Project Part 2: Unsupervised Learning (K-means)

CSE 575: Statistical Machine Learning

Arizona State University – Summer 2022

Course Instructor: Masudul Quraishi

Submitted By:

Aishwarya Baalaji Rao

ASU ID: 1222423228

Date of Submission: 19th June 2022

ABSTRACT: This document serves as the report to the second part of the project. In this part, we implement unsupervised learning using K-means clustering algorithm with two different types of initialization strategies. Two types of strategies are used in the experiment to choose the initial cluster centers. The goal of the implementation is to plot the objective function value vs the number of clusters for k ranging from 2-10 and document the results.

KEYWORDS: K-means Clustering, Objective Function, K-means++, Clusters, Centroids Sets

1. Introduction

K-Means Clustering is an unsupervised learning method whose objective is to cluster the observations using a given dataset. As an input, the number of clusters is specified. It produces clusters by reducing the sum of the distances between each point and its corresponding cluster's center. Clustering is a collection of methods used to organize data into clusters. Clusters are roughly described as groupings of data items that are more comparable to each other than to data objects in other clusters.

2. Problem Definition and Algorithm

2.1 Task Definition

In the second part of the project, k-means clustering algorithm is implemented on the given set of 2-D points dataset with 300 samples. Two strategies are chosen to select the initial cluster centers:

Strategy 1: Select the first centers at random from the supplied samples.

Strategy 2: Choose the starting center at random and for the i-th center (i>1), a sample with the greatest average distance to all previous (i-1) centers is chosen.

The experiment will involve testing the implementation with number of clusters (k) ranging from 2-10 and plotting the objective function value vs the number of clusters.

2.2 Algorithm Definition

K-means clustering follows a straightforward technique for categorizing a given data set into a predetermined number of clusters, denoted by 'k'. The clusters are then positioned as points, and all observations or data points are linked with the closest cluster, calculated, and adjusted. The procedure is then repeated utilizing the new adjustments until convergence is achieved. Given 'n' samples and 'k', the basic k-means algorithm is as follows:

Begin

initialize $\mu_1, \mu_2, \ldots, \mu_k$ (randomly selected)

do classify *n* samples according to nearest μ_i

recompute μ_i until no change in μ_i

return $\mu_1, \mu_2, \ldots, \mu_k$

End

3. Experimental Evaluation

3.1 Methodology

The algorithm's primary component operates via a two-step procedure known as expectation-maximization. In the expectation stage, each data point is assigned to its closest centroid. The maximizing stage next computes the mean of each cluster's points and establishes the new centroid. Computing the sum of the squared error (SSE) when the centroids converge or match the previous iteration's assignment determines the cluster assignment quality.

$$J_e = \sum_{i=1}^{C} \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2$$

Eq 1. Sum of Squared Error Formula

SSE is the sum of each point's squared Euclidean distance to its centroid as follows:

$$d = \sqrt{[\,(x_2 - x_1)^2 + (y_2 - y_1)^2]}$$

Eq 2. Euclidean Distance Formula

The centroid mean is calculated by averaging all the data points in that cluster and this point can change with each iteration to optimally represent the cluster points. It is computed as:

$$C_i = \frac{1}{||S_i||} \sum_{x_j \epsilon S_i} x_j$$

Eq 3. Centroid mean Formula

In K-means, the optimization criterion is to minimize the sum of squared error. This is given by the objective function which decreases as the 'K' value increases. The objective function is evaluated as the 'argmin' of the following formula in the code:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Eq 4. Objective Function Formula

3.1.1 Initialization Strategy 1

In this strategy, we are following the basic k-means algorithm by randomly choosing an initial cluster center from the given samples of the dataset. First, the number of clusters (k) in the final solution must be specified in k-means clustering. An initial center for each cluster is generated by randomly picking k items from the data set, also known as centroids or cluster means. Next, each residual object is allocated to its nearest centroid based on its Euclidean distance (Eq. 2) from the cluster mean, known as the Cluster assignment step. The next step is the centroid update where the method computes each cluster's new

mean. After recalculating the centers, every observation is rechecked to determine whether it belongs to a different cluster. Iteratively repeat cluster assignment and centroid update until cluster assignments achieve convergence.

