ECE 8110 Machine Learning Midterm 3

May 1, 2014

Salvatore Giorgi
salvatore.giorgi@temple.edu

Exercise 1. *Consider 5 data points*

$$\{(0,1),(-1,0)\} \in \textit{ Class 1},$$

$$\{(1,0),(0,-1),(-\frac{1}{2},\frac{1}{2})\} \in \textit{ Class 2}.$$

In this problem we are going to walk through the K-MEANS clustering process.

(a) Assume your initial guesses for two cluster centers are

$$\mu_1 = (0,0), \quad \mu_2 = (\frac{1}{2},\frac{1}{2}).$$

Execute an iteration of K-MEANS by computing the new cluster centers and assigning the data points to the correct cluster. Use averaging to compute the new cluster center.

(b) Assign an identity to each cluster based on a majority-voting scheme and draw the maximum likelihood decision surface.

(c) Consider two test data points:

$$(-\frac{3}{4},\frac{3}{4}) \in \textit{ Class 1}, \quad (\frac{1}{2},\frac{1}{2}) \in \textit{ Class 2}.$$

Compute the probability of error based on your K-MEANS clustering.

(d) Compute the probability of error based on a k-nearest neighbor rule. How different should this result be from (c) for large k?

Solution for (a): For two data points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ we consider the following distance metric:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$

We calculate the distance between each data point and both μ_1 and μ_2 and compare the distances:

$$d((0,1),\mu_1) > d((0,1),\mu_2),$$

$$d((-1,0),\mu_1) < d((-1,0),\mu_2),$$

$$d((1,0),\mu_1) > d((1,0),\mu_2),$$

$$d((0,-1),\mu_1) < d((0,-1),\mu_2),$$

$$d((-\frac{1}{2},\frac{1}{2}),\mu_1) < d((-\frac{1}{2},\frac{1}{2}),\mu_2).$$

From this, we get the following clusters:

$$\text{Cluster 1 } = \{(-1,0),(0,-1),(-\frac{1}{2},\frac{1}{2})\}, \quad \text{Cluster 2 } = \{(0,1),(1,0)\}$$

From this, we compute the new means $\hat{\mu}_1, \hat{\mu}_2$ by averaging the data in each cluster:

$$\hat{\mu}_1 = \big(\frac{1}{3}(-1+0-\frac{1}{2}), \frac{1}{3}(0-1+\frac{1}{2})\big) = (-\frac{1}{2}, -\frac{1}{6}),$$

$$\hat{\mu}_2 = \big(\frac{1}{2}(0+1), \frac{1}{2}(1+0)\big) = (\frac{1}{2}, \frac{1}{2}).$$

