CLUSTER ANALYSIS

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.

Cluster analysis tools based on k-means, k-medoids, and several methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS.

DATA STRUCTURES

Data matrix (or object-by-variable structure):

This represents n objects, such as persons, with p variables (also called measurements or attributes), such as age, height, weight, gender, and so on. The structure is in the form of a relational table, or n-by-p matrix (n objects X p variables):

Dissimilarity matrix (or object-by-object structure):

This stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table:

The rows and columns of the data matrix represent different entities, while those of the dissimilarity matrix represent the same entity. Thus, the data matrix is often called a two-mode matrix, whereas the dissimilarity matrix is called a one-mode matrix. Many clustering algorithms operate on a dissimilarity matrix. If the data are presented in the form of a data matrix, it can first be transformed into a dissimilarity matrix before applying such clustering algorithms.

TYPES OF DATA

1. **Interval-scaled (numeric) variables :**

Interval-scaled (numeric) variables are continuous measurements of a roughly linear scale. Examples – weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.

The measurement unit used can affect the clustering Types of Data in Cluster analysis – For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to a very different clustering structure.

2. Binary variables:

A binary variable has only two states: 0 or 1, where 0 means that the variable is absent, and 1 means that it is present.

Given the variable smoker describing a patient 1 indicates that the patient smokes and 0 indicates that the patient does not. Treating binary variables as if they are interval-scaled can lead to misleading clustering results. Therefore, methods specific to binary data are necessary for computing dissimilarities.

3. Categorical (nominal) variables:

A categorical (nominal) variable is a generalization of the binary variable in that it can take on more than two states. – Example: map_color is a categorical variable that may have five states: red, yellow, green, pink, and blue. The states can be denoted by letters, symbols, or a set Types of Data in Cluster Analysis The states can be denoted by letters, symbols, or a set of integers.

Nominal variables: Has no particular order to its categories.

Ordinal variables:

A discrete ordinal variable resembles a categorical variable, except that the M states of the ordinal value are ordered in a meaningful sequence.

Example: professional ranks are often enumerated in a sequential order, such as assistant, associate, and full for professors. Ordinal variables may also be obtained from the discretization of interval-scaled quantities by splitting the value range into a finite number of classes. The values of an ordinal variable can be mapped to ranks.

Example: Suppose that we have the sample data:

There are three states for test-2, namely fair, good, and excellent.

4. **Mixed Variable:** It is a combination of different variables.

PROPERTIES

- Clustering scalability
- Algorithm usability with multiple types of data
- Dealing with unstructured data
- Interoperability

PARTITIONING METHODS

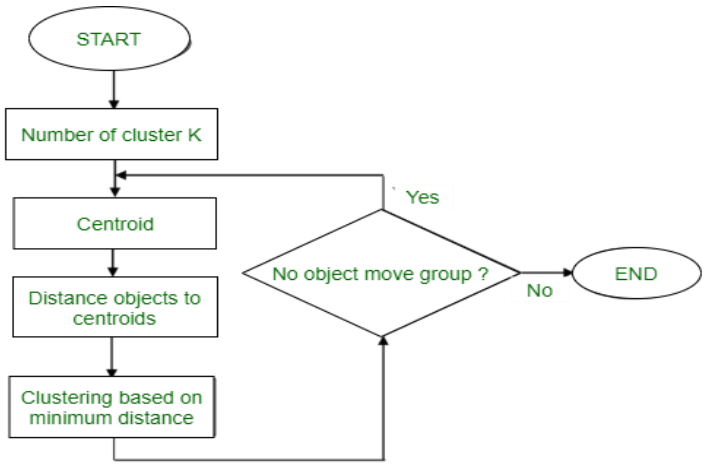
A partitioning method constructs k partitions of the data, where each partition represents a cluster and k <= n. That is, it classifies the data into k groups, which together satisfy the following requirements:

Each group must contain at least one object, and Each object must belong to exactly one group. A partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are close or related to each other, whereas objects of different clusters are far apart or very different.

There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications) etc. In this article, we will be seeing the working of K Mean algorithm in detail.

