Statistical Machine Learning – Project Part 2

Unsupervised Learning (K-Means)

Dataset Used:

The data set used is one provided for the project. The dataset contains a list of 300 co-ordinates with coordinate values in the range $0 \leq x \leq 10$ and $0 \leq y \leq 10$. The data given is not labeled and so an unsupervised learning technique is used to correctly cluster the data.

Objective:

The objective of the project is to use the K-means clustering algorithm to effectively cluster the given data set. K-means clustering is a unsupervised machine learning method used on unlabeled data. The algorithm tries to cluster the group the data by iteratively updating a centroid value that is used to group the data item into each cluster. The value k signifies the number of clusters to which the data will be separated. In the algorithm, this value is initially assumed. In order to get the ideal value of k for the given dataset, the elbow method is used. This method is explained briefly in the results section.

For this project, the algorithm is run for all values from k = 2 to k = 10. The objective function (shown below) is used for each k value and plotted to visualize the ideal k value for the dataset.

$$\sum_{i=i}^{k} \sum_{x \, \epsilon \, D_i} \|x - \mu_i\|^2$$

Two methods are used to calculate the initial k centroids.

- The first method involves picking k random centroid from the dataset.
- The second method involves picking the first centroid randomly from the dataset. After that, for every i^{th} centroid, we consider that point in the dataset that whose average distance from the $(i-1)$ centroids is maximal.

Two runs are done with each centroid method and the resulting plot of Objective function vs Number of clusters is recorded. These graphs will be shown below in the results section.

Result:

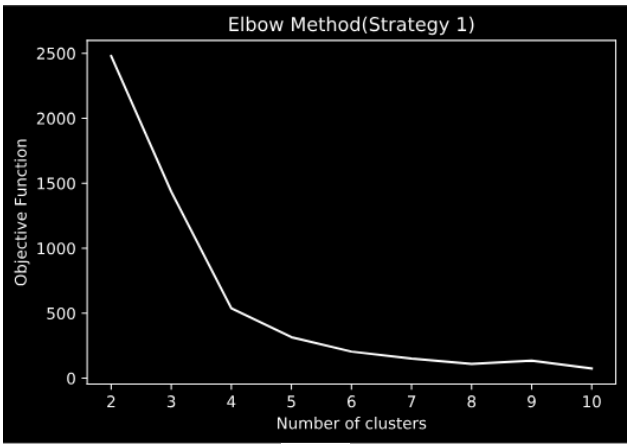
Each strategy was run twice with different initializations. Two inferences can be made from the outputs received:

- The Objective function (generally) decreases as the number of clusters increases. However, the rate of decrease is very high till one value, after which the slope decreases considerably. The point where this happens is generally considered the ideal value for k (this method is referred to as the elbow method). For this data set, across both strategies, the ideal k value is 4 (inferred from the graph).
- While both strategies seemed to give good results on average, the second strategy seemed to be the more consistent method to get the ideal results. This shows the impact the selection of initial centroid can affect the final strategy. In strategy 1, there was no relation between the points selected. This can result in some undesirable results during some initializations. Overall, strategy 2 of selecting points with some relation between them seems to give better results.

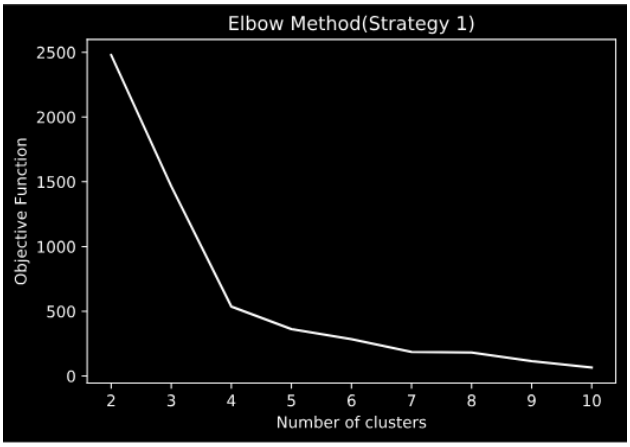
Conclusion:

In this project, K-Means clustering was used to cluster the given unlabeled dataset using two different strategies. Each strategy was run twice and the results obtained was used to identify an ideal value for k i.e the ideal number of clusters to properly separate the data.

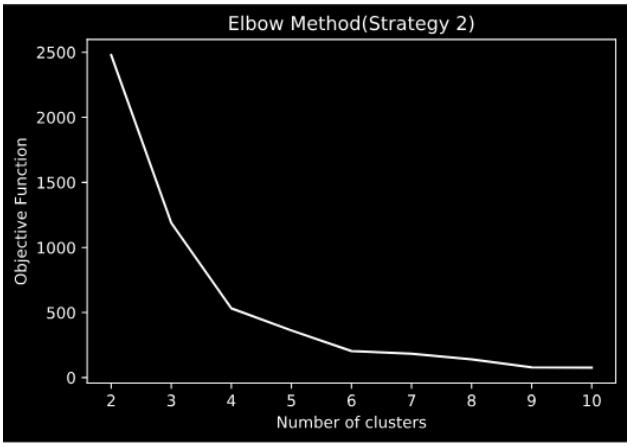
Graphs:



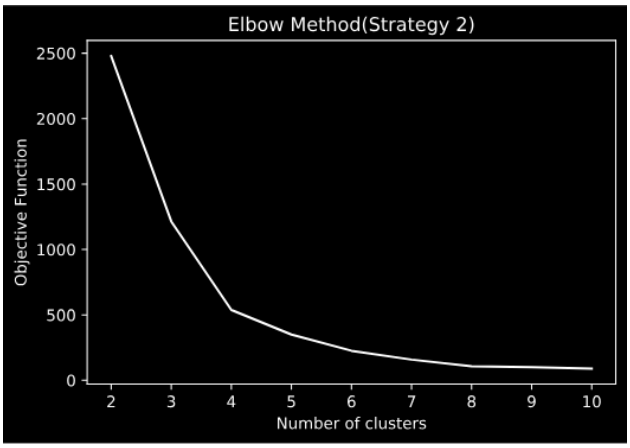
(i)



(ii)



(iii)



(iv)

Graph showing Objective function vs Number of clusters (value of k). The two above are for strategy 1 while the two at the bottom are for strategy 2.