Unsupervised Learning

Aashish Mahato

Agenda

+ Unsupervised Learning
+ K-Means Clustering
+ DBSCAN

What is Unsupervised Learning?

+ Machine learning algorithms can be broadly classified into two categories - supervised and unsupervised learning. There are other categories also like semi-supervised learning and reinforcement learning. But, most of the algorithms are classified as supervised or unsupervised learning.

+ The difference between them happens because of presence of target variable. In unsupervised learning, there is no target variable. The dataset only has input variables which describe the data. This is called unsupervised learning.

Unsupervised Learning Methods

Clustering

+ Clustering is a technique for exploring raw, unlabeled data and breaking it down into groups (or clusters) based on similarities or differences.
+ It is used in a variety of applications, including customer segmentation, fraud detection, and image analysis.
+ Clustering algorithms split data into natural groups by finding similar structures or patterns in uncategorized data.

Unsupervised Learning Methods

Association Rules

+ An association rule is a rule-based method for finding relationships between variables in a given dataset.
+ These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products.
+ Understanding consumption habits of customers enables businesses to develop better cross-selling strategies and recommendation engines. Examples of this can be seen in Amazon's "Customers Who Bought This Item Also Bought" or Spotify's "Discover Weekly" playlist.

Unsupervised Learning Methods

Dimensionality Reduction

+ Dimensionality reduction is an unsupervised learning technique that reduces the number of features, or dimensions, in a dataset.
+ Dimensionality reduction extracts important features from the dataset, reducing the number of irrelevant or random features present.
+ This method uses principle component analysis (PCA) and singular value decomposition (SVD) algorithms to reduce the number of data inputs without compromising the integrity of the properties in the original data.

Applications of Unsupervised Learning

+ Customer Segmentation: Algorithms cluster customers based on purchasing behavior or demographics, enabling targeted marketing strategies.

+ Anomaly Detection: Identifies unusual patterns in data, aiding fraud detection, cybersecurity and equipment failure prevention.

+ Recommendation Systems: Suggests products, movies or music by analyzing user behavior and preferences.

+ Image and Text Clustering: Groups similar images or documents for tasks like organization, classification or content recommendation.

K-Means Clustering

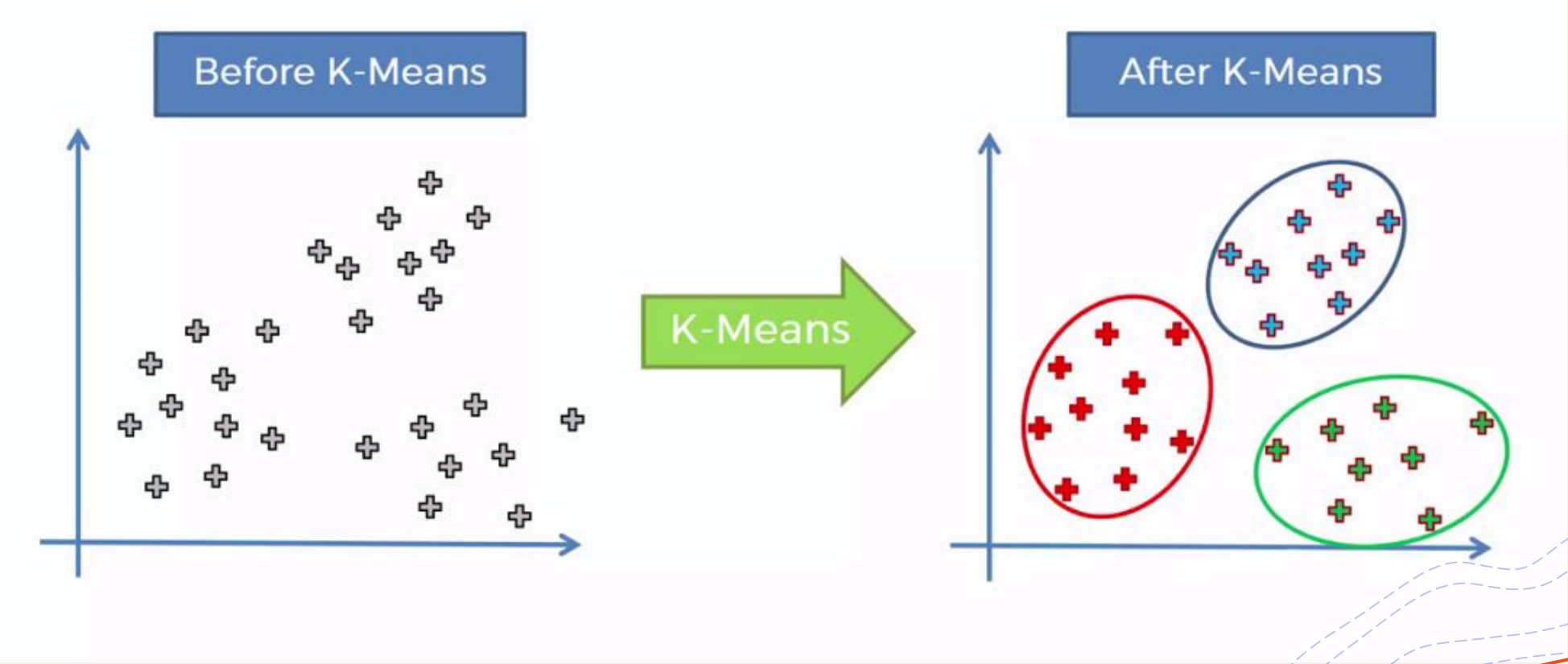
What is K-Means Clustering?