K-Mean (A centroid based Technique): The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster). The similarity of the cluster is determined with respect to the mean value of the cluster. It is a type of square error algorithm. At the start randomly k objects from the dataset are chosen in which each of the objects represents a cluster mean(centre). For the rest of the data objects, they are assigned to the nearest cluster based on their distance from the cluster mean. The new mean of each of the cluster is then calculated with the added data objects.

The K-means algorithm effectively partitions the dataset into clusters based on feature similarity, with each iteration refining the cluster centroids to minimize intra-cluster variance. This method is widely used for its simplicity and efficiency in handling large datasets.



Example:

Customer	Income	Spending Score
A	15	39
B	16	81
C	17	6
D	18	77
E	19	40
F	20	76

We want to group these customers into 2 clusters (K=2) using the K-means algorithm.

Steps:

1. **Initialization:**
   - Select K initial centroids randomly. Let's assume we randomly choose the centroids as follows:
     - Centroid 1 (C1): (15, 39)
     - Centroid 2 (C2): (18, 77)
2. **Assignment Step:**
   - Assign each point to the nearest centroid based on the Euclidean distance.
     - Distance formula: distance=(x2−x1)2+(y2−y1)2\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}distance=(x2−x1)2+(y2−y1)2

   Calculating distances:

   - For Customer A (15, 39):
     - Distance to C1: (15−15)2+(39−39)2=0\sqrt{(15-15)^2 + (39-39)^2} = 0(15−15)2+(39−39)2=0
     - Distance to C2: (18−15)2+(77−39)2=32+382=9+1444≈38.12\sqrt{(18-15)^2 + (77-39)^2} = \sqrt{3^2 + 38^2} = \sqrt{9 + 1444} \approx 38.12(18−15)2+(77−39)2=32+382=9+1444≈38.12
     - Closest to C1.
   - For Customer B (16, 81):
     - Distance to C1: (16−15)2+(81−39)2=1+1764≈42\sqrt{(16-15)^2 + (81-39)^2} = \sqrt{1 + 1764} \approx 42(16−15)2+(81−39)2=1+1764≈42
     - Distance to C2: (18−16)2+(77−81)2=4+16=20≈4.47\sqrt{(18-16)^2 + (77-81)^2} = \sqrt{4 + 16} = \sqrt{20} \approx 4.47(18−16)2+(77−81)2=4+16=20≈4.47

        - Closest to C2.
    - Repeat for other points.

Assignments:

- Cluster 1: {A, C, E}
- Cluster 2: {B, D, F}

3. **Update Step:**
    - Recalculate the centroids of the clusters by averaging the points in each cluster.
        - New centroid for Cluster 1: (15+17+193,39+6+403)=(17,28.33)\left(\frac{15+17+19}{3}, \frac{39+6+40}{3} \right) = (17, 28.33)(315+17+19,339+6+40)=(17,28.33)
        - New centroid for Cluster 2: (16+18+203,81+77+763)=(18,78)\left(\frac{16+18+20}{3}, \frac{81+77+76}{3} \right) = (18, 78)(316+18+20,381+77+76)=(18,78)
4. **Repeat Steps 2 and 3:**
    - Reassign points to the nearest centroid.
    - Update centroids based on new assignments.

Continue iterating until the centroids no longer change significantly or until a maximum number of iterations is reached.

 Initial Step:

Points: Centroids:
A(15, 39) C1(15, 39)
B(16, 81) C2(18, 77)
C(17, 6)
D(18, 77)
E(19, 40)
F(20, 76)

 First Assignment:

Cluster 1: {A(15, 39), C(17, 6), E(19, 40)}
Cluster 2: {B(16, 81), D(18, 77), F(20, 76)}

 First Update:

New Centroids:
C1: (17, 28.33)
C2: (18, 78)

HIERARICHAL METHOD

Groups the data into tree of clusters. The tree structure is called DENDOGRAM.

