Project Part 2: Unsupervised Learning (K-means)

Project Report

By Arun Deepak Chandrasekar

The project focuses on implementing an unsupervised learning technique namely K-Means Clustering on the dataset given which consists of 2d data points. Major tasks which are discussed below are as follows:

- Strategy-1- K-means is implemented by selecting the initial centres randomly from the given sample. The algorithm is repeated with different number of cluster-centers (2-10). For each iteration i.e. k value the value of objective function J(k) is calculated using equation1. The entire process is repeated with a different initialization.

- Strategy-2 – Here the initialization of centroids is done slightly different (equation 2). The first center is randomly picked ; for the i-th center (i>1), a sample is chosen (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal. The algorithm is repeated with different number of cluster-centers (2-10). For each iteration i.e. k value the value of objective function J(k) is calculated using equation1. The entire process is repeated with a different initialization.

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \quad \text{(Eq.1)}$$

$$argmax_i \ \frac{1}{n} \sum_j D(C_i, C_j) \quad \text{(Eq.2)}$$

I.Strategy-1

a) Initialization-1

The following results were obtained when objective function was calculated using equation 1 during the first iteration of strategy-1.

Table-1: Strategy1- Initialization-1

k- number of clusters	J(k)- Objective Function
2	704.0045894890272
3	578.9151004150797
4	449.2357077972007
5	405.25559703531366
6	331.88765890167144
7	309.7280510105324
8	270.9568638238103
9	256.88722755064384
10	260.5352157487646

b) Initialization-2

The following results were obtained when objective function was calculated using equation 1 during the second iteration of strategy-1.

Table-2: Strategy1- Initialization-2

k- number of clusters	J(k)- Objective Function
2	800.6496739316412
3	566.5062430966651
4	449.4788074924878
5	387.8315814191166
6	365.41034866385746
7	357.35132824395555
8	266.6500564736015
9	235.2378107238663
10	222.2388710829485

The objective function vs K graph for Strategy-1 was obtained and the values for different initialization were found to be varying.

