Machine Learning

Assignment # 4

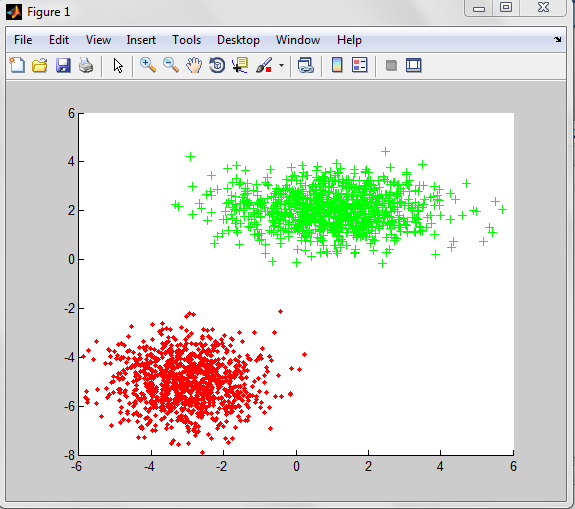
# Ahmad Tariq

# 15060004

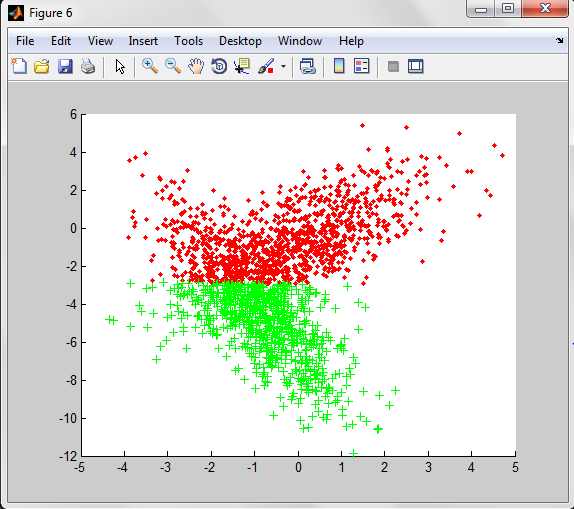
## Question 2:

We are given 2 data sets. Each dataset has 2000 data points each. The data has 2 features which makes it 2 dimensional. It is further provided that both the data (data 1 and data 2) have 2 clusters each. We are to apply K – means algorithm to associate each of the data points to one of the clusters.

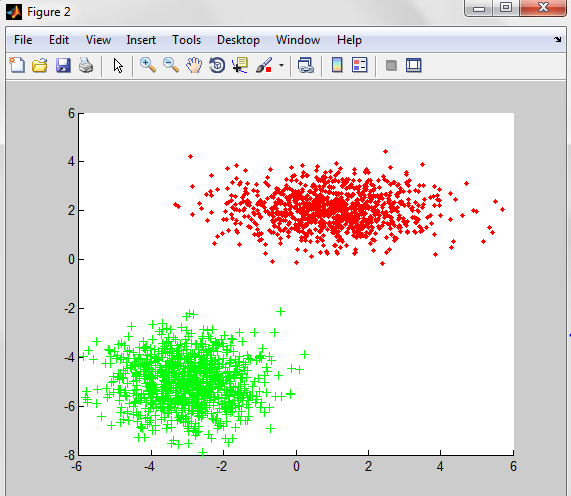
1. In the first part, we applied the K- means algorithm to data 1. Means for the each cluster were initialized by randomly assign data points to each of the means. Mean matrix becomes 1x2 for each of the two clusters. The convergence would happen when the means remain constant to two decimal places. The algorithm when applied took 3-6 iterations to converge based on where the means where initialized.



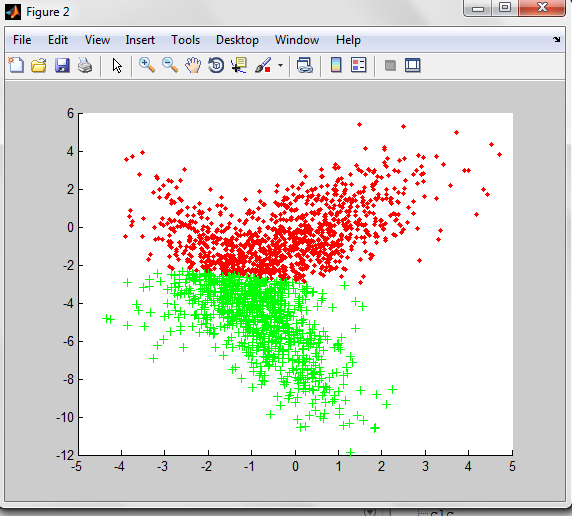
1. In the second part the K – means was applied to the second data. The initialization and convergence conditions were the same as the previous part. The convergence was obtained in 3-12 iterations depending on the initialization of the means for both the clusters



1. In this part, we were given prior data sets for each of the original data. The prior data set is divided into two, one for each feature of each data. The data in the prior datasets is labeled and we are to initialize K – means using the means of these data. K- Means was run with the prior data incorporated. The convergence was achieved in 2 iterations.



1. Now prior data for the second data was incorporated with the second data set. Convergence was achieved in 2 iterations.



## Question 3

In this question, it is said that each of the data point in both the data sets belongs to a different Gaussian distributions, being two in total. We are to estimate the association of each data point s to one of the Gaussian distributions. We are using EM algorithm. We initialize mean and covariance for both the distributions, and then calculate PDFs against each distribution. We then associate each point with a distribution and update the means and covariance matrices. We keep this until convergence. The convergence condition is that mean and covariance remain constant to two decimal places.

1. Incomplete
2. Incomplete
3. Incomplete
4. Incomplete

## Question 4

K –Means and EM Comparison

1. The iterations it takes EM to converge are more than the iteration for the convergence in K –means. This is due to the increase in number of parameters required to be estimated. In GMMs, we have to estimate the means and covariance of two distributions present in the data set.
2. Yes, it has a great effect in the reduction of the number of iterations it takes the algorithms to converge, because we have a set of labeled data points from each clusters that provides us with better initialization of the parameters, hence the convergence is achieved rapidly.
3. EM is more accurate, because as it takes the means and covariance of the data points, it can better see the direction of the centroid of each cluster, where as in K – Means it depends on where the means are actually initialized, because, it gave different results for each random initialized mean.
4. It did increase the accuracy of the K- Means algorithm. The iterations were reduced.

## Question 5

In K means, we assign the data sets hard boundaries, where as in EM, we use soft boundaries for the clusters, and assign probabilities to each data point.