Course: PMIT-6307; Class Test-03; Marks-10; Time-30 Minutes

Ch-08.

1. Illustrate the process of the 5-nearest neighbor's algorithm with an example. [4]
2. Consider a situation where you have 1000 fruits which are either 'banana' or 'apple' or 'other'. [6]
The possible attributes of any fruit are: [Long, Sweet, Yellow]. The training dataset & summary of the training dataset will look like Table-1 & Table-2.
Now, consider a case where you're given that a fruit is **long, sweet and yellow**; and you need to predict what type of fruit it is [using **Naive Bayes**].

Table 1: Training Dataset

Long	Sweet	Yellow	Fruit
0	0	1	Apple
1	0	1	Banana
0	1	0	Apple
1	1	1	Other
..	..	..	..

Table 2: Summary

Type	Long	Not Long	Sweet	Not sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Apple	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

Ch-09

3. Why do most researcher professor SVM classification? How does SVM work
4. What is the class imbalance problem? Why is it a problem? How mitigate this problem
5. Define scalable clustering algorithm with example. Where it is the most useful and why?

Ch-10.

6. Define hierarchical clustering; What is the different type of clustering
7. Write the steps of K-mean clustering; What is the different problem
8. Write some limitation of K-mean clustering. How can you overcome this problem
9. What do you mean by clustering? What are the different types of clusters?
10. How a KNN classification performed? How do you decide min Pts and Eps?
11. Is density based clustering
12. Or not

(12) $\sqrt{(161-165)^2+(61-65)^2} = 5.66$

(13) $\sqrt{(161-168)^2+(61-62)^2} = 7.07$

(14) $\sqrt{(161-168)^2+(61-63)^2} = 7.28$

(15) $\sqrt{(161-168)^2+(61-66)^2} = 8.60$

(16) $\sqrt{(161-170)^2+(61-63)^2} = 9.22$

(17) $\sqrt{(161-170)^2+(61-64)^2} = 9.49$

(18) $\sqrt{(161-170)^2+(61-68)^2} = 11.40$

Step 1: Calculate similarity based on distance function using Euclidean Distance

Step 2: Find K-Nearest Neighbors: Let k = 5. Searches for the 5 customer closest to Noreen. If 4 of them had medium T-shirt size then best guess for Noreen is medium T-shirt size.

Nearest Neighbor Classifier

Let K=5.

Give that,
Customer name = Monica.
Height = 161 Cm.
Weight = 61 kg
Target size = ?
Calculate Euclidean Distance in all attribute:—

SL	Height	Weight	Size	Distance	Position
01	158	58	M	4.24	
02	158	59	M	3.61	
03	158	63	M	3.61	
04	160	59	M	1.41	01
05	160	60	M	1.42	02
06	163	60	M	2.24	04
07	163	61	M	2.00	03
08	160	64	L	3.16	05
09	163	64	L	3.61	
10	165	61	L	4.00	
11	165	62	L	4.12	
12	165	65	L	5.66	
13	168	62	L	7.07	
14	168	63	L	7.28	
15	168	66	L	8.60	
16	170	63	L	9.22	
17	170	64	L	9.49	
18	170	68	L	11.40	

(1) $\sqrt{(161-158)^2+(61-58)^2} = 4.24$
(2) $\sqrt{(161-158)^2+(61-59)^2} = 3.61$
(3) $\sqrt{(161-158)^2+(61-63)^2} = 3.61$
(4) $\sqrt{(161-160)^2+(61-59)^2} = 1.41$
(5) $\sqrt{(161-160)^2+(61-60)^2} = 1.41$
(6) $\sqrt{(161-163)^2+(61-60)^2} = 2.24$
(7) $\sqrt{(161-163)^2+(61-61)^2} = 2.00$
(8) $\sqrt{(161-160)^2+(61-64)^2} = 3.16$
(9) $\sqrt{(161-163)^2+(61-64)^2} = 3.61$
(10) $\sqrt{(161-165)^2+(61-61)^2} = 4.00$
(11) $\sqrt{(161-165)^2+(61-62)^2} = 4.12$
(12) $\sqrt{(161-165)^2+(61-65)^2} = 5.66$
(13) $\sqrt{(161-168)^2+(61-62)^2} = 7.07$
(14) $\sqrt{(161-168)^2+(61-63)^2} = 7.28$
(15) $\sqrt{(161-168)^2+(61-66)^2} = 8.60$
(16) $\sqrt{(161-170)^2+(61-63)^2} = 9.22$
(17) $\sqrt{(161-170)^2+(61-64)^2} = 9.49$
(18) $\sqrt{(161-170)^2+(61-68)^2} = 11.40$

Step 1: Calculate similarity based on distance function using Euclidean Distance.

Step 2: Find K-Nearest Neighbors: Let K=5. Scorches for the 5 customer closest to Monica. 3 of them had medium(M) size is largely size. Then best guess for Monica is Medium T shirt size.

Naive Bayes Classifier (Bayes theorem)

Frequency Table:-

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	4

Day	Outlook	Play
0	Rain	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rain	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rain	No
9	Sunny	No
10	Sunny	Yes
11	Rain	No
12	Overcast	Yes
13	Overcast	Yes
14		

Likelyhood Table:-

Weather	No	Yes	
Overcast	0	5	5/14 = 0.35
Rainy	2	2	4/14 = 0.29
Sunny	2	3	5/14 = 0.35
All	4/14 = 0.29	10/14 = 0.71	

Here, $P_{Sunny} = 0.35$, $P(Yes) = 0.71$, $P(Sunny/Yes) = \frac{3}{10} = 0.3$

$$P(Yes|Sunny) = \frac{0.3 \times 0.71}{0.35} = 0.60$$

Here, $P_{Sunny} = 0.35$, $P(No) = 0.29$, $P(Sunny/No) = \frac{2}{4} = 0.5$

$$P(No|Sunny) = \frac{0.5 \times 0.29}{0.35} = 0.41$$

Here, $P(Yes|Sunny) > P(No|Sunny)$.

Hence, a sunny day, Player can play the game.

→ Vertical approach

Horizontal approach

$$P(Yes|Sunny) = \frac{P(Sunny/Yes) \times P(Yes)}{P(Sunny)} = \frac{0.3 \times 0.71}{0.35} \rightarrow 0.60$$

$$P(No|Sunny) = \frac{P(Sunny/No) \times P(No)}{P(Sunny)} = \frac{0.5 \times 0.29}{0.35} = 0.41$$

$P(Sunny/Yes) = \frac{\text{Sunny Yes}}{\text{Yes Total}} = \frac{3}{10} = 0.3$

$P(Sunny/No) = \frac{\text{Sunny (No)}}{\text{No Total}} = \frac{2}{4} = 0.5$

Naive Bayes Classifier (Bayes theorem)

Frequency Table:-

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	4

Day	outlook	Play.
0	Rain	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rain	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rain	No
9	Sunny	No
10	Sunny	Yes
11	Rain	No
12	Overcast	Yes
13	Overcast	Yes
14		

Likelyhood Table:-

Weather	No	Yes	
Overcast	0	5	5/14 = 0.35
Rainy	2	2	4/14 = 0.29
Sunny	2	3	5/14 = 0.35
All	4/14 = 0.29	10/14 = 0.71	

Where, $P_{sunny} = 0.35$, $P(Yes) = 0.71$, $P(Sunny/Yes) = 3/10 = 0.3$

$$P(Yes/Sunny) = \frac{0.3 \times 0.71}{0.35} = 0.60$$

Where, $P_{sunny} = 0.35$, $P_{No} = 0.29$, $P(Sunny/No) = \frac{2}{4} = 0.5$

$$P(No/Sunny) = \frac{0.5 \times 0.29}{0.35} = 0.41$$

Here, $P(Yes/Sunny) > P(No/Sunny)$.

Hence, a sunny day, Player can play the game.

→ Vertical approach

Horizontal approach

$$P(Yes/Sunny) = \frac{P(Sunny/Yes) \times P(Yes)}{P(Sunny)} = \frac{0.3 \times 0.71}{0.35} \rightarrow 0.60$$

$$P(No/Sunny) = \frac{P(Sunny/No) \times P(No)}{P(Sunny)} = \frac{0.5 \times 0.29}{0.35} = 0.41$$

$P(Sunny/Yes) = \frac{\text{Sunny Yes}}{\text{Yes Total}} = \frac{3}{10} = 0.3$

$P(Sunny/No) = \frac{\text{Sunny (No)}}{\text{No Total}} = \frac{2}{4} = 0.5$

Example

outlook

	Yes	No	P(Yes)	P(No)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temp.

	Yes	No	P(Yes)	P(No)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(Yes)	P(No)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind.

	Yes	No	P(Yes)	P(No)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Day	outlook	Temp.	Humidity	Windy	
0	Rain	Hot	High	False	
1	Rain	Hot	High	True	
2	Overcast	Hot	High	F	
3	Sunny	Mild	High	F	
4	Sunny	Cool	Normal	F	
5	Sunny	Cool	Normal	T	No
6	Overcast	Cool	Normal	T	Yes
7	Rain	Mild	High	F	No
8	Rain	Cool	Normal	F	Yes
9	Sunny	Mild	Normal	F	Yes
10	Rain	Mild	Normal	T	Yes
11	Overcast	Mild	High	T	Yes
12	Overcast	Hot	Normal	F	Yes
13	Sunny	Mild	High	T	No

P(Play (Today))

Play	P(Yes)/P(No)
Yes (9)	9/14
No 5	5/14
Total 14	100%

Probability of playing golf is given by :—

$$P(\text{Yes}|\text{Today}) = \frac{P(\text{Sunny outlook}|\text{Yes})\,P(\text{Hot Temp}|\text{Yes})\,P(\text{Norm Humidity}|\text{Yes})\,P(\text{No wind}|\text{Yes})}{P(\text{Today})}$$

$$= \frac{2}{9}\cdot\frac{2}{9}\cdot\frac{6}{9}\cdot\frac{6}{9}\cdot\frac{9}{14} \approx 0.01411$$

$$P(\text{Not}|\text{Today}) = \frac{P(\text{Sunny outlook}|\text{No})\,P(\text{Hot Temp}|\text{No})\,P(\text{Norm Humidity}|\text{No})\,P(\text{No wind}|\text{No})}{P(\text{Today})}$$

$$\propto \frac{3}{5}\cdot\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{2}{5}\cdot\frac{5}{14} = 0.0068$$

So, prediction that golf would be played is "Yes"

CLUSTERING

1. What are the different types of clustering methods used in business intelligence?

Clustering is an undirected technique used in data mining for identifying several hidden patterns in the data without coming up with any specific hypothesis. The reason behind using clustering is to identify similarities between certain objects and make a group of similar ones.

There are two different types of clustering, which are hierarchical and non-hierarchical methods. Non-hierarchical Clustering In this method, the dataset containing N objects is divided into M clusters. In business intelligence, the most widely used non-hierarchical clustering technique is K-means. Hierarchical Clustering In this method, a set of nested clusters are produced. In these nested clusters, every pair of objects is further nested to form a large cluster until only one cluster remains in the end.

The different types of clustering methods below.

