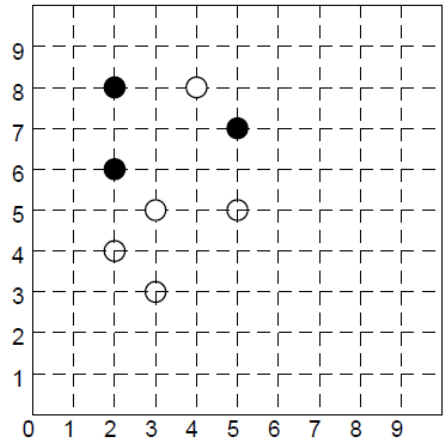
Lab 10

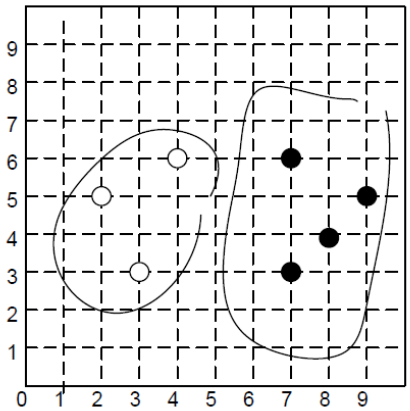
Problem 1. The k-means algorithm is being run on a small dataset and. After a certain number of iterations, we have two clusters as shown in the figure. Here, clear circles are Cluster1 objects and filled circles are Cluster2 objects.



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

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 kmeans. So, the first thing you need to do is to compute new centroids of two clusters.

Problem 2. Consider the following two clusters:



Compute the distance between the two clusters (1) using minimum distance and (2) using mean distance. Use the Manhattan distance measure.

Problem 3. Use the provided *a6-p3.arff* dataset for this problem. It has 452 instances and 2 attributes.

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.*

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
…