**Vidyavahini First Grade College**

**Fifth Semester B.C.A Degree Examination(CBCS Scheme), November 2018**

Computer Science

**DATA MINING-Model Question Paper-2**

**[Time: 3 Hours] [Max marks: 90]**

**Instructions to Candidates: Answer all the Sections**

**SECTION-A**

Answer any **TEN** questions**: [10 x 1 =10]**

1.What is knowledge in DM?

## Information read, heard or seen and understood and integrated is called knowledge in DM.

2. What is noisy data.?

Unstructured, incomplete, irrelevant, redundant data which result in programming errors, improper analysis sometimes leads to hardware failure are called noisy data.

3.What is data mart?

It is the subset of data warehouse contain only small slices of data warehouse. Eg: Data pertaining to particular department. Two types of data mart. Dependent and Independent.

4. Define logistic regression ?

***Logistic Regression*** is basically a supervised classification algorithm. Logistic Regression is one of the most used Machine Learning algorithms for binary classification. It gives you a discrete binary outcome between 0 and 1.

5.What is supervised learning.

**Supervised learning is a machine learning technique where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. Y = f(X)**

6.Expand DBSCAN?

Density Based Spatial Clustering of Applications with Noise

7.What is Pruning?

**The process of trimming the decision tree.ie, removing the branches which are not needed for the classification.**

8.What is Dendrogram?

A Dendrogram is a type of tree diagram showing hierarchical clustering .It represents the relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples and they can represent any type of grouped data.

9.Mention any 2 clustering softwares?

* ClustanGraphics3
* CMSR Data Miner
* IBM SPSS Modeler

10.What is HDFS in Big Data?

# HDFS (Hadoop Distributed File System): HDFS takes care of storage part of Hadoop applications. Map Reduce applications consume data from HDFS. HDFS creates multiple replicas of data blocks and distributes them on compute nodes in cluster. This distribution enables reliable and extremely rapid computations.

11. What is Web crawler?

It is also known as **spider** or **bots.** It is a software component that traverses the web to gather information

12.Give an example of unstructured data?

e-mail messages,word processing documents

**SECTION-B**

Answer any **FIVE** questions**: [5 x 3 = 15]**

13. Explain briefly applications of Data mining in any 3 fields.

|  |  |
| --- | --- |
| **Communications** | **Data mining techniques are used in communication sector to predict customer behavior to offer highly targeted and relevant campaigns.** |
| **Insurance** | **Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing customers.** |
| **Education** | **Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.** |

14. What is Classification ? mention its importance in DM?

Classification is a predictive Data mining task which maps the data into predefined groups or classes. **Classification can be used in direct marketing that** is to reduce marketing costs by targeting a set of customers who are likely to buy a new product.Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past.

Following are the examples of cases where the data analysis task is Classification −

* A bank loan officer wants to analyze the data in order to know which customers (loan applicant) are risky or which are safe.
* A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

**Applications of classification data mining task.**

1. Internet traffic interception - certain governments (possibly from the middle east) would like to restrict certain categories of web pages. For example, due to religious restrictions, certain movie pages may be restricted/censored. This is a clear application of classification.
2. Video classification - as and when you upload a video on youtube, the video has to be classified into appropriate categories and meta-data added to it .
3. Voice and image classification in for google assistance and other applications.
4. Weather Predictions
5. Resume sorting

15. What are the requirements of Clustering in DM?

The following points throw light on why clustering is required in data mining −

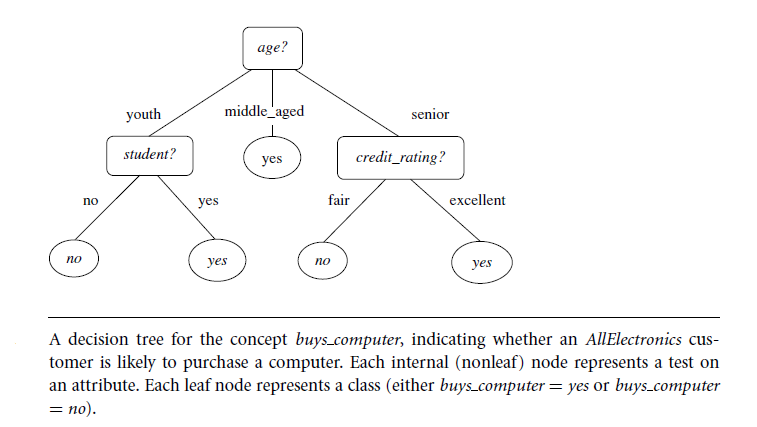
* **Scalability** − we need highly scalable clustering algorithms to deal with large databases.
* **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
* **Discovery of clusters with attribute shape** −The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
* **High dimensionality** −The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
* **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

**Interpretability** −The clustering results should be interpretable, comprehensible, and usable

16. What is Decision Tree? Explain with an example

A **decision tree** is a flowchart-like tree structure, where each **internal node** (nonleaf node) denotes a test on an attribute, each **branch** represents an outcome of the test, and each **leaf node** (or *terminal node*) holds a class label. The topmost node in a tree is the **root** node.