1. Density-Based Clustering
2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
3. OPTICS (Ordering Points to Identify Clustering Structure)
4. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)
5. Hierarchical Clustering
6. Fuzzy Clustering
7. Partitioning Clustering
8. PAM (Partitioning Around Medoids)
9. Grid-Based Clustering

2. When is Clustering used?

The primary function of clustering is to perform segmentation, whether it is store, product, or customer. Customers and products can be clustered into hierarchical groups based on different attributes. Another usage of the clustering technique is seen for detecting anomalies like fraud transactions. Here, a cluster with all the good transactions is detected and kept as a sample. This is said to be a normal cluster. Whenever something is out of the line from this cluster, it comes under the suspect section. This method is found to be really useful in detecting the presence of abnormal cells in the body. Other than that, clustering is widely used to break down large datasets to create smaller data groups. This enhances the efficiency of assessing the data.

3. What are the advantages of Clustering?

Clustering is said to be more effective than a random sampling of the given data due to several reasons. The two major advantages of clustering are: Requires fewer resources A cluster creates a group of fewer resources from

the entire sample. Due to this, there is a lesser requirement of resources as compared to random sampling. Random sampling will require travel and administrative expenses, but this is not the case over here. Feasible option Here, every cluster determines an entire set of the population as homogeneous groups are created from the entire population. With this, it becomes easy to include more subjects in a single study.

4. Why do use clustering in ML?

Cluster analysis is usually used to classify data into structures that are more easily understood and manipulated. It is an unsupervised machine learning task.

5. What is the difference between clustering and classification in ML?

Classifying the input labels basis on the class labels is classification. On the other hand, the process of grouping basis the similarity without taking help from class labels is known as clustering.

6. Why clustering is better than classification?

The machine learns from the existing data in clustering because the need for multiple pieces of training is not required. Grouping is done on similarities as it is unsupervised learning. Classification on the contrary is complex because it is a supervised type of learning and requires training on the data sets.

7. What is hierarchical clustering?

It is a form of clustering algorithm that produces 1 to n clusters, where n represents the number of observations in a data set. There are two types of hierarchical clustering, divisive (top-down) and agglomerative (bottom-up).

8. Clustering Data Mining techniques: Hierarchical Clustering

When you're on a quest to find data pieces and map them according to cluster probability, the Hierarchical Clustering method works like a **charm**. Now, the mapped data pieces may belong to a Cluster with distinct qualities in terms of multidimensional scaling, cross-tabulation, or quantitative relationships among data variables in several aspects.

Considering how to find a single cluster after merging the various clusters while retaining the hierarchy of the attributes on which they are classed in mind? The stages of the Hierarchical Clustering method mentioned below can be used to accomplish this:

- Begin by picking the data points and clustering them according to the hierarchy.
- Are you considering how the clusters will be interpreted? With a Top-down or Bottom-up method, a Dendrogram may be utilized to comprehend the hierarchy of clusters properly.
- Clusters are merged until only one remains, and we may use a variety of metrics to determine how close the clusters are when merging them,

such as Euclidean distance, Manhattan distance, or Mahalanobis distance.

- For the time being, the process has ended because the intersection point has been discovered and well mapped on the dendrogram.

What is Clustering in Data Mining?

In clustering, a group of different data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data.

Data clustering, also called data segmentation, aims to partition a collection of data into a predefined number of subsets (or clusters) that are optimal in terms of some predefined criterion function. Data clustering is a fundamental and enabling tool that has a broad range of applications in many areas.

What is clustering and its types?

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters.

Density-Based Clustering

In this method, the clusters are created based upon the density of the data points which are represented in the data space. The regions that become dense due to the huge number of data points residing in that region are considered as clusters.

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

DBSCAN groups data points together based on the distance metric. It follows the criterion for a minimum number of data points. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.It takes two parameters – *eps* and *minimum points.* Eps indicates how close the data points should be to be considered as neighbors. The criterion for minimum points should be completed to consider that region as a dense region.

5) Clustering Data Mining techniques: Density-Based Spatial Clustering

When it comes to discovering clusters in bigger geographical databases, the Density-based Spatial Clustering Algorithm with Noise (or DBSCAN) is a superior alternative to K Means when it comes to cross-examining the density of its data points. It's also more appealing and efficient than CLARANS, which stands for Clustering LARge ApplicatioNS via Medoid-based partitioning approach.

The DBSCAN Clustering algorithm approach is beneficial and comparable to the mean-shift density-based Clustering algorithm.

DBSCAN's method starts with an unvisited data point and uses distance (Epsilon) to extract the neighborhood before designating the point as visited. If two points are within a certain distance of each other, they have termed neighbors. Following are the steps of DBSCAN:

- When enough points (based on minPoints) are found, the clustering process begins, using the current data point as the initial point in the new cluster. If there aren't enough points, the algorithm flags it as visited and classifies it as noise defect clustering.
- The initial point in the new cluster utilizes the same distance to define its neighborhood, resulting in a clustered point neighborhood, and the process continues for every additional cluster point added to the group. This process is repeated until all data points have been labeled and visited.
- After all of the data points in the neighborhood have been visited, a fresh unvisited data point is chosen for clustering. As a result, all data points are labeled as noise or clustered under the visited label.

In density based clustering , How do you decide min Point an EPS

The eps should be chosen based on the distance of the dataset (we can use a k-distance graph to find it), but in general small eps values are preferable. minPoints: As a general rule, a minimum minPoints can be derived from a number of dimensions (D) in the data set, as minPoints ≥ D + 1.

Parameters: The DBSCAN algorithm basically requires 2 parameters:

***eps**: specifies how close points should be to each other to be considered a part of a cluster. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors.*

minPoints: the minimum number of points to form a dense region. For example, if we set the minPoints parameter as 5, then we need at least 5 points to form a dense region.

Parameter estimation:

The parameter estimation is a problem for every data mining task. To choose good parameters we need to understand how they are used and have at least a basic previous knowledge about the data set that will be used.

eps: if the eps value chosen is too small, a large part of the data will not be clustered. It will be considered outliers because don't satisfy the number of points to create a dense region. On the other hand, if the value that was chosen is too high, clusters will merge and the majority of objects will be in the same cluster. The eps should be chosen based on the distance of the dataset (we can use a k-distance graph to find it), but in general small eps values are preferable.

minPoints: As a general rule, a minimum minPoints can be derived from a number of dimensions (D) in the data set, as minPoints ≥ D + 1. Larger values are usually better for data sets with noise and will form more significant clusters.

The minimum value for the minPoints must be 3, but the larger the data set, the larger the minPoints value that should be chosen.

Why should we use DBSCAN?

The DBSCAN algorithm should be used to find associations and structures in data that are hard to find manually but that can be relevant and useful to find patterns and predict trends.

Clustering methods are usually used in biology, medicine, social sciences, archaeology, marketing, characters recognition, management systems and so on.

Let's think in a practical use of DBSCAN. Suppose we have an e-commerce and we want to improve our sales by recommending relevant products to our customers. We don't know exactly what our customers are looking for but based on a data set we can predict and recommend a relevant product to a specific customer. We can apply the DBSCAN to our data set (based on the e-commerce database) and find clusters based on the products that the users have bought. Using this clusters we can find similarities between customers, for example, the customer A have bought 1 pen, 1 book and 1 scissors and the customer B have bought 1 book and 1 scissors, then we can recommend 1 pen to the customer B. This is just a little example of use of DBSCAN, but it can be used in a lot of applications in several areas.

How can we easily implement DBSCAN?

As I already wrote (tip: don't believe in everything I write) the DBSCAN is a well-known algorithm, therefore, you don't need to worry about implement it yourself. You can use one of the libraries/packages that can be found on the internet. Here is a list of links that you can find the DBSCAN implementation:

I also have developed an application (in Portuguese) to explain how DBSCAN works in a didactically way. The application was written in C++ and you can find it on Github.

What is a real life example of DBSCAN clustering?

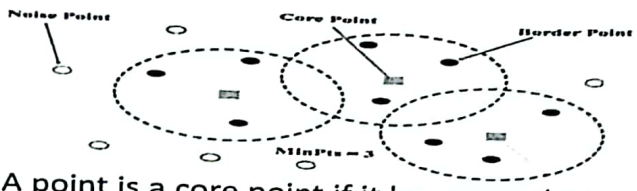
Using this clusters we can find similarities between customers, for example, if customer A has bought a pen, a book and one pair scissors, while customer B purchased a book and one pair of scissors, then you could recommend a pen to customer B.

The basic principle of dbscan clustering

The principle of DBSCAN is to find the neighborhoods of data points exceeds certain density threshold. The density threshold is defined by two parameters: the radius of the neighborhood (eps) and the minimum number of neighbors/data points (minPts) within the radius of the neighborhood. With these two thresholds in mind, DBSCAN starts from a random point to find its first density neighborhood. Then from its first density neighborhood, it starts to find the second density neighborhood. If the second density neighborhood exists, DBSCAN will merge the first and second density

neighborhoods to become a bigger density neighborhood. The same process keep going on until its neighborhood cannot meet the threshold anymore.

Pictorial example of Core Point, Noise Point and Border Point-Explain.



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 nor a border point.

What is DBSCAN useful for?

DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density.
Advantages. DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means. DBSCAN can find arbitrarily-shaped clusters. It can even find a cluster completely surrounded by (but not connected to) a different cluster.
Density-Based Clustering refers to one of the most popular unsupervised learning methodologies used in model building and machine learning algorithms. The data points in the region separated by two clusters of low point density are considered as noise.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

HDBSCAN is a density-based clustering method that extends the DBSCAN methodology by converting it to a hierarchical clustering algorithm.

Hierarchical Clustering

Hierarchical Clustering groups (Agglomerative or also called as Bottom-Up Approach) or divides (Divisive or also called as Top-Down Approach) the clusters based on the distance metrics.
In agglomerative clustering, initially, each data point acts as a cluster, and then it groups the clusters one by one. This comes under in one of the most sought-after clustering methods.
Divisive is the opposite of Agglomerative, it starts off with all the points into one cluster and divides them to create more clusters. These algorithms create a distance matrix of all the existing clusters and perform the linkage between

the clusters depending on the criteria of the linkage. The clustering of the data points is represented by using a dendrogram. There are different types of linkages: –

- **Single Linkage:** – In single linkage the distance between the two clusters is the shortest distance between points in those two clusters.
- **Complete Linkage:** – In complete linkage, the distance between the two clusters is the farthest distance between points in those two clusters.
- **Average Linkage:** – In average linkage the distance between the two clusters is the average distance of every point in the cluster with every point in another cluster.

K-Means Clustering: – K-Means clustering is one of the most widely used algorithms. It partitions the data points into k clusters based upon the distance metric used for the clustering. The value of 'k' is to be defined by the user. The distance is calculated between the data points and the centroids of the clusters.

K-means clustering is **a type of unsupervised learning used when you have unlabeled data** (i.e., data without defined categories or groups). This algorithm aims to find groups in the data, with the number of groups represented by the variable K. In this clustering method, the number of clusters found from the data is denoted by the letter 'K.'

The data point which is closest to the centroid of the cluster gets assigned to that cluster. After an iteration, it computes the centroids of those clusters again and the process continues until a pre-defined number of iterations are completed or when the centroids of the clusters do not change after an iteration. It is a very computationally expensive algorithm as it computes the distance of every data point with the centroids of all the clusters at each iteration. This makes it difficult for implementing the same for huge data sets.

2) Clustering Data Mining Techniques: K-Means Clustering

After determining the centroid value between two data points, the K-Means Clustering Algorithm repeatedly discovers the k number of clusters. It is rather useful to compute Cluster Centroids with their vector quantization observations, by virtue of which data points with changeable characteristics may be brought to Clustering.

