UNIT-1-Introduction to Business intelligence

1. TWO MARKS QUESTION WITH ANSWERS:

1. What are the uses of multi featurecubes?

Multi feature cubes, which compute complex queries involving multiple dependent aggregates at multiple granularity. These cubes are very useful in practice. Many complex data mining queries can be answered by multi feature cubes without any significant increase in computational cost, in comparison to cube computation for simple queries with standard data cubes.

2. Compare OLTP and OLAPSystems.

If an on-line operational database systems is used for efficient retrieval, efficient storage and management of large amounts of data, then the system is said to be on-line transaction processing. Data warehouse systems serves users (or) knowledge workers in the role of data analysis and decision-making. Such systems can organize and present data in various formats. These systems are known as on-line analytical processing systems.

3. What is data warehousemetadata?

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

4. Explain the differences between star and snowflakeschema.

The dimension table of the snowflake schema model may be kept in normalized

Form to reduce redundancies. Such a table is easy to maintain and saves storage space.

5. In the context of data warehousing what is datatransformation?

`In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:
Smoothing, Aggregation, Generalization, Normalization, Attribute construction

6. Define Slice and Diceoperation.

The slice operation performs a selection on one dimension of the cube resulting in A sub cube. The dice operation defines a sub cube by performing a selection on two (or) moredimensions.

7. List the characteristics of a data warehouse.

There are four key characteristics which separate the data warehouse from other major operational systems:

1. Subject Orientation: Data organized bysubject
2. Integration: Consistency of definingparameters
3. Non-volatility: Stable data storagemedium
4. Time-variance: Timeliness of data and accessterms

8. What are the various sources for datawarehouse?

Handling of relational and complex types of data: Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.

Mining information from heterogeneous databases and global information systems: Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases.

2. THREE MARKS QUESTION WITHANSWERS:

1. What is datawarehouse?

A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. (Or) A data warehouse is a subject-oriented, time-variant and nonvolatile collection of data insupport of management's decision-makingprocess.

2. Differentiate fact table and dimension table.

Fact table contains the name of facts (or) measures as well as keys to each of the related dimensional tables. A dimension table is used for describing the dimension. (e.g.) A dimension table for item may contain the attributes item_ name, brand and type.

3 Briefly discuss the schemas for multidimensionaldatabases.

Stars schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension.

Snowflakes schema: The snowflake schema is a variant of the star schema model, where

Some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

Fact Constellations: Sophisticated applications may require multiple fact tables to *share* dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

4. How is a data warehouse different from a database? How are theysimilar?

Data warehouse is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making. A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples(records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. Both are used to store and manipulate the data.

5. What is descriptive and predictive datamining?

Descriptive data mining, which describes data in a concise and summarative manner and presents interesting general properties of the data.

Predictive data mining, which analyzes data in order to construct one or a set of models and attempts to predict the behavior of new data sets. Predictive data mining, such as classification, regression analysis, and trendanalysis.

6. Differentiate data mining and datawarehousing.

Datamining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referredtoasgoldminingratherthanrockorsandmining.Thus,dataminingshouldhave

been more appropriately named "knowledge mining from data,"
A **data warehouse** is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or salesamount

3. Five-marks questions andanswers

1) Define data warehouse? Differentiate between operational database systems and data warehouses?

A) A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

operational systems	**data warehousing systems**
Operational systems are generally designed to support high-volume transaction processing with minimal back-end reporting.	Data warehousing systems are generally designed to support high-volume analytical processing (i.e. OLAP) and subsequent, often elaborate report generation.

