K-Means

Epoch IIT Hyderabad

Himani Agrawal
MA22BTECH11008

Contents

Introduction

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means.In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible.

Steps involved in Algorithm

Initialization: Choose the number of clusters k you want to classify your data into.

Centroid calculation: Initialize k points randomly, these points are called means or cluster centroids.

Assignment step: Assign each data point to its closest centroid. This “closeness” is measured using the Euclidean distance formula.

$$d(x,y)=((x1-y1)^2+(x2-y2)^2+...+(xn-yn)^2)^1/2$$

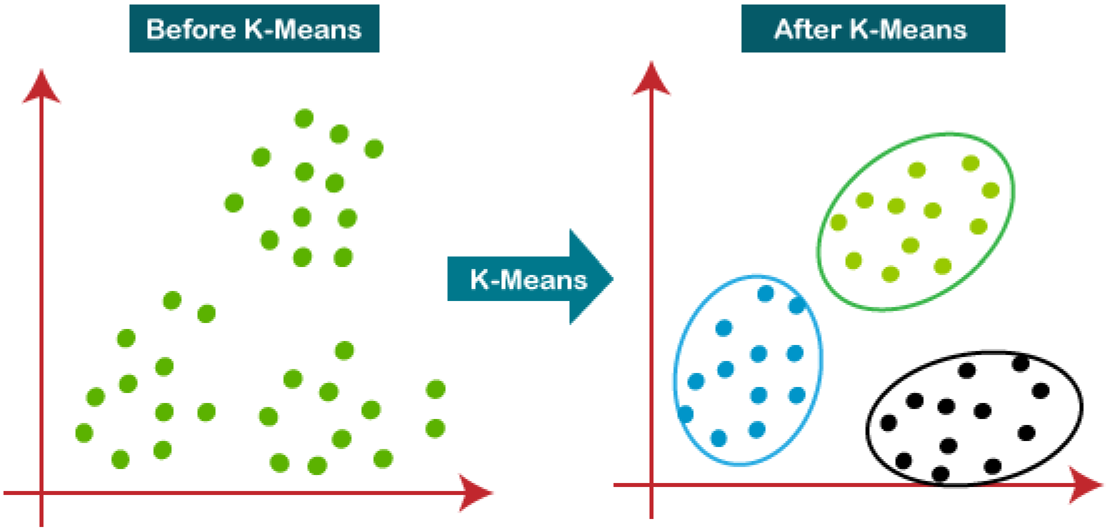
where `d(x,y)`is the Euclidean distance, and `x` and `y` are the data points in a n-dimensional space.

Update step: Calculate the new centroid (mean point) of each cluster. This is done by taking the mean of all data points assigned to that cluster's centroid:

$$\mu=1/n\sum xi$$

where μ is the new centroid, n is the total number of data points assigned to the centroid, and `x_i` is each individual data point.

Convergence: Repeat the assignment and update steps until the centroids do not change significantly, or a maximum number of iterations is reached


Before K-Means
K-Means
After K-Means