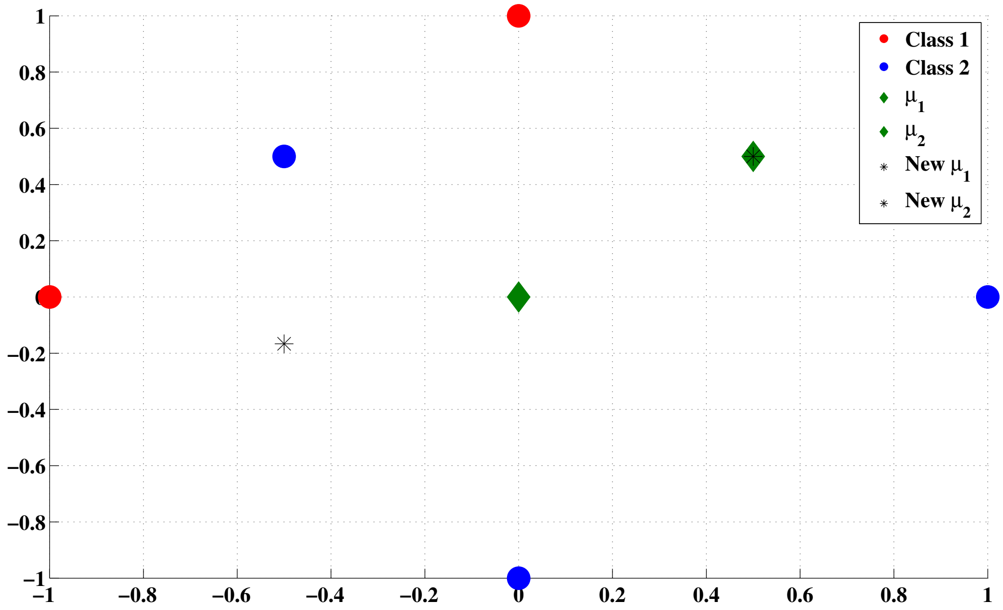

Figure 1: Data with μ_1, μ_2 and $\hat{\mu}_1, \hat{\mu}_2$

Solution for (b):

Since Cluster 1 contains two data points from Class 2 and only one data point from Class 1, from our majority voting scheme, we assign the data in Cluster 1 to Class 2. Therefore, we assign the data from Cluster 2 to Class 1. This is seen in Fig. 2, with the maximum likelihood decision surface drawn. We found this line by calculating the line between $\hat{\mu}_1$ and $\hat{\mu}_2$ and then finding the perpendicular line that passes through its center.

Solution for (c):

Next we consider the new data:

$$(-\frac{3}{4}, \frac{3}{4}) \in \text{ Class } 1, \quad (\frac{1}{2}, \frac{1}{2}) \in \text{ Class } 2.$$

As we can see in Fig. 3, based on our decision surface, both data points will be incorrectly classified.

Solution for (d):

We consider $k \in \{1, \cdots, 5\}$ and compute the probability of error for a kNN classifier using majority voting. We see that when $k = 1$ both points are incorrectly classified. When $k = 2$, the point $(\frac{1}{2}, \frac{1}{2})$ will be classified as Class 1. The other point is closest to one point from Class 2, and the next closest point is equally likely from either Class. Thus, one point is always misclassified

