Assignment 9

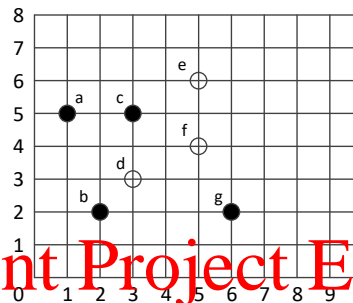**Due**: 4/8

Note: Show all your work.

Problem 1 (10 points). The k-means algorithm is being run on a small dataset and. After a certain number of iterations, we have two clusters as shown in the figure. Here, filled circles are Cluster1 objects and clear circles are Cluster2 objects.

Cluster1: {a, b, c, g}
Cluster2: {d, e, f}

Run two more iterations of the k-Means clustering algorithm and show the two clusters at the end of each iteration. You don't need to draw figures like above. It is sufficient that you indicate which objects belong to each cluster at the end of each iteration. Again, show all your work and use Manhattan distance when calculating distances. Note that this is not the beginning of the execution of k-means. You are in the middle of the execution of k-means. So, the first thing you need to do is to compute new centroids of all clusters.

Problem 2 (10 points). Consider the following two clusters:
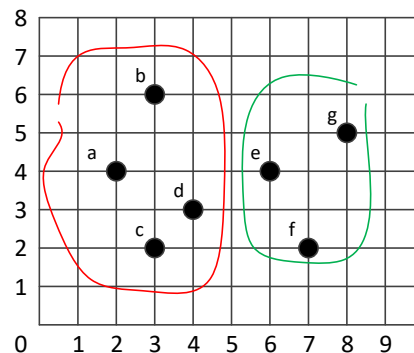
Compute the distance between the two clusters (1) using minimum distance and (2) using average distance. Use the Manhattan distance measure.

Problem 3 (20 points). Use the provided *a9-p3.arff* dataset for this problem. This dataset has calories and total fat content of 75 candy bars.

Problem 3-1 Run the *SimpleKMeans* algorithm of Weka on this dataset with *k* = 2, 3, 4, 5, 6, and 7. For each *k*, record the value of *within cluster sum of squared errors* (which you can find in Weka's cluster output window) and plot a graph where the x-axis is *k* and y-axis is *within cluster sum of squared errors*. Then, determine an optimal number of clusters using the *elbow method* which we discussed in the class (it is also described in page 486 of the textbook).

Problem 3-2 Using the optimal number of clusters which you determined in Problem 3-1, run *SimpleKMeans* again and characterize the generated clusters using the two attribute values. The following is an example of characterization of clusters:

Cluster 0:
  * Calories is mostly between 1000 and 2000, mean of Calories is 1500
  * Totalfat is mostly between 10 and 20, mean of Totalfat is 14

Cluster 1:
  * Calories is mostly between 2000 and 3000, mean of Calories is 2600
  * Totalfat is mostly between 15 and 25, mean of Totalfat is 20

…

Problem 4 (10 points). Follow the instructions in *JMP-clustering-assignment.pdf* file. Include the required screenshots and your answers to some questions in your submission.

Submission:

Include all answers in a single file and name it *lastName_firstName*_HW9.EXT. Here, "EXT" is an appropriate file extension (e.g., docx or pdf). If you have multiple files, then combine all files into a single archive file. Name the archive file *lastName_firstName*_HW9.EXT. Here, "EXT" is an appropriate archive file extension (e.g., zip or rar). Upload the file to Blackboard.