As the clustering process accelerates, a large amount of unlabeled real-world data will become more efficient as it is split into clusters of different forms and densities. Have you ever considered how the centroid distance is calculated? Take a look at the K-means steps stated below:

- First, decide on the number of clusters that will differ in shape and density. Let's call that number k, and its value can be anything from 3,4 to anything else.
- You may now assign data points to the number of the cluster. The centroid distance is then calculated using the least squared Euclidean distance once the data point and cluster have been chosen.
- The data point resembles the cluster if it is substantially closer to the centroid distance; otherwise, it does not.
- Iteratively compute the centroid distances with the selected data point until you find the largest number of clusters made up of related data points. When assured convergence (a point where data points are well clustered) is reached, the algorithm stops clustering.

What are the Applications of Data Mining Clustering Techniques?

- Clustering can assist marketers identify unique groups in their consumer bases and describe them based on purchase behaviors in the business world.
- It may be used in biology to create plant and animal taxonomies to classify genes with similar functions.
- Clustering can also assist in identifying regions of land usage in an earth observation database, as well as groupings of motor insurance customers with a high average claim cost.
- Cluster analysis may be used as a standalone data mining function to obtain insight into data distribution, notice the features of each cluster, and focus on a specific group of clusters for further study.

What are the Requirements of Clustering Data Mining Techniques?

- **Scalability:** Many clustering techniques work well on small data sets with less than 200 data objects, however, a huge database might include millions of objects. Clustering on a subset of a big dataset might result in skewed findings. Clustering methods that are highly scalable are required.
- **Usability and interpretability:** Users anticipate interpretable, thorough, and usable clustering findings. As a result, clustering may require unique semantic interpretations and applications. It's crucial to investigate how the application aim influences Clustering Data Mining technique selection.
- **High dimensionality:** A database or a data warehouse can have several dimensions or properties. Many clustering algorithms excel

at dealing with low-dimensional data (two or three dimensions). Human eyes are capable of assessing clustering quality in up to three dimensions. Clustering data items in a high-dimensional space may be difficult, especially when the data is sparse and heavily skewed (misleading data).

- **Constraint-based clustering:** Clustering may be required in real-world applications due to a variety of restrictions. Assume you're in charge of selecting locations for a certain number of new automatic cash dispensing machines (ATMs) in a city. You may decide this by clustering households while taking into account limits such as the city's waterways, highway networks, and client needs per area. Finding groupings of data with appropriate clustering behavior that fulfill stated requirements is a difficult issue.

Limitations of K-means clustering

- difficult to choose the number of clusters, k
- cannot be used with arbitrary distances
- sensitive to scaling – requires careful preprocessing
- does not produce the same result every time
- clever initialization helps finding better results
- sensitive to outliers (squared errors emphasize outliers)
- cluster sizes can be quite unbalanced (e.g., one-element outlier clusters)

Process of overcome the Limitations of K-means clustering

A common way to avoid this problem is to use the K-means++ algorithm, which selects the initial centers based on a probability distribution that favors points that are far from each other. This can reduce the number of iterations and improve the quality of the final clusters.

How a KNN Classification

KNN classifier is a machine learning algorithm used for classification and regression problems. It works by finding the K nearest points in the training dataset and uses their class to predict the class or value of a new data point.

What is an example of KNN classification?

With the help of KNN algorithms, we can classify a potential voter into various classes like "Will Vote", "Will not Vote", "Will Vote to Party 'Congress', "Will Vote to Party 'BJP'. Other areas in which KNN algorithm can be used are Speech Recognition, Handwriting Detection, Image Recognition and Video Recognition.

Why use KNN classification?

The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.

Is KNN best for classification?

The main advantage of KNN over other algorithms is that KNN can be used for multiclass classification. Therefore if the data consists of more than two labels or in simple words if you are required to classify the data in more than two categories then KNN can be a suitable algorithm.

Step of KNN Classification

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.

Why do we use SVM? Why do most researcher perform SVM Classification.

SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. This is one of the reasons we use SVMs in machine learning. It can handle both classification and regression on linear and non-linear data.

What do you mean by OLAP?

Online analytical processing (OLAP) is software technology you can use to analyze business data from different points of view. Organizations collect and store data from multiple data sources, such as websites, applications, smart meters, and internal systems.

What is an example of an OLAP?

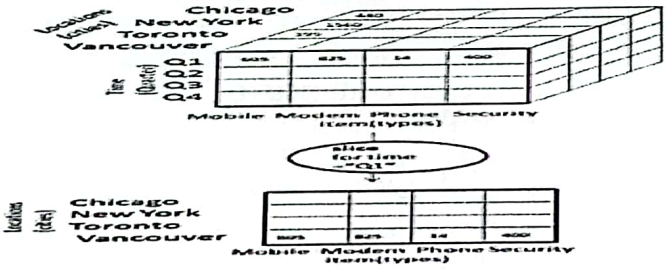
For example, a user can request that data be analyzed to display a spreadsheet showing all of a company's beach ball products sold in Florida in the month of July, compare revenue figures with those for the same products in September and then see a comparison of other product sales in Florida in the same time period.

What is Slicing and Dicing?

Slicing and Dicing refers to a way of segmenting, viewing and comprehending data in a database. Large blocks of data is cut into smaller segments and the process is repeated until the correct level of detail is achieved for proper analysis. Therefore slicing and dicing presents the data in new and diverse perspectives and provides a closer view of it for analysis. For example a report is showing annual performance of a particular product. If we want to view the quarterly performance, we can use slicing and dicing strategy to drill down to the quarterly level.

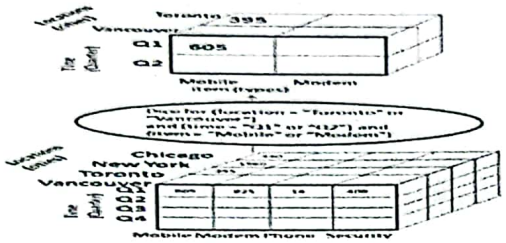
OLAP (Online Analytical Processing) is a computer process that enables user to select and extract data from different viewpoints. Slicing and Dicing term is generally used in OLAP databases that presents data to the end user in multidimensional cube format like a 3D spreadsheet (called an OLAP cube). The OLAP cube comprises numeric facts called measures which are categorized by dimensions.

Slicing refers to selecting a subset of the cube by choosing a single value for one of its dimension and creating a smaller cube with one less dimension.



In the above diagram, the slice is performed for dimension "time" using the variable "Q1" (criteria time). It generates a new cube with 2 dimensions. Thus slice operation produced a sliced OLAP cube by enabling the user to choose specific value for one of its dimension.

Dicing on the other hand generates a subcube by picking two or more values from multiple dimensions of the cube. The cube is rotated independent of its dimension.



In the above diagram, the dice operation is based on the following selection criteria involving all the three variables:

- Location = "Vancouver" or "Toronto"
- Time = "Q1" or "Q2"
- Item = " Modem" or "Mobile"

Thus dicing and slicing allows detailed analysis of data from different angles.

Hence, this concludes the definition of Slicing and Dicing along with its overview. This article has been researched & authored by the Business Concepts Team. It has been reviewed & published by the MBA Skool Team. The content on MBA Skool has been created for educational & academic purpose only.
Browse the definition and meaning of more similar terms. The Management Dictionary covers over 1800 business concepts from 5 categories.

why EDA is necessary for data mining

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Why is EDA important in data mining?

In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task.

Why should we perform EDA?

Exploratory data analysis can help detect obvious errors, identify outliers in datasets, understand relationships, unearth important factors, find patterns within data, and provide new insights.

why do we use cross validation in classification

The purpose of cross–validation is to test the ability of a machine learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.

Cross-validation is a technique that helps you assess the accuracy and generalizability of your predictive models. It involves splitting your data into multiple subsets and training and testing your model on different combinations of them.

Advantages of the learning curve model

Strategic planning to improve the output of employees or even whole departments. Motivating company staff by creating a culture of ongoing learning and progress-tracking. Identifying trends that can be used for more accurate forecasting and better business decisions.

The ROC curve is used to assess the overall diagnostic performance of a test and to compare the performance of two or more diagnostic tests. It is also used to select an optimal cut-off value for determining the presence or absence of a disease.

Advantages of the ROC curve

You say "Typically, the optimal cut-off value for a screening test is the one closest to the top left hand corner [of the ROC curve]." One could expand on that and say that the optimal point varies according to the pre-test probability (or prevalence) of the condition and the relative disadvantages (financial and/or health outcomes) of false positive and false negative test results. To me, the advantage of keeping the whole ROC curve is that the correct optimal point can be chosen for each case rather than losing information by making a somewhat arbitrary choice of optimal cut-off point.

Professional Masters in IT (PMIT-6307); Class Test-01; Marks-15; Time-30 Minutes

1. Discuss whether or not each of the following activities is a data mining task. If the answer is yes, then also specify which one of the task categories it will belong to: 4
   a. An image analyst obtains some new images and wants to automatically detect the number of distinct objects in the image. He does not have any prior information about these objects.
   b. By looking at a CT scan, a doctor wants to identify if a patient has cancer or not. There are a lot of labelled CT scans that the doctor will use for making the decision.
   c. Predicting the future stock price of a company using historical records. → Yes,
   d. Monitoring the heart rate of a patient for abnormalities.

2. Write short note on following application in terms of goals and approaches: 5
   a. Loan Default Prediction
   b. Credit Card Fraud Detection

3. Briefly describe three data quality problems. 3

4. Is temperature an interval or ratio measurement statistically? Justify your answer. 3

5. Explain data mining & its application
6. Difference bet descriptive & predictive data mining
7. Example of descriptive data mining.
8. Difference bet Classification & clustering.
9. What are the sequential step procedure classification
10. What is the association rule describe with example.
11. Is data mining is a machine learning.
12. Why do we need study more on data mining?
13. Challenges of implementation in data mining, Explain
14. Explain reasons for data affected.
15. Scalability define - Diff. bet. Data Mining & machine learning
16. categorization of Attribute
17. Same above - 4.
18. Write the different types of attributes.

(2) Discuss whether or not each of the following activities is a data mining task. If your answer is yes, then also specify which one of the task categories it will belong to:

a) An image analyst obtains some new images and want to automatically detect the number of distinct objects in the image. He does not have any prior information about these objects.

Ans: Yes,

b) By looking at a CT scan, a doctor wants to identify if a patient has cancer or not. There are a lot of labeled CT scan that the doctor will use for making a decision.

Yes,
Goal: Predict Breast dementia available
Approach: MRD report collection analysis
Grouped, the MRI report data
Screen grouped in data to detect brain donation

c) Predicting the future stock price of company using historical record.

Yes, we would attempt to create a model that can predict the certain value of the stock price. Thus this is example of the area of data mining known as predictive modeling. We could use regression for this modelling, although researchers in many fields have developed a wide variety of techniques for predicting time series.

d) Monitoring the heartrate of a patient for abnormalities.

Yes, we would build a model of the normal behavior of heart rate and raise an alarm when unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection. This could also be considered as a classification problem if we have examples of both normal and abnormal heart behavior.

(13) Briefly describe three Data quality problems.