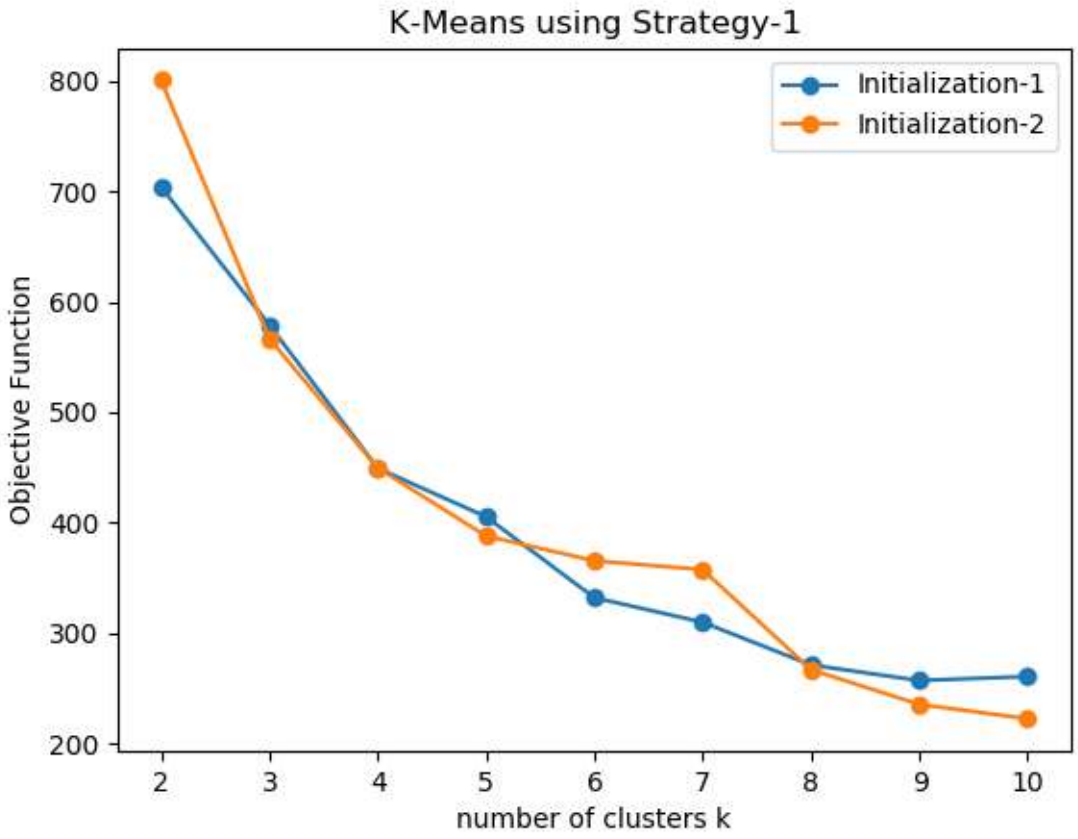

Figure-1: Represents Objective Function vs k for Strategy1, initialization1 &2

II. Strategy 2

a) Initialization-1

The following results were obtained when objective function was calculated using equation 1 during the first iteration of strategy-2.

Table-3: Strategy2- Initialization-1

k- number of clusters	J(k)- Objective Function
2	704.0045894890272
3	577.791769193311
4	449.19165364854763
5	386.5125096769727
6	342.8467558306005
7	334.29966925123836
8	285.86147937845465
9	276.3261195032295
10	241.0002965709021

b) Initialization-2

The following results were obtained when objective function was calculated using equation 1 during the second iteration of strategy-2.

Table-4: Strategy2- Initialization-2

k- number of clusters	J(k)- Objective Function
2	704.0045894890272
3	577.791769193311
4	449.19165364854763
5	386.5125096769727
6	342.8467558306005
7	334.29966925123836
8	285.86147937845465
9	276.3261195032295
10	241.0002965709021

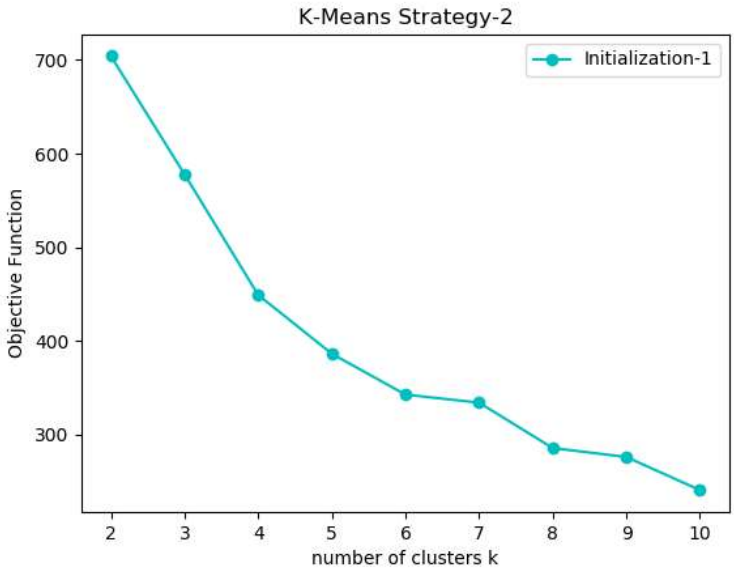

Figure-2: Represents Objective Function vs k for Strategy2, initialization1

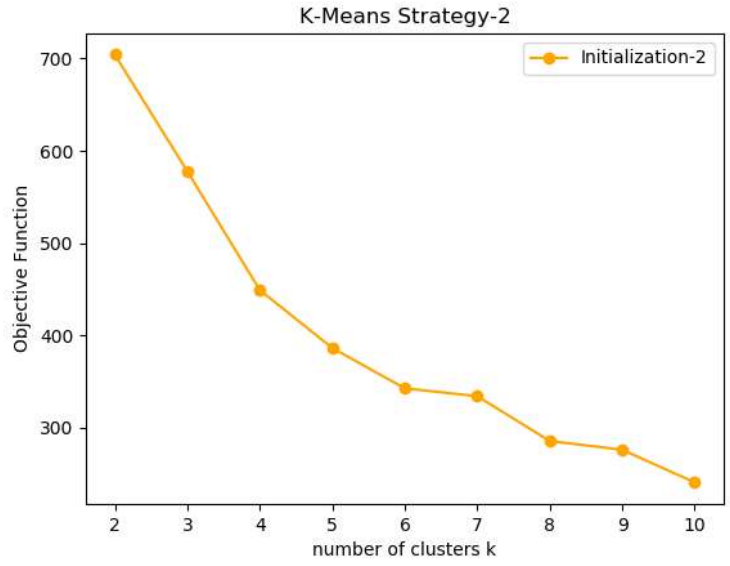

Figure-3: Represents Objective Function vs k for Strategy2, initialization2