Name: Vira Oleksyuk

Problem	Points	Score
1(a)	20	
1(b)	10	
1(c)	10	
1(d)	10	
2(a)	20	
2(b)	10	
2(c)	10	
2(d)	10	
Total	100	

Notes:

(1) The exam is closed books and notes except for one double-sided sheet of notes.

(2) Please indicate clearly your answer to the problem.

(3) If I can't read or follow your solution, it is wrong and no partial credit will be awarded.

Problem 1

Consider 5 data points: (-1,0), (0,1) – Class 1, (1,0), (0,-1), (-1/2,1/2) – Class 2. Proceed with K-MEANS clustering.

a) Assume your initial guess for two cluster centers are (0,0) and (0.5,0.5). Execute an iteration of k-means by computing the new cluster centers and assigning the data points to the correct cluster.

Figure 1 shows 5 given data points and two means as first quesses.

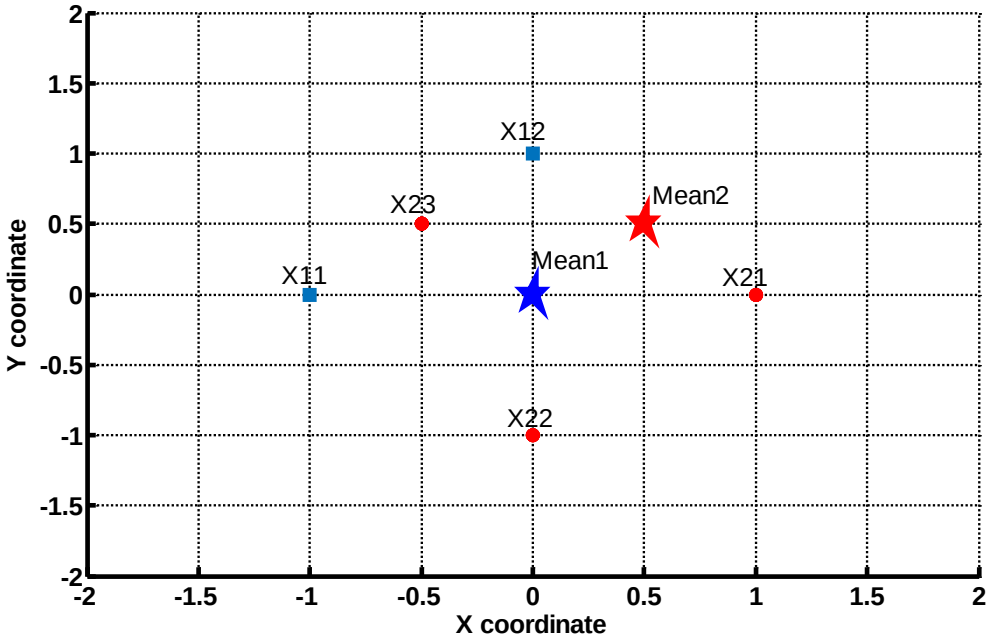

Figure 1. Data points corresponding to two classes and a first guess for their means

The simplest way to differentiate points on two classes will be to do the following. Connect two means with a line. Bisect this line with another line (Figure 2). The points above of the second line belong to one class, and points below to the line belong to another class.