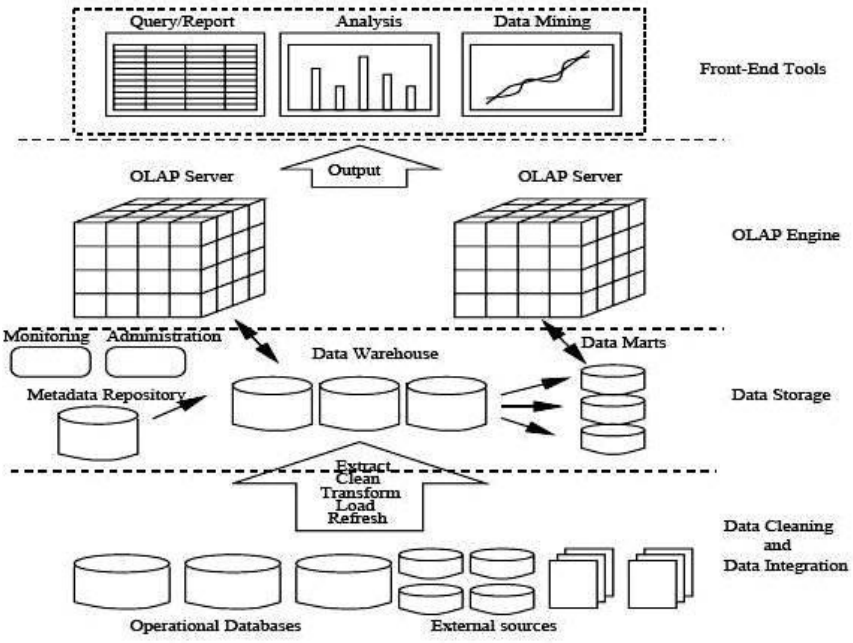
Operational systems are generally process-oriented or process-driven, meaning that they are focused on specific business processes or tasks. Example tasks include billing, registration, etc.	Data warehousing systems are generally subject-oriented, organized around business areas that the organization needs information about. Such subject areas are usually populated with data from one or more operational systems. As an example, revenue may be a subject area of a data warehouse that incorporates data from operational systems that contain student tuition data, alumnigift data, financial aid data, etc.
Operational systems are generally concerned with current data.	Data warehousing systems are generally concerned with historical data.
Data within operational systems are generally updated regularly according to need.	Data within a data warehouse is generally non-volatile, meaning that new data may be added regularly, butonce loaded, the data is rarely changed, thus preservingan

	ever-growing history of information. In short, data within a data warehouse is generally read-only.
Operational systems are generally optimized to performfast inserts and updates of relatively small volumes ofdata.	Data warehousing systems are generally optimized to performfast retrievals of relatively large volumes ofdata.
Operational systems are generally application-specific, resulting in a multitude of partially or non-integrated systems and redundant data (e.g.billing data is not integrated with payroll data).	Data warehousing systems are generally integrated at a layer above the application layer, avoiding data redundancy problems.
Operational systems generally require anon-trivial level of computing skills amongst the end-user community.	Data warehousing systems generally appeal to an end-user community with a wide range of computing skills, from novice to expert users.

2) Explain the architecture of datawarehouse.

A) The Design of a Data Warehouse: A Business AnalysisFramework:Four different views regarding the design of a data warehouse must be considered: the top- down view, the data source view, the data warehouse view, and the business query view.

- The top-down view allows the selection of relevant information necessary for the data warehouse.
- The data source view exposes the information being captured, stored and managed by operational systems.
- The data warehouse view includes fact tables and dimension tables.
- Finally the business query view is the Perspective of data in the data warehouse from the viewpoint of the end user.



Three-tier Data warehouse architecture

The bottom tier is ware-house database server which is almost always a relational database system. The middle tier is an OLAP server which is typically implemented using either (1) a Relational OLAP (ROLAP) model, (2) a Multidimensional OLAP (MOLAP) model. The top tier is a client, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

From the architecture point of view, there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse**Enterprise warehouse**: An enterprise

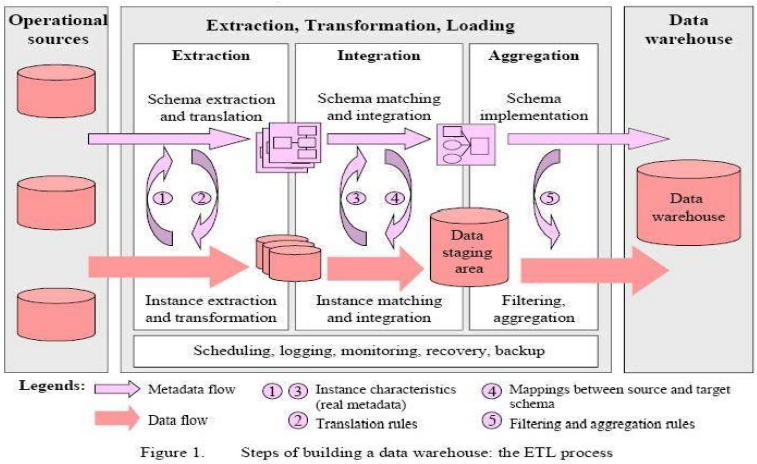


Figure 1. Steps of building a data warehouse: the ETL process

warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional

Data mart: A data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is connected to specific, selected subjects. For example, a marketing data mart may connect its subjects to customer, item, and sales. The data contained in data marts tend to be summarized. Depending on the source of data, data marts can be categorized into the following two classes:

(i).Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area.

(ii).Dependent data marts are sourced directly from enterprise data warehouses

Virtual warehouse: A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capacity on operational database servers.

3). Discuss Extraction-Transformation-loading with neatdiagram?

