Question 5 – clustering

For the implementation of this unsupervised machine learning technique, I have also used Matlab and implemented the code on my own because the algorithm is quite easy and I have already done it few times in the past for some projects.

I have implemented the algorithm much as it is described in the book but with some small changes to make it easier to implement:

* Since it was described at the course that it is better to choose initial centroids as far as possible from each other, I have taken first centroid with largest value of parameter 1 and the second centroid with smallest value of that parameter. I have done this to ensure that the centroids are found fast and without error. I have chosen to do like this because, for the past projects, when I had to choose more centroids and split data into more clusters, I have chosen the initial centroid randomly and once every 10-100 runs the results were going to be quite weird.
* I have not tested the stop condition and I have just let it run 100 times and break the operation when the centroid is no longer moving. I have found that the algorithm stops after 18 iterations but it could take longer for a lot larger datasets and it could be possible that after a lot of runs the centroids and clusters are no longer changing very much but the program does not stop running.
* Other than these I implemented the algorithm as described at the course: calculate clusters at each step by assigning each point to the cluster with closest centroid, calculate mean and update centroids position.

After running the program I have also ran the operations from Question 4 to display scatter data by first principal components and displayed centroids with 2 different colors and I also displayed the points from each cluster with 2 different colors. It can be seen that each centroid has assigned the closest points.

By viewing the results it can be seen that 2-means can identify two types of patterns and it is very likely that one cluster contains mostly patterns with label 0 and the other cluster contains mostly patterns with label 1.

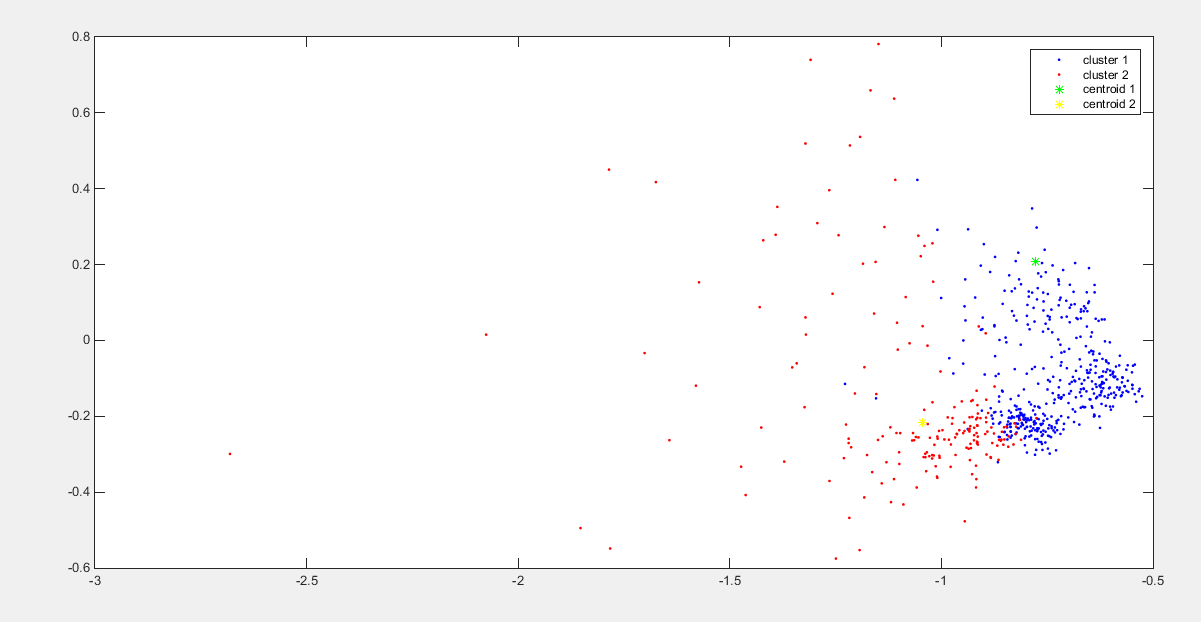


Figure Plot the results for calculating the 2 clusters of the training data by using k-means