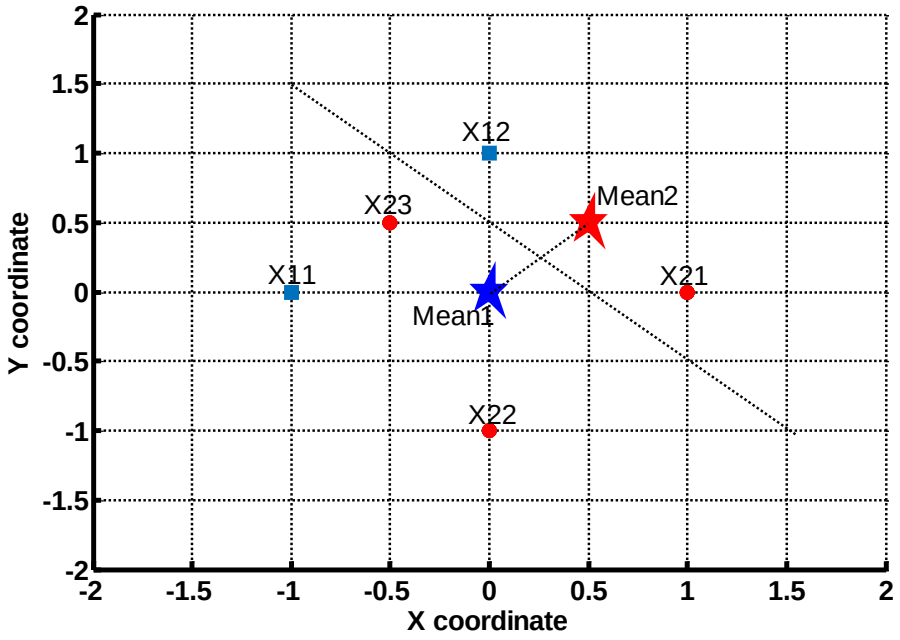

Figure 2. Clustering using initial means

Next, we recalculate the means for new clusters. One mean at (-0.5,-0.167) will be for the points X11, X23, X22, and another mean at (0.5,0.5) will be for X12 and X21 (Figure 3).

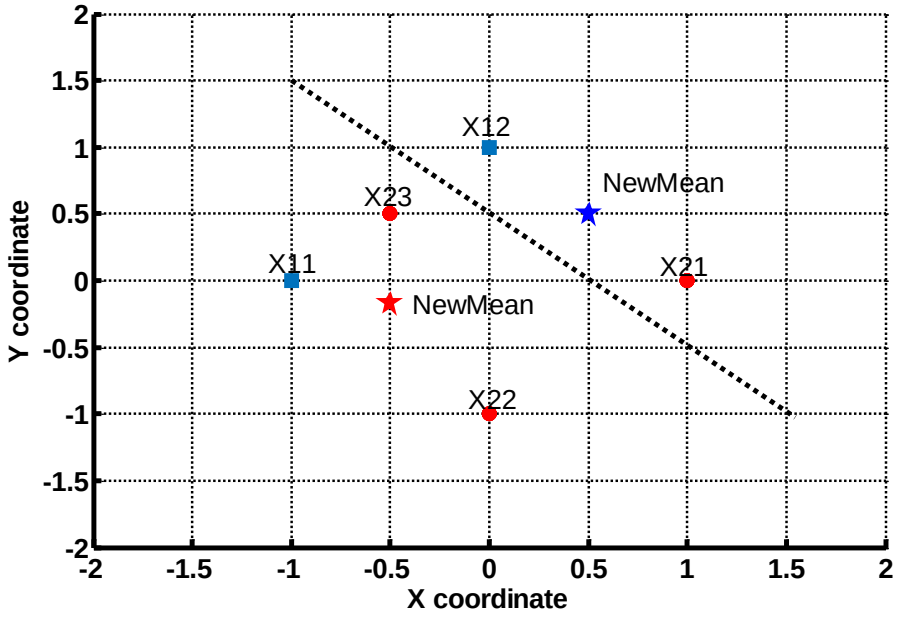

Figure 3. Recalculated means

Next, we continue with connecting of two new means and recalculating the decision surface F (Figure 4).