A) The ETL (Extract Transformation Load)process

In this section we will discussed about the 4 major process of the data warehouse. They are extract (data from the operational systems and bring it to the data warehouse), transform (the data into internal format and structure of the data warehouse), cleanse (to make sure it is of sufficient quality to be used for decision making) and load (cleanse data is put into the data warehouse).

The four processes from extraction through loading often referred collectively as Data Staging.

EXTRACT: Some of the data elements in the operational database can be reasonably be expected to be useful in the decision making, but others are of less value for that purpose. For this reason, it is necessary to extract the relevant data from the operational database before bringing into the data warehouse. Many commercial tools are available to help with the extraction process. **Data Junction** is one of the commercial products. The user of one of these tools typically has an easy-to-use windowed interface by which to specify thefollowing:

TRANSFORM

The operational databases developed can be based on any set of priorities, which keeps changing with the requirements. Therefore those who develop data warehouse based on these databases are typically faced with inconsistency among their data sources. Transformation process deals with

rectifying any inconsistency (if any).One of the most common transformation issues is 'Attribute Naming Inconsistency'. It is common for the given data element to be referred to by different data names in different databases. Employee Name may be EMP_NAME in one database, ENAME in the other. Thus one set of Data Names are picked and used consistently in the data warehouse. Once all the data elements have right names, they must be converted to common formats. The conversion may encompass the following:

Characters must be converted ASCII to EBCDIC or viceversa.
Mixed Text may be converted to all uppercase for consistency.
Numerical data must be converted in to a commonformat.
Data Format has to be standardized.
Measurement may have to convert. (Rs/ $)
Coded data (Male/ Female, M/F) must be converted into a common format.

All these transformation activities are automated and many commercial products are available to perform the tasks. **Data MAPPER** from Applied Database Technologies is one such comprehensive tool.

CLEANSING

Information quality is the key consideration in determining the value of the information. The developer of the data warehouse is not usually in a position to change the quality of its underlying historic data, though a data warehousing project can put spotlight on the data quality issues and lead to improvements for the future. It is, therefore, usually necessary to go through the data entered into the data warehouse and make it as error free as possible. This process is known as **DataCleansing**.

Data Cleansing must deal with many types of possible errors. These include missing data and incorrect data at one source; inconsistent data and conflicting data when two or more source is involved. There are several algorithms followed to clean the data, which will be discussed in the coming lecture notes.

LOADING

Loading often implies physical movement of the data from the computer(s) storing the source database(s) to that which will store the data warehouse database, assuming it is different. This takes place immediately after the extraction phase. The most common channel for data

movement is a high-speed communication link. Ex: Oracle Warehouse Builder is the API from Oracle, which provides the features to perform the ETL task on Oracle Data Warehouse.

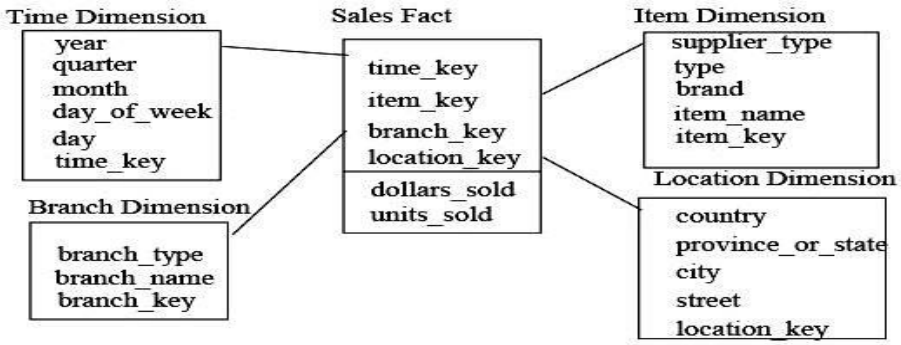
4). Discuss schemas for multi-dimensionaltables?

A) Schemas for MultidimensionalDatabases

Star schema: The star schema is a modeling paradigm in which the data warehouse contains

