Chapter 4

classmate
Date 10/04/2023
Page

Clustering: Algorithm:

Clustering is a type of unsupervised learning in ML where similar data points are grouped together into clusters based on their similarity or distance.

The aim of clustering is to find the patterns & groupings in data that are not labeled or pre-defined.

These groups or clusters are formed such that data points within the same cluster are more similar to each other than to those in other clusters.

Applications of Clustering Algorithms.

1
2
3
4
5
6

1 Text data mining:

It is a process of extracting meaningful & relevant information from textual data. Clustering helps in the process of data cleaning, preprocessing, transformation & analysis of textual data.

2 Customer Segmentation:

Clustering algorithms are used to group the customers based on their preferences, purchase history, enabling business to target specific customers.

3 Image Segmentation:

Clustering algorithms are used to partition images into multiple segments based on similarities in color, texture & other visual features.

4 Anomaly Detection:

It can be used to detect the anomalies or outliers in the data such as fraudulent transaction.

5. Document Clustering:

Clustering algorithm is used to group similar documents based on their content, enabling efficient organization & retrieval of large document collections.

6 Data mining:

The clustering algorithm simplify the data mining task by grouping a large number of features from an extremely large data set to make the analysis managable.

Types of Clustering Techniques:

1 Partitioning methods : K-Means & K-Medoid. (Centroid-based.)
2 Hierarchical methods : Hierarchical Clustering
3 Density-based methods: DBSCAN

5. Explain the Concept of K-Means Algorithm:

K-Means – A centroid-based technique:

K-Means Algorithm is a Partition-based clustering algorithm, used for clustering similar data-points together. The algorithm partitions a given data set into K-clusters, where each cluster contains data points that are more similar to each other than to the data points in other clusters.

The principle of K-Means algorithm is to assign each of the 'n' data points to one of the K clusters where, K is user defined parameter as the number of clusters desired.

The objective is to maximize the Homogeneity within the clusters & also to maximize the difference b/w the clusters. The Homogeneity & differences are measured in terms of the distance b/w the points in the data set.

Algorithm:

1. Initialize:
   Randomly select K data points as initial Centroids for the K clusters.
2. Assign each data point to the nearest centroid:
   Calculate the Euclidean distance b/w each data point & the K centroids. Assign each data point to the cluster represented by the nearest centroid.
3. Recalculate the Centroids:
   After all the data points have been assigned to the nearest cluster, recalculate the centroids of each cluster based on the mean of all the data point in that cluster.
4. Repeat Step 2 & 3:
   Repeat 2 & 3 until the centroid of each cluster no longer change or until the maximum num of iterations is reached.

8. Explain the Concept of DBSCAN Algorithm:

DBSCAN:

Density-Based Spatial Clustering of Applications with Noise i.e DBSCA is a Density-based clustering algorithm that is used to group together the data points that are close to each other in a high-density region & separate them from points in low density regions.

The density-based clustering approach provides a solution to identify clusters of arbitrary shapes. The principle is based on identifying dense area & sparse area within the data set & then run the clustering algorithm.

DBSCA is one of the popular density-based algorithm which creates clusters by using the connected regions with high density. It can discover clusters of diff shapes & sizes from a large amount of data which containing noise or outliers.

DBSCAN Algorithm operates in 2 steps:

1. Density-based clustering:

The algorithm starts by selecting a random data point & identifying all the points within its radious. The algorithm then repeates this process for each point in this newly formed group until no new points can be added.

2. Removing Noise:

Once all the groups have been formed the algorithm identifies any remaining noise points & removes them from the groups.

The output of DBSCAN algorithm is a set of clusters & noise points. Unlike the K-Means algorithm DBSCAN does not require the num of clusters to be specified beforehand.

3. Explain the Concept of Market Basket Analysis with Examples

Association Rules & Market Basket Analysis:

Association Rules in ML are type of Un-supervised learning techniques used to identify the relationship blw variables in a dataset. It is also known as Association Analysis.

A common Application of this Analysis is Market Basket Analysis, where the objective is to identify which products are frequently purchased together, that the retailers use it for the cross-selling of their products.

For example:

Every large grocery store accumulates a large volume of data about the buying patterns of the customers. On the basis of items purchased together, the retailer can push some cross-selling either by placing the items bought together in adjecent areas or creating some combo offer with those diff product types.