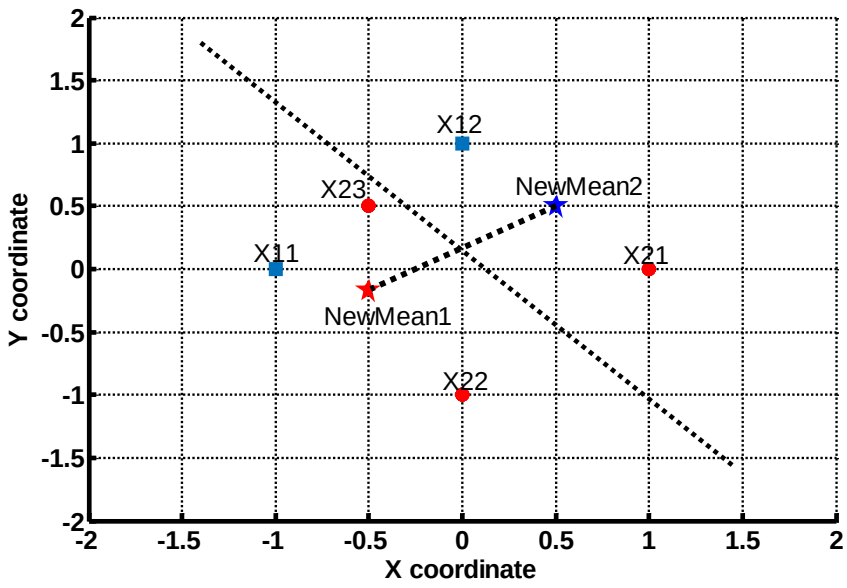

Figure 4. Recalculated means and decision surface F

b) **Assign identity to each cluster based on the majority-voting scheme and draw a ML decision surface.** Using the majority of votes, we assign classes to new clusters. In our case the means swapped between classes (Figure 5).

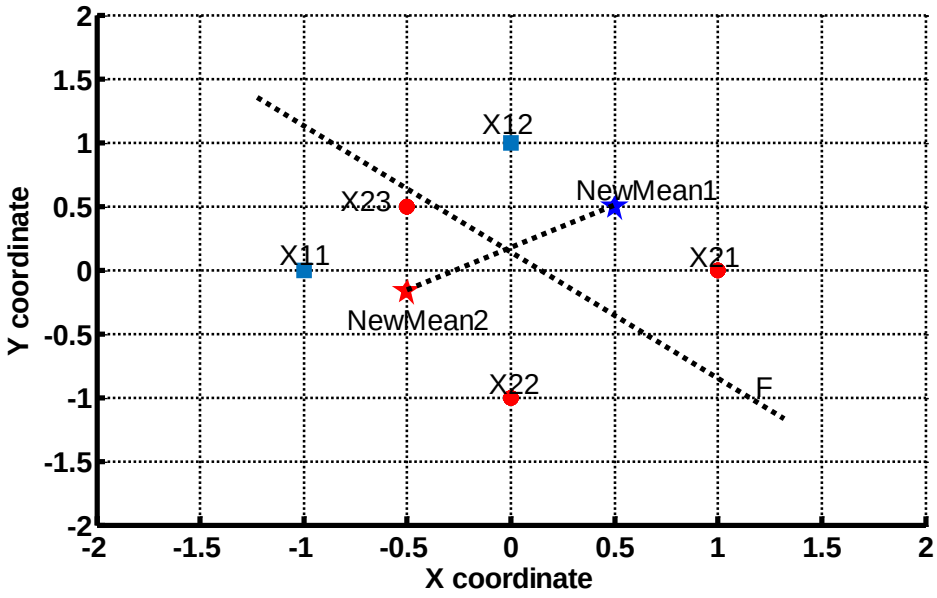

Figure 5. Swapped means with the maximum likelihood decision surface F

c) **Consider two test data points (-3/4;3/4) which belongs to class 1, and (1/2,1/2), which belongs to class 2.** Figure 6 shows test data points and the training data points.