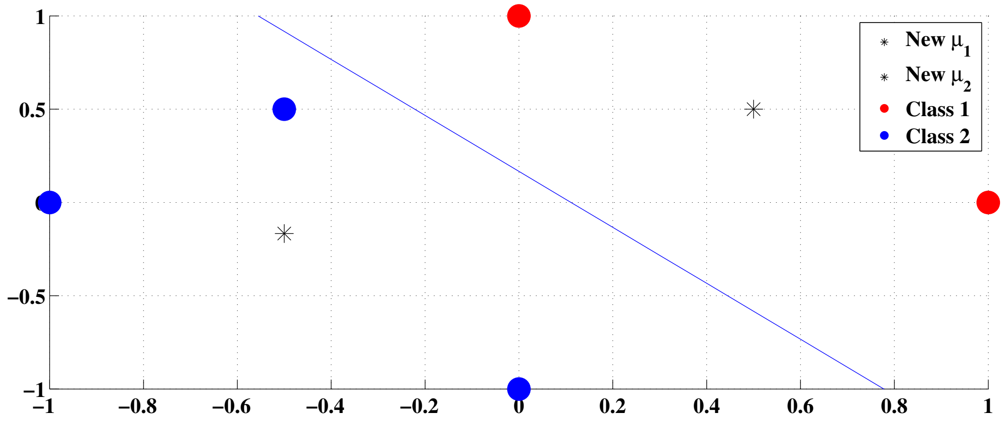

Figure 2: New Classes with $\hat{\mu}_1, \hat{\mu}_2$

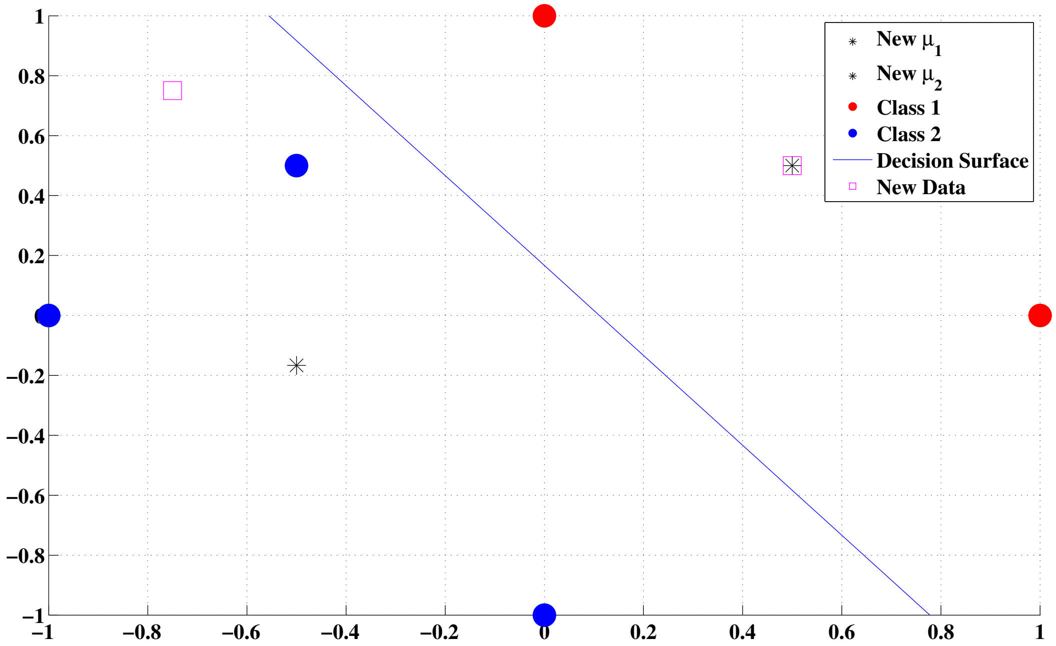

Figure 3: New Data with $\hat{\mu}_1, \hat{\mu}_2$

and another has a 50% of being misclassified. So the total probability of error is 25%. For $k = 3$ both data points will be misclassified. For $k = 4$, the point $(\frac{1}{2}, \frac{1}{2})$ is closest to two points from each class, giving us a 50% chance of misclassifying. The other point is closest to 2 points from Class 2 and 1 point from Class 1. For the remaining point, we are equally far from another point in either class. In one situation we misclassify and in the other we have a 50% chance of misclassifying. This gives us a probability of error = $0.5(0.75 + 0.5) = 0.625$. Finally, for $k = 5$, we classify both points as Class 2, since this has the most points. Thus, we have a 50% probability of error.

Table 1: Predicted and Calculated Error Rates

k	Probability of Error
1	1
2	0.25
3	1
4	0 .625
5	0.50

For kNN, with a small number of data points, we expect point closer to the decision region to be misclassified, since they are more likely to be closer to data in the wrong class. With K-MEANS we are using the centers of the clusters to classify the data, and we expect our classifications to not be swayed by points close to the decision region.

Exercise 2. *Consider the same 5 data points above.*

(a) Construct a dendogram for the data.

(b) Construct a top-down clustering (e.g., LBG) clustering (you can also think of this as a crude decision tree).

(c) If you were to use your dendogram to do unsupervised clustering of the data, what clusters would you create (specify them by the mean and the elements associated with the cluster).

(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 that impact your decision regions?

Solution for (a):

To compute the dendogram for the data, we use Matlab's built in function $dendogram$:

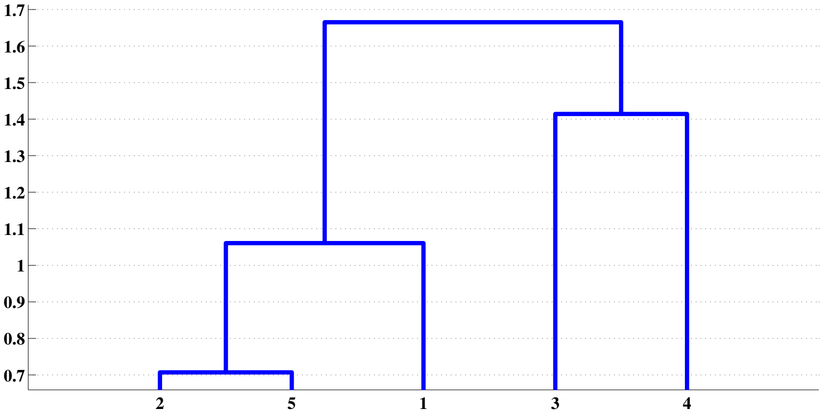

Figure 4: Dendrogram for Original Data

Solution for (b):