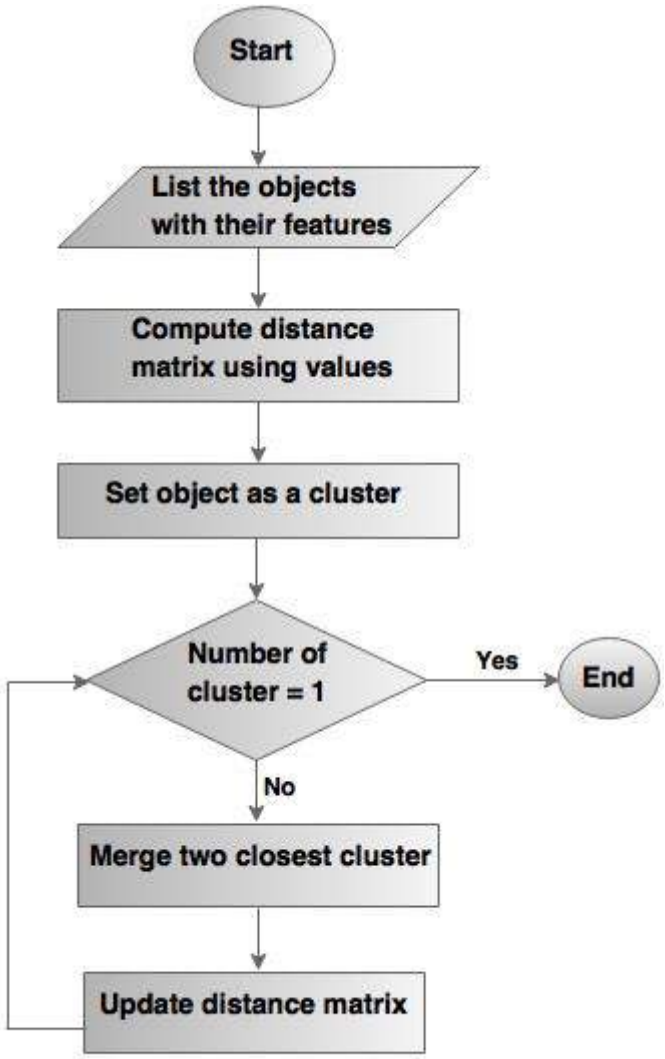
Dendogram has a sequence of all merges and splits.

We have 2 methods:

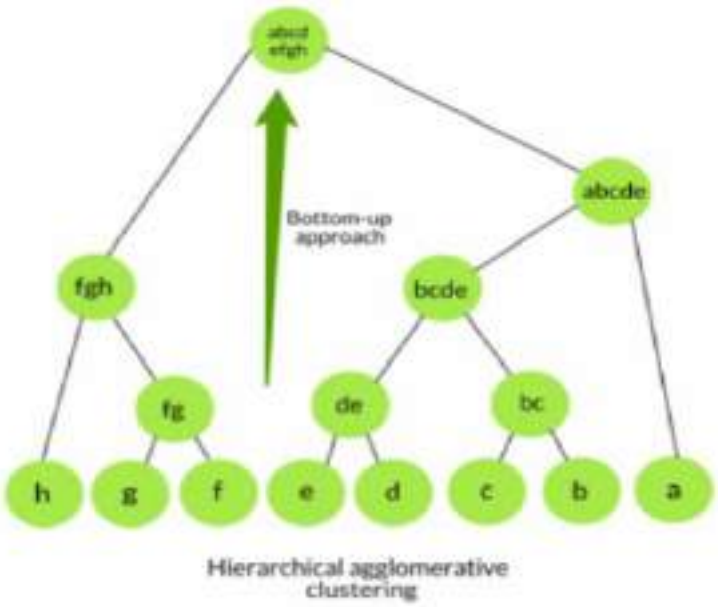
1. Agglomerative method--- Merges
2. Divisive method--- Splits

Agglomerative Method

- It is a bottom-up approach, in which clusters have sub-clusters.
- The process is explained in the following flowchart.



Agglomerative hierarchical clustering flowchart



Hierarchical agglomerative clustering

Example: For given distance matrix, draw single link, complete link and average link dendrogram.

	A	B	C	D	E
A	0				
B	2	0			
C	6	3	0		
D	11	9	7	0	
E	9	8	5	4	0

Solution: Single link: From given distance matrix.