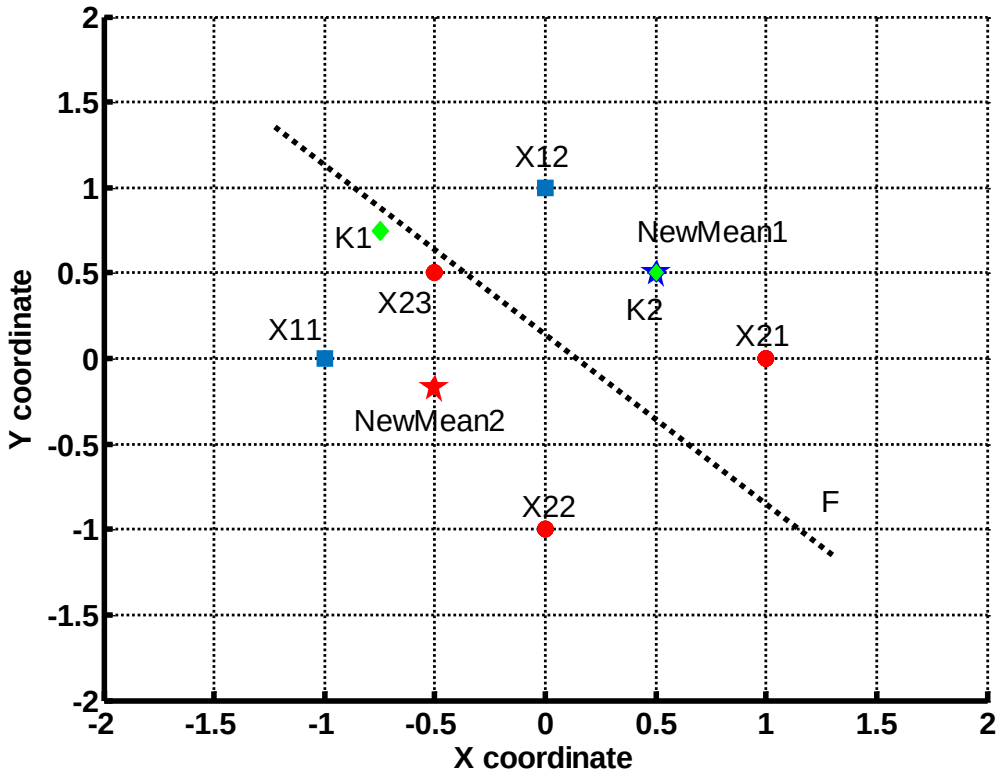

Figure 6. Visualization of the K1 and K2 test data points

d) **Compute the probability of error based on k-nearest neighbor rule. How different should this result be from (c) for large *k*.** Table 1 shows the results for the calculations.

Table 1. Results of classification with K-MEANS and k-nearest neighbor classification

Method	**Misclassified test points**	**Probability of error**
K-means clustering	2	1
A nearest neighbor rule	2 or 1	0.5~1
Two nearest neighbors rule	2 or 0	0~1
Three nearest neighbors rule	**0**	**0**
Four nearest neighbors rule	2 or 0	0~1
Five nearest neighbors rule	1	0.5

For the small *k* in the k-NN rule, the classification will have large error. For the k-means clustering and for a NN classifier, we misclassified both of test data points, which draw the error to 100%. With increase of *k* in the k-nearest neighbors rule, error improves. When in our case k=3, the misclassification error is 0%, yet it degrades and reaches 50% with further increase of k. This caused by constant misclassification of one of the test points. So the best classifier for this data set among proposed methods will be k=3 NN rule.

Problem 2

a) Consider 5 data points: X11(0,1), X12(-1,0) – Class 1, X21(1,0), X22 (0,-1), X23(-1/2,1/2) – Class 2. Construct a dendrogram for the data. Dendrogram presented in Fig.7. This is the bottom-up or agglomerative decision tree. Fig. 7 built using MATLAB tools. A dendrogram consists of many U-shaped lines that connect data points in a hierarchical tree. The height of each U represents the distance between the two data points being connected. Each leaf in the dendrogram corresponds to one data point. The agglomerative tree.

