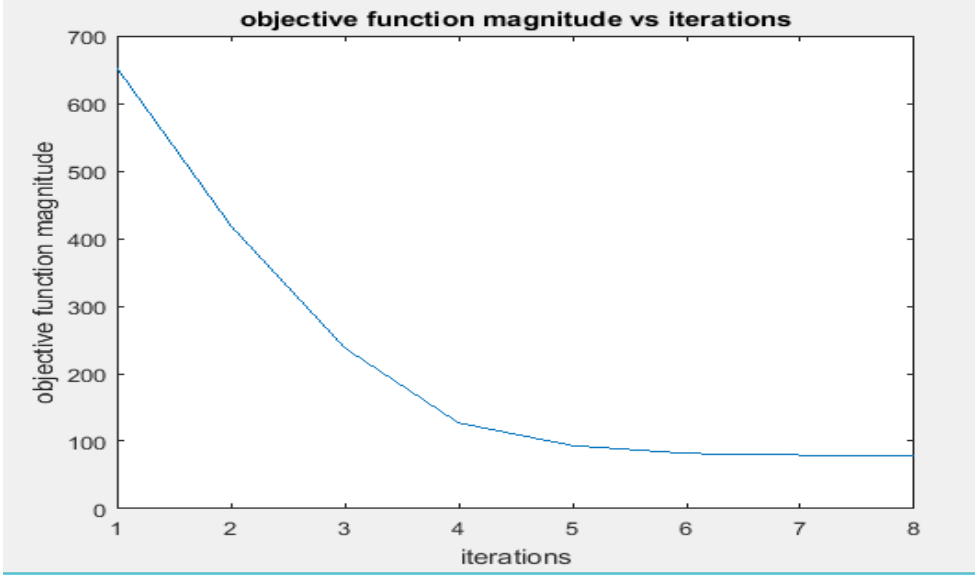
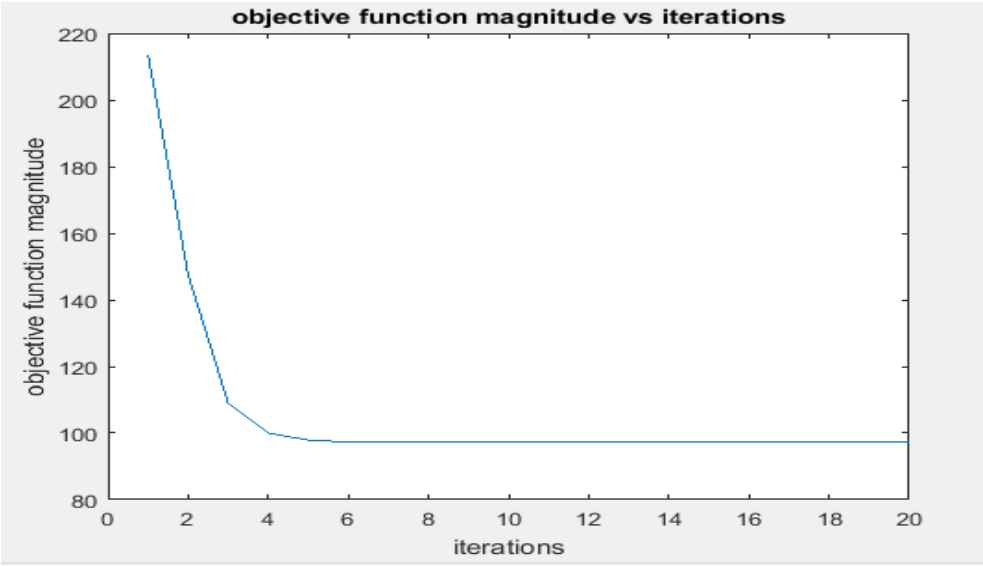
Suyog Daga
EMI-HW5
UFID:5773 5749

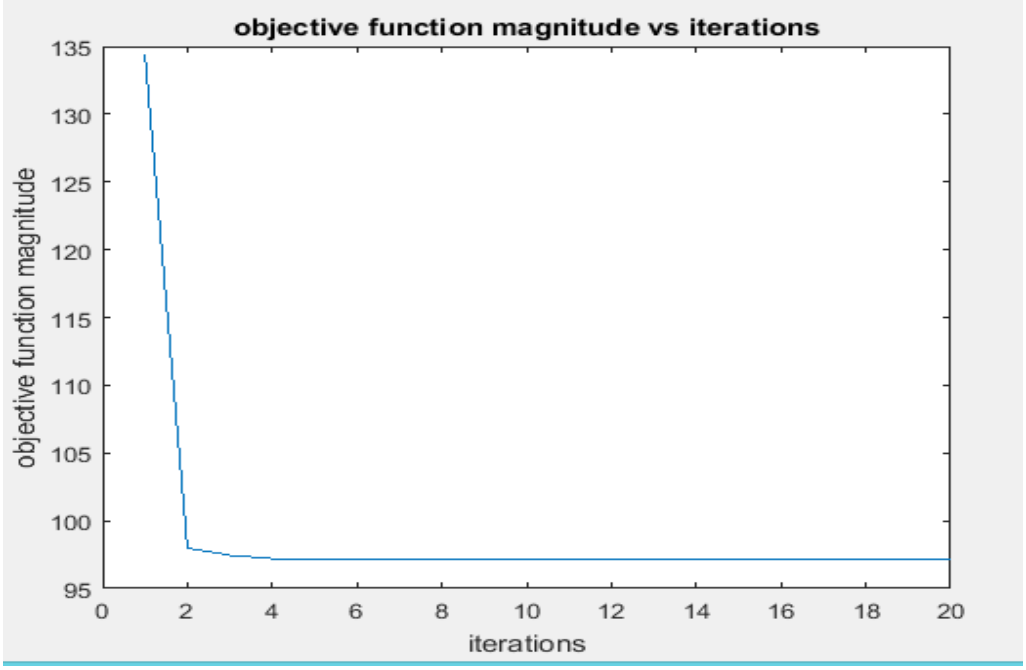
Q1)

