Tutorial 10 On-site Questions

(K-Means)

Consider the famous Iris Flower Data set which was first introduced in 1936 by the famous statistician Ronald Fisher. This data set consists of observations from flowers of Iris species. For each observation, four features were measured: the flower's length and width of the sepals and petals (in cm).

The data set is given in a file named `iris.csv`.

1. Use K-means clustering method to cluster all the flowers into k groups where $k = 1, 2, 3, ..., 10$, where for each value of k the value of WSS - the within sum of squares is obtained.

2. Write code to obtain the plot of WSS against k. Which value of k would you choose as the number of clusters for all the observations in the data set? Explain.

3. With the value of k chosen above, report the centroids of all the clusters and the number of the observations in each cluster.