**Large Scale Data Analysis I**

**Problem Set 2 (Clustering)**

**Due 3/10/2020, 4:55pm**

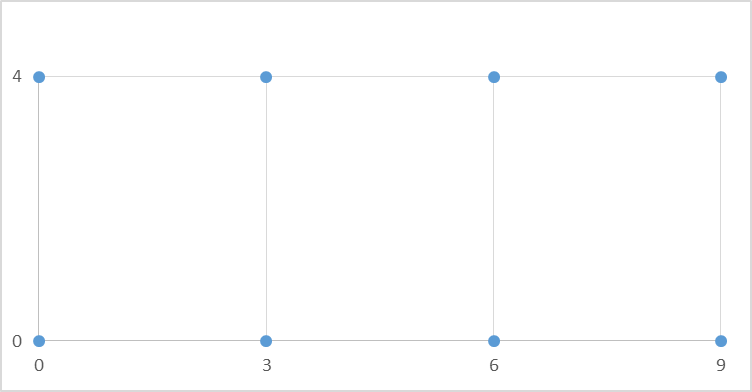
Your name:

Your e-mail address:

This assignment is worth a total of 48 points (24 + 24).

**Question 1: Clustering Methods Compared**

Given eight points in two-dimensional space, assuming a Euclidean distance metric and the following coordinates: ((0,0), (3,0), (6,0), (9,0), (0,4), (3,4), (6,4), (9,4)).



1) List a sequence of merges that could occur if you use bottom-up hierarchical clustering with the **single link** distance metric (stop when you have a single cluster; you can break ties in whatever order you’d like). (4 points)

2) If you use this approach to form k = 2 clusters, what would the resulting clusters be? (2 points) Calculate the cluster means and the distortion of this clustering. (2 points)

3) List a sequence of merges that could occur if you use bottom-up hierarchical clustering with the **complete link** distance metric (stop when you have a single cluster; you can break ties in whatever order you’d like). (4 points)

4) If you use this approach to form k = 2 clusters, what would the resulting clusters be? (2 pts)

Calculate the cluster means and the distortion of this clustering (2 pts).

[Note: there are multiple correct answers to this question, depending on the sequence of merges you chose in step 3.]

5) If you use k-means with k = 2 and initial centers of cluster 1: (0,4) and cluster 2: (9, 0), list the resulting sequence of steps as follows [use only as many iterations as necessary]: (4 pts)

Iteration 1:

Assign points \_\_\_\_\_\_\_\_\_\_\_\_\_\_ to cluster 1 and points \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ to cluster 2.

Move center of cluster 1 to \_\_\_\_\_\_\_\_\_ and center of cluster 2 to \_\_\_\_\_\_\_\_\_\_\_\_\_.

Iteration 2 (fill in only if necessary):

Assign points \_\_\_\_\_\_\_\_\_\_\_\_\_\_ to cluster 1 and points \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ to cluster 2.

Move center of cluster 1 to \_\_\_\_\_\_\_\_\_ and center of cluster 2 to \_\_\_\_\_\_\_\_\_\_\_\_\_.

….

6) Is it possible for k-means to get stuck in a suboptimal solution (as measured by distortion) for clustering these eight points into k = 2 clusters? If so, provide the initial centers that would result in the suboptimal solution, and calculate the distortion. (4 pts)

**Question 2: Spatial and Temporal Distributions of Chicago Crimes**

In this question, you will use k-means clustering in Weka to answer the question, “Do different types of crime display different trends over space and time?” The dataset “LSDA Chicago Crimes for HW 2.csv” consists of data for 119 different types of crime, each of which occurred at least 100 times in Chicago during the year 2016. For each crime type, we have various features representing the spatial and temporal distribution of crime, including:

* The proportion of all crimes of that type that occurred on each day of the week (day\_Sun, day\_Mon, …, day\_Sat).
* The proportion of all crimes of that type that occurred on each hour of the day (hour\_0 = midnight to 12:59am, hour\_1 = 1am to 1:59am, …, hour\_23 = 11pm to 11:59pm).
* The proportion of all crime of that type that occurred in each of the 77 community areas of Chicago (community\_area\_1 … community\_area\_77).

We also have, for each crime type, its categorization by the FBI:

* Category = “P1V” corresponds to Part 1 Violent Crime, i.e., serious violent crimes
* Category = “P1P” corresponds to Part 1 Property Crime, i.e., serious property crimes
* Category = “P2” corresponds to Part 2 (less serious) crimes.

For parts a-d, you should cluster the 119 crime types using k-means into k = 3 clusters using only the **hour of day** attributes (you can do this by changing the parameters of the distance function; also set dontNormalize = TRUE). Also change initializationMethod to “Farthest first”, preserveInstancesOrder to TRUE, and keep all other parameter settings at their default values.

a) Copy each cluster’s mean values for hour\_0…hour\_23 into an Excel spreadsheet and create a line graph to visualize these values by cluster. (3 pts)

b) Describe the three different hour-of-day trends represented by these three clusters (3 pts).

c) Do you notice any consistent trends about which crime types are assigned to which cluster? (Hint: this is easiest to see if you set displayStdDevs to TRUE, while part a) is easiest if you set displayStdDevs to FALSE). (3 pts)

d) Do the three clusters have different day-of-week trends as well? Do they have different spatial trends? (4 pts)

e) How well do the three groups formed by clustering hour-of-day trends correspond to the FBI’s division between P1V, P1P, and P2 crimes? To see this, re-run the same analysis, but using “Classes to clusters evaluation” (with Category as the class variable) instead of “Use training set”. Note that the resulting clusters may be a bit different than before (just by chance). Report the proportion of incorrectly clustered instances. (3 pts)

f) Next, let’s compare the clusters produced by EM to those produced by k-means. Ignore (or remove) all attributes except the crime\_type and hour of day attributes. Run EM with the default parameter settings, allowing it to choose the number of clusters. How many clusters does it produce? How do the clusters compare with the three clusters produced by k-means in part a-d? (4 pts)

g) Finally, try bottom-up hierarchical clustering, again adjusting your distance metric to only use the hour of day attributes, and using the number of clusters suggested by EM. Which works better, single-link or complete-link clustering? Describe the clusters found in each case. (4 pts)