3.1.2 Initialization Strategy 2

The second strategy is a smart and more efficient initialization strategy using the K-means++ technique and is computationally faster. The steps followed in this method are:

- Pick the initial cluster center randomly, denoted by Ci.
- For the i-th center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal.

$$d_i = max_{(j:1 \mapsto m)} ||x_i - C_j||^2$$

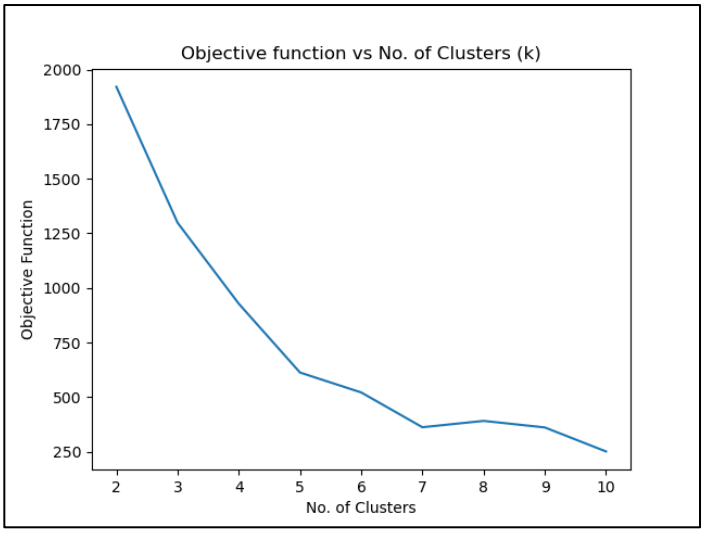
- Repeat the above steps iteratively until convergence is achieved and k-centroids are formed.

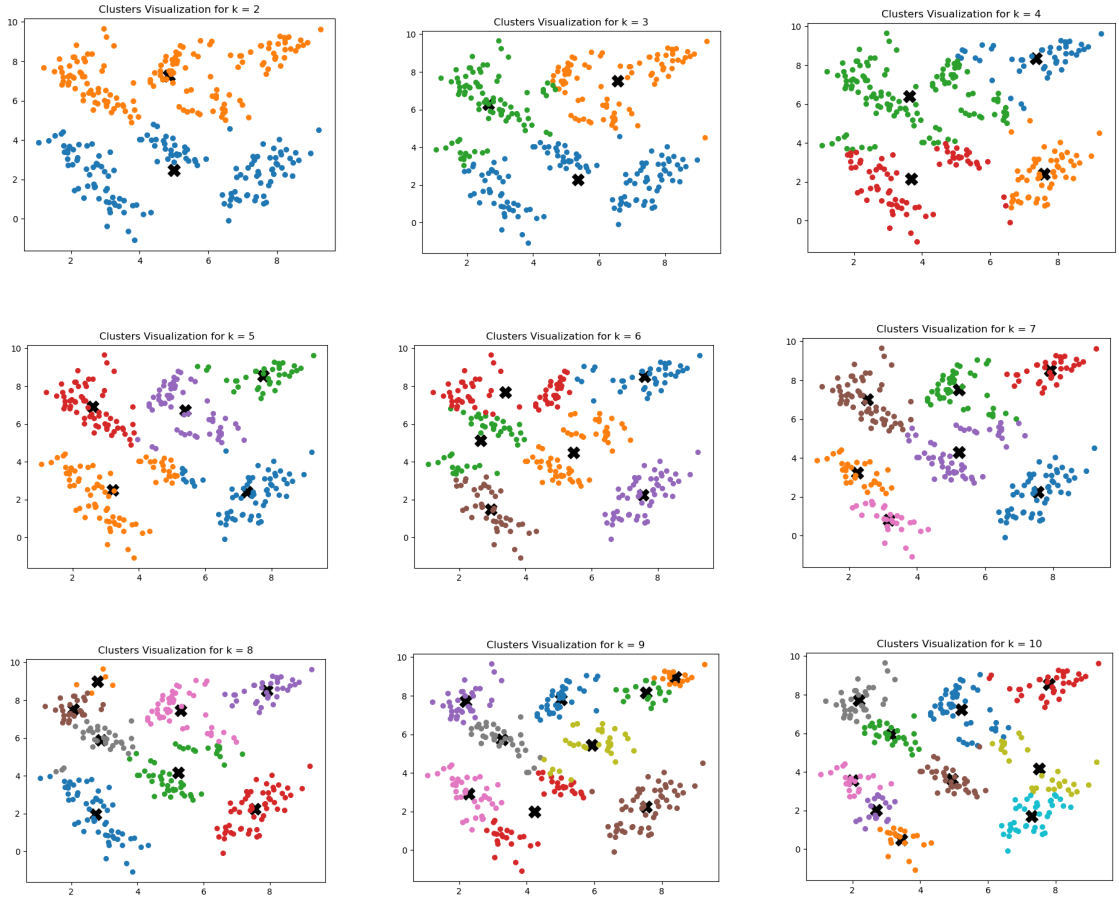
4. Results and Conclusion

The outcomes of k-Means are highly dependent on the initialization strategy, and k-means ++ (strategy 2) is computationally faster and more optimal in terms of performance and runtime. The results obtained from the two initialization strategies discussed above is shown below:

