Exam 3

Cedric Destin

May 2, 2014

Contents

1 Problem

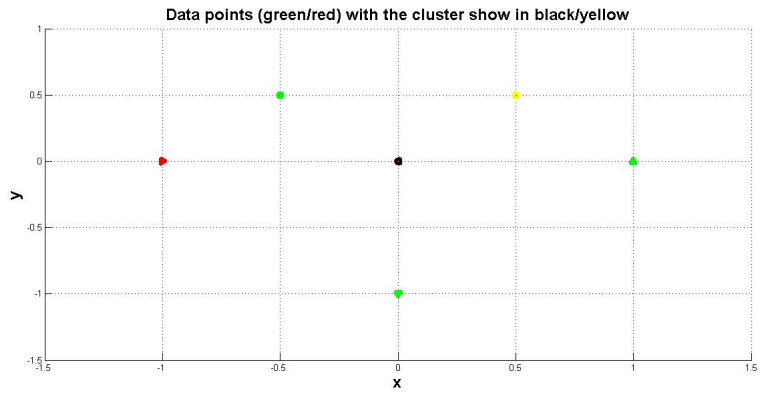


Figure 1: Scatter of the data points with the cluster centers shown in black and in yellow.

.	data	Class
1	(0,1)	class 1
2	(-1,0)	class 1
3	(1,0)	class 2
4	(0,-1)	class 2
5	(-0.5,0.5)	class 2

Table 1: Data points pre-labeled by some expert

Cluster 1	Cluster 2
(-0.5,0.5)	(1,0)

Consider 5 data points: (0,1), (-1,0), which belong to class 1, and (1,0), (0, -1), and (-1/2, 1/2), which belong to class 2. These data points are processed using the K-MEANS clustering algorithm defined as follow:

1. In the first step of the process, an assumption is made about the cluster centers *c1* and *c2* this clusters are estimated to be centered at (0,0) and (0.5,0.5) respectively, with those centers, the data set is then re-classified.

   Using the distance separating the cluster centers to the points the data is reclassified by comparing the distances of each data point to the cluster center. The distance of each data points to the cluster centers are denoted as $d_{i,j}$ where i represents one of the five data points and j represents one of the two cluster centers. The distance from the first cluster to the data point is compared to the distance of that same data point to the second cluster center, the outcome of this comparison decides which class the data point is classified to by the classifier. Finally, the process finds the location of the new cluster center by simply averaging the value of the data points in the clusters. This outcome of the K-means process is shown in Figure 2, where class 1 is shown in green and class 2 is shown in red.

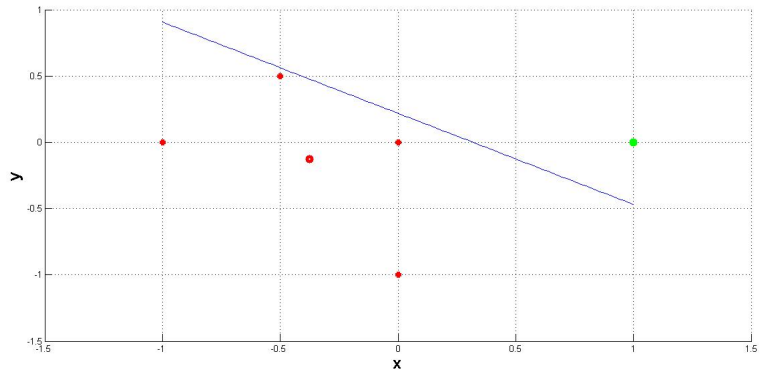


Figure 2: Outcome of the K-means algorithm. The data has been reclassified showing class 1 in red and class 2 in green, with new cluster centers

2. Based on those new clusters, the data is separated with the Decision Surface shown in Figure 3.

.	data	Class
1	(0,1)	class 1
2	(-1,0)	class 1
3	(1,0)	class 2
4	(0,-1)	class 1
5	(-0.5,0.5)	class 1

Table 2: Relabeled data using K-Mean

This surface is obtained by setting the distance of the each clusters equal to each other at an unknown point (x,y) and solving for the equation

$$\begin{aligned} y &= m \cdot x + b \\ &= \frac{x(-2 \cdot \mu_{1x} + 2 \cdot \mu_{2x}) + \mu_{1x}^2 - \mu_{2x}^2 \mu_{2y}^2 + \mu_{1y}^2}{-2\mu_{2y} + 2\mu_{1y}} \end{aligned} \tag{1}$$

3. Next we consider two new sample point into the machine, which have been previously labeled, at this stage the performance of the machine or the error performance of the K-Means algorithm.

