Unsupervised Learning

Clustering: K-means/Kernel K-means

Clustering is the assignment of objects into groups (called *clusters*) so that objects from the same cluster are more similar to each other than objects from different clusters. Often similarity is assessed according to a distance measure. Clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, search engines, image analysis and bioinformatics.

Clustering is also called **data segmentation** because it partitions large datasets into groups according to their similarity.

It can also be used for **outlier detection**. Outliers are objects that do not fall into any cluster because of too much dissimilarity with other objects. Example, credit card fraud detection.

Clustering is an unsupervised learning because the class label information is not present. Hence in unsupervised, learning is done by observation.

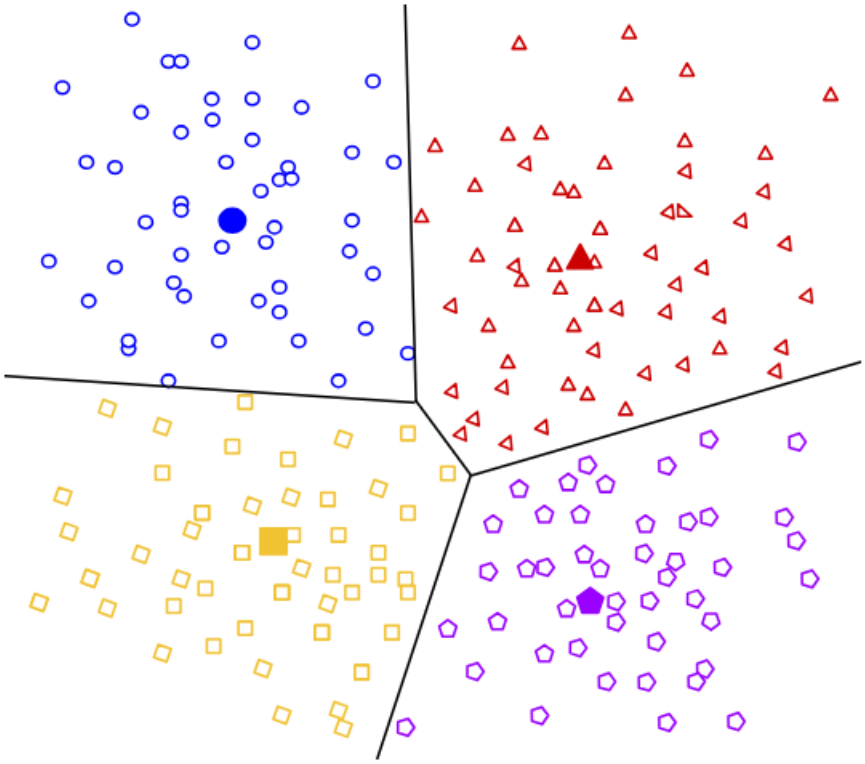
Types of Clustering Algorithm

1. Hard clustering – Each data point either belongs to a cluster completely or not.
2. Soft clustering – Instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

