Project Part 2: Unsupervised Learning (K-means)

Objective:

The requirement is to implement the k-means algorithm on the given 2D dataset using two different strategies for choosing initial cluster centers.

Dataset:

The given dataset consists a set of 300 data points and is two-dimensional. (300,2)

Procedure:

Strategy 1: Randomly picking the initial centers from the given samples. The disadvantage of this approach is that the algorithm could take much time to converge or at times might go on in worst case scenario.

Strategy 2: Picking the first center randomly and then for the i-th center (i>1), pick a sample such that the average distance of this chosen one to all previous (i-1) centers is maximal. This approach is computationally faster if k is kept small.

Of the two strategies, Strategy 2 is optimal as it is more efficient (in comparison with runtime and performance) than Strategy 1.

K means Algorithm:

Step 1: Randomly pick an initial centroid using of the given two strategies

Step 2: Calculate the Euclidean distance from each point to the centroid and assign each point to its nearest cluster using the distance calculated.

$$argmin_{C_i \in C} dist(c_i, x)^2$$

Step 3: Form new cluster by averaging all the data points of that cluster

$$\frac{\sum_{x_i \in s_i} x_i}{|S_i|}$$

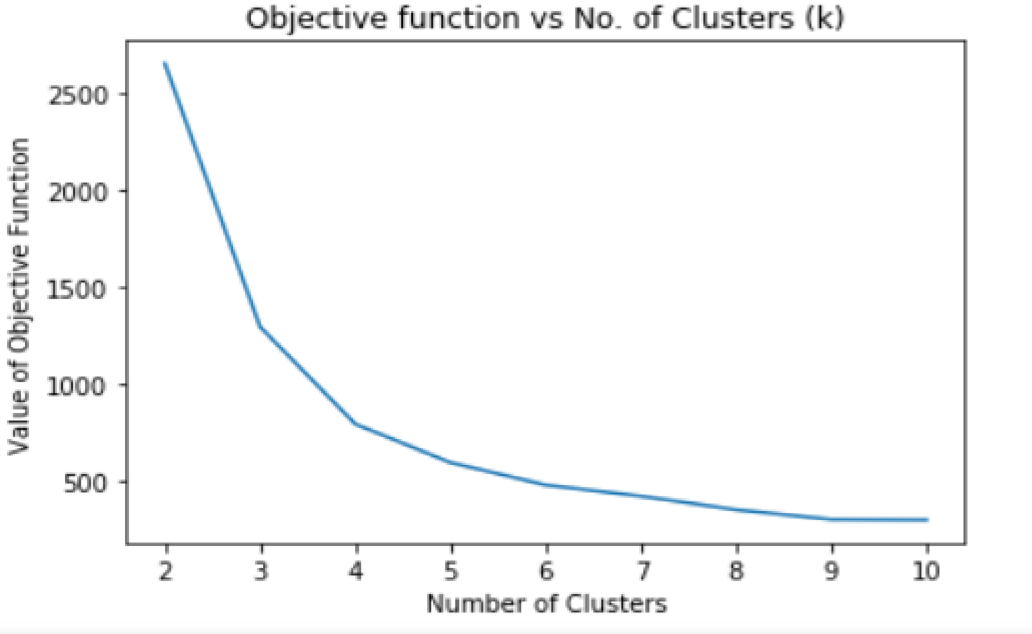
Step 4: Repeat 3 until convergence.

Objective Function is given by,

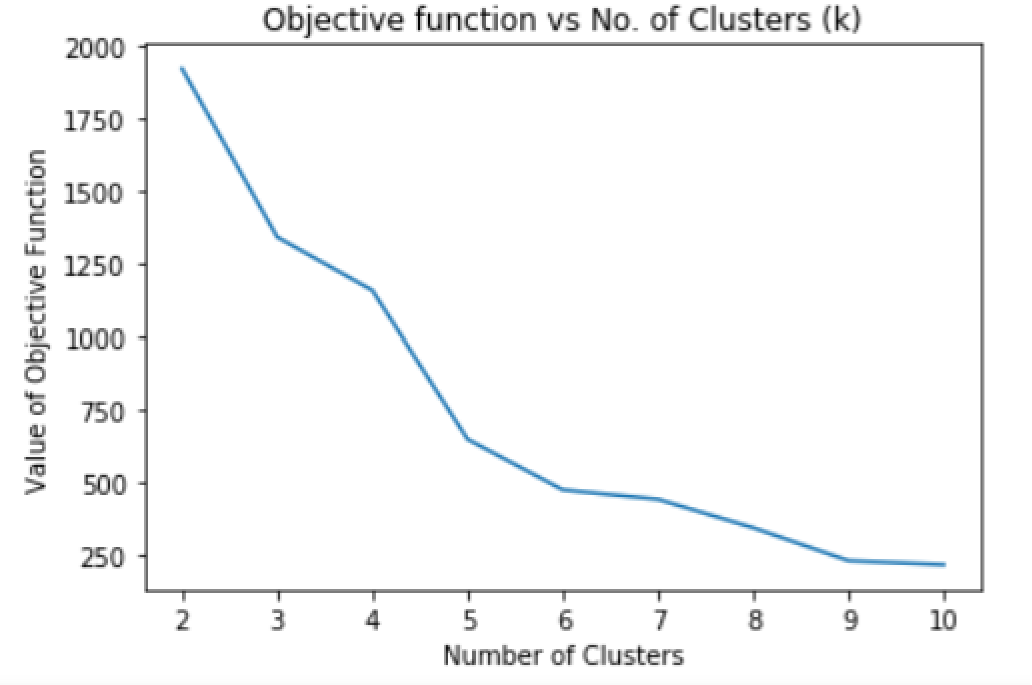
$$\sum_{k=1}^{K} \sum_{x_n \in c_k} ||x_n - \mu_k||^2$$