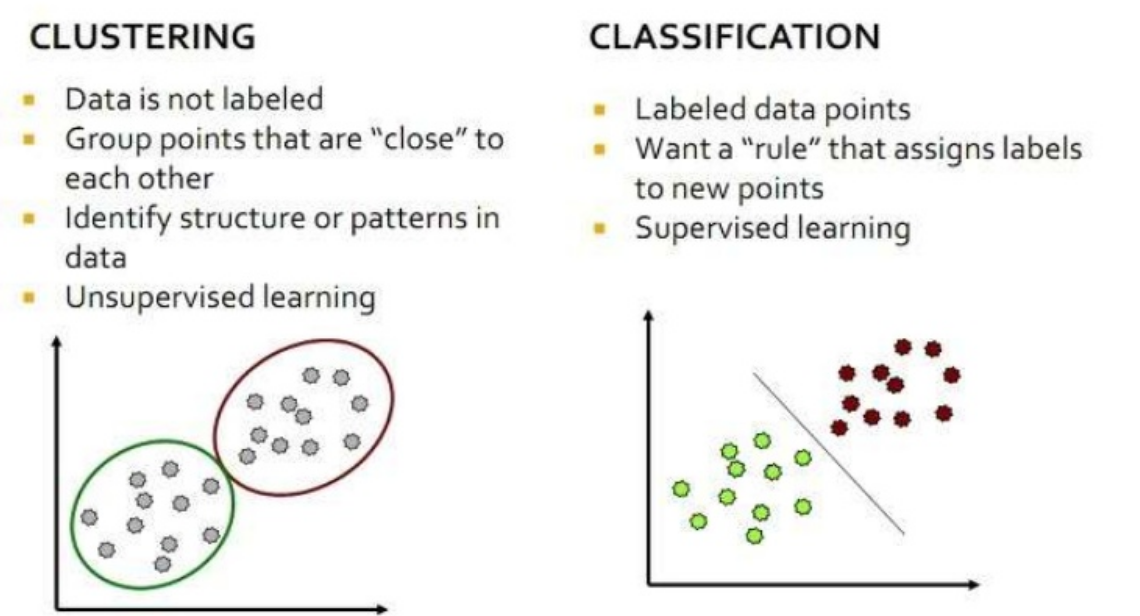
A typical decision tree for buying computer is shown in Figure:



It represents the concept *buys computer*, that is, it predicts whether a customer at *AllElectronics* is

likely to purchase a computer. Each leaf node represents a class either buys\_computer = yes or buys\_computer=no.

17. Differentiate between Clustering and Classification tasks of DM.



18. What is web data mining? mention the types .

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services.

Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content and usage data.

Web mining is classified into

1. Web content –text, image, records, etc.
2. Web structure –hyperlinks, tags, etc.
3. Web usage –http logs, app server logs, etc.

19. Briefly explain the importance of big data in an existing world?

The importance of big data is how you utilize the data which you own. Data can be fetched from any source and analyze it to solve that enable us in terms of

1) Cost reductions

2) Time reductions,

3) New product development and optimized offerings, and

4) Smart decision making.

Combination of big data with high-powered analytics, you can have great impact on your business strategy such as:

* Finding the root cause of failures, issues and defects in real time operations.
* Generating coupons at the point of sale seeing the customer’s habit of buying goods.
* Recalculating entire risk portfolios in just minutes.

Detecting fraudulent behavior before it affects and risks your organization

**SECTION-C**

Answer any **SIX** questions**: [6 x 5 = 30]**

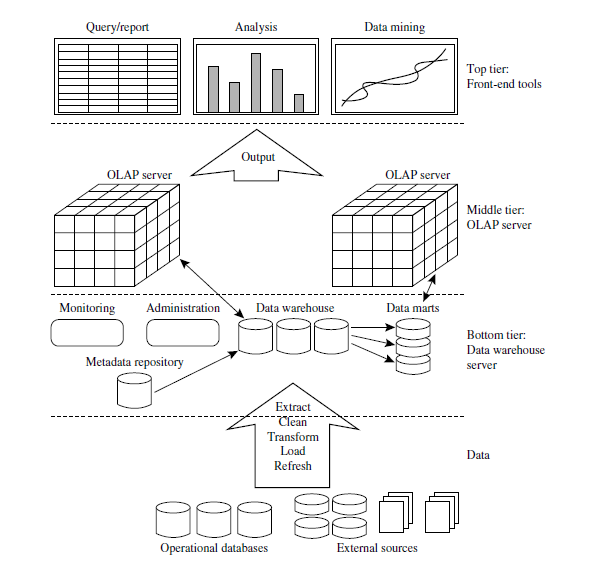
20. Explain 3-tier Architecture of Data ware house with neat diagram .