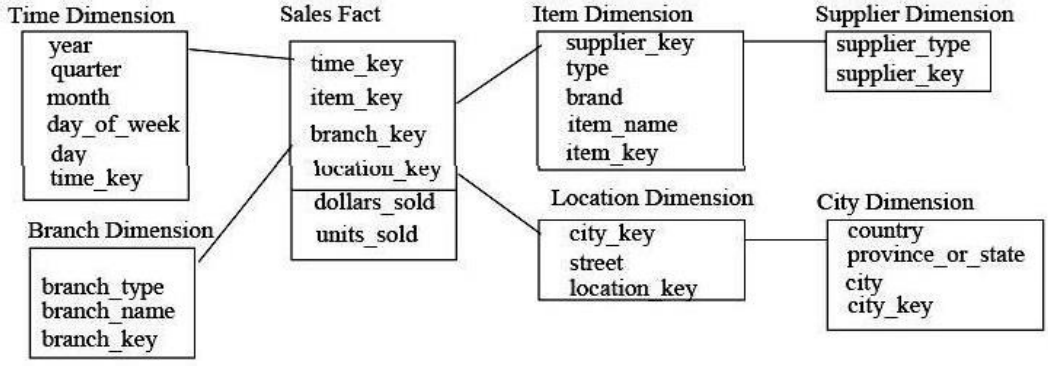
(1) a large central table (fact table), and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central facttable.

Figure Star schema of a data warehouse for sales.



Snowflake schema: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake. The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form. Such a table is easy to maintain and also saves storage space because a large dimension table can be extremely large when the dimensional structure is included as columns.

Figure: Snowflake schema of a data warehouse for sales.



Fact constellation: Sophisticated applications may require multiple facttables to share

dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a factconstellation.

5) Discuss OLAPoperations?

A) OLAP operations on multidimensionaldata.

Roll-up: The roll-up operation performs aggregation on a data cube, either byclimbing-up a concept hierarchy for a dimension or by dimension reduction. Figure shows the result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for location. This hierarchy was defined as the total order street < city < province or state <country.

Drill-down: Drill-down is the reverse of roll-up. It navigates from less detaileddata to more detailed data. Drill-down can be realized by either stepping-down a concept hierarchy for a dimension or introducing additional dimensions. Figure shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for time defined as day < month < quarter < year. Drill-down occurs by descending the time hierarchy from the level of quarter to the more detailed level of month.

Slice and dice: The slice operation performs a selection on one dimension of thegiven cube, resulting in a sub cube. Figure shows a slice operation where the sales data are selected from the central cube for the dimension time using thecriteria

time="Q2". The dice operation defines a sub cube by performing a selection on two or more dimensions.

4. Pivot (rotate): Pivot is a visualization operation which rotates the data axes in view in order to provide an alternative presentation of the data. Figure shows a pivot operation where the item and location axes in a 2-D slice are rotated.

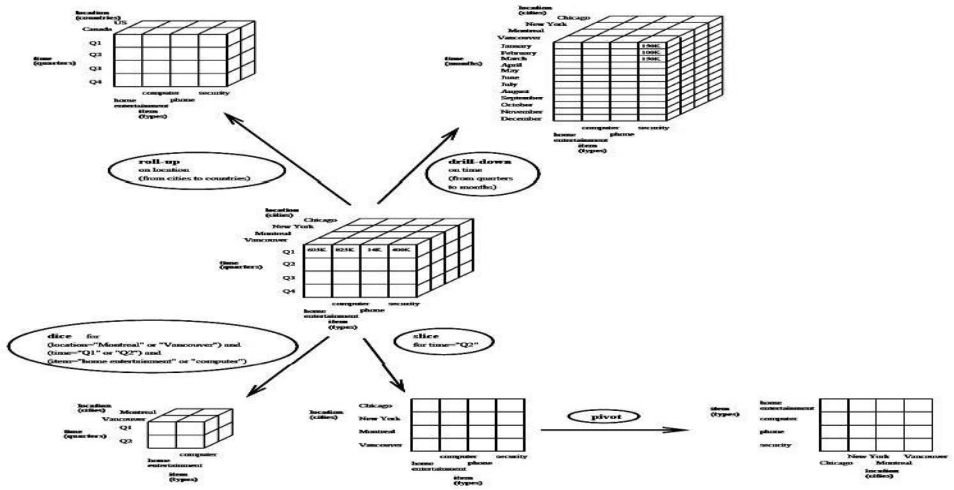


Figure: Examples of typical OLAP operations on multidimensional data.

UNIT-2-Preprocessing the data

1. TWO MARKS QUESTION WITH ANSWERS:

1. What is the need for preprocessing thedata?

Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases and data warehouses. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Other data may not be included simply because it was not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

2. What is parallel mining of concept description? (OR)Whatis conceptdescription?

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms, or (2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

3. What is dimensionalityreduction?

In dimensionality reduction, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data. If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless.

4. Mention the various tasks to beaccomplishedas part of data pre-processing. (Nov/Dec 2008)

1. Datacleaning
2. Data Integration
3. DataTransformation
4. Datareduction

5. What is data cleaning? (May/June2009)

Data cleaning means removing the inconsistent data or noise and collecting necessary information of a collection of interrelateddata.

