Unsupervised Learning (K-means)

Ramesh Babu Mannam
Arizona State University
Tempe, AZ
rmannam1@asu.edu

Akhil Maddikonda
Arizona State University
Tempe, AZ
amaddiko@asu.edu

July 2024

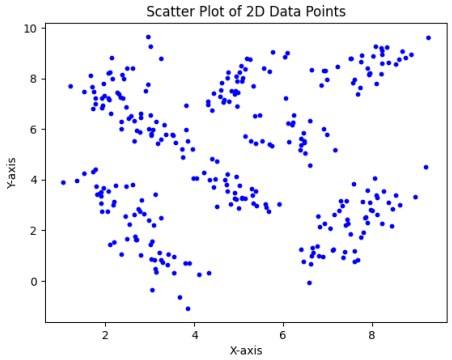
1 Introduction

The goal of this project is to apply and compare the k-means clustering method on a 2-D point dataset using two distinct initialization strategies. The first strategy involves randomly selecting starting centers from the dataset, whereas the second strategy chooses the first center at random and following centers to maximize the average distance from all previously picked centers. By applying these techniques to a dataset of 2-D points and the number of clusters (k) 2 to 10, we want to examine and compare the performance of the two different initialization approaches using the objective function values obtained.

2 Methodology

2.1 Data

The dataset consists of 300 2-D points provided in a '.mat' file named 'AllSamples.mat'.



2.2 K-means Algorithm:

The k-means algorithm is a clustering technique that organizes a dataset into k clusters, with each cluster represented by its center, or centroid. The process starts by initializing k centroids, either randomly or via specified algorithms. Each data point is allocated to the closest centroid, resulting in k clusters. The centroids are recalculated as the average of all locations inside each cluster. This process of assignment and updating is repeated iteratively until the centroids stabilize, therefore reducing the objective function, which is the sum of squared distances between points and their centroids.
The objective function for the k-means algorithm is given by:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mu_i\|^2 \tag{1}$$

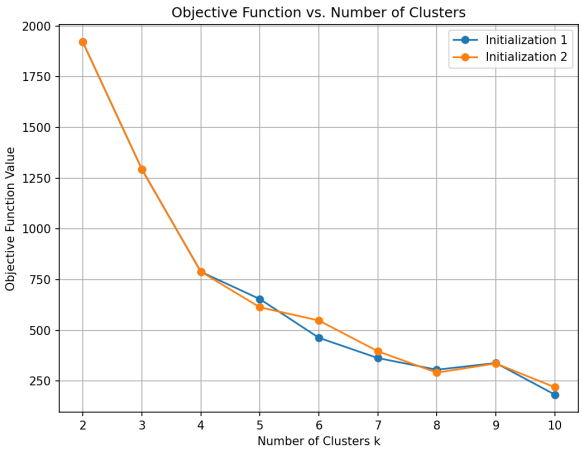
where $\mathbf{x}$ represents a data point, μ_i is the centroid of the i-th cluster, and D_i is the set of points in the i-th cluster.

2.3 Initialization Strategies:

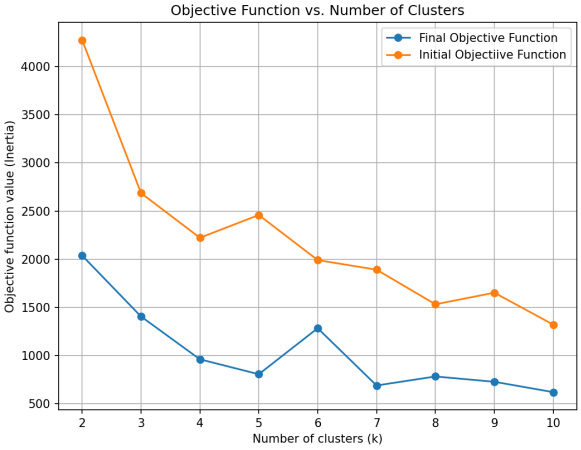
- **Strategy 1:** Randomly selects k initial cluster centers from the data points.
- **Strategy 2:** Select the first center randomly, then choose each subsequent center to be the point farthest from the existing centers. For the i-th center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous $(i-1)$ centers is maximal.

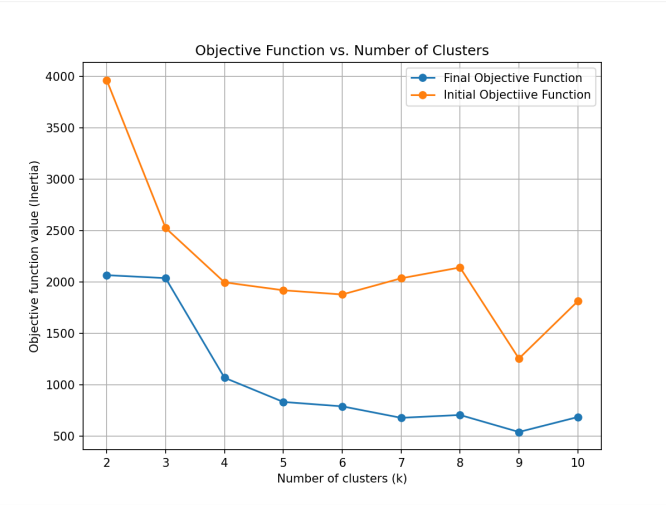
3 Results

- **Strategy 1:**



- **Strategy 2:**





4 Conclusion

Cost of the function of strategy 2 is going to be less than strategy 1 no matter what 1st cluster is since the initial clusters are going to be spaced out. In strategy 1 all the initial clusters are chosen randomly, so in some instances the initial clusters might be next to each other.