Yongbaek Cho
CSE 575: Statistical Machine Learning

Overview

K-means clustering algorithm is one type of unsupervised learning which used for cluster unlabeled data. A cluster is a collection of data that has similarities. In other words, Clustering means that when unlabeled data is given, it will be grouped into the cluster. In K-means clustering, K is the number of clusters and creating K clusters based on K centroids.

K-means algorithms works as follows:

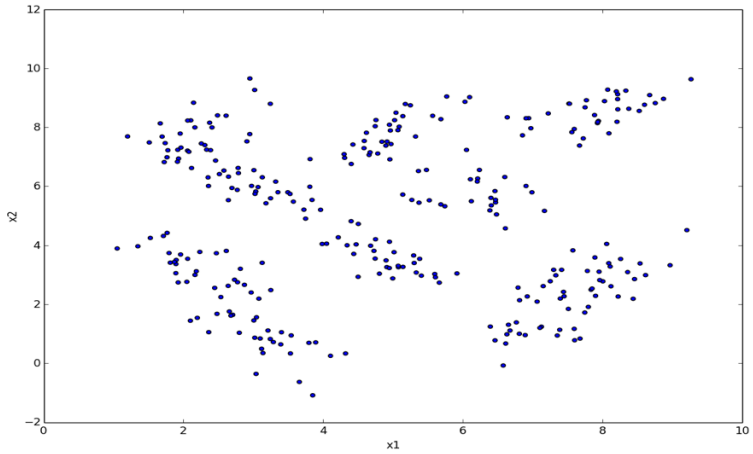
1. Specify number of clusters K.
2. Initialize centroids and randomly selecting K data points for the centroids.
3. Iterating until there is no change to the centroids.

Objectives

1. Strategy 1- Initial centroids are picked randomly from the dataset which is given.
2. Strategy 2- First centroid is picked randomly; 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. Implemented strategy1 and strategy 2 twice on the given data from K=2 to K=10.
4. Calculated the objective function: $\sum_{i=1}^{k} \sum_{x \in D} ||x - \mu||^2$
5. Plot the X-axis as number of clusters, Y-axis as the value of objective function.

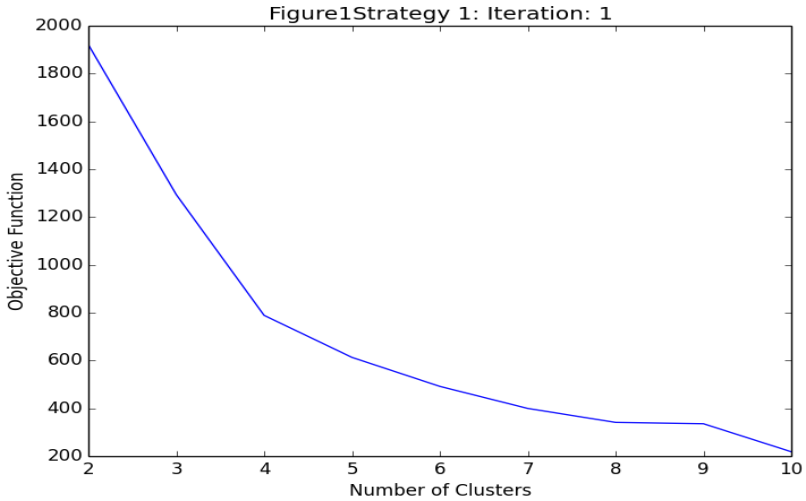
Dataset

The dataset has 2 columns and 300 rows. Here is the plot for dataset:



Strategy 1

In this strategy, Initial centroids are picked randomly for K = 2 to k = 10from the dataset which is given.
Here is the plot for ***Strategy 1; iteration 1:***



Here is the value of Objective Function:

	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Value of Objective Function	1921.03	1294.29	788.96	613.28	492.00	399.82	341.36	335.80	218.67

