Project Part 2 – K Means Algorithm

Table of Contents

Goal:

In this project, main goal is to implement the k-means algorithm for two different initialization strategies under unsupervised learning.

Implementation:

Strategy 1:

In strategy 1 initialization, I have chosen random centroids initially and calculated the cluster numbers for each data point using Euclidean distance from each centroid. And I have wrote algorithm for computing the new centroid for each cluster data point. Then recomputed the centroids of newly formed cluster. I have repeated this process of picking new centroids until centroid stop changing. After that I have calculated the objective function value using the formula given below.

$$\text{objective function} \leftarrow J = \sum_{j=1}^{\overset{\text{number of clusters}}{k}} \sum_{i=1}^{\overset{\text{number of cases}}{n}} \underbrace{\left\| \overset{\text{case } i}{x_i^{(j)}} - \overset{\text{centroid for cluster } j}{c_j} \right\|^2}_{\text{Distance function}}$$

Strategy 2:

Here only the way of initialization of centroids changes compared to strategy 1. Rest of the steps will be same. This is similar to what we do in strategy 1, but instead of randomly picking all the centroids, I have just picked one centroid, next I have computed the distance of each data point from the cluster center that has already been chosen. Then, choose the new cluster centroid from the data points such that average distance of this chosen one to all previous is maximum.

Results:

Strategy 1:

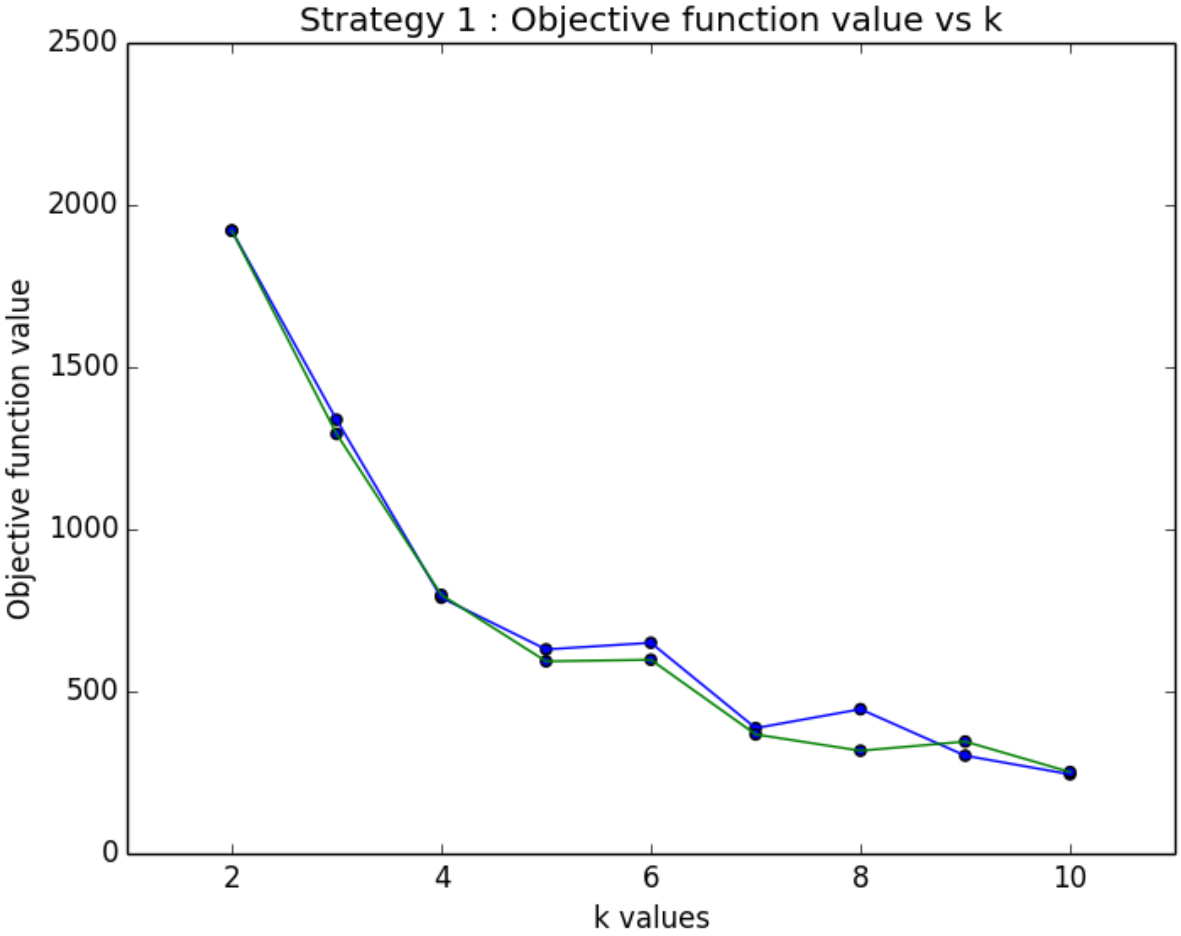
The following are the values of the results for strategy 1 with two different initializations. Below results is in the format of (K, Objective function value)

```
#### Strategy 1 ####

 Initialization 1:

    ( 2 ,  1921.03348586 )
    ( 3 ,  1338.10760165 )
    ( 4 ,  789.237972218 )
    ( 5 ,  629.762167013 )
    ( 6 ,  650.192889316 )
    ( 7 ,  387.08192944 )
    ( 8 ,  445.409633252 )
    ( 9 ,  302.116761706 )
    ( 10 ,  244.636912629 )

 Initialization 2:
    ( 2 ,  1921.03348586 )
    ( 3 ,  1294.21031543 )
    ( 4 ,  797.960184079 )
    ( 5 ,  592.937572966 )
    ( 6 ,  597.820774756 )
    ( 7 ,  367.598821264 )
    ( 8 ,  317.328917784 )
    ( 9 ,  345.581878359 )
    ( 10 ,  251.681373332 )
```

