Name: Siddharth Unny **Batch:** A4 **Sap Id:** 60004200080

Aim: Write and Explain one algorithm each on

1. Spatial Association Rules
2. Spatial Classification
3. Spatial Clustering – DBScan

Theory:

Spatial data means data related to space which can be the two-dimensional abstraction of the surface of the earth or a man-made space like the layout of a VLSI design, a volume containing a model of the human brain, or another 3d-space representing the arrangement of chains of protein molecules. The data consists of geometric information and can be either discrete or continuous. The explicit location and extension of spatial objects define implicit relations of spatial neighborhood (such as topological, distance and direction relations) which are used by spatial data mining algorithms. Therefore, spatial data mining algorithms are required for spatial characterization and spatial trend analysis.

Spatial data mining or knowledge discovery in spatial databases differs from regular data mining in analogous with the differences between non-spatial data and spatial data. The attributes of a spatial object stored in a database may be affected by the attributes of the spatial neighbors of that object. In addition, spatial location, and implicit information about the location of an object, may be exactly the information that can be extracted through spatial data mining.

Spatial Association Rules

Spatial association means connectedness or relationship between and among variables over space. A single variable may be spatially autocorrelated; that is, values of the variable are somehow connected or related spatially. Many variables may be associated one with another at one or more sites. If there is spatial interaction there is also spatial association. Maps can depict spatial association. A mathematical shorthand technique can be used to represent, in general, measures of spatial association. Scientists test or theorize about variables to determine whether spatial association, either observed or expected, can be confirmed. Statistical procedures that have been developed for identifying and measuring the existence of spatial association are outlined.

Name: Siddharth Unny **Batch:** A4 **Sap Id:** 60004200080

Algorithm : Apriori Algorithm

The Apriori algorithm uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rules, it determines how strongly or how weakly two objects are connected. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset associations efficiently. It is the iterative process for finding the frequent itemsets from the large dataset. This algorithm is mainly used for market basket analysis and helps to find those products that can be bought together. It can also be used in the healthcare field to Frequent itemsets which are those items whose support is greater than the threshold value or user-specified minimum support. It means if A & B are the frequent itemsets together, then individually A and B should also be the frequent itemset.

Steps for Apriori Algorithm:

Step-1: Determine the support of itemsets in the transactional database, and select the minimum support and confidence.

Step-2: Take all supports in the transaction with higher support value than the minimum or selected support value.

Step-3: Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.

Step-4: Sort the rules as the decreasing order of lift.

Advantages of Apriori Algorithm

- This is easy to understand algorithm
- The join and prune steps of the algorithm can be easily implemented on large datasets.

Disadvantages of Apriori Algorithm

- The Apriori algorithm works slow compared to other algorithms.
- The overall performance can be reduced as it scans the database for multiple times.
- The time complexity and space complexity of the Apriori algorithm is $O(2^D)$, which is very high. Here D represents the horizontal width present in the database.

2.Spatial Classification

Spatial classification assigns an object to a class from a given set of classes based on the attribute values of the object. It mainly considers the distance, direction, or connectivity relationships among spatial objects.

Name: Siddharth Unny **Batch:** A4 **Sap Id:** 60004200080

Algorithm: KNN Algorithm

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.
- The K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know whether it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

Working:

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.

Step-6: Our model is ready.

Selecting the value of K in the K-NN Algorithm

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples

Spatial Clustering - DBScan

Spatial Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. In spatial data sets, clustering permits a generalization of the spatial component like explicit location and extension of spatial objects which define implicit relations of spatial neighborhood. Current spatial clustering techniques can be broadly classified into three categories:

- partitional
- hierarchical
- locality-based

Name: Siddharth Unny **Batch:** A4 **Sap Id:** 60004200080

Algorithm: DBSCAN Algorithm

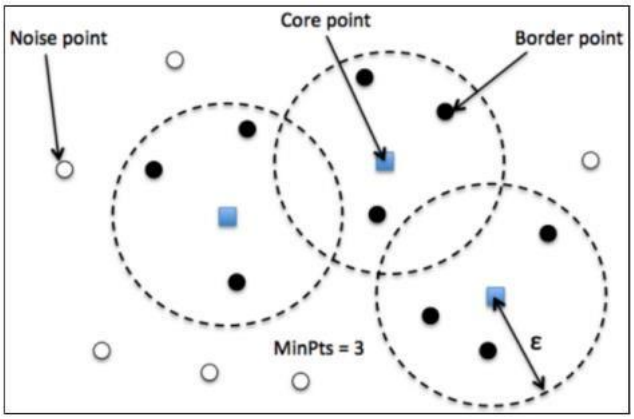
Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

The DBSCAN algorithm uses two parameters:

- minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.
- eps (ε): A distance measure that will be used to locate the points in the neighborhood of any point.

These parameters can be understood if we explore two concepts called Density Reachability and Density Connectivity. Reachability in terms of density establishes a point to be reachable from another if it lies within a particular distance (eps) from it. Connectivity, on the other hand, involves a transitivity based chaining-approach to determine whether points are located in a particular cluster. For example, p and q points could be connected if p->r->s->t->q, where a->b means b is in the neighborhood of a.

There are three types of points after the DBSCAN clustering is complete:



- Core — This is a point that has at least m points within distance n from itself.
- Border — This is a point that has at least one Core point at a distance n.

- Noise — This is a point that is neither a Core nor a Border. And it has less than m points within distance n from itself. Steps for DBSCAN clustering
- The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited).
- If there are at least 'minPoint' points within a radius of 'ε' to the point then we consider all these points to be part of the same cluster.
- The clusters are then expanded by recursively repeating the neighborhood calculation for each neighboring point

Parameter Estimation

Every data mining task has the problem of parameters. Every parameter influences the algorithm in specific ways. For DBSCAN, the parameters ε and minPts are needed.

- minPts: As a rule of thumb, a minimum minPts can be derived from the number of dimensions D in the data set, as minPts ≥ D + 1. The low value minPts = 1 does not make sense, as then every point on its own will already be a cluster. With minPts ≤ 2, the result will be the same as of hierarchical clustering with the single link metric, with the dendrogram cut at height ε. Therefore, minPts must be chosen at least 3. However, larger values are usually better for data sets with noise and will yield more significant clusters. As a rule of thumb, minPts = 2·dim can be used, but it may be necessary to choose larger values for very large data, for noisy data or for data that contains many duplicates.
- ε: The value for ε can then be chosen by using a k-distance graph, plotting the distance to the k = minPts-1 nearest neighbor ordered from the largest to the smallest value. Good values of ε are where this plot shows an "elbow": if ε is chosen much too small, a large part of the data will not be clustered; whereas for a too high value of ε, clusters will merge and the majority of objects will be in the same cluster. In general, small values of ε are preferable, and as a rule of thumb, only a small fraction of points should be within this distance of each other.
- Distance function: The choice of distance function is tightly linked to the choice of ε, and has a major impact on the outcomes. In general, it will be necessary to first
  identify a reasonable measure of similarity for the data set, before the parameter ε can be chosen. There is no estimation for this parameter, but the distance functions need to be chosen appropriately for the data set.

Conclusion:

We learnt about spatial data, spatial association rules, classification and clustering. We also studied an algorithm of each type and learnt about the working, advantages and the disadvantages of each one of them.