The Data ware house architecture mainly classified into three tiers

Bottom Tier: Data ware house server

Middle Tier: OLAP server

Top Tier: Front end tools

****

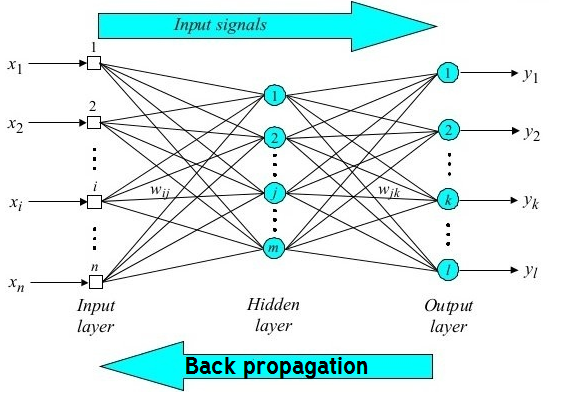
* **Bottom Tier** − The bottom tier of the architecture is the data warehouse server along with the metadata. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, clean, Transform, Load, and refresh functions.
* **Middle Tier** − In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.
  + By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
  + By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.
* **Top-Tier** − This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

21. Briefly explain Classification by back propagation.

**It is a classification method which uses a neural network learning algorithm for classifying and predicting the outcomes. It iteratively process a set of training tuples and compare networks prediction with the actual target value.**

**A neural network is a set of connected inputs/output units where each connection has a weight associated with it.** Neural networks represent a brain metaphor for information processing. These models are biologically inspired rather than an exact replica of how the brain actually functions.

Neural computing refers to a pattern recognition methodology for machine learning.

****

**Back propagation as a process**

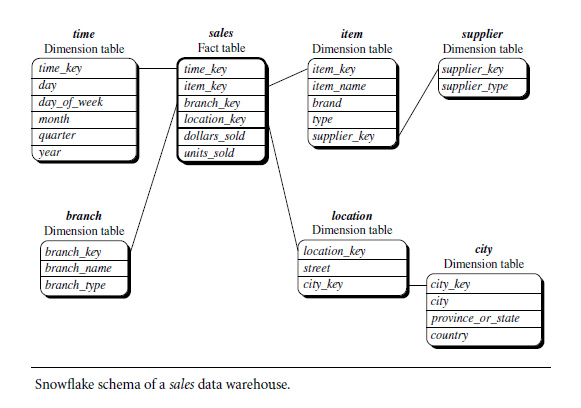
**During learning phase, the network learns by adjusting weights so as to able to predict the correct class label of the input tuples. During classification** desired outputs are compared to achieve system outputs, and then the systems are tuned by adjusting connection ie,  **for the expected class label the weights are updated with the connections backwards from output to input, hence the name back propagation.**

Neural networks have been used in many business applications for pattern recognition (radar systems, face identification, object recognition and more), forecasting, prediction, and classification.

22. Explain snow flake schema model with an example.

Snowflake is a variant of star schema where dimension tables are organized into hierarchy of tables by normalizing them which looks like snowflake.

It is represented by centralized fact table which are connected to multiple dimensions, hierarchy only for dimension tables not for fact table.



23. Briefly explain rule based method of classification with an example.

Rule-based classifier makes use of a set of IF-THEN rules for classification. e can express a rule in the following from −