x	y	
-0.75	0.75	class 1
0.5	0.5	class 2

Table 3: New data points

With the new data points the K-means algorithm is assessed in terms of the error rate. Figure 3 shows that the classifier is able to classify the data with an error rate of 50 percent, in other words, the classified data does not match the labeling that was previously given and one of those two data is not classified in the correct class.

x	y	
-0.75	0.75	class 1
0.5	0.5	class 1

Table 4: Classified data using K-Means

4. Finally, the error rate based on the K- nearest neighbor was also calculated and is compared to the K-Mean algorithm.

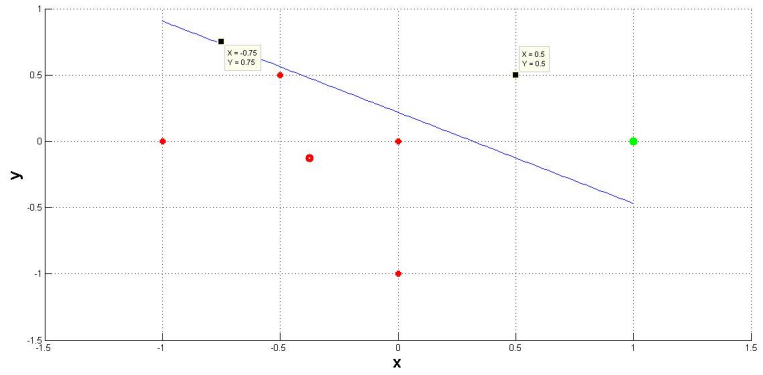


Figure 3: K-means algorithm with two new data points

First, this needs to be stated about K-Nearest Neighbor **It is the easiest Machine Learning algorithm.** But of course, it is difficult to explain it mathematically, the KNN algorithm observes a number of **labeled** data points that surround the new sample and decides the new sample's class based on the majority of data belonging in that class. In the problem the data set contains 5 labeled samples and based on the previous steps, this samples were pre-labeled data as to what is shown in Table 1. Unlike the K-Means algorithm this algorithm ends up classifying the new data points with an error of 0 % when k = 3. Figure 4 and Table 4 both corroborate to this conclusion. In contrast, when k = 5, the decision can not be made that easily due the limited amount of data points that we have in this set. With a k = 5, the k nearest neighbor demands for all the data to be analyzed to make a decision and of course, only one decision will be made from this, both data points belong to class 2 which has the most data points. Hence, this produces an error rate of 50 % as well.

x	y	
-0.75	0.75	class 1
0.5	0.5	class 2

Table 5: Classified data using KNN with k = 3

x	y	
-0.75	0.75	class 2
0.5	0.5	class 2

Table 6: Classified data using KNN with k = 5

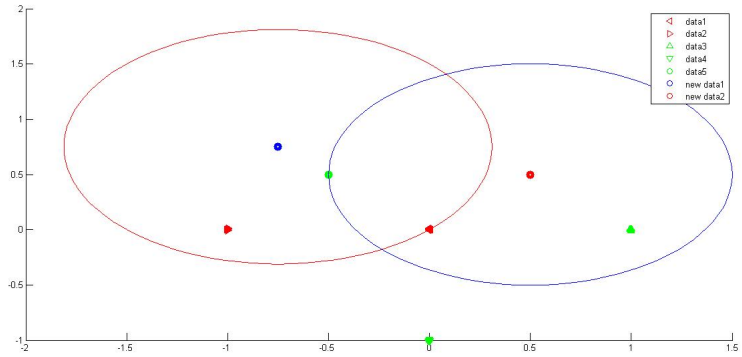


Figure 4: Data points classified using k = 3

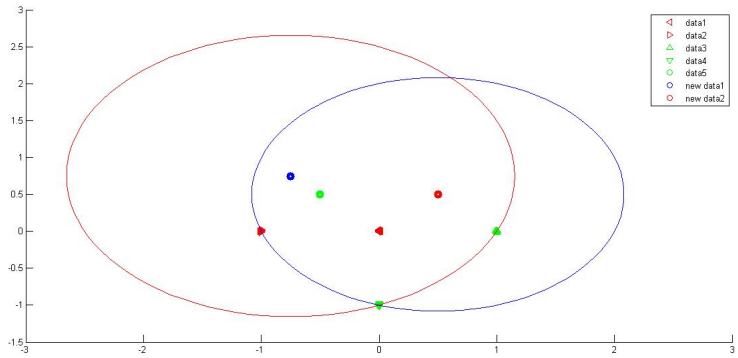


Figure 5: Data points classified using k = 5

2 Problem 2

1. Still using the same data points from problem 1, the data is now processed through a dendogram algorithm to obtain the cluster the data points that have the highest similarity. With the dendogram, the similarity of the data is seen as each data point is sorted into new clusters. The similarity is obtained as the inverse of distance for each data point away from each other, hence a cluster with a high similarity signifies that the samples are fairly close to each other or the distance is fairly close.