Clustering algorithms can also be classified as follows –

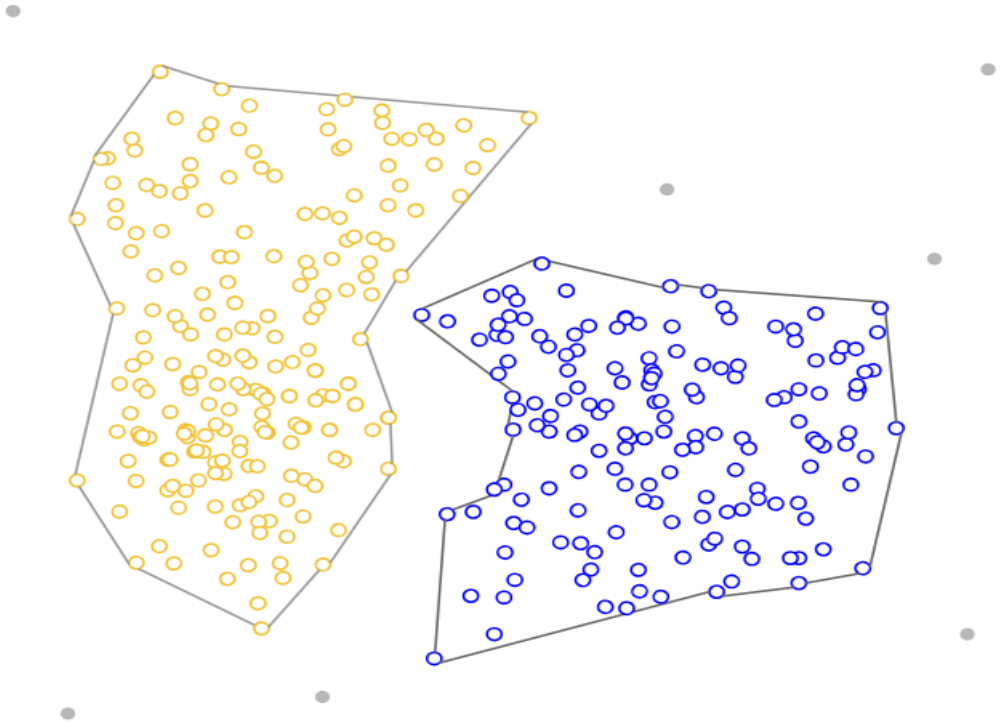
1. **Partitioning method** – Partitioning means division. Suppose we are given a database of 'n' objects and we need to partition this data into 'k' partitions of data. Within a partition, there exists some similarity among the items. So, each partition will represent a cluster and k<=n. That means it will classify the data into k groups, each group contains atleast one object and each object must belong to exactly one group. Most partitioning methods are distance-based. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid. Hence, it is also known as **centroid-based method**.

Examples
K-Means Algorithm
K-Medoid Algorithm



2. **Hierarchical method** – Hierarchical clustering is an alternative approach to partitioning method for identifying groups in a dataset. The result of a hierarchical clustering is a tree-based representation of the objects, which is also called **dendrogram**. We classify hierarchical methods on the basis of how the hierarchical decomposition is formed. 2 approaches are there –
   - **A) Agglomerative approach** – It is a bottom-up approach. Here we start with an object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing until all the groups are merged into one or until the termination condition holds.
   - **B) Divisive approach** – It is top-down approach. Here we start with all the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is done until each object is in one cluster or termination condition holds.
     Example - CURE (Clustering Using Representatives)

3. **Density-Based method -** Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.
   Example - DBSCAN algorithm



4. **Grid-Based method –** The grid-based approach differs from the conventional clustering algorithms in that it is concerned not with data points but with the value space that surrounds the data points. The data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects.
   Example - *STING* (Statistical Information Grid)

Applications of Clustering

1. **Business Intelligence** – Cluster analysis helps in target marketing, where marketers discover groups and categorize them based on the purchasing patterns. The information retrieved can be used in market segmentation, product positioning, new product development, grouping of shopping items.
2. **Pattern recognition** – Here clustering methods group similar patterns into clusters whose members are more similar to each other.
3. **Image Processing** - Clustering can be used to group similar images together, classify images based on content, and identify patterns in image data.
4. **Bioinformatics** - It can be used for classification among different species of plants and animals.
5. **Traffic analysis**: Clustering is used to group similar patterns of traffic data, such as peak hours, routes, and speeds, which can help in improving transportation planning and infrastructure.
6. **Fraud detection**: Clustering is used to identify suspicious patterns or anomalies in financial transactions, which can help in detecting fraud or other financial crimes.
7. **Search Engines** – Whenever a query is fired in Google search engine, the search engine provides the result of the nearest similar object which are clustered around the data to be searched. The speed and accuracy of the retrieved resultant is dependent on the use of the clustering algorithm.
8. **Text Mining** – It involves the process of extracting high quality information from text. This can be used for sentiment analysis and document summarization.

K-MEANS CLUSTERING ALGORITHM

Let $X=\{x_1,x_2,x_3,....x_n\}$ be the set of data points and $V=\{v_1,v_2,v_3,....v_c\}$ be the set of centers. The steps of the algorithm are –

1. Randomly select c cluster centers.
2. Calculate the distance (Euclidean distance or Manhattan distance as the metric) between each data point and cluster centers.
3. Assign the data point to the cluster having minimum distance from it and the cluster center.