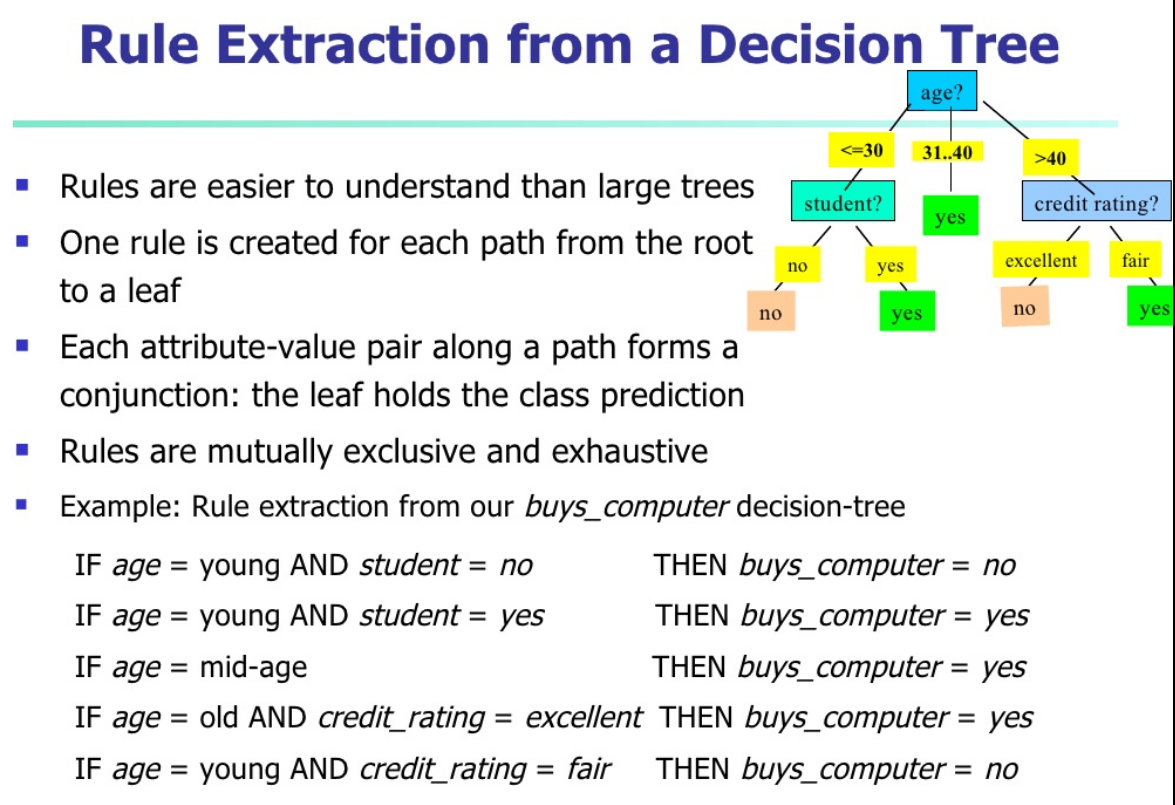
IF condition THEN conclusion

Let us consider a rule R1,

R1: IF age = youth AND student = yes

THEN buy\_computer = yes

* The IF part of the rule is called **rule antecedent** or **precondition**.
* The THEN part of the rule is called **rule consequent**.
* The antecedent part, the condition consist of one or more attribute tests and these tests are logically ANDed.
* The consequent part consists of class prediction.
* − We can also write rule R1 as follows −
* R1: (age = youth) ^ (student = yes))(buys computer = yes)
* If the condition holds true for a given tuple, then the antecedent is satisfied.
* Rule Extraction
* Here we will learn how to build a rule-based classifier by extracting IF-THEN rules from a decision tree.



24. What is Hierarchical Clustering? Explain AGNES type with an example .

A hierarchical clustering method works by grouping data objects into a tree of clusters.

It uses distance (similarity) matrix as clustering criteria.

We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed.

There are two approaches here −

* Agglomerative Approach
* Divisive Approach

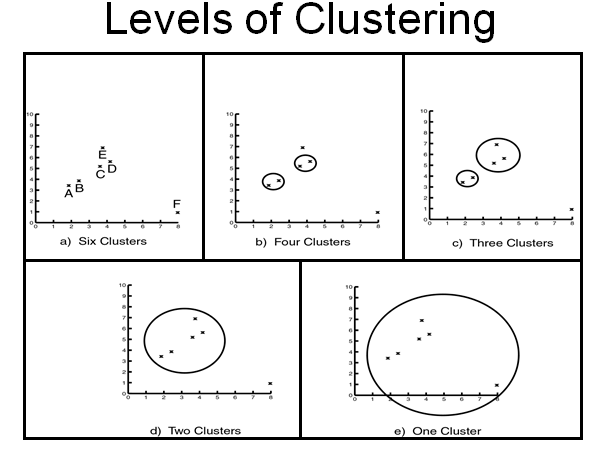
### Agglomerative Approach

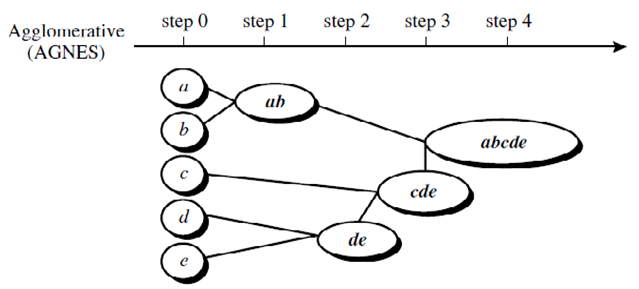
This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds.