+ K-Means clustering is the most popular unsupervised learning algorithm. It is used when we have unlabelled data which is data without defined categories or groups.

+ The algorithm follows an easy or simple way to classify a given data set through a certain number of clusters, fixed apriori.

+ K-Means algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

K-Means Clustering

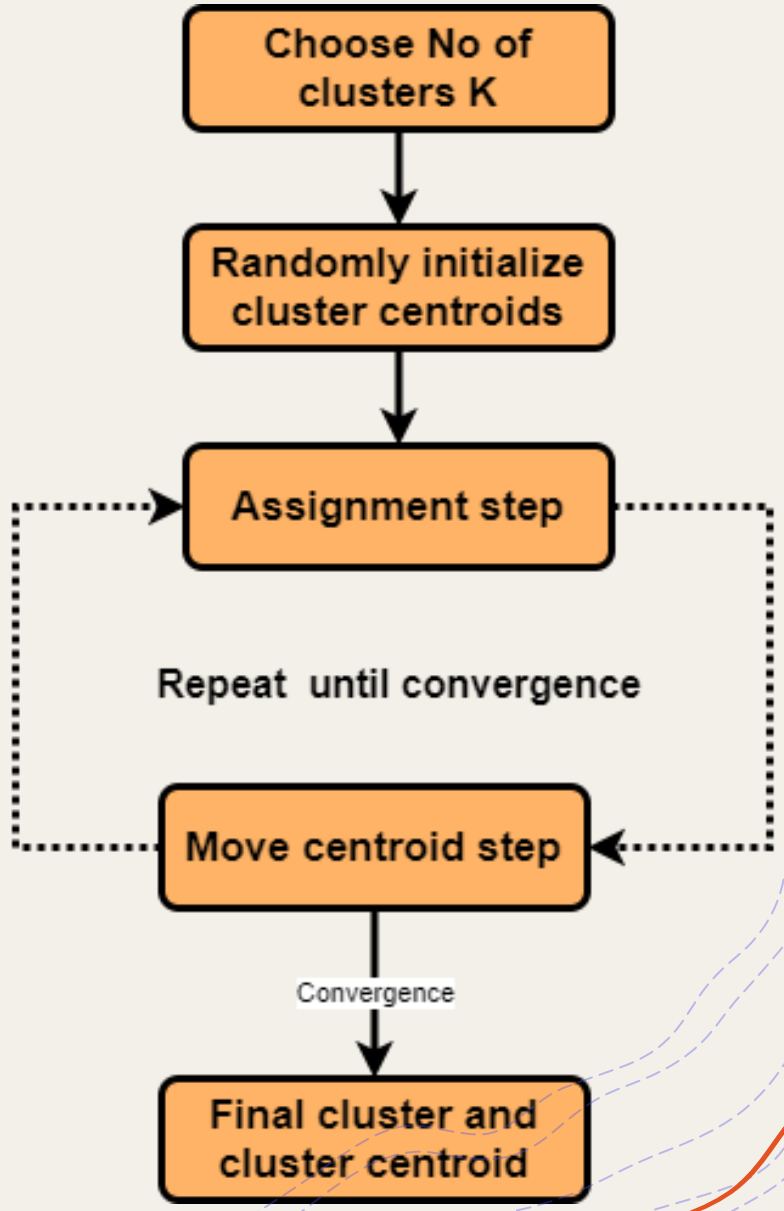


How K-Means Clustering Works?