4. Recalculate the new cluster center using following equation.

   $V_i = (1/C_i) \sum_{j=1}^{ci} x_i$

   where C_i represents the number of data points in the i^{th} cluster.

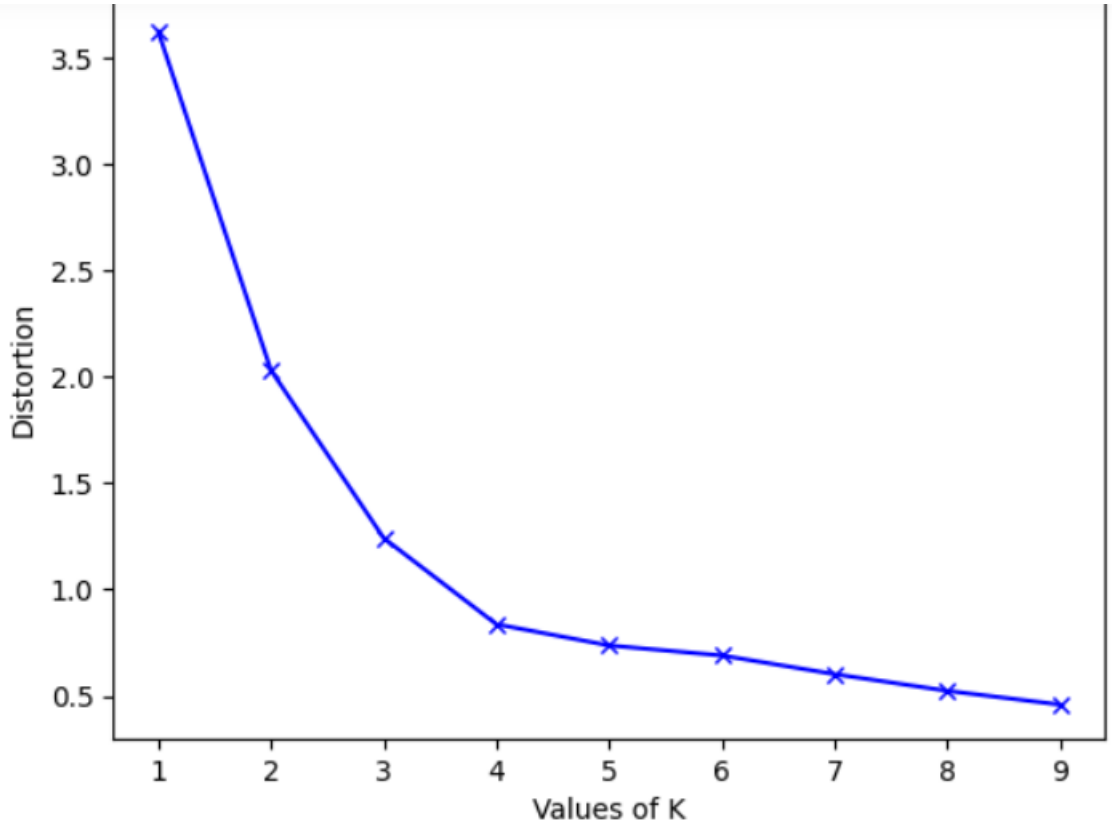
5. Recalculate the distance between each data point and the newly obtained cluster centers.
6. If no data was reassigned then Stop, otherwise repeat steps 3 to 5.

We can implement the K-Means clustering machine learning algorithm using the elbow method in the scikit-learn library in Python. Elbow method is commonly used method for finding the optimum K value.

K Means Clustering Using the Elbow Method

In the Elbow method, we are actually varying the number of clusters (K) from 1 to 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyse the graph, we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph moves almost parallel to the X-axis. The K value corresponding to this point is the optimal value of K or an optimal number of clusters.

A sample graph to plot elbow method


Distortion
3.5
3.0
2.5
2.0
1.5
1.0
0.5
1
2
3
4
5
6
7
8
9
Values of K