Agglomerative (AGNES) Hierarchical Clustering.

Given-a set of N objects to be clustered an N\*N distance matrix is constructed to make cluster,

* The basic process of clustering id this:
* Step1: Assign each object to a cluster so that for N objects we have N clusters each containing just one Object.
* Step2: Let the distances between the clusters be the same as the distances between the objects they contain.
* Step3: Find the most similar pair of clusters and merge them into a single cluster so that we now have one cluster less.
* Step4: Compute distances between the new cluster and each of the old clusters.
* Step5: Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.





25.Write short notes on Data quality and Validation measures of clustering.

Data in the real world trends to be ‘dirty’. Database users frequently report errors, unusual values & inconsistencies in the stored data. Then, it is usual in the real world for the analyzed data to be:

**Incomplete**☺: i.e. lacking attributes values, lacking certain attributes of interest or containing only aggregate data.

**Noisy**☺: containing error or outliers

**Inconsistent**☺ : Containing discrepancies in codes used to categorize items or the names used to refer to the same data items.

**Measures of Cluster Validity**

Numerical measures that are applied to judge various aspects of cluster validity are classified into the following three types.

 External Index: Used to measure the extent to which cluster labels match externally supplied class labels.

 Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.

 Relative Index: Used to compare two different clustering or clusters.

26. Explain the architecture and working of Search engine.

**SEARCH ENGINE ARCHITECTURE**:

**Search Engine** refers to a huge database of internet resources such as web pages, newsgroups, programs, images etc. It helps to locate information on World Wide Web.

User can search for any information by passing query in form of keywords or phrase. It then searches for relevant information in its database and return to the user.

Search Engine Components

Generally there are three basic components of a search engine as listed below:

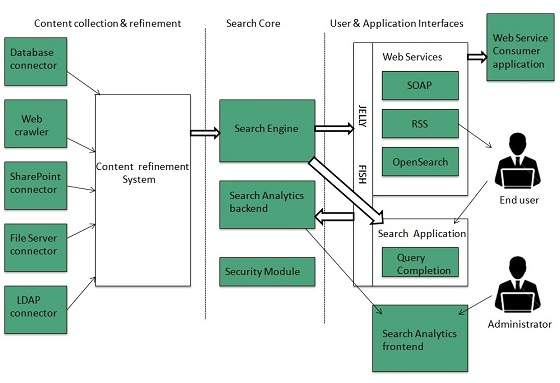
1. Web Crawler
2. Database
3. Search Interfaces

Primary Goals of Search Engines

• Effectiveness (quality): to retrieve the most relevant set of documents for a query – Processing text and storing text statistics to improve relevance

• Efficiency (speed): processing queries from users as fast as possible – Using specialized data structures

Specific goals usually fall into the above primary goals – Example: handling changing document collections – both an effectiveness issue and an efficiency issue



*Two Major Functions*

Search engine components support two major functions

• The index process: building data structures that enable searching

• The query process: using those data structures to produce a ranked list of documents for a user’s query

27. Explain briefly 5V characteristics of Big data .

## Characteristics Of 'Big Data'(V5)



(i)Volume – The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with 'Big Data'.

(ii)Variety – The next aspect of 'Big Data' is its **variety**.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analysing data.

(iii)Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors,[Mobile](https://www.guru99.com/mobile-testing.html)devices, etc. The flow of data is massive and continuous.

(iv)Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

(V) ***Value***: having access to Big data is no good unless we can turn it into value. Companies are starting to generate amazing values for big data.

**SECTION-D**

Answer any **FIVE** questions**: 5 x 7 =35**

**28.a** Differentiate between OLTP and OLAP



**b.** Define OLAP cube. Briefly explain some OLAP operations.

**OLAP Cube or Data CubeAn OLAP cube is a**[**multidimensional database**](https://searchoracle.techtarget.com/definition/multidimensional-database)**that is optimized for data warehouse and**[**online analytical processing (OLAP)**](https://searchdatamanagement.techtarget.com/definition/OLAP)**applications.**

**OLAP Operations**

In multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. It provides flexibility to view data from different perspective

**Roll-up:** The roll-up operation (also called the *drill-up* operation by some vendors) performs aggregation on a data cube, either by *climbing up a concept hierarchy* for a dimension. *country*.” The roll-up operation shown aggregates the data by ascending the *location* hierarchy from the level of *city* to the level of *country*.

