CSCI 5080 Assignment 7

Use Microsoft **WORD** to write your answer and submit it to the Dropbox in D2L.

Please indicate **how much time** you spend on each problem.

Part I (70 points)

1. (30 points) Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are

$$A_1(2, 10), A_2(2, 5), A_3(8, 4), B_1(5, 8), B_2(7, 5), B_3(6, 4), C_1(1, 2), C_2(4, 9).$$

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*

1) (15 Points) the three cluster centers after the first round of execution.

2) (15 Points) the final three clusters.

2. (40 points) The support vector machine (SVM) is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with a large set of data tuples. Read the handout on the micro-clustering technique in "Classifying large data sets using SVM with hierarchical clusters" by Yu, Yang, and Han, in Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03), and describe how to overcome this difficulty with a scalable Cluster-Based SVM (CB-SVM) method for efficient SVM classification in large datasets.

Part II (30 Points)

1. Review SQL Server Data Mining Algorithms.
   http://technet.microsoft.com/en-us/library/ms175595(v=sql.110).aspx

   Study Microsoft Clustering Algorithm
   http://technet.microsoft.com/en-us/library/ms174879(v=sql.110).aspx

2. Complete the clustering model of SQL Server 2012 Basic Data Mining Tutorial Lesson 3 through Lesson 6.
   http://technet.microsoft.com/en-us/library/ms167167(v=sql.110).aspx

   Turn in completed projects for lesson 3 and lesson 4 only.