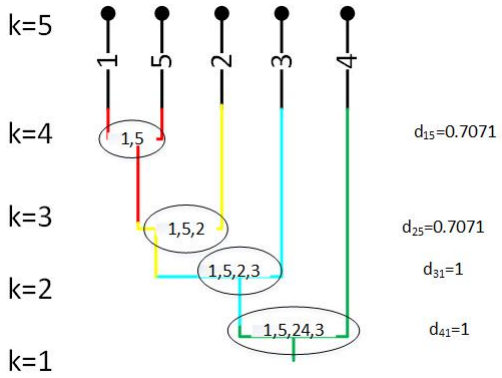


Figure 6: Dendogram for the data set, the colors mark the similarity for the data points

Figure 6 shows the dendogram obtained for the five data points in the data set, the distance shown on the right defines how similar two data points are, the first clusters consist of data points that are located within a fair distance and as this distance increases, those points are included into the final cluster.

2. While the dendogram classified the data from k clusters to 1 cluster, the data can also be classified from 1 cluster to an arbitrary number of n clusters this approach is observed below for the data.

Figure 7 shows the data classified starting from one cluster. Based on the distance from the data, the cluster is divided into more cluster based on the average distance between the data. The algorithm measures the average distance and limits the average below 1, which explains the results illustrated in Figure 7.

3. With the dendogram shown in Figure 6, it would be preferable to use clusters that would make fair decisions on the data. In other words, the amount of labeled data within the cluster is crucial when creating the clusters since the average distance is a metric for decision. In this particular, clusters, a k=3 would be preferable with the following average distance from each other as:

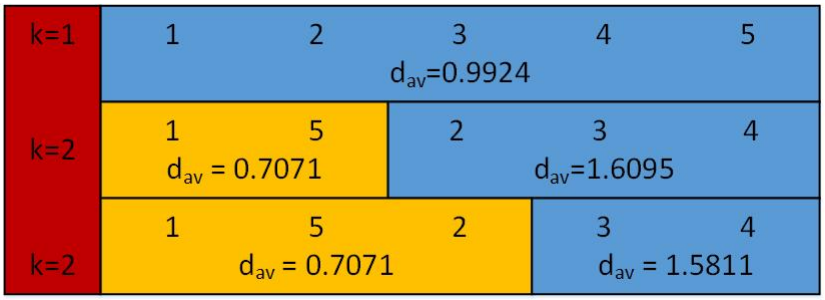
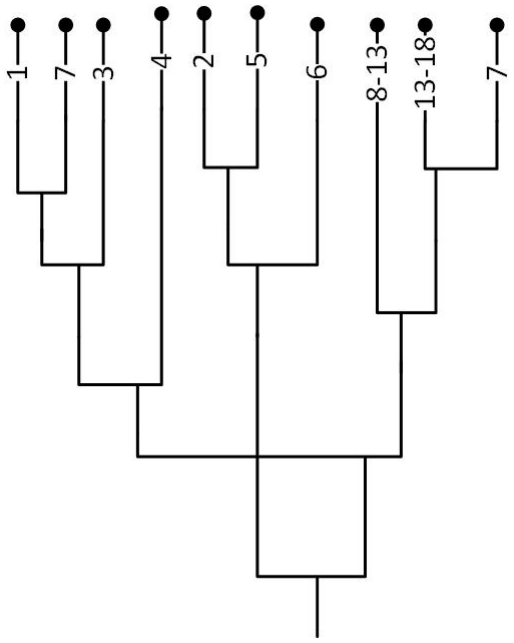


Figure 7: Classified data using a top - down approach 1 - k

Clusters	Data	Average distance
Cluster 1	(0,0), (-1,0), (-0.5,0.5)	0.7071
Cluster 2	(1,0), (0,-1)	1

4. Finally, we analyze the algorithm for the dendogram under the condition that the same data points (0,1) and (1,0) are each repeated 5 times.

   In this case, the previous dendoram algorithm computed in (1) will have the same effect on the clusters, the clusters will be made of the same data points. However, the algorithm in (2) will differ the clustering, since some data points are occurring more often, the average distance from the points will vary within some clusters.



(a) Dendogram with with (1,0) and (0,1) each repeated 5 time

(b) Top-down classifier with (1,0) and (0,1) each repeated 5 time

Figure 8: Case where some data points occur more often than other