Kumar Rohit Malhotra (5697-1748)

EEL5840
Elements of Machine Intelligence
Assignment – 5

(All the solutions have been programmed by me using MATLAB.**)**

Section 1: Implementing K – Means Classifier

- The dataset is split into X matrix, containing the feature data, and Y matrix, containing the species of the flowers for the corresponding samples.
- Initial mean is taken to be zero for all the three clusters.
- The membership matrix U is created using the given formula, and the mean matrix V is updated then on the basis of the membership matrix.
- These updates are made alternately until a consequent pair of membership matrices across a pair of consecutive iterations doesn't change.
- Fir. 1.shows the plot observed for the objective function magnitude versus the number of iterations.

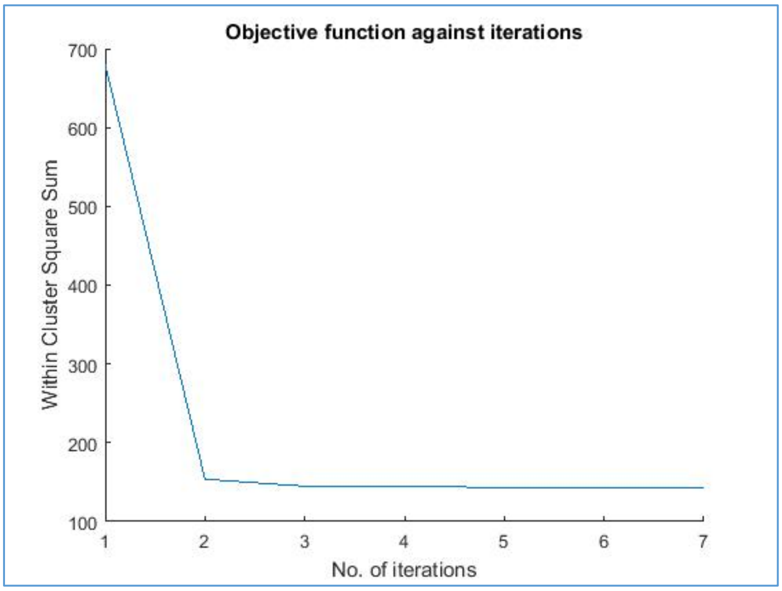

Fig. 1. Objective function versus number of iterations

- The plot can be seen to be monotonically decreasing, i.e. the Within-Cluster-Square-Sum is decreasing with number of iterations.
- Three more tests are done on the data using random cluster means, in the neighborhood of data, in each case. A monotonically decreasing plot is obtained in the three cases as well.
- The comparison of actual clusters and the clusters obtained from K-Means clustering can be seen in Fig. 2, Fig. 3 and Fig. 4, for different combinations of features.

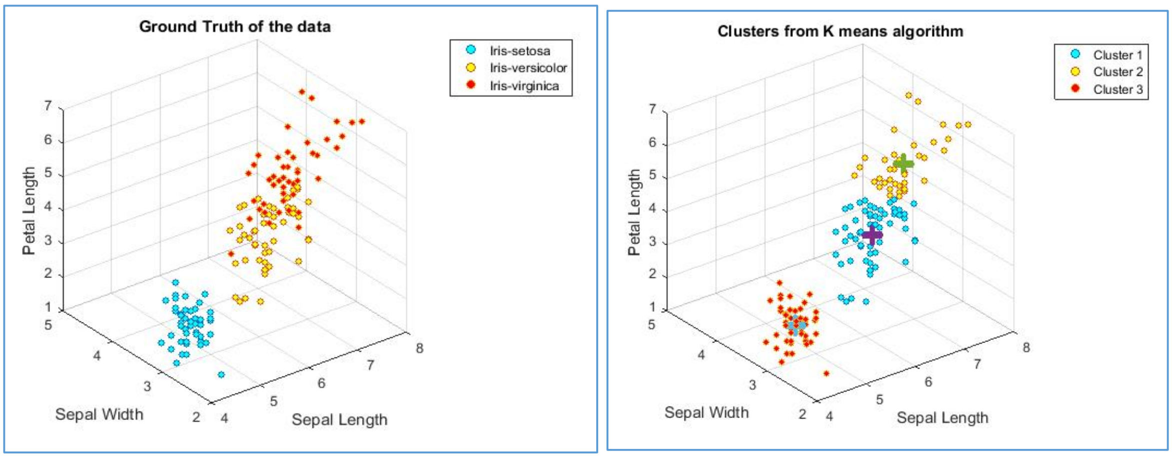

Fig. 2. Ground truth versus K-means clusters for Sepal Width, Sepal Length and Petal Length