For the LBG algorithm, we make initial guesses for the Class means, which we assume to be the same initial guesses as in Problem 1:

$$\mu_1 = (0,0), \quad \mu_2 = (\frac{1}{2}, \frac{1}{2}).$$

Computing the distance of each point from these two means, we get the following clusters

$$\text{Cluster 1} \ = \{(-1,0),(0,-1),(-\frac{1}{2},\frac{1}{2})\}, \quad \text{Cluster 2} \ = \{(0,1),(1,0)\}$$

and assign the following class assignments:

$$\text{Cluster 1} \ = \text{Class 1} \ , \quad \text{Cluster 2} \ = \text{Cluster 2}.$$

When we recompute the new means based on centroids of these clusters, we will get the same $\hat{\mu}_1, \hat{\mu}_2$ as above, except we have different class assignments with this algorithm:

$$\hat{\mu}_1 = (-\frac{1}{2}, -\frac{1}{6}) \in \text{Class 1},$$

$$\hat{\mu}_2 = (\frac{1}{2}, \frac{1}{2}) \in \text{Class } 2.$$

Solution for (c):

From the above dendrogram, we get the following Classes:

$$\begin{aligned} \text{Class } 1 &= \{(0,1), (-1,0), (-\frac{1}{2}, \frac{1}{2})\}, \\ \text{Class } 2 &= \{(1,0), (0,-1)\}. \end{aligned}$$

Computing the means by averaging we get

$$\begin{aligned} \mu_1 &= (-\frac{1}{2}, \frac{1}{2}), \\ \mu_2 &= (\frac{1}{2}, -\frac{1}{2}). \end{aligned}$$

Solution for (d):

Since we are classifying according to data distance, we don't expect the extra data to change the class. What we do expect is that this extra data will change our priors. As seen above, often we will have points that are the same distance from points in multiple classes. With no prior information, we can choose either point and get the same probability of error. In this situation, if one of those points occurs multiple times, we can assume that data point is more likely to occur.

Constructing the dendrogram for the augmented data, as seen in Fig. 5, we see that we essentially have the same classifier. The classes $(1,2,5)$ and $(3,4)$ in Fig. 4 correspond to $\big((1,2,3,4,5),6,13\big)$ and $\big((7,8,9,10,11),12\big)$ in Fig. 5.

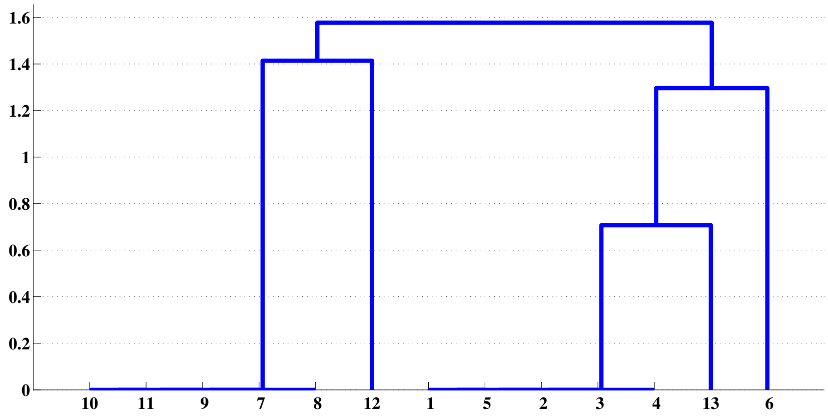

Figure 5: Dendrogram for Augmented Data