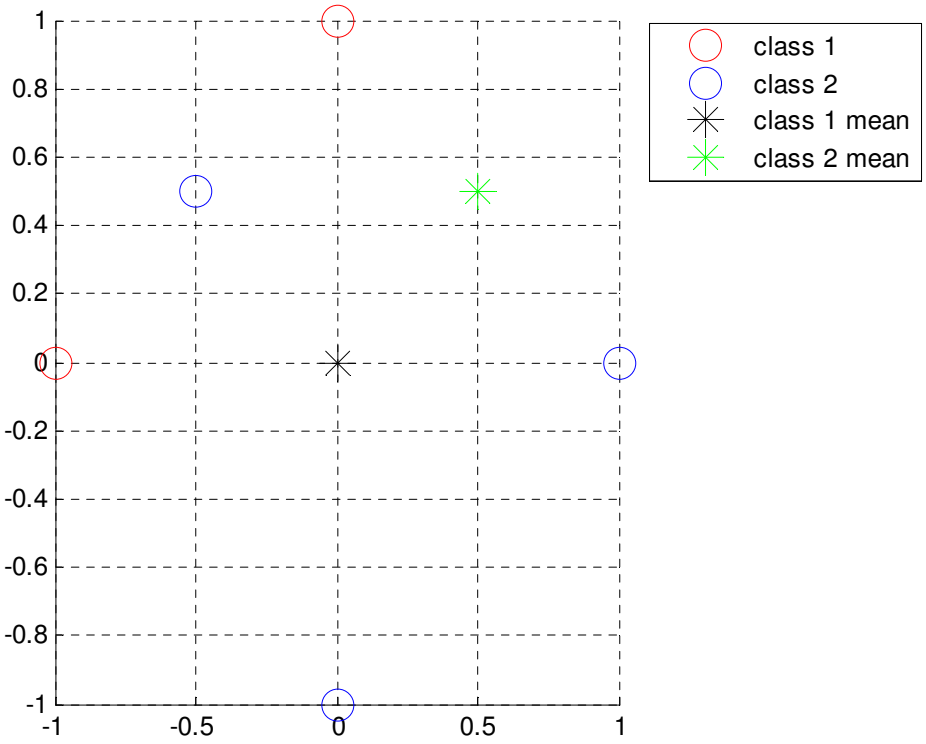
Exam Three Rework

One. Given data points (0,1),(-1,0) for class 1 and (1,0),(0,-1),(-1/2,1/2) for class 2.

One-A. Initial guess of cluster centers A: (0,0) and B: (1/2,1/2) execute an iteration of K-MEANS. Our initial distances are as follows, using mean squared:

