**EMI Assignment 5**
**Harshit Vijayvargia**
**( UFL ID -19355645)**

Question 1:

**Solution:** K means algorithm is a way to know how many categories are in our data and how many members are there in each category. In the problem, we are asked to implement K-means algorithm on given data which contains different classes of flowers. The distance metric we will use is Euclidean distance.
The biggest challenge in K- Means algorithm is:

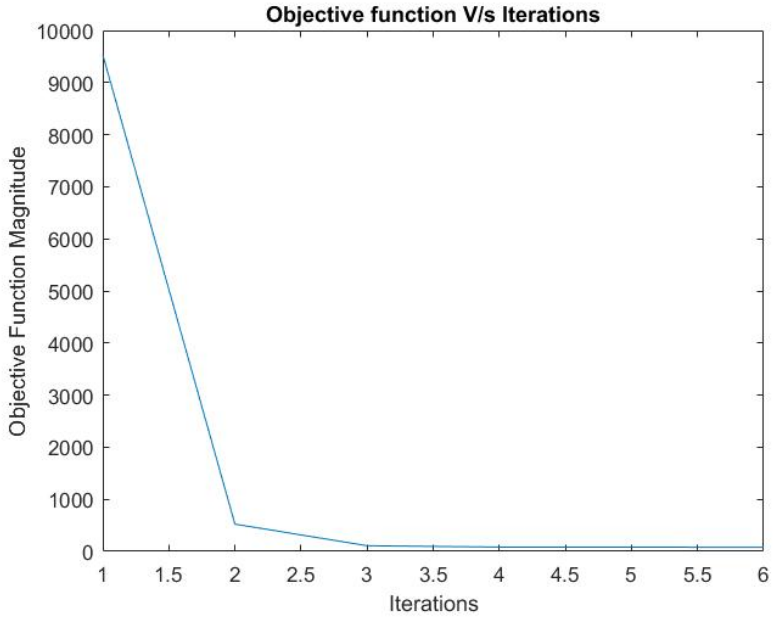
1) Determining the value of K (Number of clusters in our data): There are methods to determine K value like scree plot or analyzing the given data. In our case, we will assume that there are three classes of data.

2) Choosing the seed or the first point based on which clustering will start: The second step of K-means algorithm is an iterative step in which we must keep updating the mean based on members of clusters. This process goes on till there is a negligible difference in mean or binary membership matrix. Choosing a random seed or zero vector seed changes the number of iterations it will take for objective function to converge.

Initially I Initialized the seed to be a zero vector and applied K- Means algorithm with k=3. For my first iteration, I divided my data set or evaluated binary membership matrix randomly. After first Iteration, I assigned samples to clusters based on minimum Euclidean distance from cluster mean. For samples with equal distance value I divided data based on first occurrence in comparisons with V matrix.

Thus, I was able to classify data into three classes and objective function converged after 14 iterations. The objective function plot for this is:

1) When Seed is taken as zero vector ( Objective function converged after 14 Iterations):



