K-means Clustering

Subhash Chandra Bose Vuppala
svuppal6@asu.edu

***Abstract*—We implement the k-means algorithm and apply the implementation on the given dataset, which contains a set of 2-D points. We implement two different strategies for choosing the initial cluster centers. Then we assign the cluster numbers to the given samples and we find the mean of the clusters to find the new cluster centroids. We repeat the same steps until the centroids do not change and find the final cluster centroids. Finally, we compute the objective function as a function of K and by using the elbow method we get the optimum K value.**

***Index Terms*—K-means algorithm, cluster centroids, elbow method**

I. Introduction

Unsupervised learning is the practice of teaching a machine to work without control on data that hasn't been categorized or labelled. Without any prior data experience, the machine's task is to sort unsorted data into groups based on similarities, trends, and differences. Since there is no teacher present, unlike supervised learning, the machine will not be educated. As a result, the computer's ability to find hidden structure in unlabeled data on its own is restricted. Unsupervised learning can be divided into two types. Unsupervised learning is classified into two categories. They are clustering and association.

Clustering is the method of putting similar things together in a group. The aim of this unsupervised machine learning technique is to look for similarities in data points and group them together. This will allow us to see the underlying dynamics of various classes. You can classify different groups/segments of customers and market each category in a different way to optimize sales by grouping unlabeled data. Association rule learning is a technique for determining whether one data item is dependent on another and then mapping accordingly to make it more profitable. It looks for interesting relationships or correlations between the dataset's variables.

There are many types of Clustering, and the popular ones are k-means Clustering and hierarchical Clustering. The k-means clustering is a vector quantization process that aims to divide n observations into k clusters. Each observation belongs to the cluster with the closest mean (cluster centers or cluster centroid), which serves as the cluster's prototype. As a consequence, the data space is partitioned into Voronoi cells. Within-cluster variances (squared Euclidean distances) are minimized by k-means clustering, but not normal Euclidean distances, which is the more difficult Weber problem. There are two different strategies used for choosing the initial cluster centers. They are

- From the given samples randomly picking the initial centers.
- Picking the first center randomly and for the i-th center where i¿1, we choose a sample such that the average distance of this chosen one to all previous (i-1) centers is maximal.

There are three different stopping criteria for k-means clustering. They are

- If the centroids of the clusters are not changing.
- If the data points remain in the same cluster.
- If the maximum number of iterations is achieved.

II. Description of Solution

In this project, we should implement the k-means algorithm and apply the implementation to the given dataset, which contains a set of 2-D points.

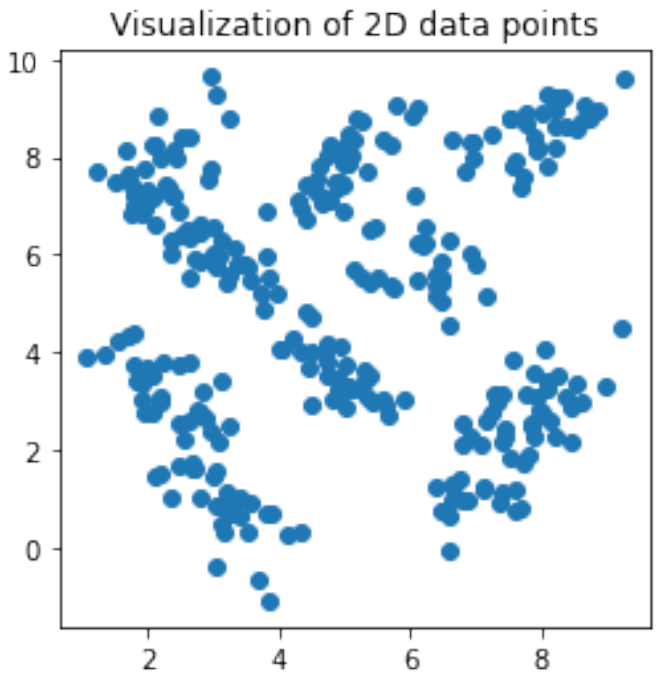

Fig. 1. Visualization of data points.

We implement two different strategies for choosing the initial cluster centers.