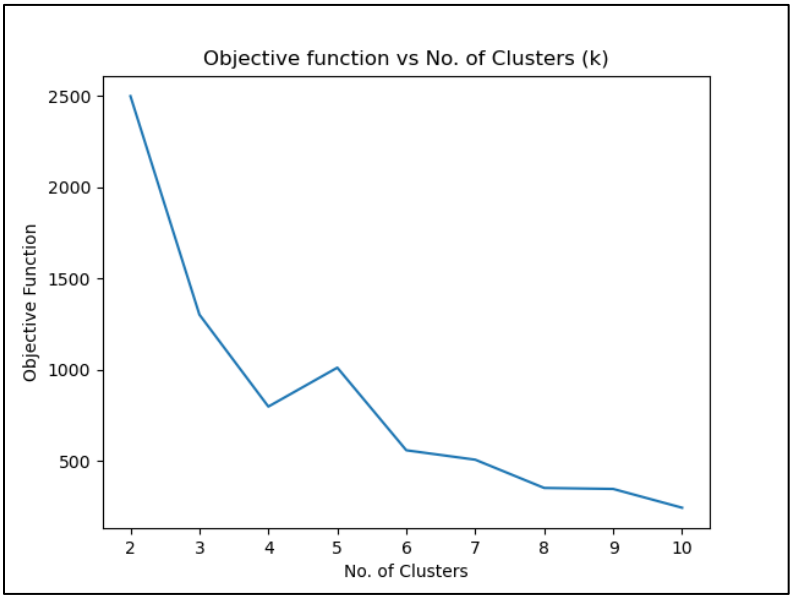
4.1 K-Means Clustering results using Strategy 1

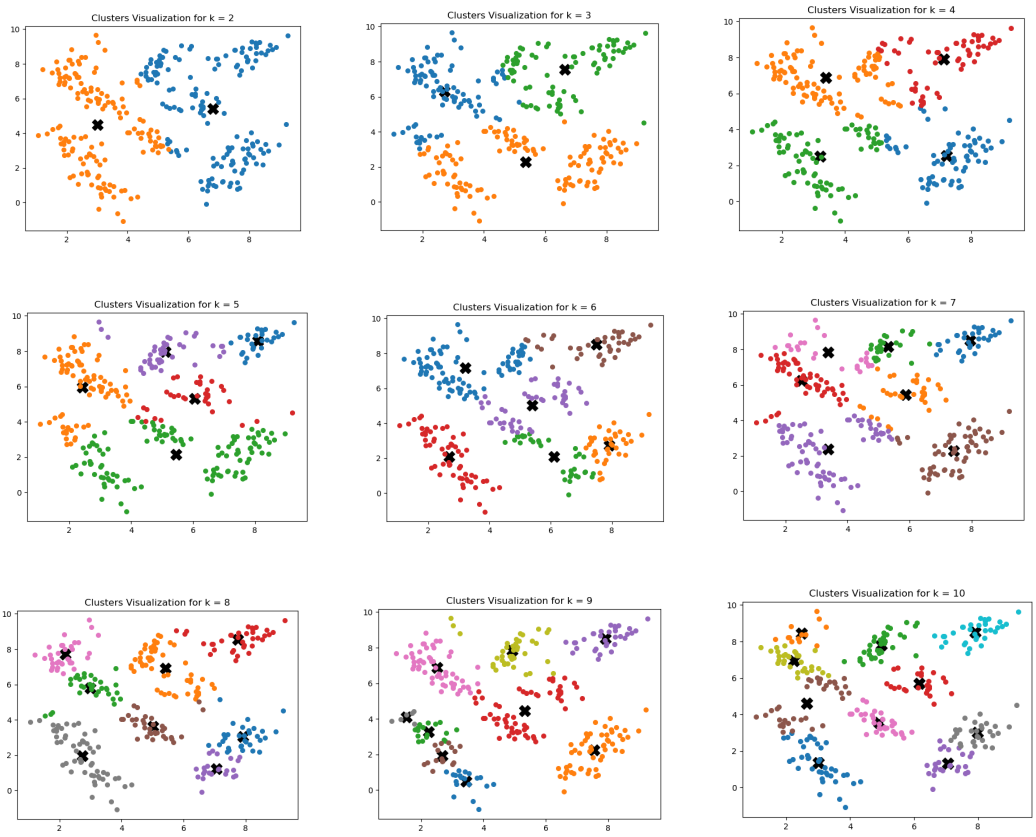
4.1.1 First run with random initialization 1





