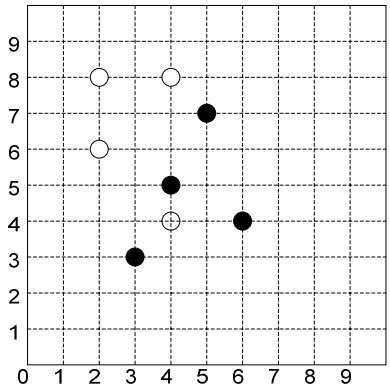
Assignment 9

Due: 4/10

Note: Show all your work.

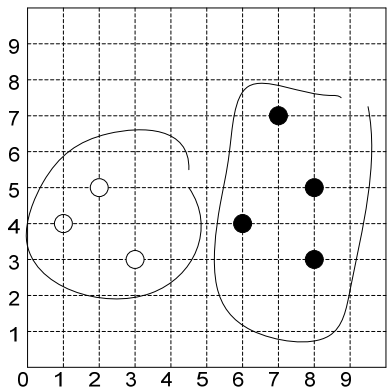
Problem 1 (10 points). The k-means algorithm is being run on a small dataset and. After a certain number of iterations, we have two clusters as shown in the figure. Here, clear circles are Cluster1 objects and filled circles are Cluster2 objects.



Cluster1: (2,6) (2,8) (4,4) (4,8)
Cluster2: (3,3) (4,5) (5,7) (6,4)

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 running of k-means. You are in the middle of the running 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. These distance measures are defined in page 461 of the textbook. Use the Manhattan distance measure.

Problem 3 (10 points). Use the provided *a9-p3.arff* dataset for this problem.

Problem 3-1 Run the *SimpleKMeans* algorithm of Weka on this dataset with k = 2, 3, 4, 5, and 6. 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 is described in page 486 of the textbook (it is also described in Module 6 slides).

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. For example, if two attributes were age and income, characterization of clusters would look like:

> Cluster 0: Mostly younger than 21 and income between 15K and 35K
> Cluster 1: Mostly ages between 21 and 45 and income between 35K and 90K
> …

Problem 4 (10 points). This problem has two sections. Problem 4-1 is for Oracle and Problem 4-2 is for JMP. Choose one of the two.

Problem 4-1 (Oracle)

Follow the instructions in *oracle-clustering-assignment.pdf* file. The submission requirements are indicated with **"Required."**

Problem 4-2 (JMP)

Follow the instructions in *JMP-clustering-assignment.pdf* file. Include the required screenshots and your answers to some questions in your submission.

Submission:

Submit all solutions in a single Word or PDF document and upload it to course assignments section. Please make sure that there are no spaces in the file name. Use *lastName_firstName*_HW9.doc or *lastName_firstName*_HW9.pdf as the file name.