K Means

Epoch IIT Hyderabad

Arin Aggarwal
MA22BTECH11006

Contents

1 Introduction

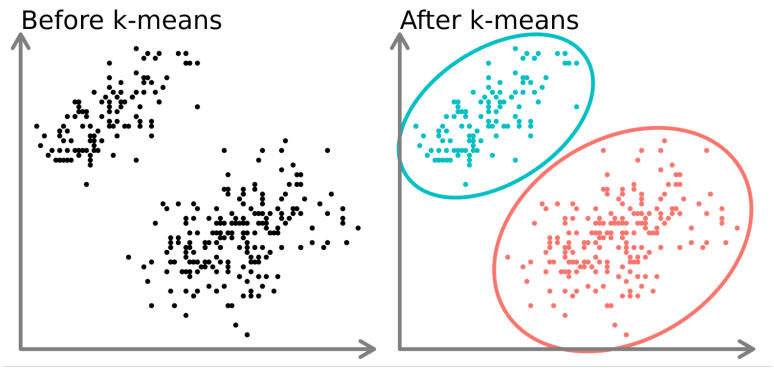
K means clustering is an unsupervised classification algorithm that classifies the given data points into k groups(clusters).

2 Working

1) Select the 'k' to decide the number of clusters.
2) Select random k points/centroids.
3) Assign each data point to their closest centroid (using distances), which will form the predefined k clusters.
4) Calculate the variance and place a new centroid of each cluster based on average of the cluster points.
5) Keep on repeating the procedure.

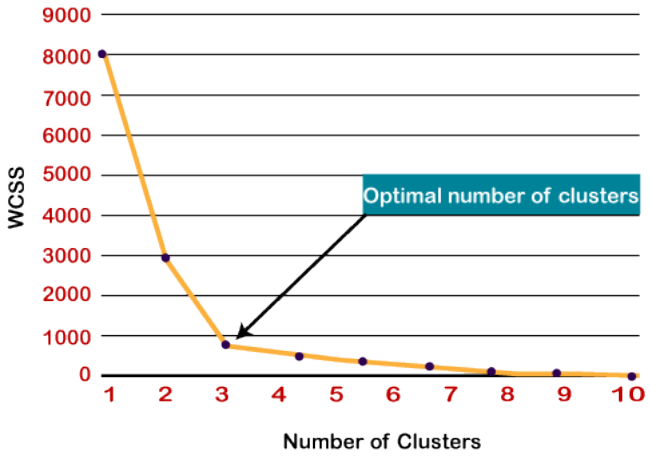
This process may stop in 2 ways:

1) When the clusters formed stabilizes.
2) When the number of maximum iterations specified is reached.



3 Choosing Optimal k Value

If k is not pre-defined, the best value for it can be determined using the **Elbow Method**. It uses the concept of Within Cluster Sum of Squares (WCSS) which is nothing but sum of square of distances of all points from their respective centroids. Upon plotting the WCSS v/s k graph, the value of k where we see a sharp point/bend is chosen as the best value for k.



4 Evaluation Metrics

1) **Inertia** : Inertia calculates the sum of distances of all the points within a cluster from the centroid of that cluster. The distance between them should be as low as possible. A low inertia value is considered good. Hence Inertia focuses on minimizing intra-cluster distance.
2) **Dunn Index** : Different clusters should be as different from each other as possible. For this, Along with the distance between the centroid and points, the Dunn index also takes into account the distance between two clusters (inter-cluster distance). It is the ratio of the minimum of inter-cluster distances and maximum of intra-cluster distances and hence we need to maximize it.

$$Dunn\ Index\ =\ \frac{max(Inter-cluster\ distance)}{max(Intra-cluster\ distance)}$$

3) **Silhouette Score** : The silhouette score and plot are used to evaluate the quality of a clustering. It measures the similarity of each point to its own cluster compared to other clusters, and the plot visualizes these scores for each sample. A high silhouette score indicates that the clusters are well separated, and each sample is more similar to the samples in its own cluster than to samples in other clusters. A silhouette score close to 0 suggests overlapping clusters, and a negative score suggests poor clustering solutions and score of 1 denotes best clustering.