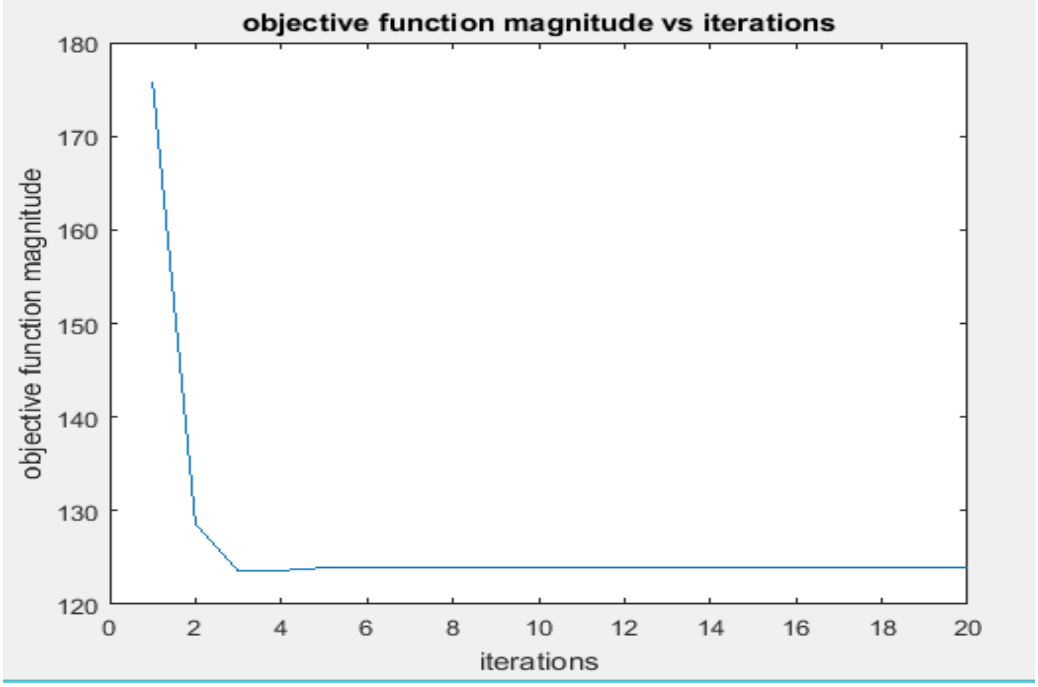
- Initially checking the data, we see that there are 3 varieties of flowers: setosa, versicolor and virginica. This gives an idea that, k=3 might be a good guess for clusters in such data set.
- Starting with k=3 and setting initial means to zero vectors, we get following graph for objective function magnitude versus iterations:



- Note the cost is approximately 650 when we start with means initialized to zero.
- Now, we try to plot the above graph for some random mean with respect to data.
  Here are the three plots of Objective Function Magnitude versus the number of Iterations for 3 different set of means:







- In all the three above plots, we see that function converges fast when means are assigned in the neighborhood of data, as compared to mean which were initialized to zero.
- We clearly can distinguish Iris-Setosa when compared to Iris-Versicolor and Iris-Virginica.
- Iris-Versicolor and Iris-Virginica have overlapping, where there is misclassification and therefore we don't have well separated clusters. Still majority of flowers in both the categories are classified properly.

Q2)

- Here is the table for Dunn's Index and Davies-Bouldin Index.

Index/Clusters (k)	2	3	4	5	6	7	8	9	10
Dunn (Vd)	0.076506	0.109435	0.082339	0.062378	0.085126	0.073922	0.08528	0.071307	0.101015
Davies-Bouldin (Vbd)	0.404293	0.661972	0.775701	0.805965	0.914158	0.969308	0.968103	0.944728	1.077259

For Dunn's index, the number of clusters that maximizes Vd is taken as best solution, which in this case is k =3
For David-Bouldin's index, optimal number of clusters corresponds to minimum value of Vdb, which in this case is k=2.

- Dunn's index gives us correct number of clusters(k=3), while David-Bouldin's index gives one less cluster (k=2) for data. The overlap in two flowers (versicolor and virginica) is too strong for Davies-Bouldin index to enumerate the cluster.

Dendrogram:
The clusters below for k=3 could be 6-8 , 14 to 26 and 1 to 3 . This clusters seems to have highest Euclidean distance between each other.