**Drill Down:** Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either *stepping down a concept* *hierarchy* for a dimension or *introducing additional dimensions*. Drill-down occurs by descending the *time* hierarchy fromthe level of *quarter* to the more detailed level of *month*.

**Slice:** The *slice* operation performs a selection on one dimension of the given cube, resulting in subcube.

**Dice:** The *dice* operation defines a subcube by performing a selection on two or more dimensions.

**Pivot:**(also called *rotate*) is a visualization operation that rotates the data axes in view to provide an alternative data presentation.

**29.a** Write the Bayesian formula for classification and explain the terms

i. likelihood ii. Prior probability iii. Posterior probability

* [](https://www.analyticsvidhya.com/wp-content/uploads/2015/09/Bayes_rule-300x172.png) *P*(*c|x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
* *P*(*c*) is the prior probability of *class*.
* *P*(*x|c*) is the likelihood which is the probability of *predictor* given *class*.
* *P*(*x*) is the prior probability of *predictor*.

**b.** explain the steps involved in construction of decision tree for classification**.**

**Steps for drawing Decision Trees**

**1. Write the main decision.**

Begin the decision tree by drawing a box (the root node) on 1 edge of your paper

Write the main decision on the box.

**2. Draw the lines**

Draw line leading out from the box for each possible solution or action. Make at least 2,

but better no more than 4 lines. Keep the lines as far apart as you can to enlarge the tree later.

**3. Illustrate the outcomes of the solution at the end of each line.**

**4. Continue adding boxes or cicles and**set

**5. Finish the tree with final outcome**

**30. a**. What is Prediction ? Mention some applications of it in DM?

**Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in**

For example, prediction models in data mining are used by a marketing manager who predict that how much amount a particular customer will spend during a sale, so that upcoming sale amount can be planned accordingly. The prediction in [data mining](https://www.cogneesol.com/data-mining-services) is known as Numeric Prediction. Generally regression analysis is used for prediction.

Typical applications

* Credit approval
* Target marketing
* Medical diagnosis
* Fraud detection

**b**. Briefly explain statistical based classification.

**Statistical classification mainly works on the principle of Logistic regression.**

***Logistic Regression*** is basically a supervised classification algorithm. Logistic Regression is one of the most used Machine Learning algorithms for binary classification. Like many other machine learning techniques, it is borrowed from the field of statistics. Regression used for classification deals with estimation (prediction) of an output (class) value based on input values from the database. It gives you a discrete binary outcome between 0 and 1.

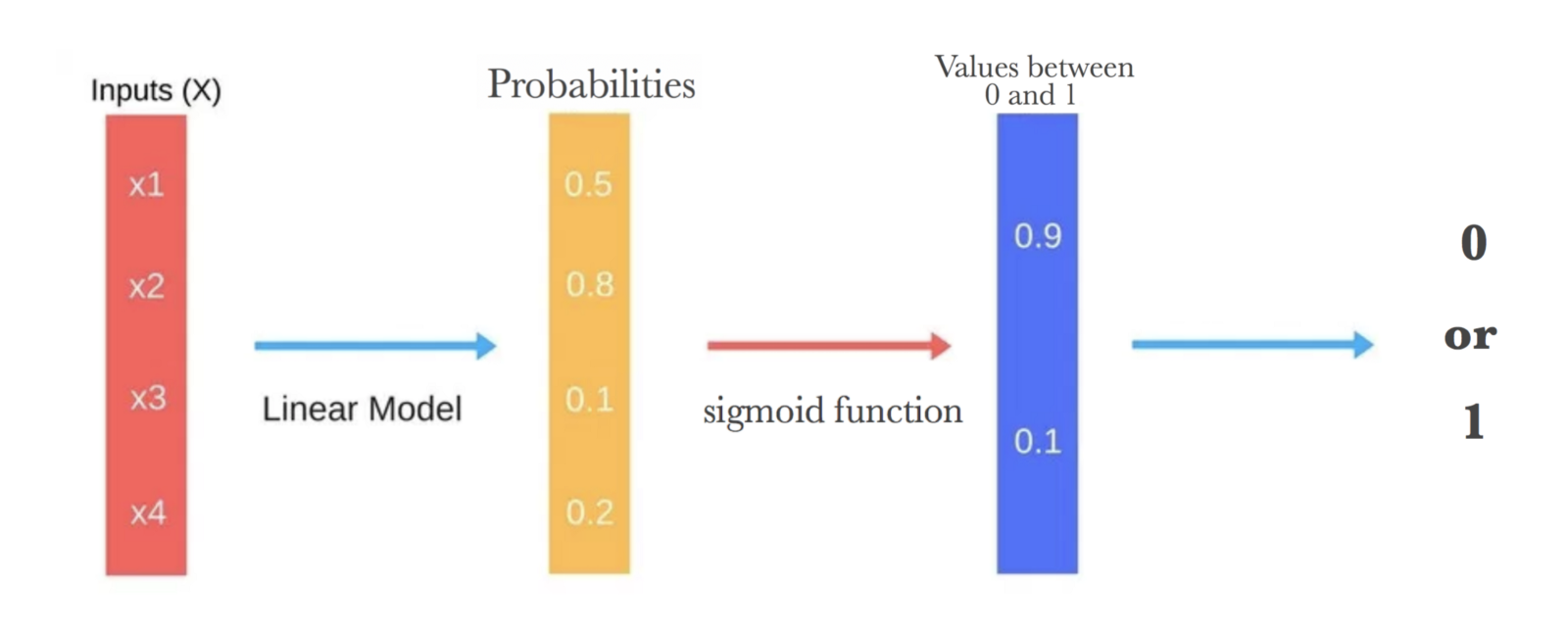
A simple example of a Logistic Regression problem would be an algorithm used for cancer detection that takes screening picture as an input and should tell if a patient has cancer (1) or not (0).

