K means
Epoch IIT Hyderabad

Chakka Surya Saketh
AI22BTECH11005

Introduction

Unsupervised Machine Learning learning is the process of teaching a computer to use unlabelled, unclassified data and enabling the algorithm to operate on that data without supervision. Without any previous data training, the machine's job in this case is to organize unsorted data according to parallels, patterns, and variations.

K-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabelled data into a predetermined number of clusters based on similarities (K).

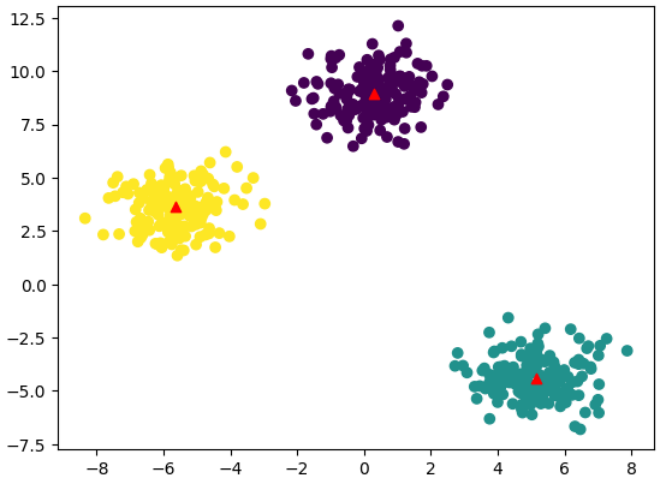
The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

Implementation

It is suggested to normalize the data while dealing with clustering algorithms such as K-Means since such algorithms employ distance-based measurement to identify the similarity between data points.

1) We need to provide the number of clusters, K, that need to be generated by this algorithm.
2) Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.
3) The cluster centroids will now be computed.
4) Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.

a) The sum of squared distances between data points and centroids would be calculated first.
b) At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).
c) Finally, compute the centroids for the clusters by averaging all of the cluster's data points.



Choosing the optimal K

1) Elbow Method: In this method, we plot the WCSS (Within-Cluster Sum of Square)against different values of the K, and we select the value of K at the elbow point in the graph, i.e., after which the value of WSCC remains constant (parallel to the x-axis).
2) Silhouette method: In this method, we calculate the silhouette coefficient of each data point. The silhouette coefficient measures how well the data point fits in the assigned cluster as compared to the other cluster. The average Silhouette coefficient for different K is calculated to find the optimal value of K with the highest coefficient value.
3) Gap statistic method: In this method, we compare the WCSS for different values of K with the expected sum of squares values randomly generated from a uniform distribution. The optimal value of K is the one with the largest gap between the observed and expected sum of squares.

Conclusion

K-means clustering is a widely used approach for clustering. Generally, practitioners begin by learning about the architecture of the dataset. K-means clusters data points into unique, non-overlapping groupings. It works very well when the clusters have a spherical form. However, it suffers from the fact that clusters geometric forms depart from spherical shapes.

Additionally, it does not learn the number of clusters from the data and needs that it be stated beforehand.