Implementing K-Means Clustering

Rahat Bin Osman
160204083
dept. of CSE
Ahsanullah university of Science and Technology
Dhaka, Bangladesh

***Abstract*—K-means-clustering is one of the most popular unsupervised machine learning algorithm. Every data points are allocated among the number of k it has, which refers to the number of centroids that we need in the dataset. A centroid is the imaginary or real location representing the center of the cluster, while keeping the centroids as small as possible.**

***Index Terms*—Clustering, Centroids, Euclidean Distance, Mean Distance, Unsupervised Learning**

I. Introduction

Clustering is a set of techniques used to partition data into groups, or clusters. Clusters are loosely defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. In other words, the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset." The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

II. Experimental Design / Methodology

A. Plotting the data

We used 3000 unsupervised data from the dataset "data-k-mean.txt" to plot all the data in red o markers.

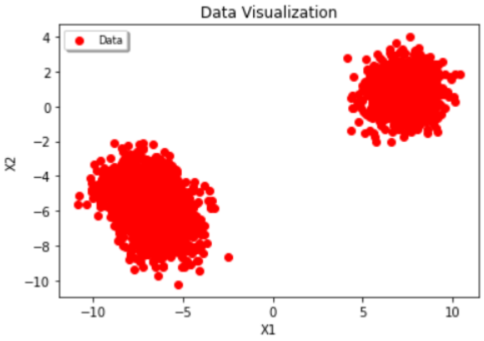

Fig. 1. Plotting the sample data

B. Performing the Algorithm

First, we need to randomly select k = 2 data points from the dataset, to initialize k centroids as a start. Then using the euclidean distance for two data point we will measure the nearest centroid for each data and cluster them. The euclidean distance for 2D space is given below.

$$Euclidean\ distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

After calculating the distance and assigning all the data point to cluster, we update the cluster centroid based on assigned data points. Cluster centroid update is simply the average of those assigned cluster data points.
This process repeats until the cluster centroids are remain same for previous iteration.

III. Result

The clustered data is plotted using different colors and markers in python using matplotlib library.

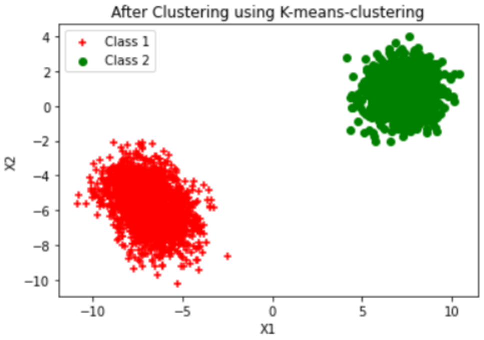

Fig. 2. Clustered data using K-means Clustering. k

IV. Conclusion

K-means clustering is an extensively used technique for data cluster analysis. Furthermore, it delivers training results quickly. However, its performance is usually not as competitive as those of the other sophisticated clustering techniques because slight variations in the data could lead to high variance.