The below association rule signifies that people who have bought bread & milk have often bought egg also; so, for retailer, it makes sense that these items are placed together for cross-selling.

The problem of deriving associations from data is of utmost importance in unsupervised learning. This problem is often to refered to as Market Basket Problem. In this problem we are given a set of items & a large collection of transactions that are subsets of these items. The task here is to find the relationship blw the presence of various items within the basket.

Ex: Consider the following 5 Transactions in a Super Market

t_1 = { Pears Body wash, Lux Soap, Hair oil }

t_2 = { Pears Body wash, Bay Lotion }

t_3 = { Bay Lotion, Lux Soap, Hair oil }

t_4 = { Pears Body Wash, Hair oil, Body Lotion }

t_5 = { Baby Lotion, Lux Soap, Body Lotion }

Here A = { Peas Body wash, Lux Soap, Baby Lotion, Hair oil, Body Lotion }

Transactional Database T is,

T = { t_1, t_2, t_3, t_4, t_5 }

Now,

Here t_2 supports { Peas Body wash, Baby Lotion } & 'Lux Soap' is supported by 3 Transactions out of 5.

Hence support of Lux Soap is : 60%

Support :

* A Transaction t is said to support an item i_1 if i_1 is present in T
* Support can be represented as a percentage or a number.

Here, Let A = { i_1, i_2, ... i_n } be a set of items.

Let T the transaction database be a set of transactions where each transaction t is set of Items in the Basket. Hence we can say t is a subset of A.

4. Explain the Apriori Algorithm with its Application?

Apriori Algorithm:

The Apriori Algorithm is an association rule mining algorithm that identifies the frequent itemsets in a given data set & generates the association rules from those frequent item sets. The algorithm works on the principle of the "Apriori Property" which states that any subset of a frequent itemset must also be frequent.

The Apriori Algorithm is used to overcome the problem of finding frequent itemsets from a large data set efficiently. It helps to identify a strong associations btw items in a data set by finding frequent item sets & generating association rules.

Algorithm has 3 main steps:

1 Support Counting:

In this step the algorithm scans the data set to count the number of occurances of each item. It then calculates the support of each item,

2 Generating Frequent itemsets:

It generates Frequent item sets by combining the items with a high support value.

3 Generating Association Rules:

Finally the algorithm generates association rules from the frequent item sets.

Applications of Apriori:

1 Market Basket Analysis
2 Fraud Detection
3 Healthcare
4 Recommender Systems
5 Social Network Analysis.

1 Market Basket Analysis:

Apriori Algorithm helps retailers to understand the relationships blw products & to identify which products are more likely to be purchased together by the customer. This info can be used to optimise product pricing & placement.

2 Fraud Detection:

The Apriority Algorithm can be used to detect fraudulent transactions by identifying patterns of behaviour that are unusual or suspicious.

3. Healthcare:

It is used to analyze the patient data, including symptoms, medical history & test results to identify patterns that could indicate the presence of diseases.

4 Recommender System:

Apriori algorithm can be used in filtering, which is used to build, Recommender system for online, shopping, streaming services & other apps.

5. Social Network Analysis

This Algorithm can be used to identify the patterns of behaviour & relationships blw users on social media platforms, which can be used for targeted marketing or to detect fraudulent activity.

6. Explain Reinforcement Learning with Example.

Reinforcement Learning:

Reinforcement Learning is a type of Machine Learning where an agent learns to make decisions in an environment by performing certain actions & receiving feedback in the form of rewards or punishments.

More precisely, it is a method where an Intelligent agent interacts with the environment & learns to act within that.

The basic component of reinforcement learning include an agent, an environment & a set of actions, states & rewards. The agent is the decision maker, the enviroment is the world in which the agent operates, and the actions, states & rewards are the building blocks of the learning process.

For example:

The best example for reinforcement learning is the game of chess. The agent is the chess player, the environment is the chessboard, & the actions are the moves that the player can make. The state of the game is determined by the current configuration of the chess pieces on the board. The reward is the score that the player receives based on the outcome of the game.

Initially, the agent has no knowledge of how to play chess & make random moves. The agent receives feedback in the form of reward or punishments, depending on whether the move made was good or bad. If the move is good, the agent is rewarded with a positive score, & if the move is bad, the agent is punished with a -ve score. The agent then updates its strategy based on the rewards & punishments received.

Over time, the agent learns from its mistakes & improves it's strategy by making better moves.