A. STRATEGY 1

There are seven steps in this strategy. They are shown below:

a) step-1: We randomly pick the initial centers from the given samples.

b) step-2: We assign the cluster numbers to the given samples.

c) step-3: By forming the clusters we find the mean of the clusters to find the new cluster center.

d) step-4: We repeat this until the centers don't change.

e) step-5: We compute the objective function as a function of k (k = 2, 3, . . . , 10).

f) step-6: We plot a graph between k and objective function and find the optimal k-value by using elbow method.(Elbow method means there will be a sharp elbow in the graph of explained variation versus clusters.)

g) step-7: We repeat all these steps with another initialization.

B. STRATEGY 2

There are nine steps in this strategy. They are shown below:

a) step-1: We pick the first center randomly.

b) step-2: Then we choose the point which is farthest from the first center and make it as the second center.

c) step-3: Based on k value we select the center based on the maximum distance between the average of the centers.

d) step-4: Then we assign the cluster numbers to the given samples.

e) step-5: After forming the clusters we find the mean of the clusters to find the new cluster center.

f) step-6: We repeat this until the centers don't change.

g) step-7: We compute the objective function as a function of k (k = 2, 3, . . . , 10).

h) step-8: We plot a graph between k and objective function and find the optimal k-value by using elbow method.(Elbow method means there will be a sharp elbow in the graph of explained variation versus clusters.)

i) step-9: We repeat all these steps with another initialization.

The mean is computed using the formula

$$\mu = \frac{1}{n}\sum_{i=i}^{n} x_i \tag{1}$$

where 'n' is number of samples

The objective function of K-means clustering is defined as

$$\sum_{k=1}^{K}\sum_{x\in D_i} ||x - \mu_k||^2 \tag{2}$$

where 'k' is number of clusters and cluster centroids are $\mu_1, \mu_2, ..., \mu_k$.

III. RESULTS

A. STRATEGY 1

a) FIRST INITIALIZATION: Fig 2 shows the graph obtained by plotting the objective function against the number of clusters k when strategy 1 is run for the first time.

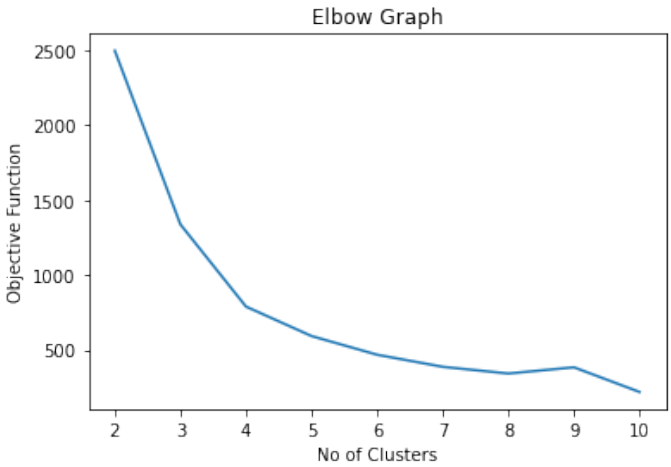

Fig. 2. Strategy-1: First initialization

Looking at the graph, we can see that at k = 4, there is a sharp decrease, i.e. an elbow, and the objective function's rate of decrease is decreasing from 4. As a consequence, we can conclude that k = 4 is the optimal cluster size.

b) SECOND INITIALIZATION: Fig 3 shows the graph obtained by plotting the objective function against the number of clusters k when strategy 1 is run for the second time.

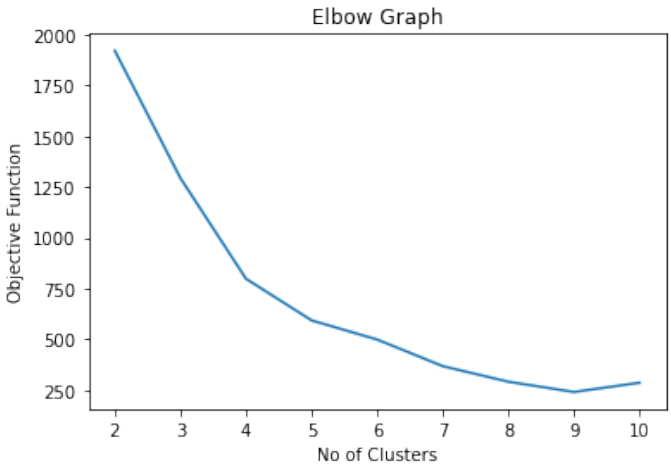

Fig. 3. Strategy-1: Second initialization

Looking at the graph, we can see that at k = 4, there is a sharp decrease, i.e. an elbow, and the objective function's rate of decrease is decreasing from 4. As a consequence, we can conclude that k = 4 is the optimal cluster size.