Results:

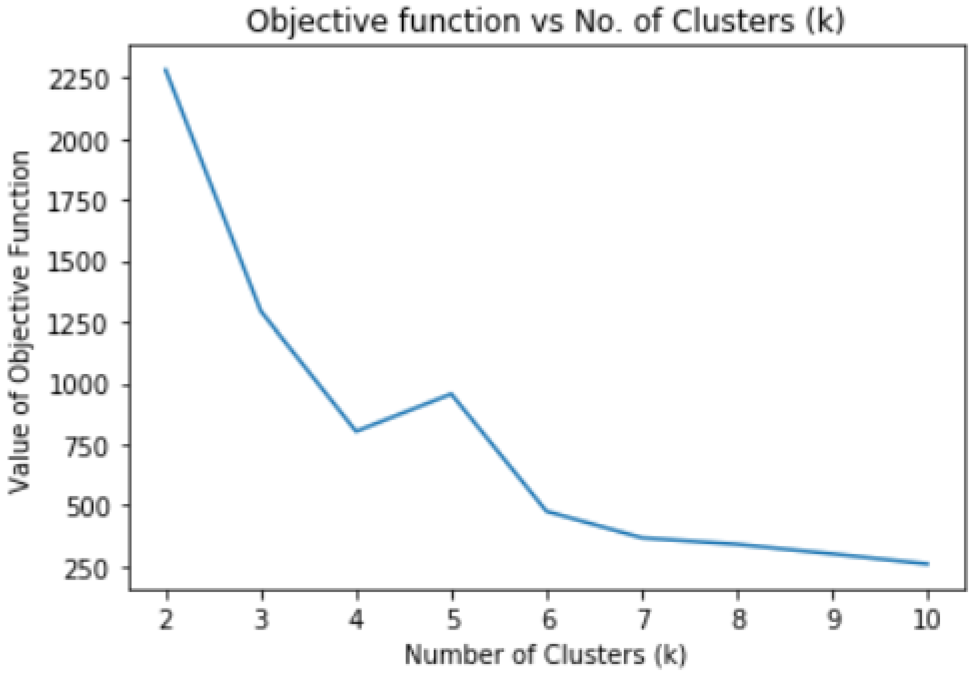
1. Plots for the objective function values vs Number of Clusters (k): Strategy 1, Initialization 1



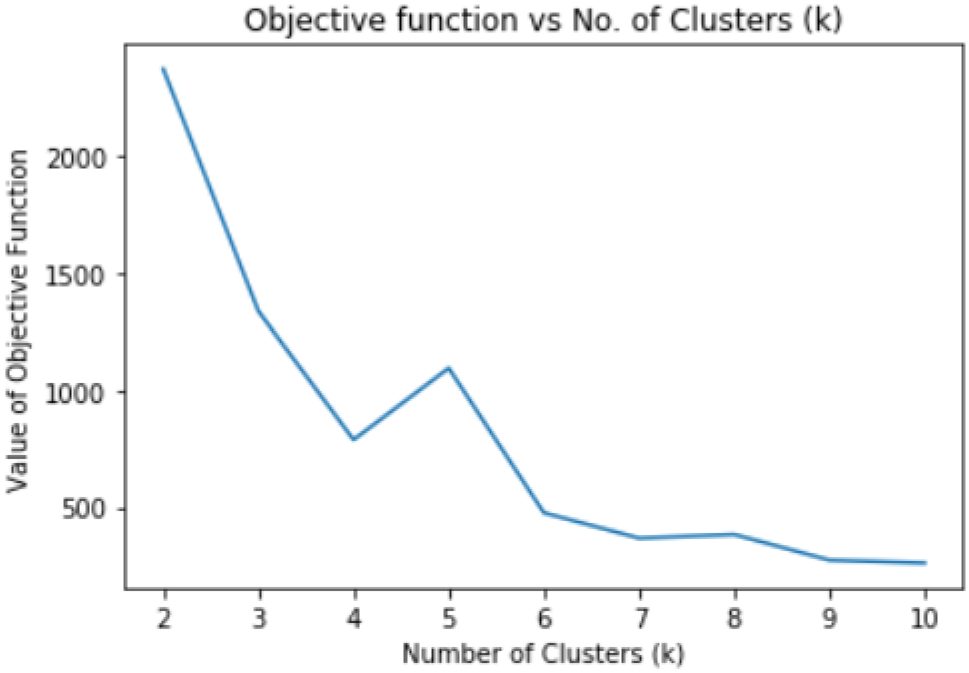
2. Plots for the objective function values vs Number of Clusters (k): Strategy 1, Initialization 2



3. Plots for the objective function values vs Number of Clusters (k): Strategy 2, Initialization 1



4. Plots for the objective function values vs Number of Clusters (k): Strategy 2, Initialization 2


Objective function vs No. of Clusters (k)
Value of Objective Function
Number of Clusters (k)