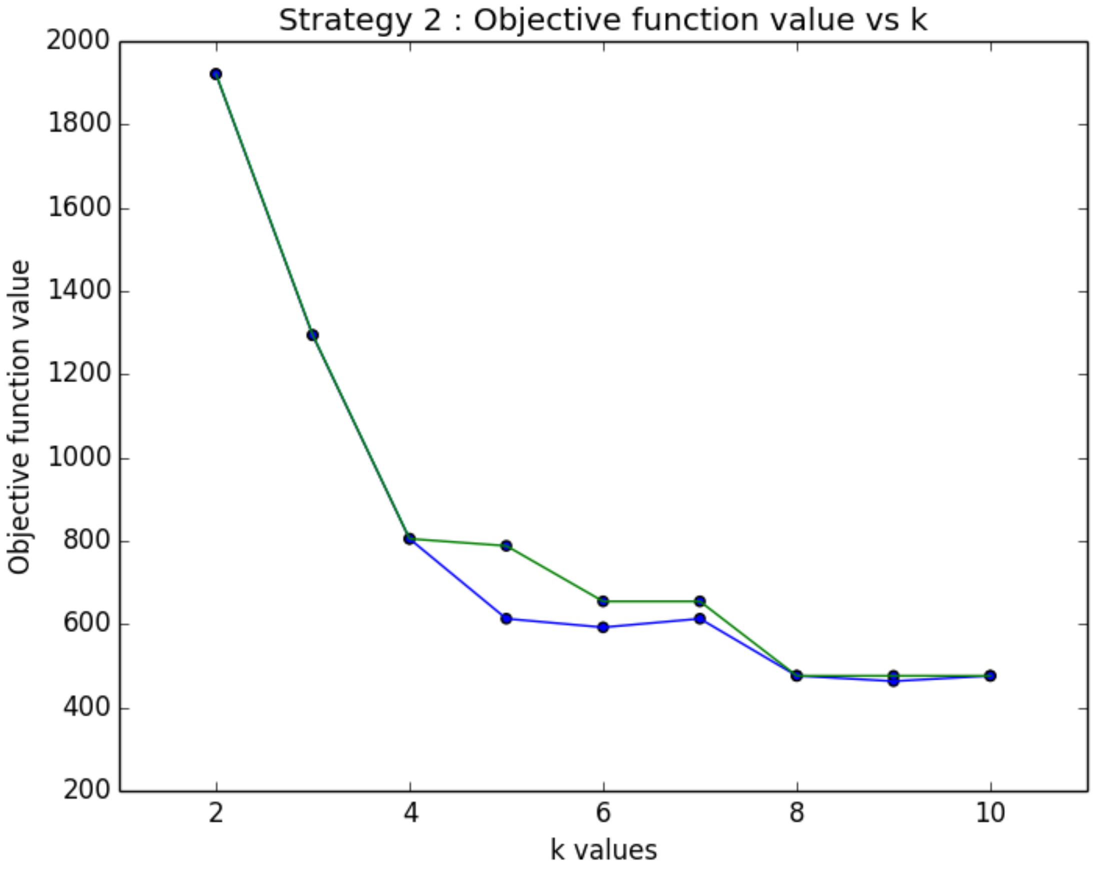


Strategy 2:

```
#### Strategy 2 ####

 Initialization 1:
    ( 2 ,  1921.03348586 )
    ( 3 ,  1293.77745239 )
    ( 4 ,  805.116645747 )
    ( 5 ,  613.282439206 )
    ( 6 ,  592.528384259 )
    ( 7 ,  613.282439206 )
    ( 8 ,  476.296570527 )
    ( 9 ,  463.213686166 )
    ( 10 ,  476.118751676 )

Initialization 2:
    ( 2 ,  1921.03348586 )
    ( 3 ,  1294.29841749 )
    ( 4 ,  805.116645747 )
    ( 5 ,  788.271640049 )
    ( 6 ,  654.877909067 )
    ( 7 ,  654.877909067 )
    ( 8 ,  476.118751676 )
    ( 9 ,  476.118751676 )
    ( 10 , 476.118751676 )
```



Analyzing Results:

1. On every initialization of k-values, the objective function value stayed the same or decreased for both the strategies.

2. End results are insensitive to the initial centroids. In each strategy, with both initializations, the initial centroids may be different, but objective function values are following the same pattern, which is: as K increases, objective function values decrease. The graphs above clearly explains this pattern.

3. When the K value varied from 2 to 10, the objective function value reduced sharply at first and then this decrease in the objective function value reduces and eventually becomes constant. So even if we increase the K value above 10, the objective function value might not decrease further.

4. As an anomaly, sometimes the objective function value might increase slightly for a single step. In strategy 1, I have noticed the result of the objective function value increased slightly when k value changed from 5 to 6 for both of the initializations. For the rest of the k values, it followed expected pattern, which is: as K increases, objective function value decrease. The same anomaly happened for the strategy 2 as well. In strategy 2, the objective function value stayed constant when k value changed from 4 to 5 in Initialization 2.