Step 1:

	A, B	C	D	E
A, B	0			
C	3	0		
D	9	7	0	

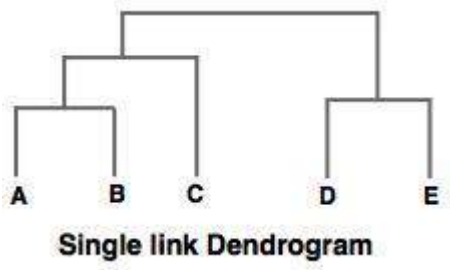
E	8	5	4	0

Step 2:

	A, B, C	D	E
A, B, C	0	0	
D	7	4	
E	5	0	0

Step 3:

	A, B, C	D, E
A, B, C	0	
D, E	5	0



Single link Dendrogram

2. Complete link: Find maximum distance to draw complete link.

Step 1:

	A, B	C	D	E
A, B	0			
C	6	0		
D	11	7	0	
E	9	5	4	0

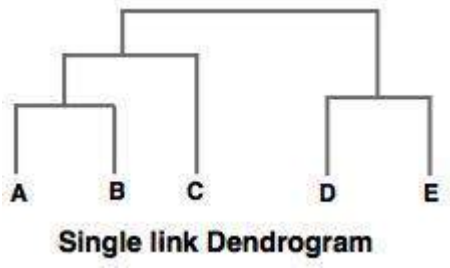
Step 2:

	A, B	C	D, E

A, B	0		
C	6	0	
D, E	11	7	0

Step 3:

	A, B, C	D, E
A, B, C	0	
D, E	11	0



Single link Dendrogram

3. Average link: To draw average link, compute the average link.

Step 1:

I) 6+3/2 = 4.5

ii) 11+9/2 = 10

iii) 9+8/2 = 8.5

	A, B	C	D	E
A, B	0			
C	4.5			
D	10	7	0	
E	8.5	5	4	0

Step 2:

I) 11+9+9+8/2 = 9.25

II) 7+5/ 2 = 6

	A, B	C	D, E
A, B	0		

C	4.5	0	
D, E	9.25	6	0

Step 3:

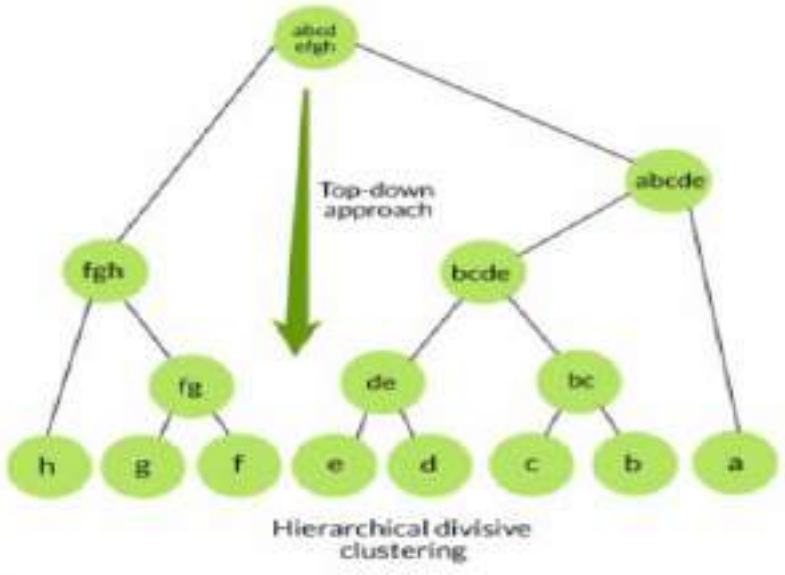
11+9.25+9.25+8+7+5/6 = 8.20

	A, B, C	D, E
A, B, C	0	
D, E	8.20	0

Divisive Method

It is a Top-Down approach

Divisive clustering **starts with one, all-inclusive cluster**. At each step, it **splits a cluster until each cluster contains a point** (or there are k clusters).



DENSITY-BASED CLUSTERING METHOD

Density-based clustering refers to **unsupervised ML approaches that find discrete clusters in the dataset**, based on the notion that a cluster/group in a dataset is a continuous area of high point density that is isolated from another cluster by sparse regions. Typically in data points in the dividing, sparse zones are regarded as noise or outliers.