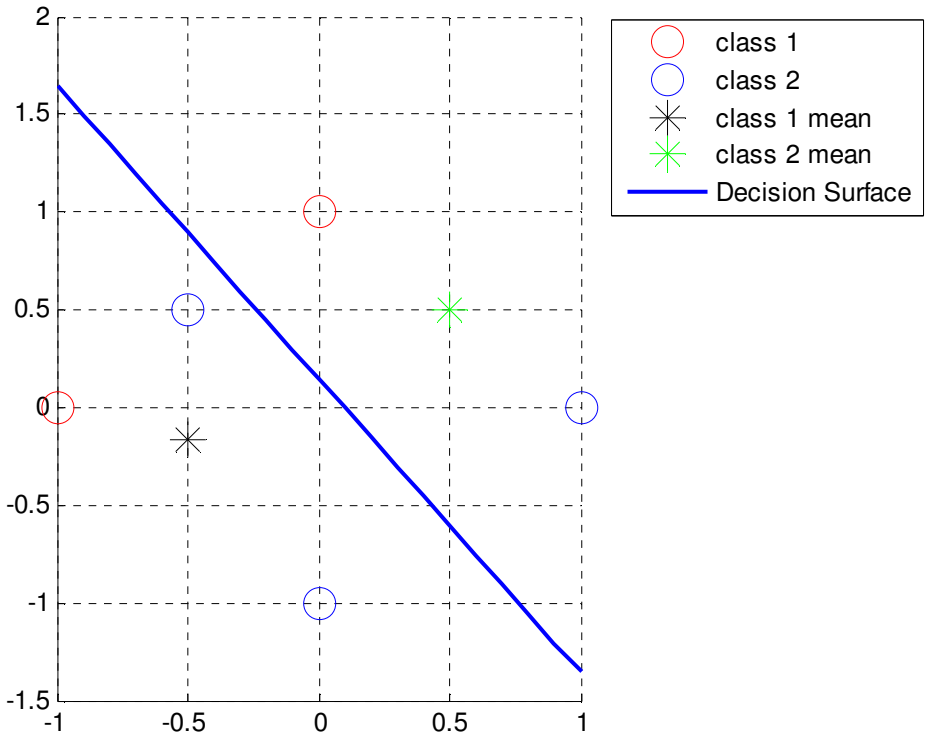


Distances from centers to data set.

	(0,0)	(1/2,1/2)	Nearest Cluster Center
(0,1)	**1**	**0.7071**	**B**
(-1,0)	**1**	**1.5811**	**A**
(1,0)	**1**	**0.7071**	**B**
(0,-1)	**1**	**1.5811**	**A**
(-1/2,1/2)	**1/4**	**1**	**A**

From these group assignments, K-MEANS should move the cluster centers to (-1/2,-1/6) and (1/2,1/2).

One-B. The initial map of the points and the provided cluster centers. After one iteration of K-MEANS the field now looks as follows.



The new centers of the clusters have shifted as expected. A decision surface has been overlaid perpendicular to the line between the two cluster centers.

Distances from cluster centers.

	A: (-1/2,-1/6)	B: (1/2,1/2)	Nearest Cluster Center
(0,1)	**1.2693**	**0.7071**	**B**
(-1,0)	**0.527**	**1.5811**	**A**
(1,0)	**1.5092**	**0.7071**	**B**
(0,-1)	**0.9718**	**1.5811**	**A**
(-1/2,1/2)	**0.667**	**1**	**A**

One-C. Given test points of (-3/4,3/4) which belongs to class 1 and (1/2,1/2) which belongs to class 2. What is the probability of error based on the K-MEANS clustering?

Based from the cluster centers the distances between the test points will provide classification.

	(-3/4,3/4)	(1/2,1/2)
(-1/2,-1/6)	**0.9501**	**0.667**
(1/2,1/2)	**1.2748**	**0**

Point (-3/4,3/4) becomes classified as class 1 and point (1/2,1/2) becomes classified as class 2. This matches up correctly with the provided information for 0% error.

One-D. However, if the error is computed using the K Nearest-Neighbor approach the results could be different. In this instance comparing the test points to the presently classified data points to assign each a class results in the following:

	(-3/4,3/4)	**(1/2,1/2)**
B: (0,1)	**0.7906**	**0.7071**
A: (-1,0)	**0.7906**	**1.5811**
B: (1,0)	**1.9039**	**0.7071**
A: (0,-1)	**2.4749**	**1.5811**
A: (-1/2,1/2)	**0.3536**	**1**