Ans:- Data has become the heart of all business across the world. Organizations heavily rely on data analysis for the decision making process, but unfortunately 100% clean, accurate and quality data doesn't exist. Data is impacted by numerous factors that deteriorate its quality. Data quality refers to the measurement of the current state of data against traits like completeness, accuracy, reliability, relevance and timeliness. While the data quality issue indicate the presence of defect that harms the above mentioned traits. Data is beneficial only, if it is of high quality, some of the consequences of poor quality data as as follows:—

① Human Error ② Data Duplication ③ Inconsistent Data ④ Inaccurate and Missing Data ⑤ Using the wrong formula. ⑥ Data Overload ⑦ Data Downtime ⑧ Hidden Data ⑨ Outdated data ⑩ Data illiteracy ⑪ Noise and outliers ⑫ Fake data ⑬ Duplicate data

Human Error:- Even with all the automation data is still typed on various web interfaces, hence there is a high possibility of typographical mistakes leading to inaccurate data. This data entry can be done both by customer and the employee. Customer may write the correct data into the wrong data field. Similarly employee may take a mistake while handling or migrating the data.

Data Duplication:- Now a days, data comes from multiple channels giving rise to duplicate data when merged. It result in multiple variations of the same record providing skewed analytical results and incorrect insights. The budget is also wasted on this duplicate records. We can make use of Data Duplicate Tools to find similar type of records within the

Inaccurate & Missing Data:- Inaccurate data can seriously impact the decision making ability of the business to achieve their goals. It is tough to identify because the former entries can be just some spelling mistake.

(2) Classification:

Write short notes on following application in terms of goals and approaches:—

a) Loan default prediction
b) Credit card fraud Detection.

Loan default prediction:-

Goal:- Predict loan defaulter.

Approach:-

i) Analysis bank transaction/statement.
ii) Regularly payment/irregularly payment history
iii) Irregularly payment history continuously or sudden period.
iv) If sudden period it is not detected.
v) If irregular payment history in continuum to detect loan defaulter.

Credit card fraud Detection:-

Goal:- Predict fraudent ones in credit card transaction.

Approach:-

i) Use credit card transaction and the information on its account hold as all item
ii) Previously transaction history as fraud or fair. This form the class attribute.
iii) Learn a model for the class of transaction
iv) Use this data

Pattern.
↓
Learning
↓
Loan defaulter | Loan well payer

(1) What is data mining?

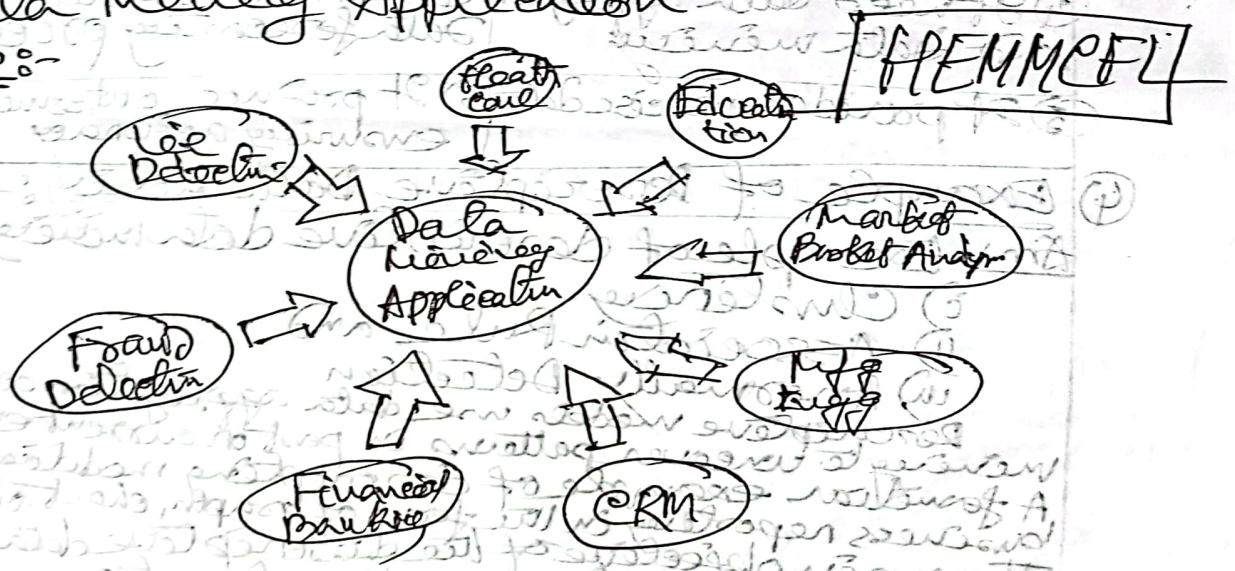
Ans:- The process of extracting information to identify patterns, trends and useful data that would allow the business the business to take the data-driven decision from huge sets of data is called Data Mining.

Input data → Data Processing → Data Mining → Post processing → Information

The process of knowledge discovery in database (KDD)

(2) Data Mining Application

Ans:-



③ Different Bet^n Descriptive & Predictive Data Mining

Descriptive Data Mining	Predictive Data Mining
① Descriptive mining is usually used to provide correlation, cross-tabulation, frequency etc.	① The term predictive means to predict something. So predictive data mining is the analysis done to predict the future event or other data or trend.
② It is based on the reactive approach.	② It is based on the proactive approach.
③ It specifies the characteristics of the data in a target data set.	③ It executes the induction over the current and past data, so that prediction can happen.
④ It need data aggregation and data mining	④ It need statistics and data forecasting procedure.
⑤ It provides precise data	⑤ It produce outcomes which ensuring accuracy

④ Example of Descriptive Data Mining:

Ans: Example of descriptive data mining include

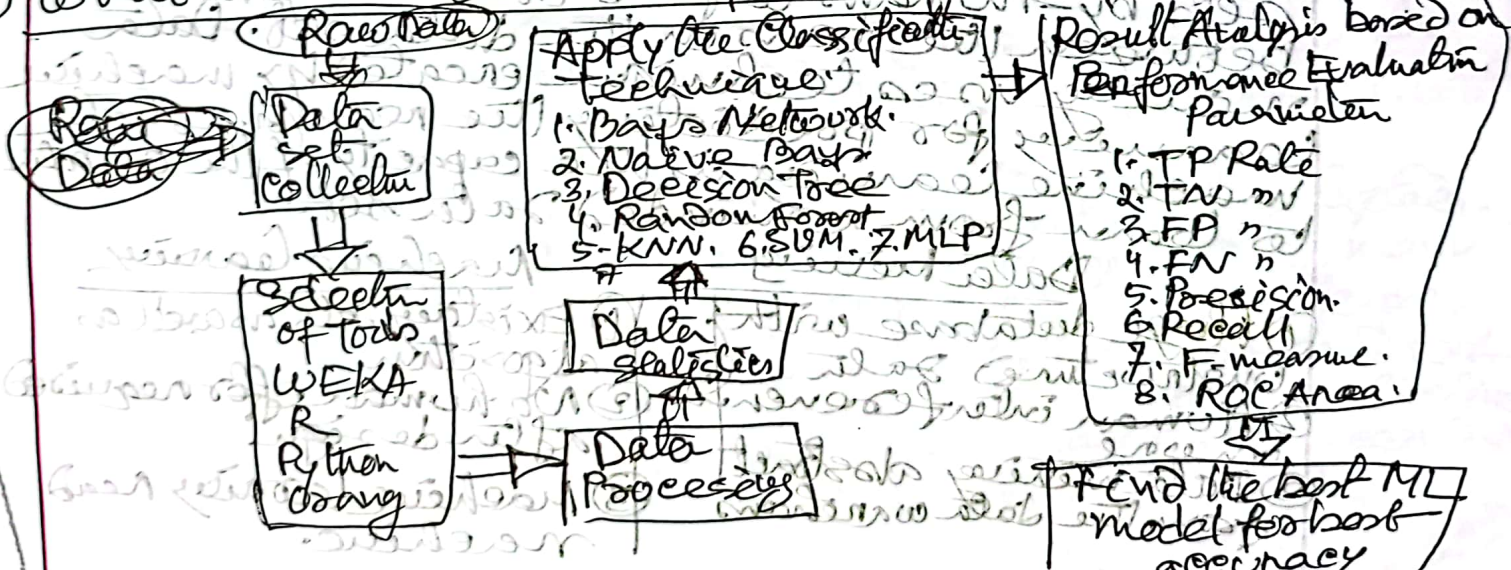
i) Clustering
ii) Association Rule and
iii) Anomally Detection

Descriptive words use data aggregation and data mining to uncover patterns in past or current event. A familiar example of descriptive modeling is business reporting in the form of graph, chart & dashboard. The main objective of the descriptive data mining is the summarization and transformation of data into useful information for monetoring & reporting purpose.

(5)

Classification	Clustering
(1) Classification is used to label data	(1) Clustering is used to group similar data instance together
(2) 	(2) 
(3) The no. of classes is known.	(3) The no. of classes is unknown.
(4) Training data is required	(4) No training data is required
(5) Based on the training data, the classification model is used to classify future instance into already defined classes.	(5) Clustering is used to make sense of existing data
(6) Popular algorithm for classification include Naive Bayes classifier, Decision Tree and Random Forests.	(6) Popular algorithm used for clustering include K-means, Means shift clustering and Density based Spatial Clustering of Application with Noise

(6) What are the sequential step in doing classification of set of data.



(7) What is the Associate Rule Discovery with example

Ans:- Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large database. It is intended to identify strong rules discovered in database using some measures of interestingness.

Associate rules are useful data mining for analyzing and forecasting customer behaviour. For example: peanut butter and jelly are frequently purchased together because a lot of people like to make PB&J sandwiches.

(8) In data mining is a machine learning

Ans:- Data mining is performed on certain data sets by humans to find interesting patterns between the items in the data set. Data mining uses techniques created by machine learning for predicting the results. while machine learning is the capacity of the computer to learn from a minded data set.

Data mining	Machine learning
(1) Huge database with unstructured data	(1) Existing data as well as algorithm
(2) Human interference more	(2) No human effort required after design
(3) Data mining abstract from the data warehouse	(3) Machine learning need machine.

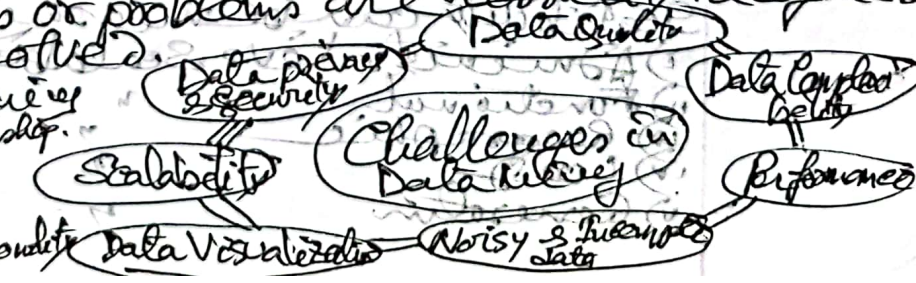
Why do we need study more on data mining.

Ans:- We need more study on data mining in the following reasons:-

i) The data mining techniques enables organization to obtain knowledge-based data.
ii) Data mining enables organizations to make lucrative modifications in operation and production.
iii) Compared with other statistical data applications, data mining is a cost-effective.
iv) Data-mining helps the decision-making process of an organization.
v) It facilitate the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
vi) It can be induced in the new system as well as the existing platforms.
vii) It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

(b) Challenges of implementation data mining, Explain.

Ans:- Data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods and techniques etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequate resolved.