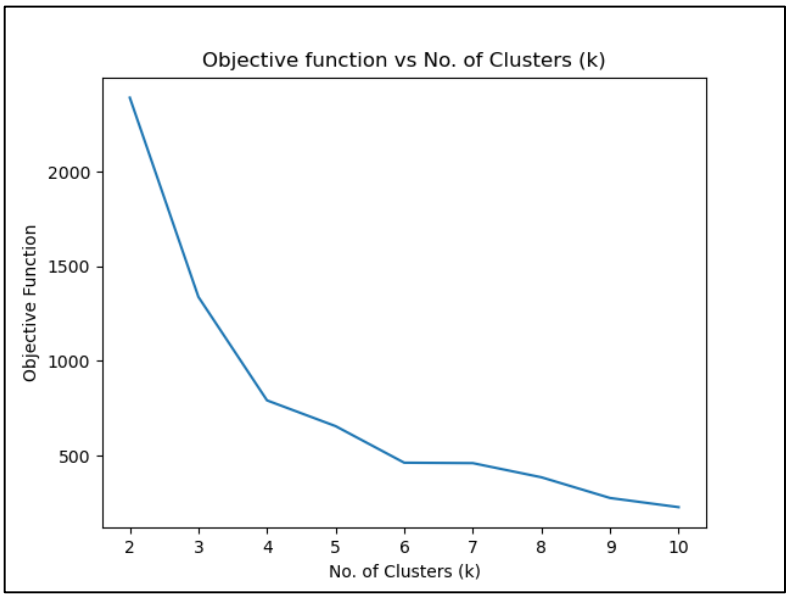
4.1.2 Second run with random initialization 2

