K means

It is a clustering algorithm that aims to group similar entities in one cluster.

Steps

1. Number of clusters (k) is chosen

2. It randomly initialize the cluster centers of each cluster i.e. the algorithm starts with a first group of randomly selected centroids i.e. randomly selecting few points considering it to be the centroids of the clusters.

3. For each data point, compute the Euclidian distance from all the centroids and assign the cluster based on the minimal distance to all the centroids. A centroid is the imaginary or real location representing the center of the cluster

4. This step is move centroid step. K means moves the centroid of each cluster by taking the average of all the data points in the cluster. In other words, the algorithm calculates the new centroid (average of all the points in a cluster) of each cluster

It halts creating and optimizing clusters when either:

- The centroids have stabilized i.e. there is no change in their values

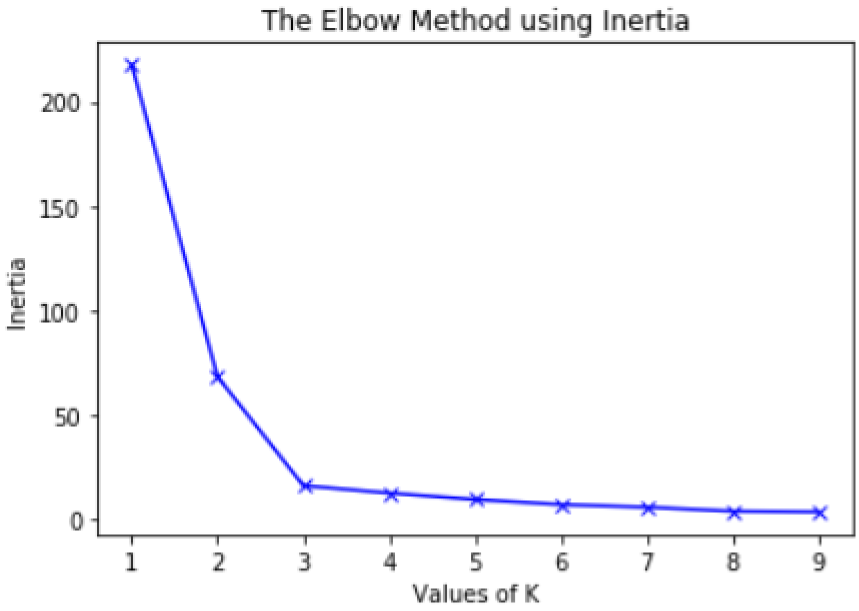
- The defined number of iterations has been achieved.

ELBOW METHOD

It is a method to know the best value of number of clusters.

We'll plot values for K on the horizontal axis and the Inertia on the Y axis where Inertia is sum of distances of samples to their closest cluster center

For each k value, we will identify the sum of squared distances of samples to the nearest cluster centre. To determine the optimal number of clusters, we have to select the value of k at the elbow i.e. the point after which the distortion/inertia start decreasing in a linear fashion


The Elbow Method using Inertia
Inertia
Values of K