Challenges of Data Mining
1) Data Quality
2) Data Complexity
3) Data Privacy & Security
4) Scalability.
5) Data Ownership.
6) High Dimensionality

(11) Explain reasons for Data affected:-

Ans: There are various types of reasons for affected or loss of Data is following:-

i) Hardware feature
ii) Human Error
iii) Malware
iv) Misplacing the Devices or theft
v) Software causes.
vi) Natural Disasters
vii) Hackers.
viii) Miscellaneous Issues

(12) Scalability:- Scalability is the property of a system handle a growing amount of work. One definition for software system specifies that this may be done by adding resources to the system.

In an economic context, a scalable business model implies that company can increase sales given increase resources. For example: a package delivery system is scalable because more packages can be delivered by adding more delivery vehicles. However, if all packages had to first pass through a single warehouse for sorting, the system would not be scalable, because one warehouse can handle only a limited number of packages.

Scalibility can be measured over multiple Dimensions:-

i) Administrative Scalability
ii) Functional
iii) Geographic
iv) Load
v) Generation

Categorization of Attributes :-

Quantitative Data :- Quantitative means we can count it and its numerical

i) Discrete data - Involved whole number. Example - People in class, Finger on your hand, No. of children someone etc.

ii) Continuous Data - It can be divided up as much as we want, and measure to many decimal places. Example - Weight of the car, temperature, speed of an Airplane etc.

Qualitative Data :- Qualitative means we can't and its not numerical

i) Nominal Data - It used to label variables without any quantitative values. Example (=, ≠).

a) Male/Female (sex)
b) Race Color
c) Nationalities
d) Name of people.

ii) Ordinal Data - It used for measure of satisfaction, happiness, etc. Example: How likely are you recommend our services (likely, Very likely, unlikely). (<, >) Good, Better, Best

iii) Interval Data - It allows to measure standard deviation and central tendency. Example - temperature in degree celcius. 20°C warmer than 10 and the difference between 20°C & 10°C is 10°C. The difference between 10 and 0 is also 10°C. (+, -)

iv) Ratio Data - Ratio data is very equal interval data, except zero means none. For ratio data, it is not possible to have negative values. Example - Height is ratio data. It is not possible to have negative height. If an object height is zero, then there is no object. This is different than something like temperature. Both 0° an -5° are completely valid and meaningful temperature (*, /). current/age, mass, Kelvin Temp in Kelvin

(a) Is temperature an interval or ratio measurement statistically? Justify your answer.

Ans:- Temperature can be either an interval or ratio attribute measurement statistically, depending on its measurement scale. When measured on the Kelvin scale, a temperature of 2 is, in a physically meaningful way, twice that of a temperature of 1. This is not true when temperature is measured on either the Celsius or Fahrenheit scale. It is an interval scale. Because zero is not the lowest possible temperature. In the Kelvin scale, a ratio scale, zero represents a total lack of thermal efficiency.

Fahrenheit temperature is not physically meaningful.

(b) List the different types of attributes with properties.

Ans:- Different types of Attributes with their properties follow is below:-

i) Nominal Attributes → Distinctness (=, ≠) → (ID, eye color, zip)

ii) Ordinal Attributes → Order (<, >) → (Rating, Grade, Height)

iii) Interval Attributes → Meaningful diff (+, −) → (Calendar date, Temp (C/F))

iv) Ratio → Ratio (*, /) → (Length, Count, Time)

CH-4,5,6,7 (+CT-2)

1. How a tree based classification works? What are the advantages of tree classification.
2. Define the terms: True positive, and false negative & F-measure.
3. Write the procedures of tree based classification. How does one can check the validity of a tree?
4. When do one can stop constructing classification tree?
5. Among tree based classification & Rule based ones, which one will you prepare, why?
6. When do we use learning curve and ROC curve?
7. Explain the Ensemble method of classification.
8. What is OLAP? Why do researcher use Slicing and Dicing in data analysis.
9. Data may affected by various kinds of reason - These reason are may defend in data quality problem.
10. Why EDA is necessary for Data Mining. Why do we use cross validation in classification.
11. Define the terms EDA, Why EDA is necessary for data mining? have to use Gini index in splitting.
12. What is the main principle of Gini index?, explain. When we
13. Why do we use data aggregation and dimension reduction
14. Give an example of stratified and write its purpose of use in data mining
15. When do we need descretization and binarization of data.

Chapter-3.
Data Processing.

What do we use data aggregation?

Ans:- We use data aggregation for below purposes:-

i) Data reduction - Reduce the number of attributes or objects.

ii) Change of Scale - Cities aggregated into regions, states, countries etc. Days aggregated into weeks, months, or years.

iii) More stable data - Aggregated data tends to have less variability.

Types of Samplings:-

A) Simple Random Sampling
   i) Sampling without replacement.
   ii) Sampling with replacement.

B) Stratified Sampling.

Dimensionality Reduction:-

Purposes:-
i) Avoid curse of dimensionality
ii) Reduce amount of time and memory required by data mining algorithm
iii) Allow data to be more easily visualized
iv) May help to eliminate irrelevant features or reduce noise.

Techniques:-
i) Principle Component Analysis (PCA)
ii) Singular value decomposition
iii) Others: supervised and non-linear techniques

Step for PCA Algorithm:-

① Getting the dataset ② Representing data into a structure ③ Standardizing the data ④ Calculating the covariance of ⑤ Calculating the Eigen values and Eigen vectors ⑥ Sorting the Eigen vectors ⑦ Calculating the new features or Principal components.

Chapter-4.

Similarity and Dissimilarity Measures.

Similarity Measures:-

- Numerical measure of how alike two data objects are
- Is higher when objects are more alike.
- Often falls in the range [0,1]

Dissimilarity Measures:-

- Numerical measure of how different two data objects are
- Lower when objects are more alike.
- Minimum dissimilarity is often 0.
- Upper limit varies.
- Proximity refers to a similarity or dissimilarity.

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = 0$ if $x = y$, 1 if $x \neq y$	$s = 1$ if $x = y$, 0 if $x \neq y$
Ordinal	$d = \{x-y\}/(n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = 1/(1+d)$, $s = 1 - d$ / $s = 1 - \frac{d - \min d}{\max d - \min d}$

Dissimilarity - Euclidean Distance:-

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the kth attributes (components) of data objects x and y.

Dissimilarity - Minkowski Distance:-

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r} = \left(|x_{j1} - x_{i1}|^p + |x_{i2} - x_{j2}|^p + (x_{i3} + x_{j3})^3 \right)^{1/3}$$

where r is the order parameter, n is the no. of dimensions (attributes) and x_k and y_k are respectively, the kth attributes (compound) of data objects.

Manhattan Distance:- $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |$

Examples:-

Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):-

a) Compute the Euclidean distance between two objects.

b) " " Manhattan " " " "

c) " " Minkowski " " " "

using $p = 3$.

Soln:-

a) Euclidean distance:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2 + (x_{i4} - x_{j4})^2}$$

$$= \sqrt{|22-20|^2 + |1-0|^2 + |42-36|^2 + |10-8|^2} = 6.71$$

b) Manhattan Distance:

$$d(i,j) = |x_{i1} - x_{i2}| + |x_{i2} - x_{j2}| + |x_{i3} - x_{j3}| + |x_{i4} - x_{i4}|$$

$$= |22-20| + |1-0| + |42-36| + |10-8|$$

$$= 11$$

c) Minkowski Distance

$$d(i,j) = \left(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + |x_{i3} - x_{j3}|^p + |x_{i4} - x_{j4}|^p\right)^{1/p}$$

$$= \left(|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3\right)^{1/3}$$

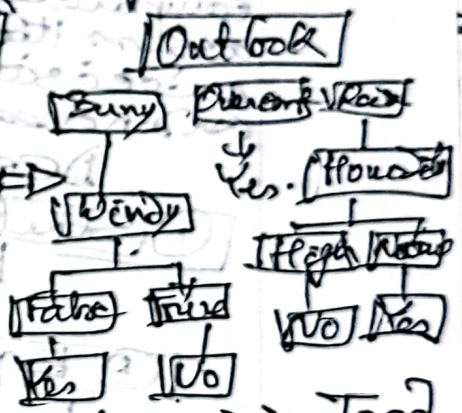
$$= 6.15$$

Chapter-5.

What is decision tree Classification?

Ans:- Decision tree builds classification or regression models in the form of a tree structure. It breakdown a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Example: A decision node (e.g. Outlook) has two or more branches (e.g. Sunny, Overcast, Rain). Leaf node (e.g. Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Outlook	Temp	Humidity	Windy	Play golf



- In what aspect you may like decision tree classification?

Ans:- A decision tree could be used to help a company decide which city to move its headquarter to or whether to open a satellite office. Decision trees also a popular tools in machine learning as trees can be use to build predictive models.

How does decision tree Works?

Ans:- A decision tree is a flow chart, that starts with one main idea and then branches out based on the consequences of your decision. It's called a decision tree because the model typically looks like a tree with branches.

Those trees are used for decision tree analysis, which involves visually outlining potential outcomes, costs and consequences of a complex decision. You can use a decision tree to calculate the expected value of each outcome based on the decision and consequence that led to it. Then by comparing the outcomes to one another, you can assess the best course of action. You can also use a decision tree to solve problems, manage cost and reveal opportunities.

Decision tree analysis work in five step:-

1. Start with your idea
2. Add chance and decision nodes
3. Expand until you reach end point
4. Calculate Tree Value
5. Evalute outcomes

How does the Decision Tree algorithm work?

Ans:- In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node the algorithm again compare the attribute value with the other sub-nodes and move further. It continues the process until it reaches the ~~node~~ leaf node of the tree. The complete process can be better understood using the below algorithm.

Step-1 - Begin the tree with the root node. Which contains the complete dataset.

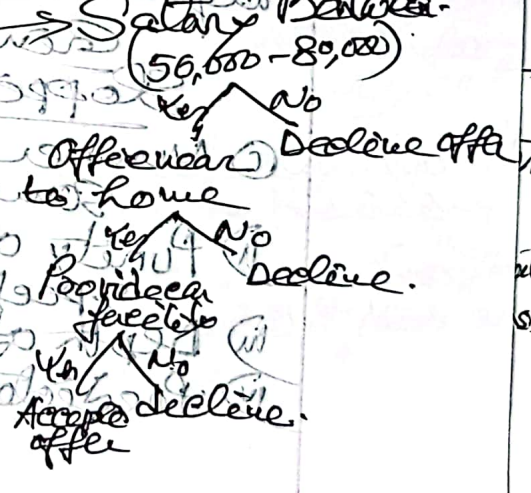
Step-2 - Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3 - Divide the S into subset that contains possible values for the best attribute.

Step-4 - Generate the decision tree node which contains the best attribute.

Step-5 - Recursively make new decision trees using the subset created in step-3

Examples:-



Advantages of Decision Tree:-

1. It is simple to understand.
2. It can be use very useful for solving decision-related problem.
3. It helps to think about all the possible outcomes for problem.
4. There is less requirement of data cleaning compared to other algorithm.

Attributes Selection measures:- There are two popular techniques for Attribute selection measure (ASM), which are -

i) Information Gain
ii) Gini Index.

Information Gain:- Information Gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It can be calculated as below formula -

Information Gain = Entropy(S) - [(Weighted Avg) * Entropy(each feature)]

Entropy:- Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as

Entropy(S) = $-P_{yes}\log_2 P_{yes} - P_{no}\log_2 P_{no}$

Gini Index:- Gini Index is a measure of impurity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm.

Gini Index = $1 - \sum P_j^2$

Stopping Criteria:-

i) Number of cases in the node is less than some pre-specified limit.
ii) Purity of the node is more than some pre-specified limit.
iii) Depth of the node is more than some limit.
iv) Predictor value for all records are identical

÷ Classification :- Lect 05.

