K-Means Clustering
EPOCH IIT HYDERABAD

Rishitha Pallala

18 August 2023

1 Introduction

It is an unsupervised machine learning algorithm , it's also known as Flat-clustering algorithm and 'k' is the number of clusters in the data set. Although the clustering algorithm works for a predetermined k-value, the k-value can be determined by elbow plot.

2 Step-Wise Analysis

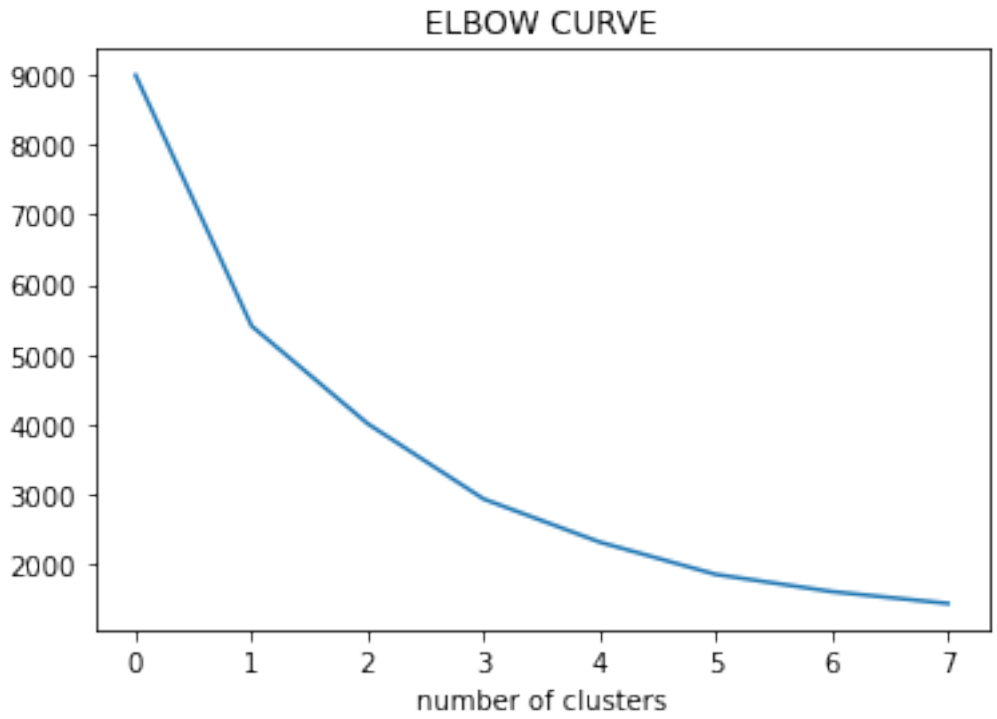
For a given k-value , the following steps are followed for clustering the data into k-groups.

1. Randomly choose k data points and assign the data points to these random k points based on the least euclidean distance.
2. calculate the centroid of the clusters and reassign the data points to each of the cluster based on the minimum euclidean distance.
3. Iterate the process until the centroid doesn't change. this is the most optimal clustering for the data.

3 Optimal k-value and Elbow Plot

It is necessary to find the optimal k-value and elbow plot is most commonly used to determine k. Firstly, WCSS(Within cluster sum of squares , given by equation (1) for 2-d data) is calculated for each value of k, and plotted.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2 \tag{1}$$



The blend is seen at the ideal k-value. Hence, k-means can be useful for clustering data with high number of variables, as calculating distance is faster and easy.