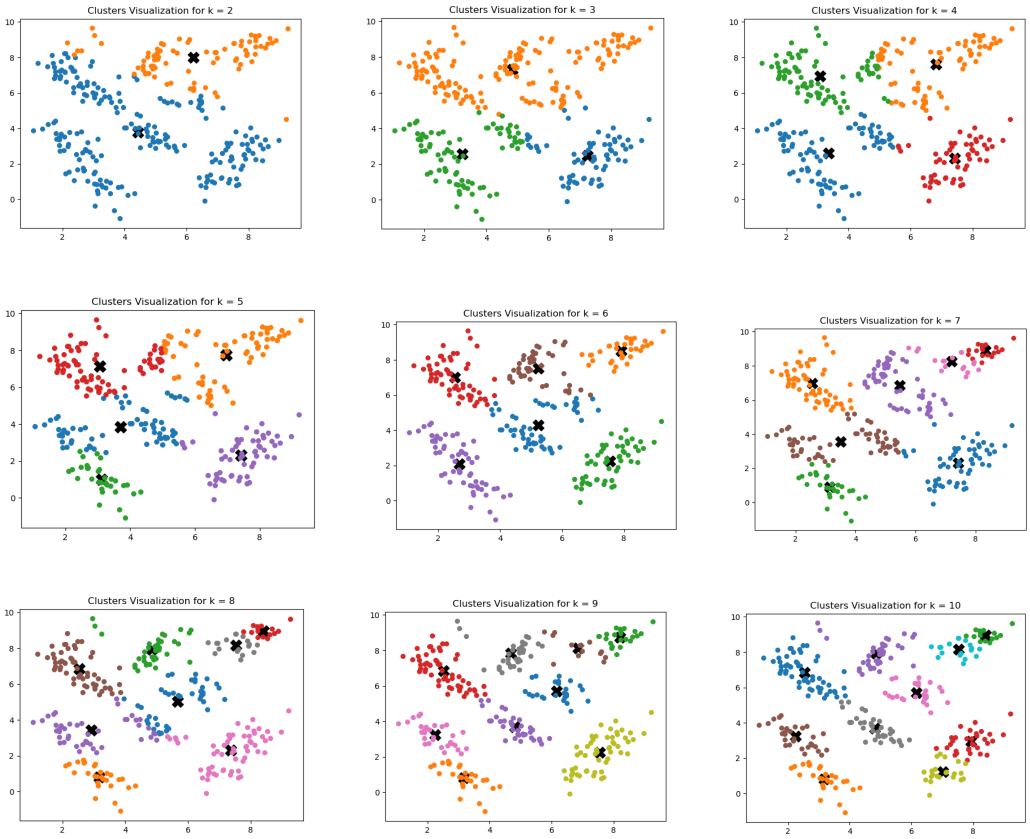




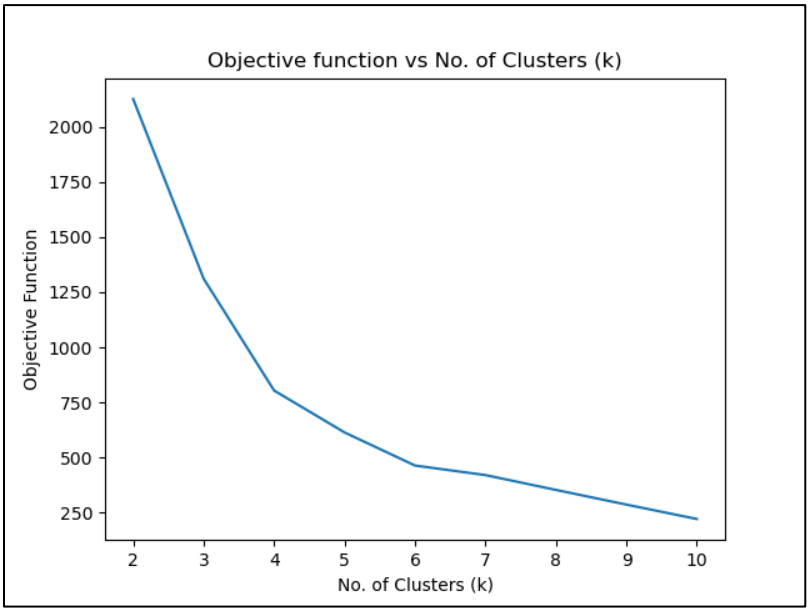
4.2 K-Means Clustering results using Strategy 2

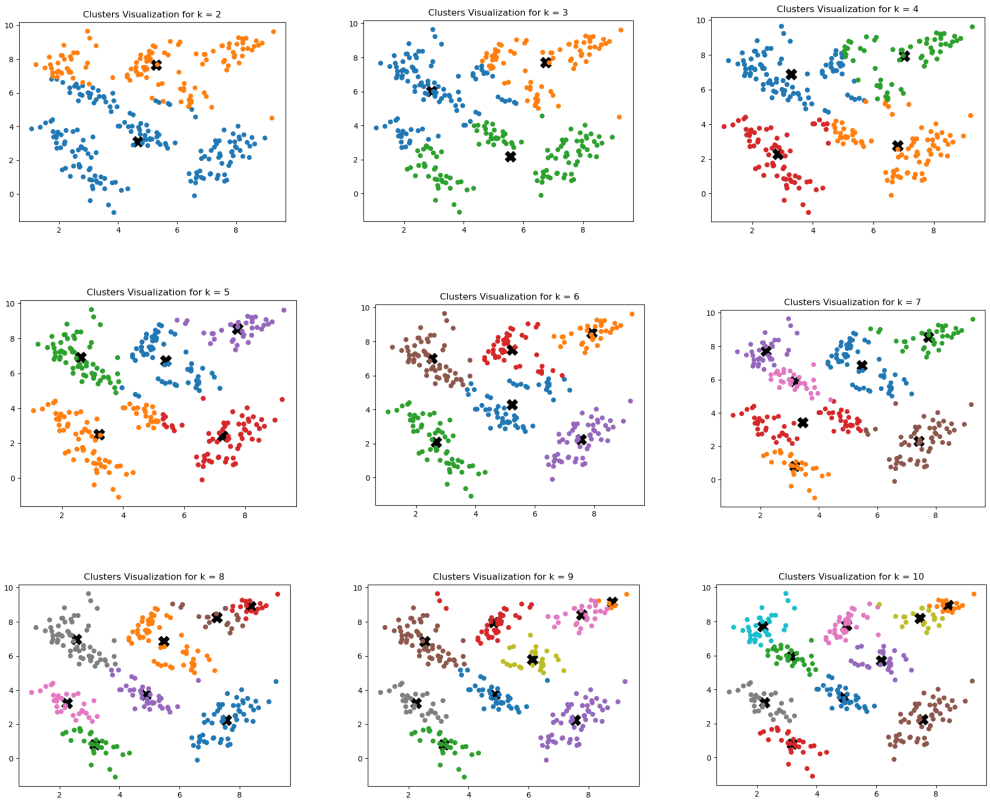
4.2.1 First run with random initialization 1





4.2.2 Second run with random initialization 2





5. Software Specifications

1. Python version: 3.9.12 **2. Python IDE used**: Sublime Text Editor **3. Run environment**: Terminal

References

1. https://en.wikipedia.org/wiki/K-means_clustering

2. https://en.wikipedia.org/wiki/K-means%2B%2B