Definition :- Classification is the process of identifying and grouping object or ideas into predetermined categories.

Input → Classification model → Output.

Attribute set ⇒ Classification model ⇒ Class label

Classification Task :- Binary Classification can be categorized into one of two classes.

i) Binary classification (No. of class is 2).
ii) Multiclass classification (No. of class is longer than 2).

A classification model serves two important roles in data mining. It must provide accurate prediction with a fast response time.

i) Predictive model (To classify previously unlabeled instances)
ii) Descriptive model (To identify the characteristics that distinguish instances from different classes).

It is particularly used for practical application, such as medical diagnosis. It help to determine characteristics.

Training set :- A classifier model is created using a given set of instances, known as the training set.

Learning Algorithm :- The systematic approach for learning a classification model given a training set is known as a learning algorithm.

Induction :- The process of using learning algorithm to build a classification model from the training data is known as induction.

Deduction :- The process of applying a classification model on unseen test instances to predict their class labels is known as deduction.

Classification Techniques:- A classification technique refers to a general approach to classification:-

(A) Base Classification
   i) Decision Tree Based Methods
   ii) Rule Based Methods
   iii) Nearest Neighbor
   iv) Naive Bayes and Bayesian Belief Network.
   v) Support Vector Machines.
   vi) Neural & Deep Neural Network.

(B) Ensemble Classifier
   i) Boosting
   ii) Bagging.
   iii) Random Forests.

Confusion Matrix:- The performance of a classifier/model can be evaluated by comparing the predicted labels against the true labels of instances. This information can be summarized in a table is called a confusion matrix. A confusion matrix is nothing but a table with two dimensions
   i) Actual and
   ii) Predicted

And furthermore, both the dimensions have "True Positive (TP)", True Negative (TN), false Positive (FP) and False Negative (FN).

	1	0
1	TP	FP
0	FN	TN

TP- If is the case when both actual class & predicted class of data point is 1.

TN- If is the case when both actual class & predicted class of data point is 0.

FP- If is the case when actual class of data point is 0 & predicted class of data point is 1.

FN- If is the case when actual class of data point is 1 & predicted class of data point is 0.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN}$$

$$\text{Error Rate} = \frac{\text{No. of wrong predictions}}{\text{Total No. of predictions}}$$

$$\text{Precession} = \frac{TP}{TP + FP}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precession} \times \text{Recall}}{\text{Precession} + \text{Recall}}$$

F1 Score is the weighted average of Precession & Recall

$$\text{Recall or Sensitivity} = \frac{TP}{TP + FN}$$

Decision Tree Induction:

Many Algorithms

i) Hunt's Algorithm

ii) CART

iii) ID3, C4.5

iv) SLIQ, SPRINT

Design Issues of Decision Tree Induction

> How should training records be split?
- Method for expressing test condition
  - depending on attribute types.
- Measure for evaluating the goodness of a test condition

> How should the splitting procedure stop?
- Stop splitting if all the records belong to the same class or have identical attribute values
- Early termination

Test Condition for Nominal/Ordinal/Continuous Attributes

(A) Multiway split: Use many partitions as distinct values.
(B) Binary split: Divides values into two subsets.

Measures of Node Impurity

(1) Gini Index $= 1 - \sum_{i=0}^{c-1} P_i(t)^2$ — where $P(t)$ is the frequency of class i at node t, i.e. $\frac{\text{total number of class } i}{\text{total number of records}}$

(2) Entropy $H(S) = -\sum_{i=0}^{c-1} P_i(t) \log_2 P_i(t) = -P_\oplus \log_2 P_\oplus - P_\ominus \log_2 P_\ominus$

(3) Misclassification Error $= 1 - \max[P_i(t)]$

i) Entropy is zero (minimum) when all the examples same class.
ii) Entropy is 1 (maximum) when n i are equal no. of $+$ & $-$ examples

(4) Average Entropy (I)
(5) Information Gain (IG)

What are the steps in ID3 algorithm?

The steps in ID3 algorithm are as follows:

1. Calculate entropy for dataset.
2. For each attribute/feature.
2.1. Calculate entropy for all its categorical values.
2.2. Calculate information gain for the feature.
3. Find the feature with maximum information gain.
4. Repeat it until we get the desired tree.

Use ID3 algorithm on a data We'll discuss it here mathematically and later see it's implementation in Python. So, Let's take an example to make it more clear.

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Here,dataset is of binary classes(yes and no), where 9 out of 14 are "yes" and 5 out of 14 are "no".

Complete entropy of dataset is:

$$H(S) = -p(yes) * \log2(p(yes)) - p(no) * \log2(p(no))$$
$$= -(9/14) * \log2(9/14) - (5/14) * \log2(5/14) = -(-0.41) - (-0.53) = 0.94$$

For each attribute of the dataset, let's follow the step-2 of pseudocode : -First Attribute - Outlook

Categorical values - sunny, overcast and rain

$H(Outlook=sunny) = -(2/5)*\log(2/5) - (3/5)*\log(3/5) = 0.971$

$H(Outlook=rain) = -(3/5)*\log(3/5) - (2/5)*\log(2/5) = 0.971$

$H(Outlook=overcast) = -(4/4)*\log(4/4) - 0 = 0$

Average Entropy Information for Outlook -

$I(Outlook) = p(sunny) * H(Outlook=sunny) + p(rain) * H(Outlook=rain) + p(overcast) * H(Outlook=overcast)$

$= (5/14)*0.971 + (5/14)*0.971 + (4/14)*0 = 0.693$

$Information\ Gain = H(S) - I(Outlook) = 0.94 - 0.693 = 0.247$

I suggest you to do the same calculations for all the remaining attributes without scrolling down....

Second Attribute – Temperature, Categorical values - hot, mild, cool

$H(Temperature=hot) = -(2/4)*\log(2/4) - (2/4)*\log(2/4) = 1$

$H(Temperature=cool) = -(3/4)*\log(3/4) - (1/4)*\log(1/4) = 0.811$

$H(Temperature=mild) = -(4/6)*\log(4/6) - (2/6)*\log(2/6) = 0.9179$

Average Entropy Information for Temperature -

$I(Temperature) = p(hot)*H(Temperature=hot) + p(mild)*H(Temperature=mild) + p(cool)*H(Temperature=cool) = (4/14)*1 + (6/14)*0.9179 + (4/14)*0.811 = 0.9108$

Information Gain = H(S) - I(Temperature) = 0.94 - 0.9108 = 0.0292

Third Attribute – Humidity, Categorical values - high, normal

H(Humidity=high) = -(3/7)*log(3/7)-(4/7)*log(4/7) = 0.983

H(Humidity=normal) = -(6/7)*log(6/7)-(1/7)*log(1/7) = 0.591

Average Entropy Information for Humidity -

I(Humidity) = p(high)*H(Humidity=high) + p(normal)*H(Humidity=normal)

= (7/14)*0.983 + (7/14)*0.591 = 0.787

Information Gain = H(S) - I(Humidity) = 0.94 - 0.787 = 0.153

Fourth Attribute - Wind

Categorical values - weak, strong

H(Wind=weak) = -(6/8)*log(6/8)-(2/8)*log(2/8) = 0.811

H(Wind=strong) = -(3/6)*log(3/6)-(3/6)*log(3/6) = 1

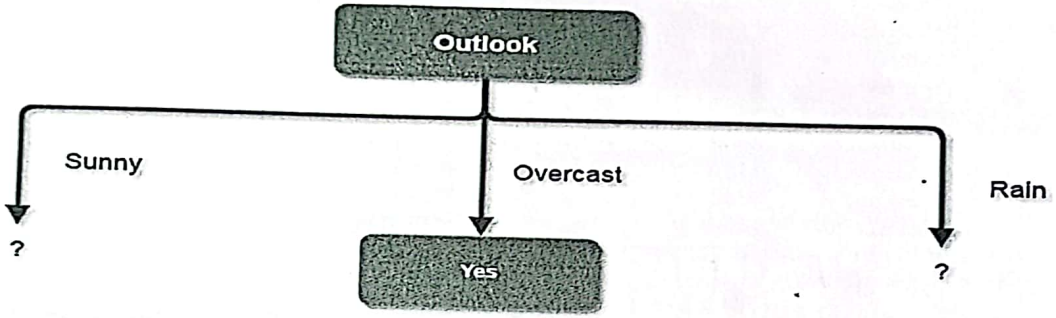
Average Entropy Information for Wind -

I(Wind) = p(weak)*H(Wind=weak) + p(strong)*H(Wind=strong)

= (8/14)*0.811 + (6/14)*1 = 0.892

Information Gain = H(S) - I(Wind) = 0.94 - 0.892 = 0.048

Here, the attribute with maximum information gain is Outlook. So, the decision tree built so far -



Here, when Outlook == overcast, it is of pure class(Yes).

Now, we have to repeat same procedure for the data with rows consist of Outlook value as Sunny and then for Outlook value as Rain.

I again recommend you to perform further calculations and then cross check by scrolling down...

Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [1, 2, 8, 9, 11]}.

Complete entropy of Sunny is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (2/5) * log2(2/5) - (3/5) * log2(3/5) = 0.971

First Attribute - Temperature

Categorical values - hot, mild, cool

H(Sunny, Temperature=hot) = -0-(2/2)*log(2/2) = 0

H(Sunny, Temperature=cool) = -(1)*log(1)- 0 = 0

H(Sunny, Temperature=mild) = -(1/2)*log(1/2)-(1/2)*log(1/2) = 1

Average Entropy Information for Temperature -

I(Sunny, Temperature) = p(Sunny, hot)*H(Sunny, Temperature=hot) + p(Sunny, mild)*H(Sunny, Temperature=mild) + p(Sunny, cool)*H(Sunny, Temperature=cool)= (2/5)*0 + (1/5)*0 + (2/5)*1 = 0.4

Information Gain = H(Sunny) - I(Sunny, Temperature) = 0.971 - 0.4 = 0.571

Second Attribute - Humidity

Categorical values - high, normal

H(Sunny, Humidity=high) = - 0 - (3/3) *log(3/3) = 0

H(Sunny, Humidity=normal) = -(2/2) *log(2/2)-0 = 0

Average Entropy Information for Humidity -

I(Sunny, Humidity) = p(Sunny, high) *H(Sunny, Humidity=high) + p(Sunny, normal) *H(Sunny, Humidity=normal) = (3/5) *0 + (2/5) *0 = 0

Information Gain = H(Sunny) - I(Sunny, Humidity) = 0.971 - 0 = 0.971

Third Attribute - Wind

Categorical values - weak, strong

H(Sunny, Wind=weak) = -(1/3) *log(1/3) - (2/3) *log(2/3) = 0.918

H(Sunny, Wind=strong) = -(1/2) *log(1/2) - (1/2) *log(1/2) = 1

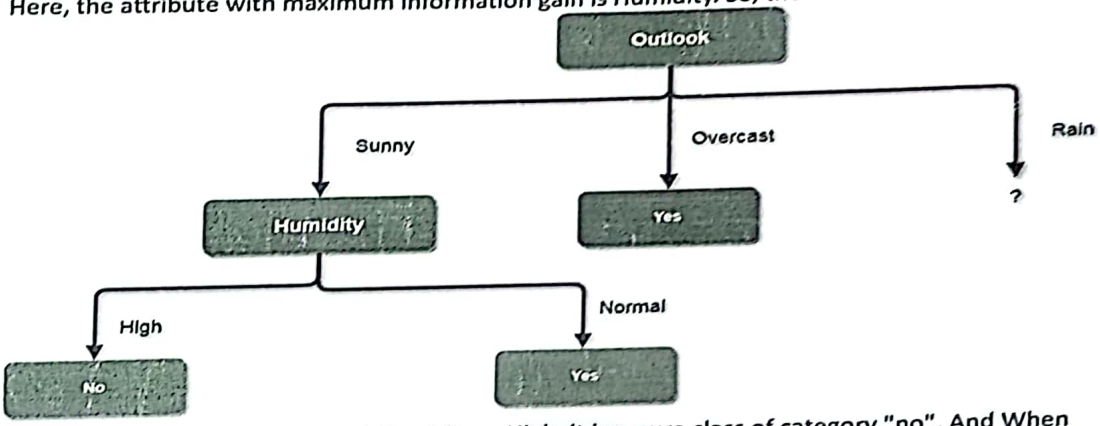
Average Entropy Information for Wind -

I(Sunny, Wind) = p(Sunny, weak) *H(Sunny, Wind=weak) + p(Sunny, strong) *H(Sunny, Wind=strong)

= (3/5) *0.918 + (2/5) *1 = 0.9508

Information Gain = H(Sunny) - I(Sunny, Wind) = 0.971 - 0.9508 = 0.0202

Here, the attribute with maximum information gain is Humidity. So, the decision tree built so far -



Here, when Outlook = Sunny and Humidity = High, it is a pure class of category "no". And When Outlook = Sunny and Humidity = Normal, it is again a pure class of category "yes". Therefore, we don't need to do further calculations.

