides a partition of n points into at most K sets. We will denote $C_k \subseteq [n]$ to be denote indices of all the points belonging to cluster k. Now say we have a function $mathrmdissimilarity$ that measures dissimilarity between two points x_t and x_s as dissimilarity(x_t, x_s).

Question 1: Show that the following two objectives are equivalent:

1. Minimize total within-cluster dissimilarity :

$$M_1 = \sum_{k=1}^{K} \sum_{s,t \in C_k} \text{dissimilarity}(x_t, x_s)$$

2. Maximize total between cluster dissimilarity:

$$M_2 = \sum_{s,t:c(x_s) \neq c(x_t)} \text{dissimilarity}(x_t, x_s)$$

That is show that maximizing M_2 is same as minimizing M_1

Question 2: Let c be the cluster assignment given by single link clustering algorithm. As discussed, single link algorithm works by repeatedly merging the closest two clusters. Specifically on every iteration, distance between two clusters is defined by the shortest distance between pair of points such that one point is in one cluster and the other point is in the second cluster. So on every iteration, the closest two clusters is picked and the points are merged to be in one cluster. On iteration i, if say clusters C and C' are merged, let us define d_i as the distance between between these two clusters, that is

$$d_i = \min_{s \in C, t \in C'} d(x_s, x_t)$$

Show that if c is the cluster assignment returned by single link clustering, then for any i indexing iterations of the single link clustering algorithm run,

$$\min_{t,s:c(x_s) \neq c(x_t)} d(x_s, x_t) \geq d_i$$