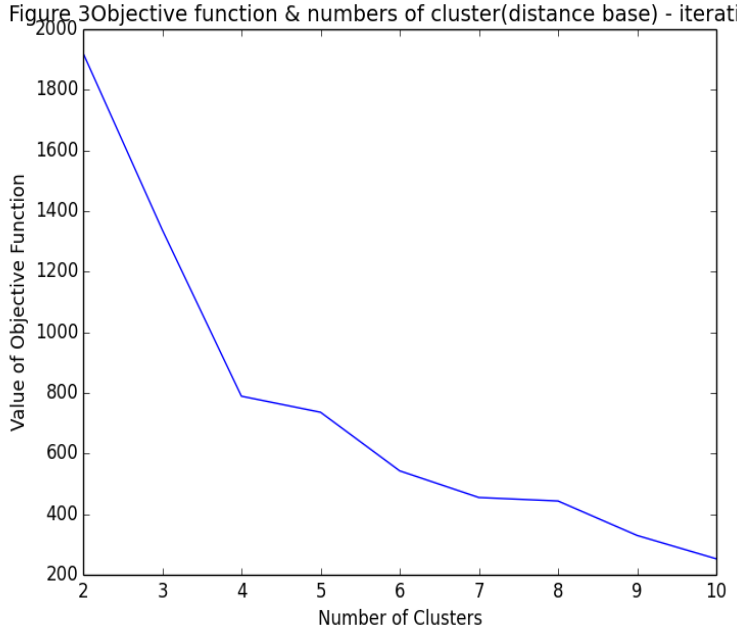
Strategy 1; iteration 2

	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Value of Objective Function	1921.03	1338.07	788.96	592.06	561.20	390.91	426.93	313.74	283.69

Strategy 2

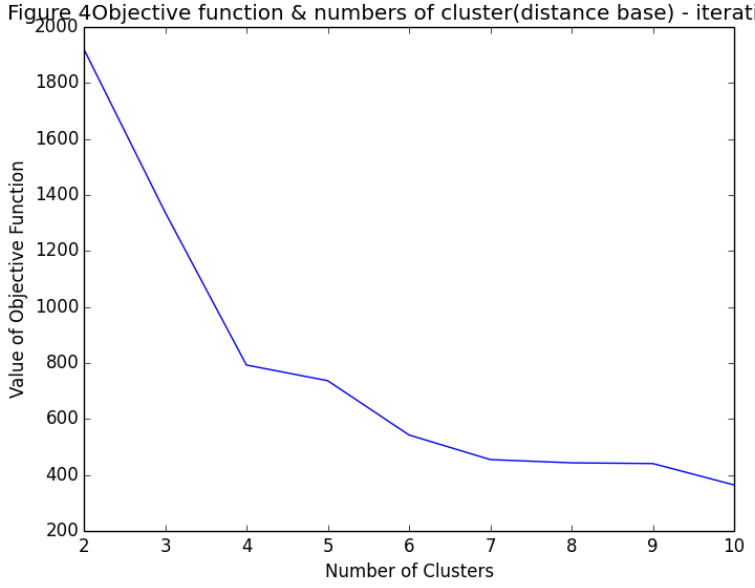
First centroid is picked randomly; for the i-th center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal.

Figure 3Objective function & numbers of cluster(distance base) - iteration: 1



	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Value of Objective Function	1921.03	1338.13	789.23	736.25	542.94	454.99	443.477	329.78	252.70

Figure 4Objective function & numbers of cluster(distance base) - iteration: 2



	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Value of Objective Function	1921.03	1338.07	792.71	736.76	542.94	454.99	443.47	440.85	364.47

Conclusion

Strategy1 and Strategy2 was run twice and some results can be made from each strategy.

1. The figures prove that the slope shows steeper decline from the value of k is 2 to 4 and after that slow decline from the value of k is 4 to 10. The value of k will be the ideal value for the showing this occurrence. From the dataset and each strategy, the ideal value of k is 4.

2. Both strategies seemed enough to show the result, but the strategy 2 seemed to be better way to get the result and ideal value of k. Also, this shows how the selection of the initial centroids is important to implement k-means Algorithm and has impact on the strategy. To sum up, the selection of strategy 2 seems to better than the strategy 1 to get a good result.