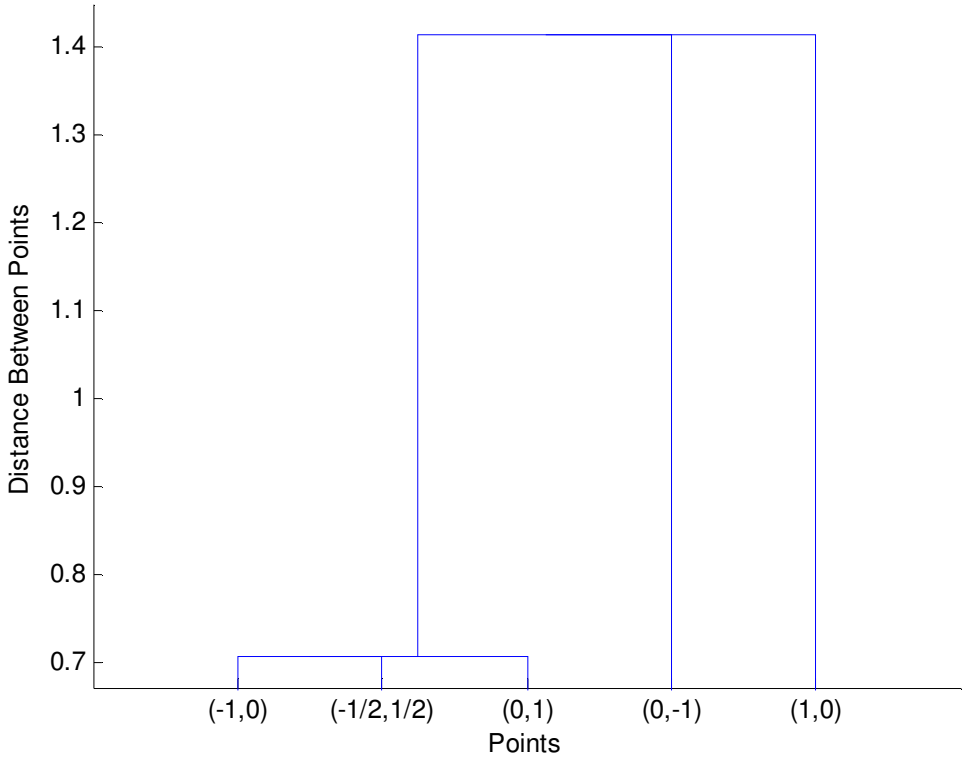
The point (1/2,1/2) present with two equidistant neighbors, both of which are class two data points. The point (-3/4,3/4) aligns with a class one data point. In this case the 1 nearest-neighbor results match the results from the cluster center comparison, which each point being correctly classified. This is slightly humorous given that not all of the initial data points are accurately classified yet themselves after the first iteration of K-MEANS.

For larger values of K the KNN results should match those found from the clustering algorithm, as long as K does not exceed the total number of points divided by the total number of classes. In our case, once K is set as 2 there is a bit of error from the second point for (-3/4,3/4) are two points of equal distance, but classified differently. If K reached 4 both points would have equal votes for either point so the results would only have 50% accuracy for any given classification attempt.

Two. Using the same points from before (0,1),(-1,0),(1,0),(0,-1),(-1/2,1/2).

Two-A. Construct a dendrogram for the data.

Matlab has excellent tools to link the data and build the dendrogram. The expectation is that points with the closest distance to each other will be grouped together. Naturally, the three points at and around (-1/2,1/2) should be the first level and (0,-1),(1,0) will have the greatest distance between the other points. The resulting dendrogram supports the analysis.



Two-B. Construct a top-down clustering graph.

The initial mean of the give points presents as (-1/10,1/10) and provides the following distances from each of the given data points.

	(0,1)	**(-1,0)**	**(1,0)**	**(0,-1)**	**(-1/2,1/2)**
(-1/10,1/10)	**0.9055**	**0.9055**	**1.1045**	**1.1045**	**0.5657**

Splitting this into two clusters removes the point (-1/2,1/2) from the rest to build a cluster of four and a cluster of one. The next iteration would take the mean of the cluster of 4 to compute new distances to keep splitting it down. However, the new mean is (0,0) which puts the remaining four points equidistant apart. This results in a two tiered graph where point one is (-1/2,1/2) and 0.5657 distance from the group mean and the other points are all labeled as 1 distance from the group mean.

Two-C. If you were to use your dendrogram to do an unsupervised clustering of the data, what clusters would you create (specify them by the mean and the elements associated with the cluster)?

Given unsupervised clustering the one mean should be (-1/2,1/2) that associates with the points (-1,0),(-1/2,1/2),(0,1). The other mean should be (1/2,1/2) and associate with the points (0,-1),(1,0).

Two-D. Suppose (0,1) and (1,0) occur five 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?

If those two points appeared five times more frequently than the other points they would begin to bias the average calculation when finding a new mean. To compensate for this all points would need to be preprocessed via a weighting algorithm that takes the two more common points at 20% strength and the remaining three points at 100% strength. This should enable the mean to remain unbiased as the points are generated and will not adversely impact the decision regions.

However, it may be more wise to introduce a new decision region that directly focuses on the most common points as they exhibit variation not only in their distance from the group center, but also in frequency of appearance. This would allow another dimension of analysis to the data because it would not be clustering only in the distance domain, but also the frequency domain. This is suggested because repetition of two specific values at a rate of five times the average of the others will throw off data integrity when the overall data size is small even with the best weighting correction tools.