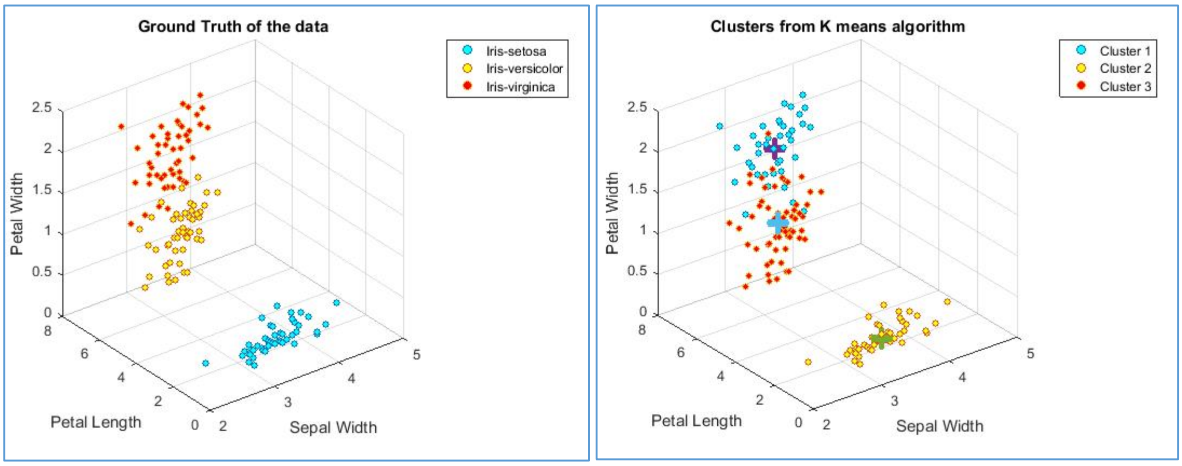

Fig. 3. Ground truth versus K-means clusters for Sepal Width, Petal Length and petal Width

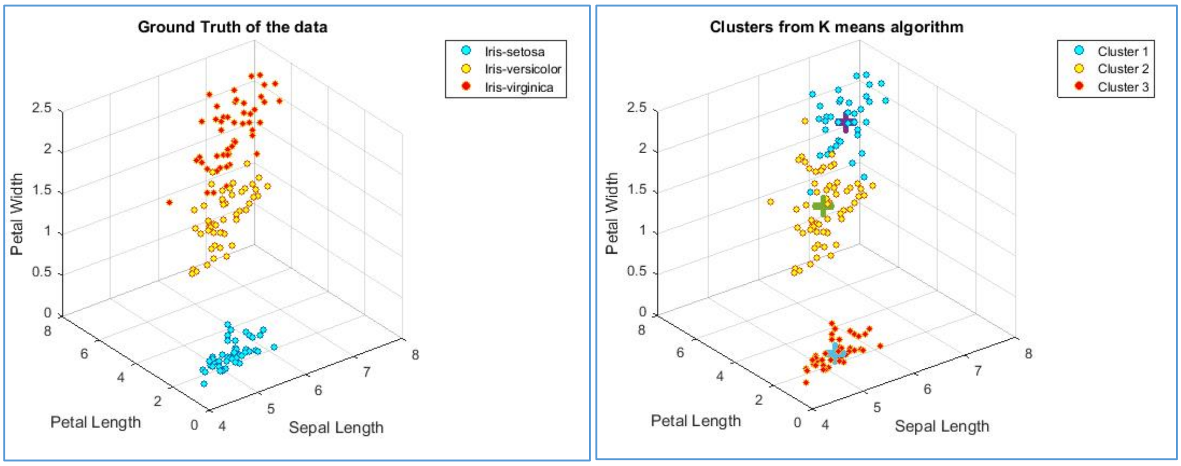

Fig. 4. Ground truth versus K-means clusters for Sepal Length, Petal Length and Petal Width

- As can be seen from the three figures above, we can distinguish between the three classes of flowers based upon the cluster structure, though some of the samples from one actual cluster move into some other cluster when found by k-means algorithm.

