# CSCI 6333 Data Mining & Warehousing

**Module 4: Cluster Analysis**

**Homework Assignment Three**

**All problems are equal-weighted with 20 points each.**

1. Estimate the time complexity of the k-means clustering algorithm in terms of the dimension , the number of data points and the number of iterations ?
2. Suppose that a set of -dimensional points has pairwise disjoint clusters . Show how to find the centroids , to minimize the sum of squared errors SSE below

where is the Euclidean distance and .

1. It seems that the k-means clustering algorithm is simple and runs linearly in the size of the data set at each iteration. Explain why the k-mean clustering problem, when regarded as optimization problem to minimize the sum of squared errors, is computationally hard?
2. In hierarchical clustering, a proximity matrix is maintained for all clusters. Initially, each single point is viewed as a cluster. In addition to the proximity matrix, we also maintain a sorted list of all entries in the matrix. Explain in detail at each step of merging two closest clusters, how the proximity matrix can be updated in time? How the sorted list of the matrix entries can be updated in time? Here, is the dimension and is the number of points. Assume we use the MAX distance as the distance between two clusters.
3. Given two parameters Eps and MinPts, give a simple algorithm to find all core points, border points and noise points. Estimate the time complexity of your algorithm in terms of dimension and the number of points . Can you find a more efficient algorithm?