1) Initialization: Start by randomly selecting K points from the dataset. These points will act as the initial cluster centroids.

2) Assignment: For each data point in the dataset, calculate the distance between that point and each of the K centroids. Assign the data point to the cluster whose centroid is closest to it. This step effectively forms K clusters.

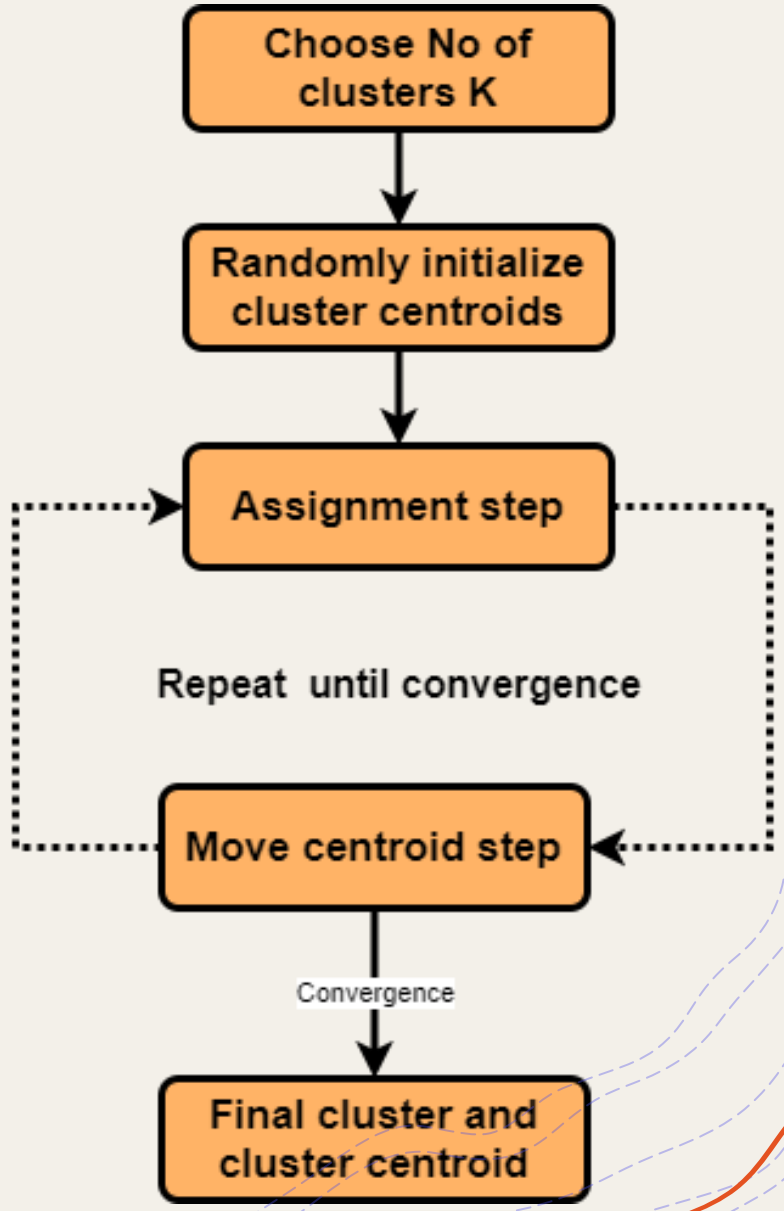
3) Update centroids: Once all data points have been assigned to clusters, recalculate the centroids of the clusters by taking the mean of all data points assigned to each cluster.