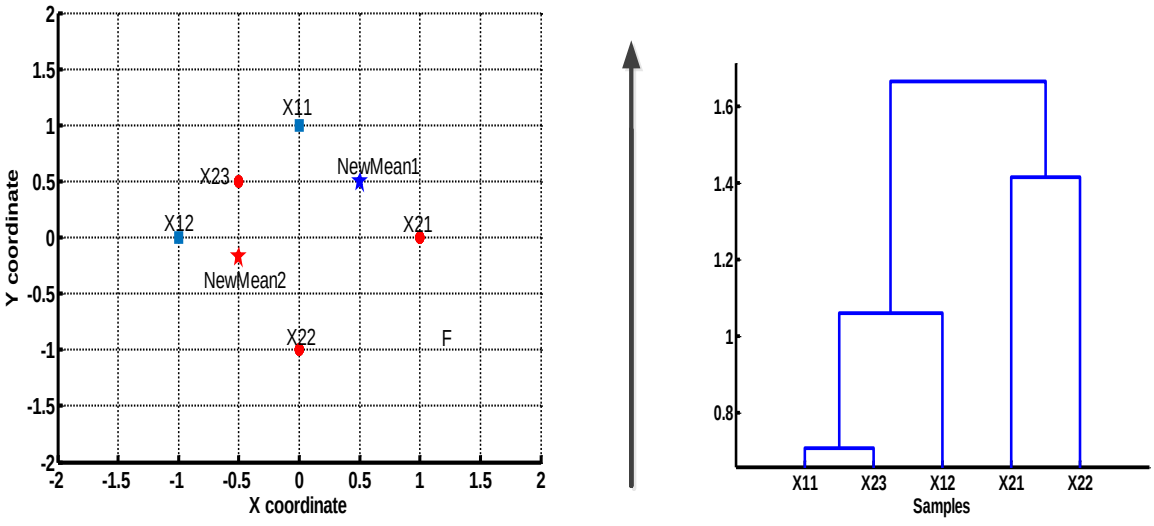

Figure 7. Data points and a dendrogram with average linkage (Unweighted average distance (UPGMA))

b) Construct a top-down clustering (based on k-mean). Figure 8 show the results for top-down clustering using k-mean.

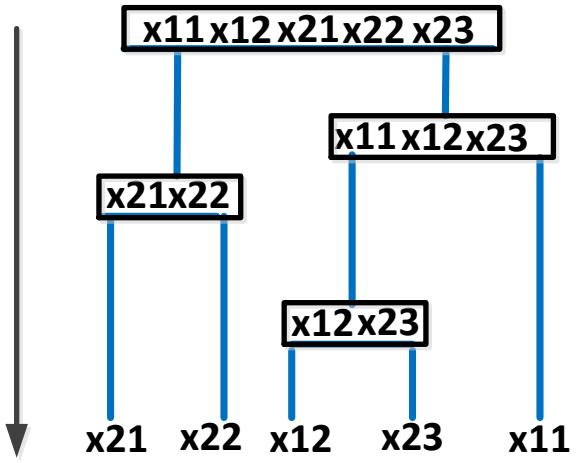

Figure 8. Top-down clustering using k-mean

Top-down clustering is a more complex than down-up procedure, yet it tends to be more accurate. It benefits from the learning of the global distribution.

c) If you were to use your dendogram to do unsupervised clustering of the data, what clusters would you create? (mean and elements)