DBSCAN (Density Based Spatial Clustering Of Application With Noise) clustering is used to differentiate between dense clusters and sparser noise. The DBSCAN algorithm is the fastest of the clustering algorithms, but it can only be used if there is a clear Search Distance that applies to all candidate clusters and performs effectively. This implies that all significant clusters possess comparable densities. The Search Time Interval and Time Field parameters allow you to locate spatiotemporal groups of points.

Density = number of points within a specified radius (Eps)

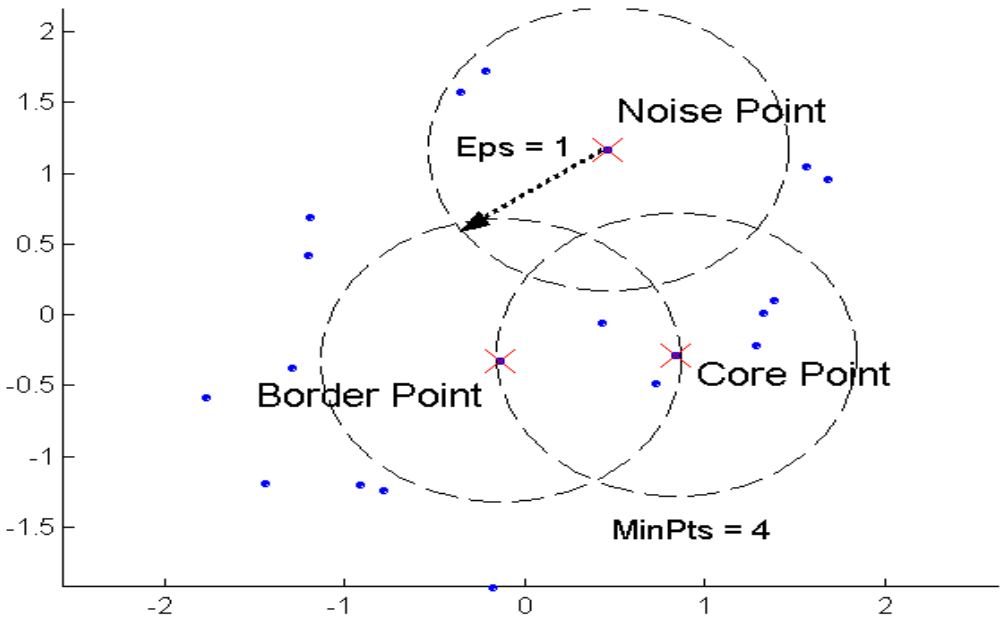
– A point is a core point if it has more than a specified number of points (MinPts) within Eps

- These are points that are at the interior of a cluster

– A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point

– A noise point is any point that is not a core point or a border point.

Core Points: It should satisfy the condition of minpts
Border Points: Neighbor of core
Noise Points: Nor core nor boundary



DBSCAN ALGORITHM

Eliminate noise points

Perform clustering on the remaining points

```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← current_cluster_label + 1
    Label the current core point with cluster label current_cluster_label
  end if
  for all points in the Eps-neighborhood, except i^th the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label current_cluster_label
    end if
  end for
end for
```

GRID-BASED CLUSTERING METHOD

Grid-Based Clustering method uses a multi-resolution grid data structure.

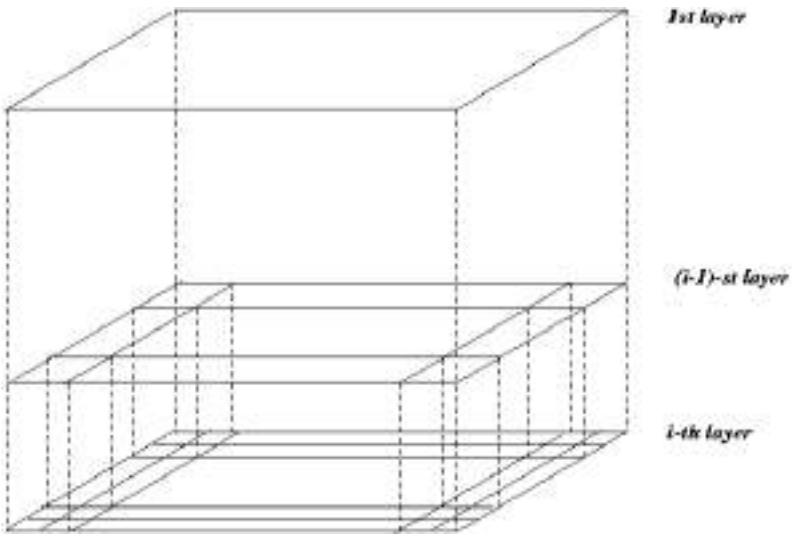
- **STING** (a **ST**atistical **IN**formation **G**rid approach) by Wang, Yang, and Muntz (1997)

STING - A Statistical Information Grid Approach

STING was proposed by Wang, Yang, and Muntz (VLDB'97).