B. STRATEGY 2

a) FIRST INITIALIZATION: Fig 4 shows the graph obtained by plotting the objective function against the number of clusters k when strategy 2 is run for the first time.

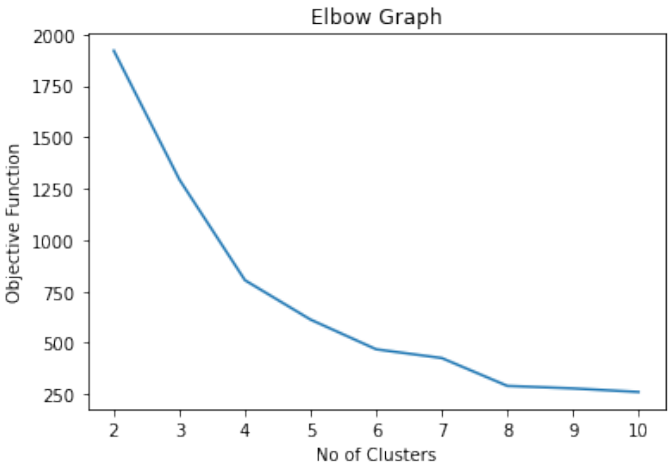

Fig. 4. Strategy-2: First initialization

Looking at the graph, we can see that at k = 4, there is a sharp decrease, i.e. an elbow, and the objective function's rate of decrease is decreasing from 4. As a consequence, we can conclude that k = 4 is the optimal cluster size.

b) SECOND INITIALIZATION: Fig 5 shows the graph obtained by plotting the objective function against the number of clusters k when strategy 2 is run for the second time.

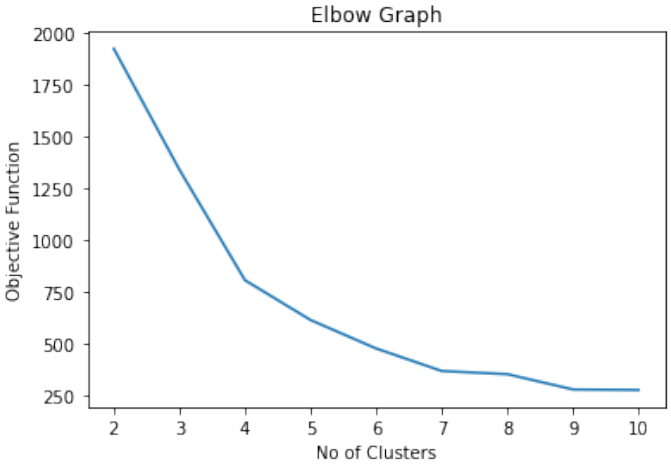

Fig. 5. Strategy-2: Second initialization

Looking at the graph, we can see that at k = 4, there is a sharp decrease, i.e. an elbow, and the objective function's rate of decrease is decreasing from 4. As a consequence, we can conclude that k = 4 is the optimal cluster size.

IV. LESSONS LEARNED

- I got to know about unsupervised learning and its usage in the current world.
- I got to know about the background of Clustering and its usage in finding the segments of people in the real-world market.
- When we have little knowledge about the data, K-means clustering is a quick and efficient algorithm for classifying data points into categories.
- I learned different strategies for choosing initial cluster centers and different stopping criteria for K-means clustering.
- Any new data can be easily allocated to the correct group once the algorithm has been run and the groups have been identified.
- If only one cluster forms naturally, the k-means algorithm is unlikely to produce the desired results.
- I got to know that k-means is sensitive to initialization of initial cluster centers.
- If the initialization is not done correctly, the cluster centroids will take local maximum.
- I learned about the elbow method and how it is used to find the optimal K value.
- From all the graphs, I have learned that k=4 is the optimal number of cluster values.

REFERENCES

[1] Aristidis Likas, Nikos Vlassis, Jakob J.Verbeek on The global k-means clustering algorithm. Amsterdam: May 2002.
[2] D T Pham, S S Dimov, C D Nguyen on Selection of K in K-means clustering. Uk: Jan 2005.
[3] Paul S. Bradley and Usama M. Fayyad, on Refining Initial Points for K-Means Clustering. May 1998.
[4] Kiri Wagstaf, Claire Cardie , Seth Rogers and Stefan Schroedl on Constrained K-means Clustering with Background Knowledge.