I would build a top-down tree, shown in Fig.9

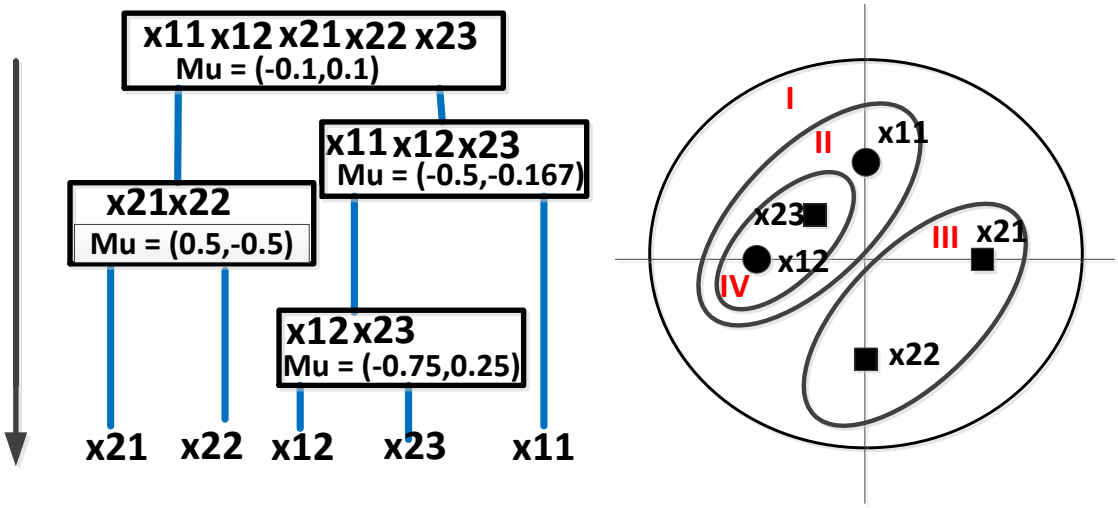

Figure 9. Dendrogram and clusters corresponding to the given data.

d) Suppose (0,1) and (1,0) occur 5 times more often than the rest of the data points. How would you adjust your strategy for clustering the data? How would it impact the decision regions?

If those points will be repeated multiple times, it will increase priors for a particular cluster. In our case, the given repetitive points belong to two different clusters. The cluster with X21 points and X22 will be unchanged. However, the cluster with X23, X11 repetitive points, and X12 will have some corrections. Point X23 will first more likely cluster with X11, than with X12 as shown on Fig.10. In our case, the decision regions will just slightly change to a particular path in clustering, yet in the case of large data sets it may lead to significant changes in decision regions.

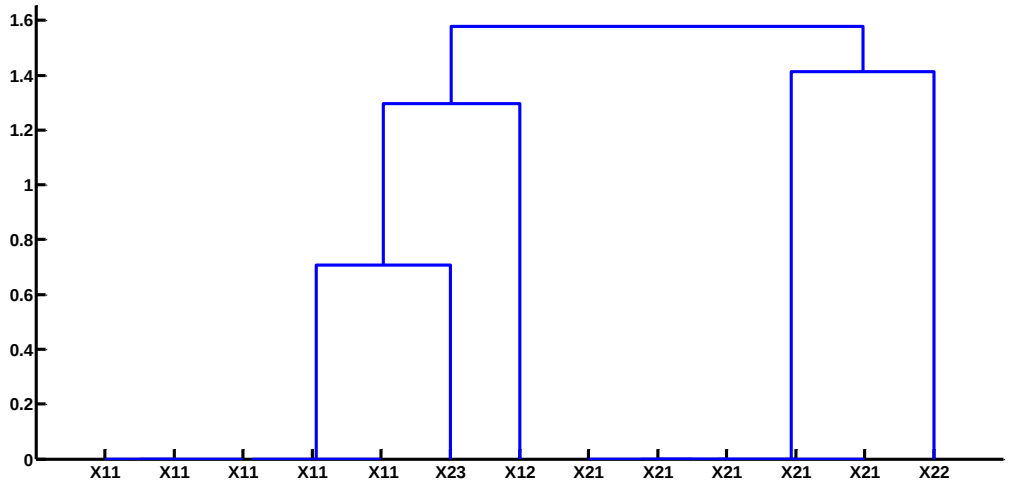

Figure 10. Clustering with repetitive data points X11 and X21.

References

[1] Retrieved from: http://cgm.cs.mcgill.ca/~soss/cs644/projects/simard/nn_prob_err.html