CS 378 **Introduction to Data Mining** **Spring 2009**

Homework 5

Instructor: Inderjit Dhillon Date Due: April 16, 2009

Keywords: *K-means, Agglomerative Clustering*

1. (15 points) K-means

   Step 1: Implement the k-means algorithm. Your implementation should initialize points by assigning each point to a cluster at random. Your implementation should iterate until the algorithm converges (i.e. no cluster assignments change from one iteration to the next). Submit the documented source code of your implementation.

   Step 2: Cluster the Iris dataset that you used in the second homework. Run your algorithm and compute the accuracy of your clustering. Unlike classification, the cluster ids of your k-means clusters will in general not correspond with the class labels. You must first compute the optimal mapping of cluster ids to class labels (for example, associate class 1 with cluster 3, class 2 with cluster 1, and class 3 with cluster 2). You should also compute the confusion matrix for your clustering results. Submit results for one sample run giving the initial clustering, final clusterings, and confusion matrix.

   (The Iris dataset is at `http://www.cs.utexas.edu/~wtang/cs378/iris.tar.gz`)

   Step 3: Run your algorithm 50 times over the Iris dataset, each time using a different initialization. For each run, compute both the accuracy of clustering, as well as the k-means objective function value of the final clustering. Plot the distribution of these accuracies and also plot the distribution of the objective function values. Finally, give a scatter plot showing the correlation (if any) between the clustering accuracy and the clustering objective function value. Write a paragraph to interpret these graphs - what can you conclude?

2. (15 points) Agglomerative Clustering

   Step 1: Implement single-link agglomerative clustering using Euclidean distance. Your implementation should take in the number of agglomerated clusters k as an input argument. Your algorithm should run until there are only k agglomerated clusters. Submit documented code.

   Step 2: Run your algorithm on the Iris dataset for $k = 3$. Compute the accuracy of the clustering. How does the accuracy compare to k-means?

   Step 3: Give the theoretical running time for both the agglomerative clustering algorithm as well as k-means. Also, compute the observed running time of a sample run over the Iris dataset of each algorithm.