Now, finding the best attribute for splitting the data with Outlook=Sunny values{ Dataset rows = [4, 5, 6, 10, 14]}.

Complete entropy of Rain is -

H(S) = - p(yes) * log2(p(yes)) - p(no) * log2(p(no))

= - (3/5) * log(3/5) - (2/5) * log(2/5) = 0.971

First Attribute - Temperature

Categorical values - mild, cool

H(Rain, Temperature=cool) = -(1/2) *log(1/2) - (1/2) *log(1/2) = 1

H(Rain, Temperature=mild) = -(2/3) *log(2/3) - (1/3) *log(1/3) = 0.918

Average Entropy Information for Temperature -

I(Rain, Temperature) = p(Rain, mild) *H(Rain, Temperature=mild) + p(Rain, cool) *H(Rain, Temperature=cool) = (2/5) *1 + (3/5) *0.918 = 0.9508

Information Gain = H(Rain) - I(Rain, Temperature) = 0.971 - 0.9508 = 0.0202

Second Attribute - Wind

Categorical values - weak, strong

H(Wind=weak) = -(3/3)*log(3/3)-0 = 0

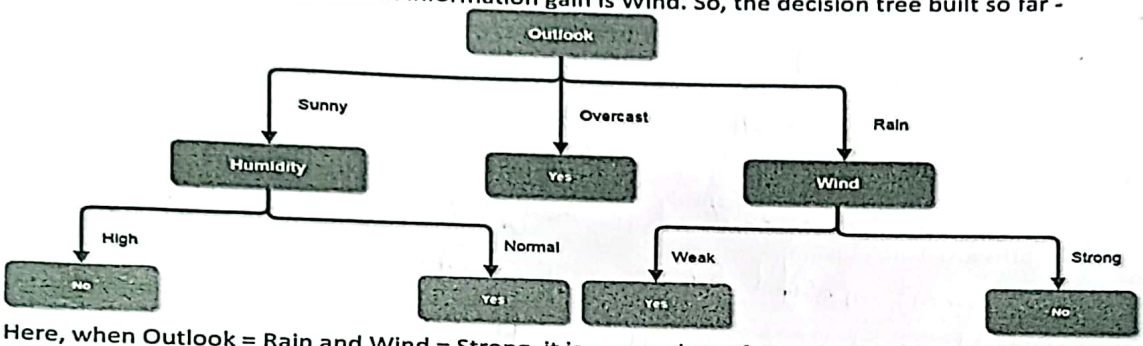
H(Wind=strong) = 0-(2/2)*log(2/2) = 0

Average Entropy Information for Wind -

I(Wind) = p(Rain, weak)*H(Rain, Wind=weak) + p(Rain, strong)*H(Rain, Wind=strong) = (3/5)*0 + (2/5)*0 = 0

Information Gain = H(Rain) - I(Rain, Wind) = 0.971 - 0 = 0.971

Here, the attribute with maximum information gain is Wind. So, the decision tree built so far -



Here, when Outlook = Rain and Wind = Strong, it is a pure class of category "no". And When Outlook = Rain and Wind = Weak, it is again a pure class of category "yes".

And this is our final desired tree for the given dataset.

What are the characteristics of ID3 algorithm?

Finally, I am concluding with Characteristics of ID3.

Characteristics of ID3 Algorithm are as follows:

ID3 uses a greedy approach that's why it does not guarantee an optimal solution; it can get stuck in local optimums.

ID3 can overfit to the training data (to avoid overfitting, smaller decision trees should be preferred over larger ones).

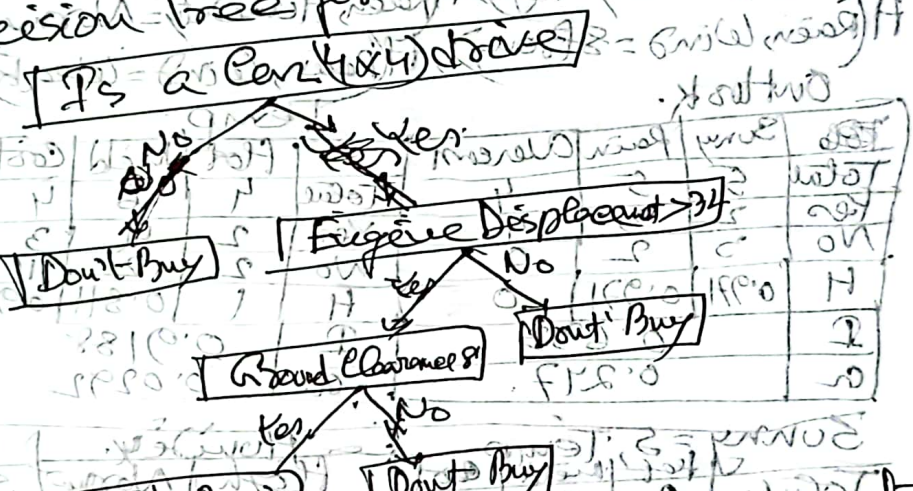
This algorithm usually produces small trees, but it does not always produce the smallest possible tree.

ID3 is harder to use on continuous data (if the values of any given attribute is continuous, then there are many more places to split the data on this attribute, and searching for the best value to split by can be time consuming).

With this technical article at OpenGenus, you must have a strong idea of ID3 Algorithm and how to use ID3 Algorithm to build a Decision Tree to predict the weather.

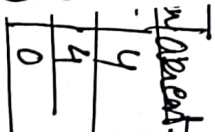
Decision Tree:- Decision Tree is a popular machine learning algorithm mainly used for classification. Concept of entropy and information gain are required to apply decision tree for a data set. It is used for non-linear classification and regression model. The main idea of a decision tree is to split our data set into many smaller and smaller datasets in many steps based on features until you reach a node which is small enough to contain data points that fall under the same label.

Decision Tree for Classification and Prediction



Entropy:- Entropy is a measure of the amount of uncertainty in the dataset. Mathematical representation of Entropy is $H(S) = \sum_{i=1} -P_i \log_2 P_i$

Information Gain IG(A):- Information Gain IG(A) tells us how much uncertainty in S was reduced after splitting set S on attribute A. $IG(A,S) = H(S) - \sum_{t \in T} P(t) H(t)$.



PD-3.

(A) Complete Entropy of H(S).

1st Attributes → H(Outlook=Sunny), H(Outlook=Rain), H(Outlook=Overcast), I, G.
2nd " → H(Temp=Hot), H(Temp=Mild), H(Temp=Cool), I, G.
3rd " → H(Humidity=High), H(Humidity=Normal), I, G.
4th " → H(Wind=Weak), H(Wind=Strong), I, G.

(B) Split Tree: Outlook (Sunny), Entropy

H(Sunny, Temp=Hot), H(Sunny, Temp=Mild), H(Sunny, Temp=Cool), I, G.
H(Sunny, Humidity=High), H(Sunny, Humidity=Normal), I, G.
H(Sunny, Wind=Weak), H(Sunny, Wind=Strong), I, G.

(C) Split Tree: Outlook (Rain), Entropy

H(Rain, Temp=Hot), H(Rain, Temp=Mild), H(Rain, Temp=Cool), I, G.
H(Rain, Humidity=High), H(Rain, Humidity=Normal), I, G
H(Rain, Wind=Strong), H(Rain, Wind=Weak), I, G

Outlook.

Outlook	Sunny	Rain	Overcast
Total	5	5	4
Yes	2	3	4
No	3	2	0
H	0.971	0.971	0
I	0.693		
G	0.247		

Temp.

Temp	Hot	Mild	Cool
Total	4	6	4
Yes	2	4	3
No	2	2	1
H	1	0.811	0.917
I	0.9188		
G	0.0292		

Humidity.

Humidity	High	Normal
Total	7	7
Yes	3	6
No	4	1
H	0.98	0.59
I	0.787	
G	0.153	

Wind.

Wind	Strong	Weak
Total	8	6
Yes	6	3
No	2	3
H	0.81	
I	0.89	
G	0.04	

Sunny = S, Temp.

Temp	Hot	Mild	Cool
Total	2	2	1
Yes	0	1	1
No	2	1	0

Humidity.

Humidity	High	Normal
	3	2
	0	2
	3	0

Wind.

Wind	Strong	Weak
	2	3
	1	1
	1	2

Rain = S. Temp.

Temp	Hot	Mild	Cool
Total	0	3	2
Yes	0	2	1
No	0	1	1

Humidity

Humidity	High	Normal
	1	2
	1	1

Wind

Wind	Strong	Weak
	0	3
	2	0

Continue...

$H(S) = -P_{yes}\log_2 P_{yes} - P_{no}\log_2 P_{no}$.

	Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	D1	Sunny	Hot	High	Weak	No
2	D2	Sunny	Hot	High	Strong	No
3	D3	Overcast	Hot	High	Weak	Yes ✓
4	D4	Rain	Mild	High	Weak	Yes ✓
5	D5	Rain	Cool	Normal	Weak	Yes ✓
6	D6	Rain	Cool	Normal	Strong	No
7	D7	Overcast	Cool	Normal	Strong	Yes ✓
8	D8	Sunny	Mild	High	Weak	No
9	D9	Sunny	Cool	Normal	Weak	Yes ✓
10	D10	Rain	Mild	Normal	Weak	Yes ✓
11	D11	Sunny	Mild	Normal	Strong	Yes ✓
12	D12	Overcast	Mild	High	Strong	Yes ✓
13	D13	Overcast	Hot	Normal	Weak	Yes ✓
14	D14	Rain	Mild	High	Strong	No

Total = 14
Yes = 09.
No = 05

Categorical Value:-

Outlook = Sunny	Rain	Overcast
Total 5	5	4
Y = 2	3	4
N = 3	2	0

Round-1:- Outlook: $H(S) = -9/14 \log_2 9/14 - 5/14 \log_2 5/14 = 0.94$

Outlook				Temperature				Humedity			Wind.		
~~Total~~	Sunny	Rain	overlook		Hot	mild	Cool.		High	Normal		Week	Strong
Total	5	5	4	Total	4	6	4	Total	7	7	Total=	8	6
Yes	2	3	4	Yes	2	4	3	Yes	3	6	Yes=	6	3
No	3	2	0	No	2	2	1	No	4	1	No =	2	3
H	0.971	0.971	0	H	1	0.811	0.9179	H	0.983	0.591	H	0.811	1
I	0.693			I	0.9188			I	0.787		I	0.892	
G.	0.247			G	0.0292			G	01153		G.	0.048	

Round-2:- Sunny Values:- $H(S) = -2/5 \log_2 2/5 - 3/5 \log_2 3/5 = 0.971$

Round-3:- Rain Values:-

Round-4:- Wind Values:-

Roll-2321337.
Md. Monir Ullah

Ans. to the qes No-1.