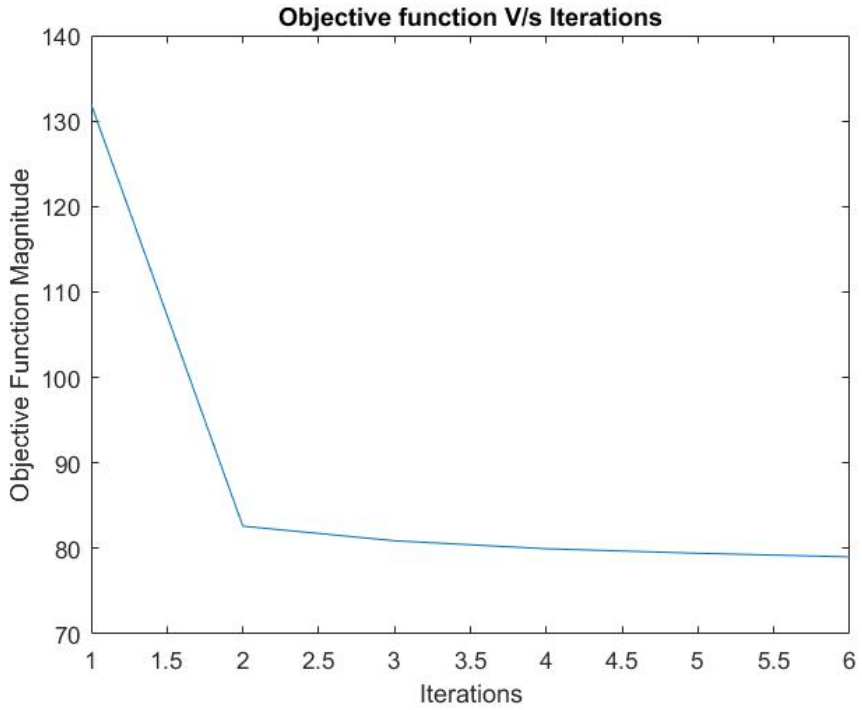

The magnitude of objective values I obtained are:

| Iteration | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Obj function Magnitude | 9536.2 | 567.0275 | 98.5443 | 87.38958 | 84.80173 | 84.10218 | 83.13638 | 81.839 | 80.89578 | 79.96298 | 79.43376 | 79.01071 | 78.94507 | 78.94507 |

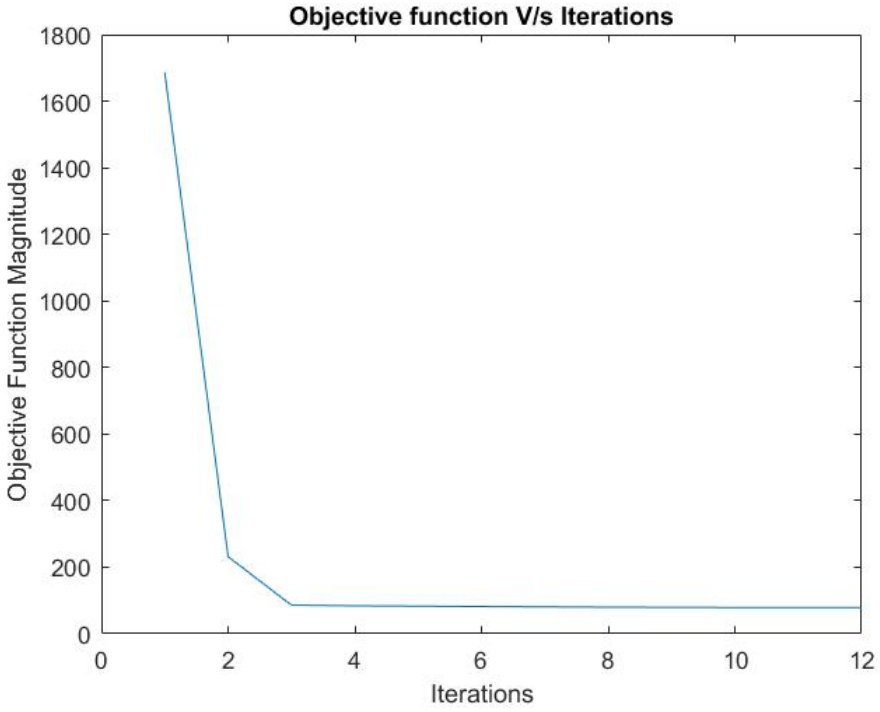
**The observation can be made that the curve is monotonically decreasing or at each iteration we are reducing the sum of distance of all samples from there cluster means.**

When I choose three random seeds. The objective function converged in different iterations since our process is randomized now but what I observed is that initial magnitude of objective function decreased. The reason behind this is that in first case distances were taken from origin so it was very large. The plots for objective function magnitude vs Iterations for three instances when I took random seeds are:

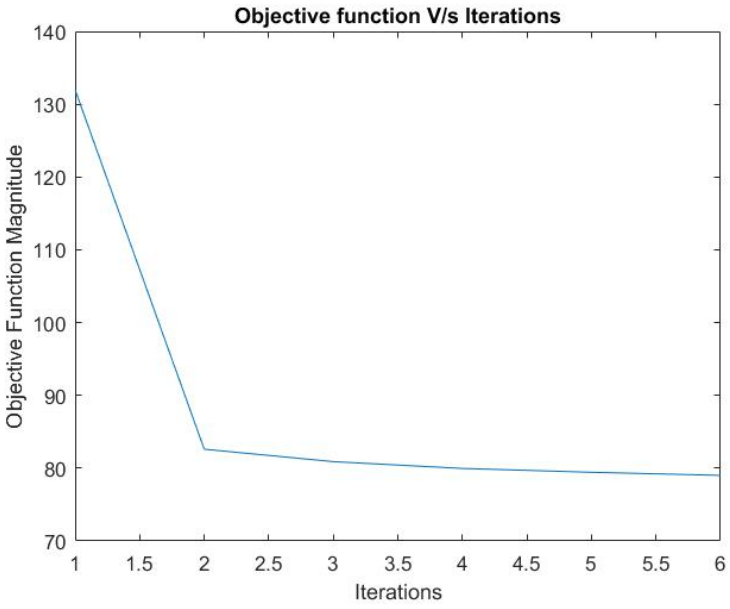
1)Converged in 4 Iterations:




2)Converged in 12 Iteration



3) Converged in 6 Iterations




**Finally, I was able to classify three classes based on binary membership matrix.** The number of samples which I was able to categorize:
Class 1 : 61 samples
Class 2: 39 samples
Class 3 : 50 samples
This data I evaluated on the basis of number of ones in binary membership matrix.

**The observation which can be made from this is that data is overlapping and two of the classes are overlapping since in the given data each class was having 50 samples.**

Question 2:

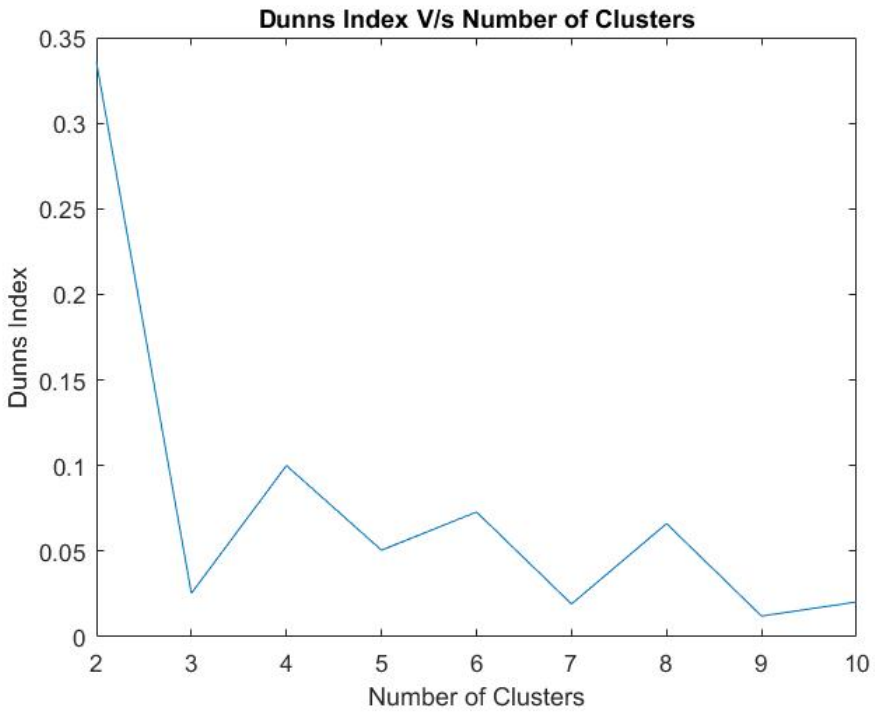
For second problem, we have to evaluate cluster validity based on Dunn index and Davies-Bouldin index.
These both indices are internal validity indices which evaluate clustering results using only features and information from the data set. The table and plot for both indices are shown below:

| Number of Clusters | Dunn's Index |
|---|---|
| **2** | **0.335652174** |
| 3 | 0.025423729 |
| 4 | 0.1 |
| 5 | 0.050583658 |
| 6 | 0.072881356 |
| 7 | 0.019157088 |
| 8 | 0.066101695 |
| 9 | 0.012106538 |
| 10 | 0.020338983 |

Table.1

Maximum value of Dunn's index obtained at K=2

The plot for Dunn index is:



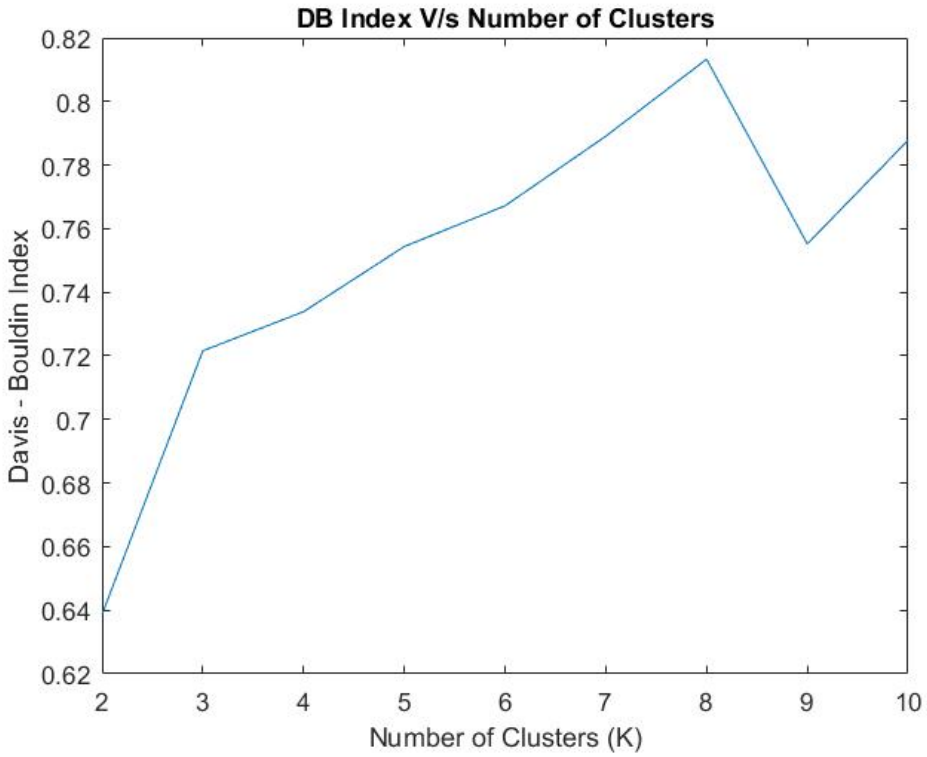

**The Maximum value of Dunn's index is obtained at K=2. This does not satisfy the correct number of clusters in data which is 3. The only reason can be that 2 classes are overlapping or two classes are not linearly separable.**

2) DB Index V/s Number of clusters :

