Data Mining and Warehousing [CS-2376]
Prof. Anirban Mondal
Spring 2021, Ashoka University

K-Means Clustering Assignment

Rishabh Goswami, Mohammed Zahir Ali, Shreyash Naman

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid. The k-means algorithm takes a dataset of ‘n’ points as input, together with an integer parameter ‘k’ specifying how many clusters to create. The output is a set of ‘k’ cluster centroids and a labeling of the dataset that maps each of the data points to a unique cluster based on least distance.

The first step is to randomly initialize a few points. These points are called cluster centroids. For our implementation of K-means clustering, we select the first ‘k’ points in the dataset as the initial cluster centroids. You can choose any number of cluster centroids. But the number of cluster centroids has to be less than the total number of data points.

The second step is the cluster assignment step. In this step, we need to loop through each of the ‘g’ dots. We calculate the distance between every point and all the centroids. The point is assigned to that cluster, to whose centroid it is the closest.

The next step is to move the cluster centroids. For this, we take an average of the coordinates of all the points in a cluster, and the new centroid for this cluster is placed at the average coordinate. This is done for all ‘k’ clusters and cluster centroids.

Upon getting new cluster centroids, we repeat step two - reassigning the clusters based on the new centroids.

The same steps are repeated until we reach a stopping criteria.There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations are reached

In the code submitted, we implemented the K-means algorithm in Python. We used the libraries Numpy, Pandas (for importing and working with datasets), and Matplotlib (for plotting the clusters). We used the maximum number of iterations as a stopping criterion.
We have included a sample dataset along with the submission - ‘crime.csv’ wherein we used the ‘latitude’ and ‘longitude’ data points from the dataset to test our algorithm.