Ans:. Given that

TP = 6.
TN = 2
FP = 1
FN = 1

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FP}.$$

$$= \frac{6+2}{6+2+1+1}$$

$$= 0.8$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{6}{6+1} = \frac{6}{7} \; 0.857$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{6}{6+1} = \frac{6}{7} = 0.857$$

$$F_1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$= 2 \times \frac{0.857 \times 0.857}{0.857 + 0.857} = 2 \times \frac{0.7344}{1.714}$$

$$= \frac{1.4688}{1.714} = 0.856$$

1) To illustrate, suppose S is a collection of 14 examples of some boolean concept, including 9 positive and 5 negative examples (we adopt notation [9+, 5−] to summarize such a sample of data). Then the entropy of S relative to this boolean classification is

Data Mining.

08/09/23.

Association Analysis:

Association Rules Mining :- { Support, Confidence

Support Count (σ), Eg σ({Milk, Bread, Diapers}) = 2

Support, Eg s({Milk, Bread, Diapers}) = 2/5

Min sup threshold

Rule Evaluation Metrics:- → Support (s), → Confidence (c).

Rule ... Accepted ... Support ... Confidence.

TID	Item
1	Bread, Milk.
2	Bread, Diaper, Beer, Eggs.
3	Milk, Diaper, Beer, Coke.
4	Bread, Milk, Diaper, Beer.
5	Bread, Milk, Diaper, Coke.

Brute-force approach:-

$R = 3^d - 2^{d+1} + 1$

... practical acceptable ...

... Acceptable ... Support ...

$s = \frac{\sigma(\text{Milk, Diaper, Bread})}{|T|} = \frac{2}{5} = 0.4$

$\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}$

Apriori Algorithm: ... Support based pruning. $c = \alpha$

Threshold 50% ...

Frequent Item

Min Support = Total No. of Items

= 3

Min Support threshold = 60%

Example-9:

Q. Support threshold = 50%, Confidence = 60%

Solution:-

Support threshold 50%
= 6 items × 50% = 3.

Step-1:- k=1.

I_5 deleted, because min threshold is 3 but $I_5 = 2$.

Step-2:- k=2

Table-1.

Transaction	List of Items
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

Table-2.

Item	Count
I_1	4
I_2	5
I_3	4
I_4	3
I_5	2

Table-3

Item	Count
I_1	4
I_2	5
I_3	4
I_4	3

Table-4

Item	Count
I_1, I_2	4
I_1, I_3	3
I_1, I_4	2
I_2, I_3	3
I_2, I_4	3
I_3, I_4	2

Table-5.

Item	Count
I_1, I_2	4
I_1, I_3	3
I_2, I_3	4
I_2, I_4	3

Table-6.

Item	Count
I_1, I_2, I_3	3
I_1, I_2, I_4	2
I_1, I_3, I_4	1
I_2, I_3, I_4	2

Only I_1, I_2, I_3 is frequent.

Generate Association Rule:-

$I_1, I_2 \Rightarrow I_3$
$I_1, I_3 \Rightarrow I_2$
$I_2, I_3 \Rightarrow I_1$

$I_1, I_2, I_3 = 3$.

Support = $\frac{\text{no. of item A}}{\text{Total no. of item}}$

Confidence $(A \to B) = \frac{\text{Support}(A \cup B)}{\text{Support } A}$.

$I_1, I_2 \Rightarrow$
I_2, I_1
I_3, I_2

$I_1 \Rightarrow I_2, I_3 = \frac{3}{4} = 75\%$
$I_2 \Rightarrow I_3, I_1 = \frac{3}{5} = 60\%$
$I_3 \Rightarrow I_1, I_2 = \frac{3}{4} = 75\%$

Confidence = $\frac{\text{Support}(I_1, I_2, I_3)}{\text{Support}(I_1, I_2)} = \frac{3}{4} \Rightarrow 75\%$

$= \frac{3}{3} = 100\%$.

$= \frac{3}{4} = 75\%$.

This shows that, all the above association rules are strong if minimum confidence threshold is 60%.

Example:-02

Milk, Bread, Butter = 3

Item	Support (Frequency)
Milk	8
Bread	7
Sugar	5
Butter	7

TID	Items
T1	Milk, Bread
T2	Bread, Sugar
T3	Bread, Butter
T4	Milk, Bread, Sugar
T5	Milk, Bread, Butter
T6	Milk, Bread, Butter
T7	Milk, Sugar
T8	Milk, Sugar
T9	Sugar, Butter
T10	Milk, Sugar, Butter
T11	Milk, Bread, Butter

Milk, Bread	Butter, C=60%
Bread, Butter	Milk, C=100%
Milk, Butter	Bread, C=60%

Item	Support
Milk, Bread	5
Milk, Sugar	3
Milk, Butter	5
Bread, Sugar	2
Bread, Butter	3
Sugar, Butter	2

Based on Association rules mentioned in the above table, we can recommended products to the customer to optimize product placement in retail stores

Milk, Bread, Sugar	1
Milk, Bread, Butter	3
Milk, Sugar, Butter	1

Example:-03

Step-1

Product	Frequency
Rice (R)	4
Pulse (P)	5
Oil (O)	4
Milk (M)	4

Step-2: Threshold = 50%

Item	Frequency
RP	4
RO	3
RM	2
PO	4
PM	3
OM	2

Tran. ID	Rice	Pulse	Oil	Milk	Apple
T1	1	1	1	0	0
T2	0	1	1	1	0
T3	0	0	0	1	1
T4	1	1	0	1	0
T5	1	1	1	0	1
T6	1	1	1	1	1

Step-3: We get RP, RO, PO & PM.

Step-4: RP, RO & RPO
PO, PM & POM.

Step-5:

Itemset	Frequency
RPO	4
POM	3

Example 04:

Min Support Count = 2
Table-1 c = 60%

TID	Items.
T1	I_1, I_2, I_5
T2	I_2, I_4
T3	I_2, I_3
T4	I_1, I_2, I_4
T5	I_1, I_3
T6	I_2, I_3
T7	I_1, I_3
T8	I_1, I_2, I_3, I_5
T9	I_1, I_2, I_3

Step-1 : K = 1.

K = 1 (Table-2).

Itemset	Sup count
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

Step-2.

K = 2 (Table-3).

Itemset	Sup count
I_1, I_2	4
I_1, I_3	4
I_1, I_4	1
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2
I_3, I_4	0
I_3, I_5	1
I_4, I_5	0

Itemset	Sup count
I_1, I_2	4
I_1, I_3	4
I_1, I_5	2
I_2, I_3	4
I_2, I_4	2
I_2, I_5	2
I_2, I_5	2

Step3, K=3.

Itemset	Sup-count
I_1, I_2, I_3	2
I_1, I_2, I_5	2

We stop here, because no frequent itemset are found further.

Confidence:- We will show the rule generation:-

$I_1, I_2 \Rightarrow I_3 \Rightarrow \frac{2}{4} \times 100 = 50\%$

$I_1, I_3 \Rightarrow I_2 \Rightarrow \frac{2}{4} \times 100 = 50\%$

$I_2, I_3 \Rightarrow I_1 \Rightarrow \frac{2}{4} \times 100 = 50\%$

$I_1 \Rightarrow I_2, I_3 \Rightarrow \frac{2}{6} \times 100 = 33\%$

$I_2 \Rightarrow I_1, I_3 \Rightarrow \frac{2}{7} \times 100 = 27\%$

$I_3 \Rightarrow I_1, I_2 \Rightarrow \frac{2}{6} \times 100 = 33\%$

So if minimum confidence is 50%. The first 3 rules can be considered as strong association.

• Example-05:

Support Threshold = 60%, so Min Support = 60% × 5 = 3

(Table-1)

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

⇒ K=1 (Table-2):

Item	Support Count
Bread	4
Coke	2 ✗
Milk	4
Beer	3
Diaper	4
Eggs	1 ✗

Table-3 -

Item	Support Count
Bread	4
Milk	4
Beer	3
Diaper	4

Table-4 (K=2).

Itemset	Count
Bread, Milk	4
Bread, Beer	2 ✗
Bread, Diaper	3
Milk, Beer	2 ✗
Milk, Diaper	3
Beer, Diaper	3

Table-5:

Itemset	Count
Bread, Milk	4
Bread, Diapa	3
Milk, Diaper	3
Beer, Diaper	3

Table-6 (K=3).

Itemset	Count
Beer, Diaper, Milk	2
Beer, Bread, Diaper	2

Example-06:- Minimum Support is 60%; Min support = 5 × 60% = 3

TID	Itemset
T1	6, 7, 8, 5, 4, 10
T2	3, 8, 7, 5, 4, 10
T3	6, 1, 5, 4
T4	6, 9, 2, 5, 10
T5	2, 8, 5, 4

K=1

Item	Supp
1	1 ✗
2	2 ✗
3	1 ✗
4	4
5	5
6	3
7	2 ✗
8	4
9	1 ✗
10	3

Item	Supp.
4	4
5	5
6	3
8	4
10	3

K=2

Item	Sup.
4,5	4
4,6	2 ✗
4,8	3
4,10	2 ✗
5,6	3
5,8	3
5,10	3
6,8	1 ✗
6,10	2 ✗
8,10	2 ✗

4,5	4
4,8	3
5,6	3
5,8	3
5,10	3

K=3

Item	Sup
4,5,8	3
4,5,6	2 ✗
4,5,10	2 ✗
5,6,8	1 ✗
5,6,10	2 ✗
5,8,10	2 ✗

Item	Supp
4,5,8	3

Support :- Support is an indicator of how frequently the items appear in the data.

Confidence :- Confidence indicates the number of times the if-then statements are found true.

Lift :- A third metric called lift, can be used to compare confidence with expected confidence or how many times an if-then statement is expected to be found true.

$$\text{Support}(X) = \frac{\text{Number of transactions containing } (X)}{\text{Total Number of Transaction}}$$ [X is the itemset]

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{Number of transaction containing } X \text{ \& } Y}{\text{Number of transaction containing } X}$$

Support	Confidence
① It is measure of the number of times an itemsets appears in a dataset.	① Confidence is a measure of the likelihood that an itemset will appear if another itemset appears.
② Support is calculated by dividing the number of transaction containing of item by the total number of transaction	② Confidence is calculated by dividing the number of transaction containing both itemsets by the number of transaction containing the first itemset
③ Support is used to identify itemsets that occur frequently in the dataset.	③ Confidence is used to evaluate the strength of a rule.
④ Support is often used with a threshold to identify itemset that occur frequently enough to be of interest.	④ Confidence is interpreted as the percentage of transaction in which the second itemset appears given that the first itemset appears.
⑤	⑤