Section 2: Validation of Clusters

- Dunn Index and Davis – Bouldin Index for various number of clusters is calculated. Fig. 5 and Fig. 6 show the Dunn Index and Davis – Bouldin Index, for number of clusters 2 to 10, respectively.

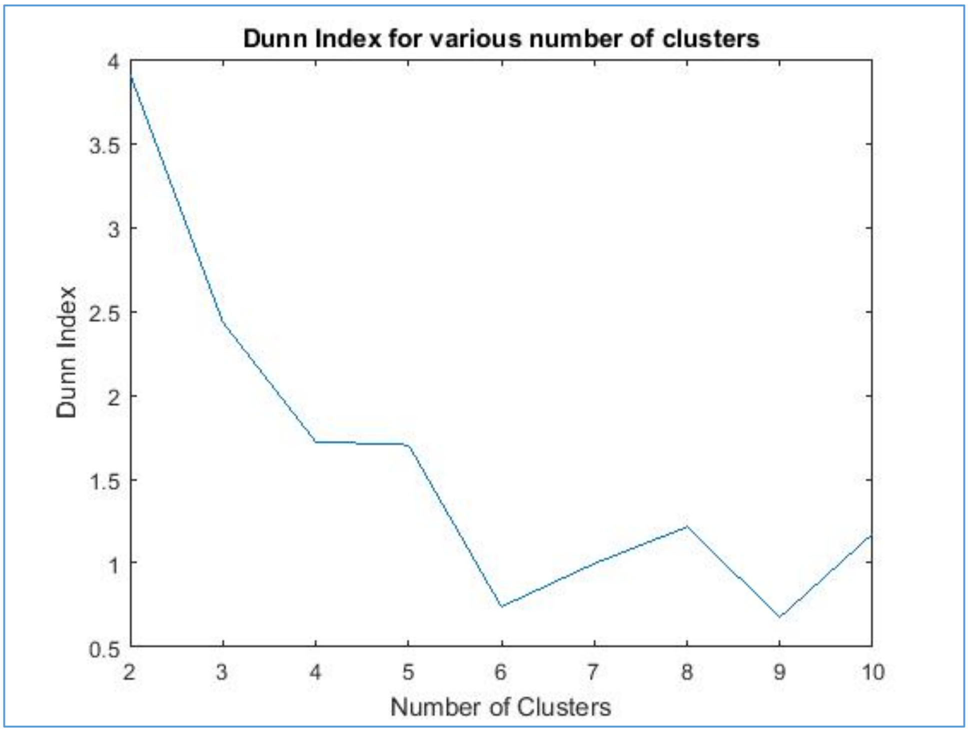

Fig. 5. Dunn Index for various number of clusters

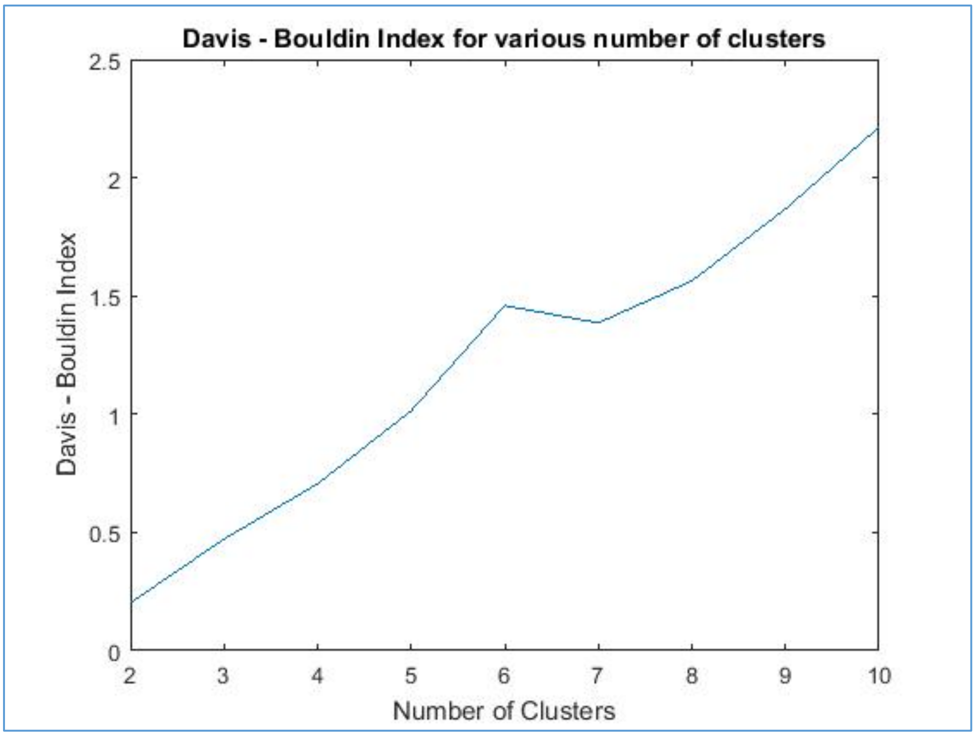

Fig. 6. Davis – Bouldin Index for various number of clusters

- Dunn Index is found to be maximum for 2 clusters, and Davis – Bouldin Index is found to be minimum for 2 clusters. Thus, according to the two validations, 2 clusters partition the data in the best way possible. This is because, as can be seen in Ground truth in Fig. 2, 3, and 4, and as said in last point of section 1, some of the samples are not exactly separable into different