In this method, the spatial area is divided into rectangular cells.

There are several levels of cells corresponding to different levels of resolution.



For each cell, the high level is partitioned into several smaller cells in the next lower level.
The statistical info of each cell is calculated and stored beforehand and is used to answer queries.
The parameters of higher-level cells can be easily calculated from parameters of lower-level cell

- Count, mean, s, min, max
- Type of distribution—normal, uniform, etc.

Then using a top-down approach we need to answer spatial data queries.
Then start from a pre-selected layer—typically with a small number of cells.
For each cell in the current level compute the confidence interval.
Now remove the irrelevant cells from further consideration.

When finishing examining the current layer, proceed to the next lower level.
Repeat this process until the bottom layer is reached.

OUTLIERS ANALYSIS

Outliers in Data mining is a very hot topic in the field of data mining. Let's discuss the outliers.

The data which deviates too much far away from other data is known as an outlier. The outlier is the data that deviate from other data.

The outlier shows variability in an experimental error or in measurement. In other words, an outlier is a data that is far away from an overall pattern of the sample data.

Outliers can indicate that the population has a heavy-tailed distribution or when measurement error occurs.

Outliers can be categorized as;

1. Collective outliers.
2. Point outliers
3. Contextual outliers

Collective outliers can be subsets of outliers when we introducing the novelties in data. For example, a signal that may indicate the discovery of a new phenomenon for the data set.
Point outliers are the data points that are far from the other distribution of the data.
Contextual outliers are the outliers just like noisy data. One example of noise data is when data have a punctuation symbol and suppose we are analyzing the background noise of the voice when doing speech recognition.
There are two types of Outliers.

1. Univariate outliers
2. Multivariate outliers

A univariate outlier is a data outlier that differs significantly from one variable. A multivariate outlier is an outlier when a combination of values on two or more than two variables have a significant difference. The univariate outlier and Multivariate outliers can influence the overall outcome of the data analysis.

Outliers can have many different causes. Some of these causes are mentioned below.

- Ther instruments used in the experiments for taking measurements suddenly malfunctioned.
- The error in data transmission.
- Due to changes in system behavior.
- Due to fraudulent behavior
- Due to human error
- Due to natural deviations in populations.
- Due to flaws in the assumed theory.
- Incorrect data collection.

Algorithm to Detect Outlier in data mining.

1. Calculate the mean of each cluster of the data.
2. Initialize the Threshold value of the data.
3. Calculate the distance of the test data from each cluster mean
4. Find the nearest cluster to the test data
5. Now, if we found that Distance is greater than Threshold, then it is a signal of Outlier.

There are many methods of outlier detection. Some of the outlier detection methods are mentioned below;

- Z-Score Normalizatoin
- Linear Regression Models (PCA, LMS)
- Information Theory Models
- High Dimensional Outlier Detection Methods (high dimensional sparse data)
- Proximity Based Models (non-parametric)
- Probabilistic and Statistical Modeling (parametric)
- Probabilistic and Statistical Modeling (parametric)
- Numeric Outlier

Numeric Outlier

Numeric Outlier is the nonparametric outlier detection technique in a one-dimensional feature space. TheNumeric outliers calculation can be performed by means of the InterQuartile Range (IQR).

Z-Score

Z-score is a data normalization technique and assumes a Gaussian distribution of the data. Outliers detection can be performed by Z-Score.

DBSCAN

The DBSCAN technique is based on the DBSCAN clustering algorithm. DBSCAN is a density-based, nonparametric outlier detection technique in a 1 or multi-dimensional feature space. In DBSCAN, all the data points are defined in the following points.

1. Core Points
2. Border Points
3. Noise Points.