How K-Means Clustering Works?

4) Repeat: Repeat steps 2 and 3 until convergence. Convergence occurs when the centroids no longer change significantly or when a specified number of iterations is reached.

5) Final Result: Once convergence is achieved, the algorithm outputs the final cluster centroids and the assignment of each data point to a cluster.



Advantages of K-Means Clustering

+ One of the simplest algorithm to understand
+ Since it uses simple computations it is relatively efficient
+ Gives better results when there is less data overlapping

Disadvantages of K-Means Clustering

+ Number of clusters need to be defined by user
+ Doesn't work well in case of overlapping data
+ Unable to handle the noisy data and outliers
+ Algorithm fails for non-linear data set

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

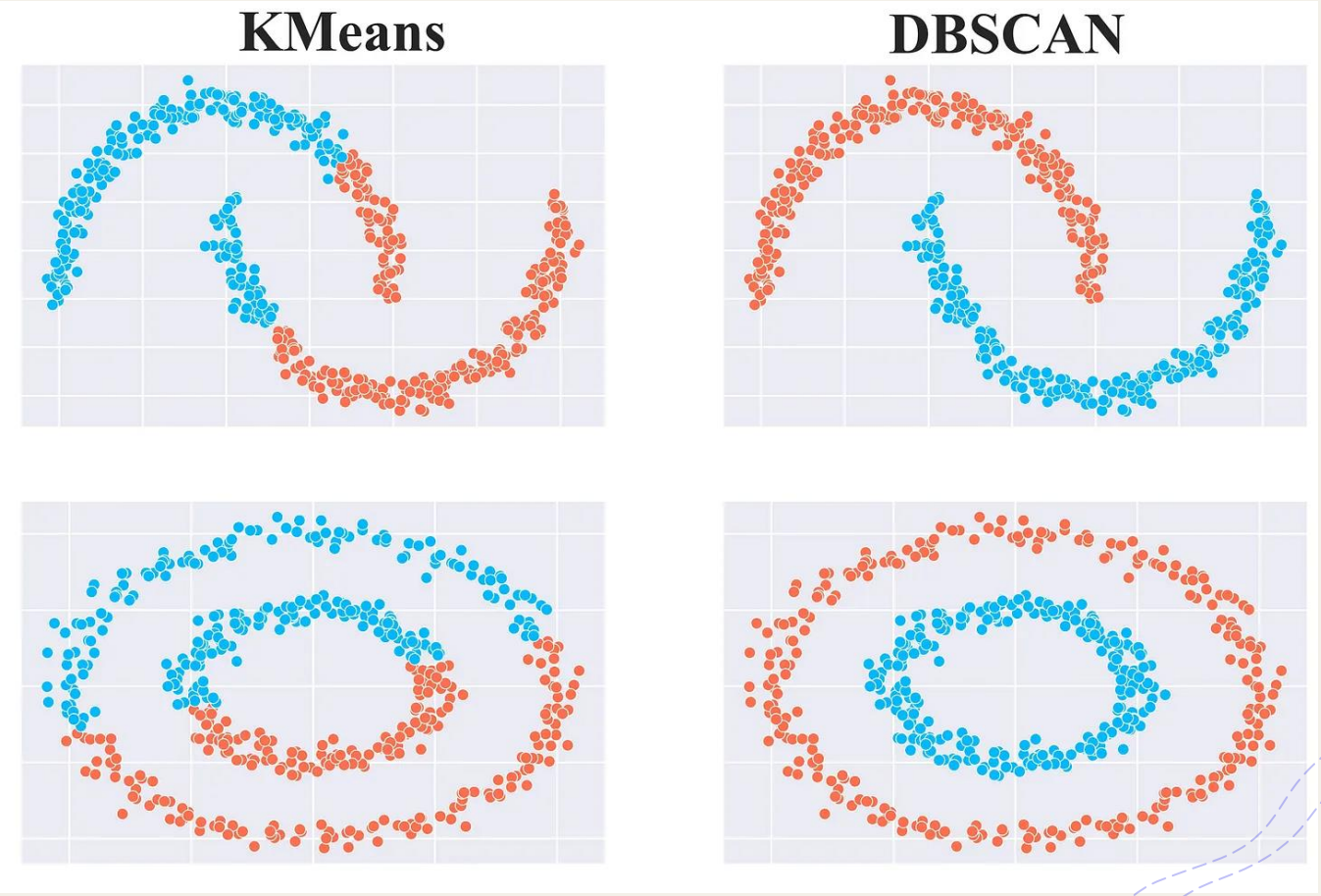
What is DBSCAN Clustering?