clusters when number of clusters is taken to be 3. In other words, two clusters can be seen to overlap for a few samples. Hence, such two clusters are better taken as a single cluster.
- The dendrogram with 30 leaves can be seen in Fig. 7.

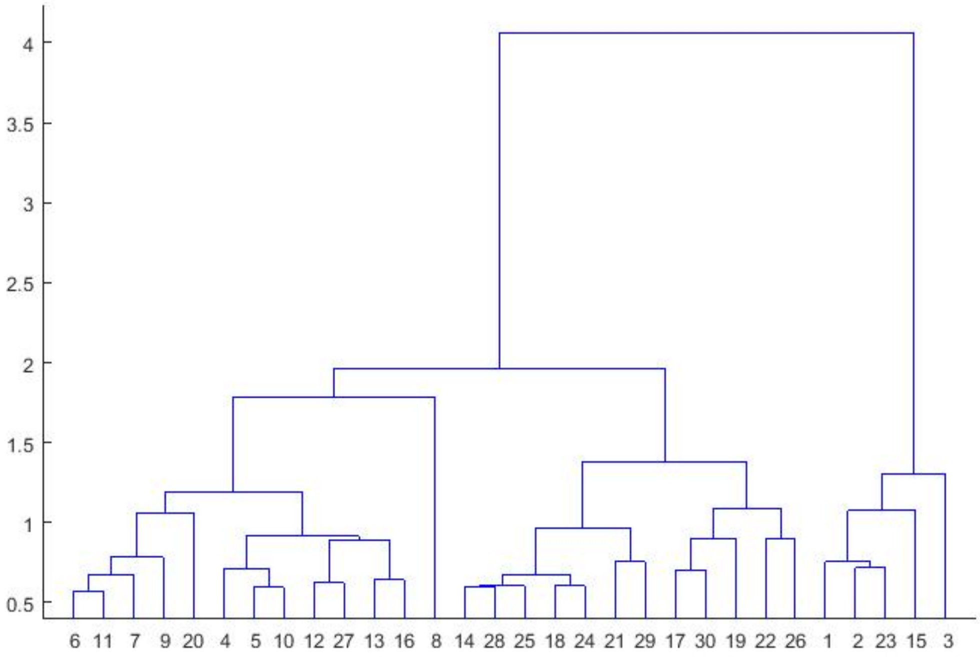

Fig. 7. Dendrogram with unweighted, average Euclidean distance

- The best number of clusters as per Dunn Index and Davis – Bouldin Index is 2. The arrangements of the leaves in the dendrogram denote the similarity of the samples. The heights of the nodes show how different the chunks are. We find the tallest structure for the chunk from leaf number 1 to 3. The above given dendrogram can be partitioned before leaf numbered 1. This partition will also give us high average, unweighted Euclidean distance between the clusters.