Name: Parth Bhatt
Student ID: 1220528816

Project 2 K-means Clustering

Introduction

The objective of this project was to develop a K-means clustering algorithm based on two strategies on the data provided. The dataset was of length 300, meaning we were provided with 300 X and Y coordinate values.

Approach

The following steps were performed on the initial points and their respective cluster provided in Strategy 1:

- Find the euclidean distance between the data points and initial centroids and cluster then by assigning the closed initial centroids.
- Take the mean of data points in those clusters and change the centroid points.
- Perform the first two steps until the newly calculated centroids are the same as the previous selected centroids.

Last four of ID: 8816	K=3	K=5
Initial Centroids	[[6.79251832 2.56208095] [7.56399709 7.83135288] [8.527899 8.55183237]]	[[6.39627447 1.24125663] [5.07250754 7.89834048] [6.79251832 2.56208095] [4.95728696 6.90897984] [7.85355511 2.53104656]]

The following steps were performed on the initial points and their respective cluster provided in Strategy 2:

- Find the average euclidean distance between the data points and previous centroids and find the next initial centroids repeat until you have found all initial centroids based on the number of clusters.
- Once you have the initial centroids, find the euclidean distance between the data points and initial centroids and cluster them by assigning the closed initial centroids.
- Take the mean of data points in those clusters and change the centroid points.
- Perform the first two steps until the newly calculated centroids are the same as the previous selected centroids.

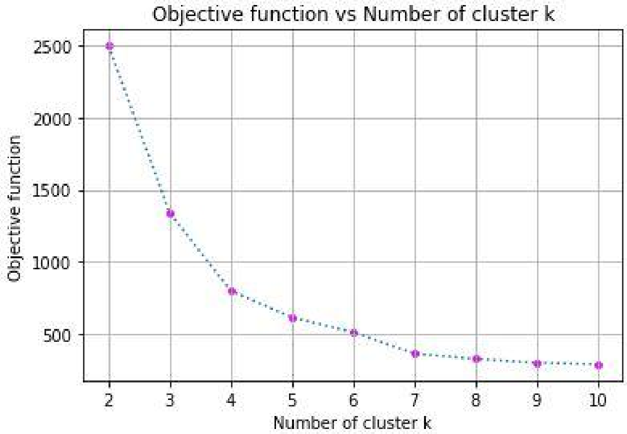
Last four of ID: 8816	K=4	K=6
Initial Centroids	[3.35409838 5.79603723]	[4.95185958 4.11756694]

Conclusion

- Strategy 1:

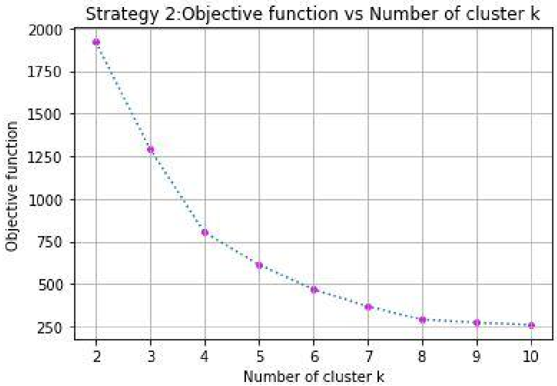
Last four of ID: 8816	K=3	K=5

Final Centroids	[[5.47740039,2.25498103],[2.56146449,6.08861338],[6.49724962,7.52297293]]	[[2.68198633,2.09461587],[6.7786424,8.07967641],[5.22321274,4.22502829],[2.87490813,7.01082281],[7.55616782,2.23516796]]
Loss-Objective Function	1293.777452391135	598.5546443663119



- Strategy 2:

Last four of ID: 8816	**K=4**	**K=6**
Final Centroids	[[3.30296804,2.55443267],[6.85658333,7.6614342],[7.34802851,2.35222497],[3.153427 , 6.9129207]]	[[5.23053667,4.2793425],[7.91430998,8.51990981],[2.68198633,2.09461587],[5.24028296,7.53131029],[7.55616782,2.23516796],[2.54165252,7.00267832]]
Loss-Objective Function	792.5378104413305	462.9263558248373



- **Strategy 2 will perform better than Strategy 1 as K-means clustering depends on initial centroids. Systematically finding the centroids as in Strategy 2 will provide less loss. The optimal number of clusters is 7 as we can see from the graph of objective function vs number of clusters that after the k=7 the loss function is not changing drastically.**