### How it works

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using its underlying logistic function.

These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the ***sigmoid function***. The Sigmoid-Function is an *S-shaped* curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. These values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier.

The picture below illustrates the steps that logistic regression goes through to give you your desired output.

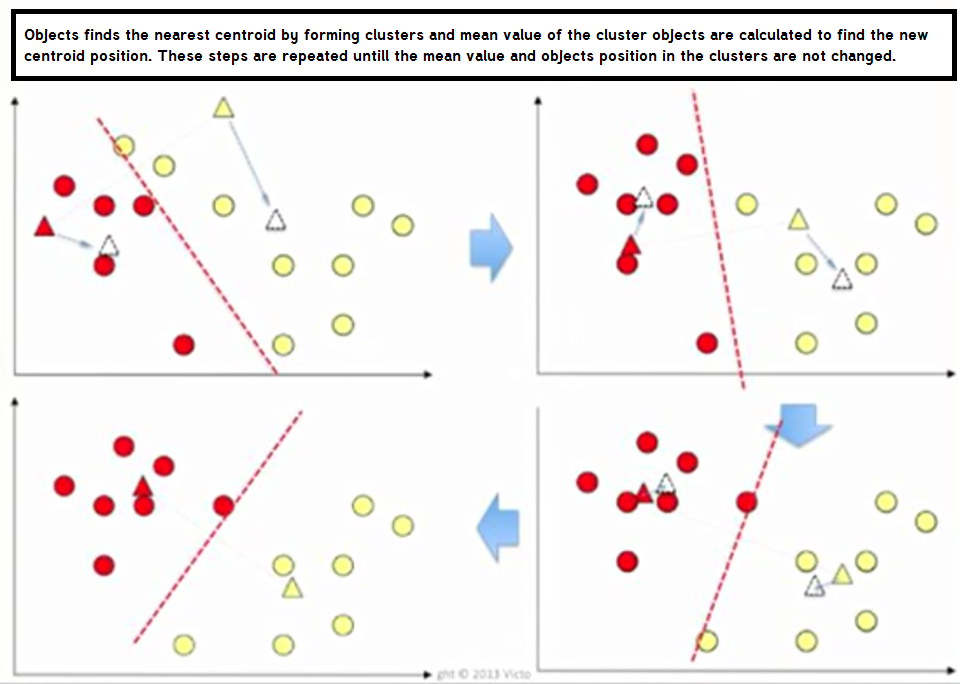


31.a. Explain partition method of clustering with an example

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements –

* For a given number of partitions (say k), the partitioning method will create an initial partitioning.
* Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

In a basic iterative algorithm, such as K-means- or K-medoids, convergence is local and the globally optimal solution cannot be guaranteed. Because the number of data points in any data set is always finite and, thereby, also the number of distinct partitions is finite, the problem of local minima could be avoided by using exhaustive search methods.



* Each group contains at least one object.
* Each object must belong to exactly one group

**b**. Given a Dataset D={2,3,4,8,10,12,15,20,22,25} make to 2 clusters using K-means

**32. a**. What is cluster analysis? Mention some applications.

• “A cluster is an aggregation of points in the space such that the distance between two points in the cluster is less than the distance between any point in the cluster and any point not in it.”

* Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
* Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
* In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
* Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
* Clustering also helps in classifying documents on the web for information discovery.
* Clustering is also used in outlier detection applications such as detection of credit card fraud.
* As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

b. Briefly explain density method of clustering with an example

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of **density reachability** and **density connectivity**

Density Reachability- A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbors which are within distance ε.  
  
Density Connectivity - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance. This is chaining process. So, if "q" is neighbor of "r", "r" is neighbor of "s", "s" is neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p".

**Advantages**

1) Does not require a-priori specification of number of clusters.  
2) Able to identify noise data while clustering.  
3) DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters.

**Disadvantages**

1) DBSCAN algorithm fails in case of varying density clusters.

2) Fails in case of neck type of dataset.

3)Does not work well in case of high dimensional data.

33.a.Write a short notes on web content mining .

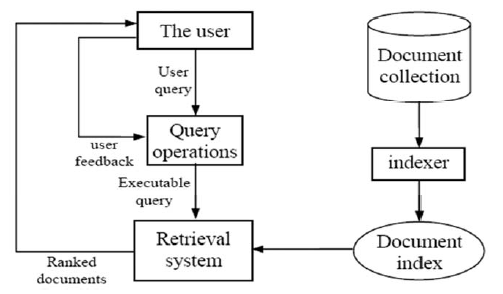
Web Content mining refers to the discovery of useful information from the contents of the webpage using text mining techniques.

Web content mining identifies the useful information from the Web contents. Web content mining could be differentiated from two points of view the agent-based approach or the database approach.

Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables.

Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

**Information Retrieval**



**Natural Language Processing:**

Text mining refers to process to deriving high quality information from text.

NLP turn text into data for analysis via application of NLP & analytical methods.

**Web Content Mining Applications**

* Identify the topics represented by a Web Documents
* Categorize Web Documents
* Find Web Pages across different servers that are similar
* Applications related to relevance
* Queries –Enhance standard Query Relevance with User, Role, and/or Task Based Relevance
* Recommendations –List of top “n” relevant documents in a collection or portion of a collection.
* Filters –Show/Hide documents based on relevance score

b. Briefly explain the ranking of web pages in Search engine.

***Ranking*:** taking the transformed query from the user interaction component and generating a ranked list of documents using scores based on a retrieval model.

The core of a search engine

Both effectiveness and efficiency matter

*Evaluation*: measuring and monitoring the effectiveness and efficiency.

Recording and analyzing user behavior using log data

Evaluation results are used to tune and improve the ranking component

***Web Mining Tools Around the Web***

**34.** Write short notes on Big Data Tools and Techniques

a. Hadoop

# Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.

# Some of the features of the Hadoop platform are that it can be efficiently used for data storage, processing, access, analysis, governance, security, operations and deployment.

# It is at the center of a growing ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and [machine](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML) learning applications.

**Features Of 'Hadoop'**

**Suitable for Big Data Analysis**

* As Big Data tends to be distributed and unstructured in nature, HADOOP clusters are best suited for analysis of Big Data. Since, it is processing logic (not the actual data) that flows to the computing nodes, less network bandwidth is consumed. This concept is called as **data locality concept** which helps increase efficiency of Hadoop based applications.
* **Scalability :**HADOOP clusters can easily be scaled to any extent by adding additional cluster nodes, and thus allows for growth of Big Data. Also, scaling does not require modifications to application logi
* **• Fault Tolerance**
* HADOOP ecosystem has a provision to replicate the input data on to other cluster nodes. That way, in the event of a cluster node failure, data processing can still proceed by using data stored on another cluster node

**b**. NoSQL.

NoSQL is a class of database management systems (DBMS) that do not follow all of the rules of a relational DBMS and cannot use traditional SQL to query data. The term is somewhat misleading when interpreted as "No SQL," and most translate it as "Not Only SQL," as this type of database is not generally a replacement but, rather, a complementary addition to RDBMSs and SQL.

**c**. Map Reduce

# Map Reduce is a computational model and software framework for writing applications which are run on Hadoop. These MapReduce programs are capable of processing enormous data in parallel on large clusters of computation nodes.

# Image result for mapreduce

* 1. **The Map Task**: This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples (key/value pairs).
  2. **The Reduce Task**: This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.
  3. **Shuffle:**redistributing data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node

# HDFS (Hadoop Distributed File System): HDFS takes care of storage part of Hadoop applications. Map Reduce applications consume data from HDFS. HDFS creates multiple replicas of data blocks and distributes them on compute nodes in cluster. This distribution enables reliable and extremely rapid computations.