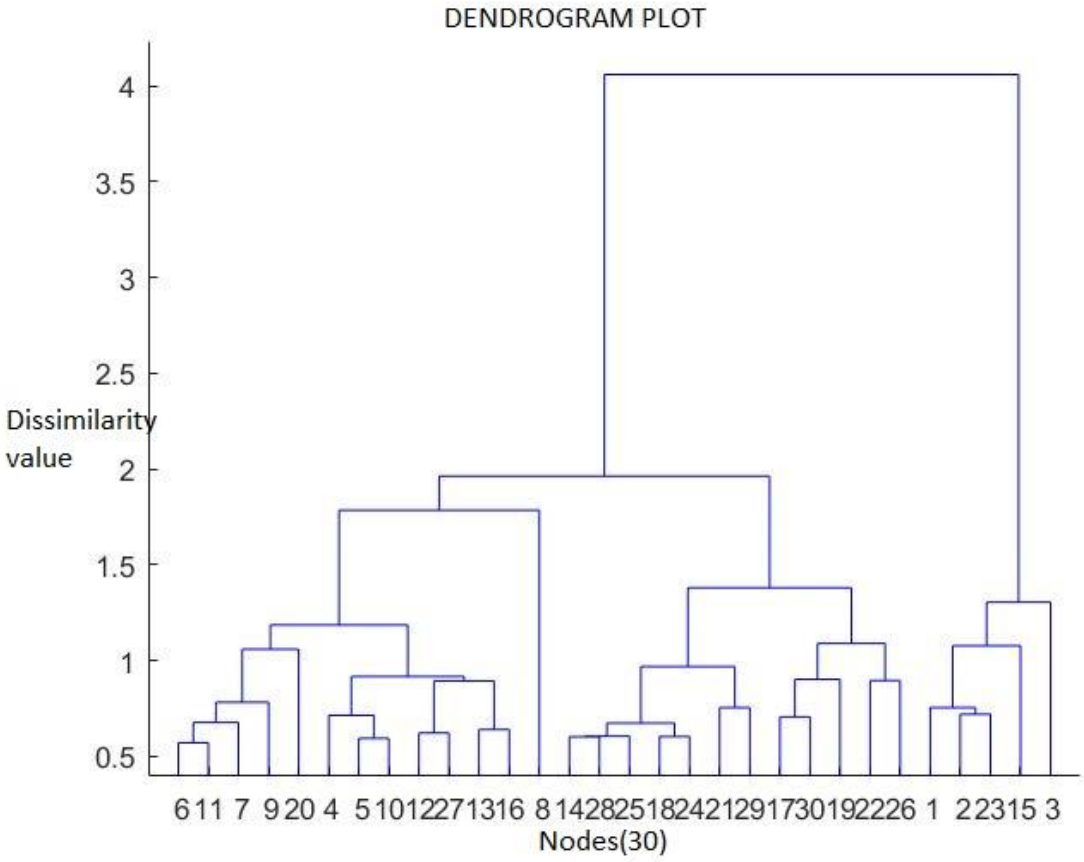
| Number of Clusters | DB Index |
|---|---|
| **2** | **0.6387** |
| 3 | 0.7216 |
| 4 | 0.7339 |
| 5 | 0.7544 |
| 6 | 0.7672 |
| 7 | 0.7891 |
| 8 | 0.8133 |
| 9 | 0.7552 |
| 10 | 0.7878 |

**Table for DB index**

The Minimum value of DB Index is obtained at K = 2 which is contradicting that number of clusters in the data set which are 3. The reason for this is same that 2 classes are strongly overlapping. This thing can be also observed from the dendrogram plot drawn on next page.



3) Dendrogram Plot:




The vertical axis represents the dissimilarity between clusters. The horizontal axis represents the objects and clusters.

The best DB index and Dunn index we evaluated was for 2 clusters.
To produce 2 clusters we can determine the cutoff point from above dendrogram. For distance/dissimilarity above 2 there are two branching taking place so we can take 2 clusters for distance above 2 (d>2).

This cutoff produces the clusters which have high unweighted average Euclidean distance. As height of the leaves is maximum for partition above this cutoff(d>2) and so will be the average.