+ The DBSCAN clustering algorithm is a density-based clustering method that is commonly used in machine learning and data mining applications.

+ Instead of assuming that clusters are spherical like K-Means, DBSCAN can identify clusters of arbitrary shapes.

+ The algorithm works by grouping together points that are close to each other based on a distance metric (**ε**) and a minimum number of points (MinPts) required to form a cluster.

DBSCAN Clustering

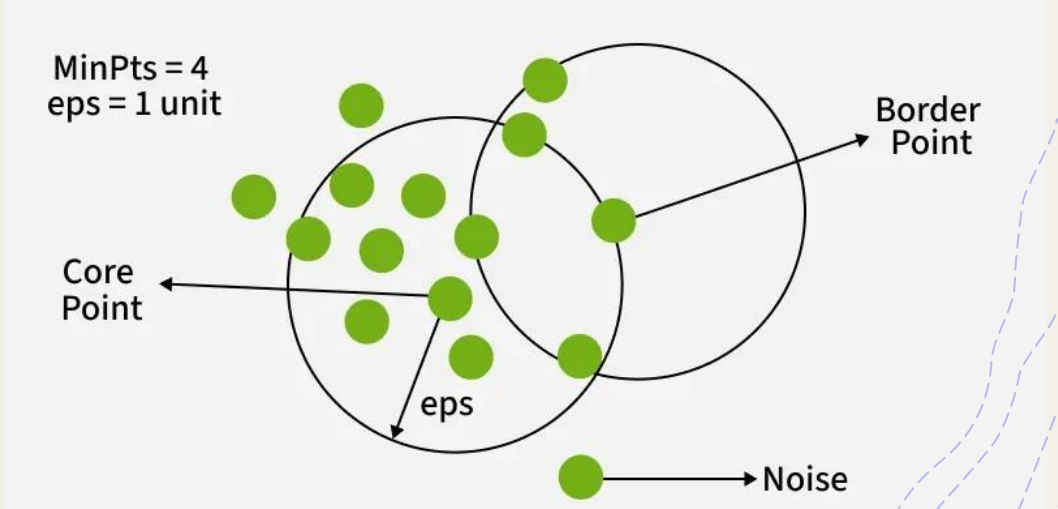


How Does DBSCAN Work?

DBSCAN works by categorizing data points into three types:

+ Core points which have a sufficient number of neighbors within a specified radius (epsilon)
+ Border points which are near core points but lack enough neighbors to be core points themselves
+ Noise points which do not belong to any cluster.

By iteratively expanding clusters from core points and connecting density-reachable points, DBSCAN forms clusters without relying on rigid assumptions about their shape or size.



DBSCAN vs. K-Means

+ DBSCAN is a density-based clustering algorithm, whereas K-Means is a centroid-based clustering algorithm.

+ DBSCAN can discover clusters of arbitrary shapes, whereas K-Means assumes that the clusters are spherical.

+ DBSCAN does not require the number of clusters to be specified in advance, whereas K-Means requires the number of clusters to be specified.

+ DBSCAN is less sensitive to initialization than K-Means.

Thank You