6. Define Data mining.

Data mining refers to extracting or "mining" knowledge from largeamounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have beenmore

appropriately named "knowledge mining from data,"

7. What are the types of concept hierarchies?

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Concept hierarchies allow specialization, or drilling down, where by concept values are replaced by lower-level concepts.

2. THREE MARKS QUESTION WITH ANSWERS:

1. List the three important issues that have to be addressed during dataintegration.

(OR)

List the issues to be considered during data integration.

There are a number of issues to consider during data integration. **Schema integration** and **object matching** can be tricky. How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identificationproblem.

Redundancy is another important issue. An attribute (such as annual revenue, for instance) may be redundant if it can be "derived" from another attribute or set of attributes.

Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

A third important issue in data integration is the **detection and resolution of data value conflicts.** For example, for the same real-world entity, attribute values from different sources may differ. This may be due to differences in representation, scaling, or encoding. For instance, a weight attribute may be stored in metric units in one system and British imperial units inanother.

2. Write the strategies for data reduction.

1. Data cubeaggregation
2. Attribute subsetselection
3. Dimensionalityreduction
4. Numerosityreduction
5. Discretization and concept hierarchygeneration.

3. Why is it important to have data mining query language?

The design of an effective data mining query language requires a deep understanding of the power, limitation, and underlying mechanisms of the various kinds of data mining tasks.A data mining query language can be used to specify data mining tasks. In particular, we examine how to define data warehouses and data marts in our SQL-based data mining query language,DMQL.

4. List the five primitives for specifying a data mining task.

The set of *task-relevant data* to be mined the *kind of knowledge* to be mined:
The *background knowledge* to be used in the discovery process the *interestingness measures and thresholds* for pattern evaluation
The expected *representation for visualizing* the discovered pattern

5. What is data generalization?

It is process that abstracts a large set of task-relevant data in a database from relatively low conceptual levels to higher conceptual levels 2 approaches for Generalization.

1) Data cube approach 2) Attribute-oriented induction approach

6. How concept hierarchies are useful in data mining?

A concept hierarchy for a given numerical attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged, or senior). Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret.

7. How do you clean the data?

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. For Missing Values

1. Ignore thetuple
2. Fill in the missing valuemanually
3. Use a global constant to fill in the missingvalue
4. Use the attribute mean to fill in the missingvalue:
5. Use the attribute mean for all samples belonging to the same class as the giventuple
6. Use the most probable value to fill in the missing value For NoisyData

1. Binning: Binning methods smooth a sorted data value by consultingits "neighborhood," that is, thevalues aroundit.
2. Regression: Data can be smoothedbyfitting the data to a function, such as with Regression
3. Clustering: Outliers may be detected by clustering, where similar values are

"clusters.

3. Five-marks question and answers

1. What is datamining?

Data mining refers to extracting or mining "knowledge from large amounts of data. There are many other terms related to data mining, such as knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in

Databases or KDD

Essential step in the process of Knowledge Discovery in Databases.

Knowledge discovery as a process is depicted in following figure and consists of an iterative sequence of the following steps:

- Data cleaning: to remove noise or irrelevant data
- Data integration: where multiple data sources may be combined
- Data selection: where data relevant to the analysis task are retrieved from the database
- Data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations
- Data mining :an essential process where intelligent methods are applied in order to extract data patterns
- Pattern evaluation to identify the truly interesting patterns representing knowledge based on some interestingness measures
- Knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2. Describe the Architecture of a typical data mining system/MajorComponents?

Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Based on this view, the architecture of a typical data mining system may have the following major components:

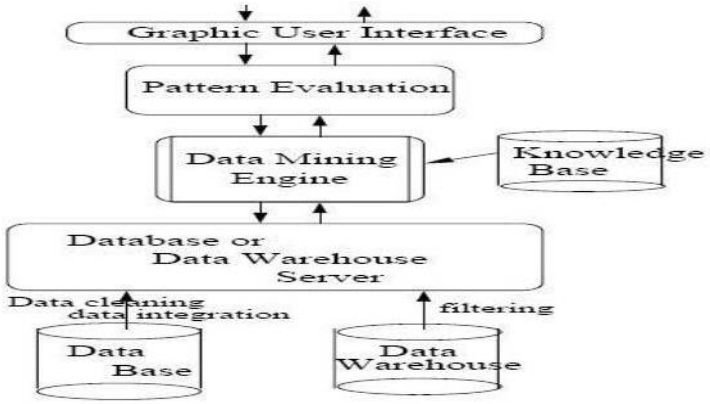
- A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.

  mining requests.
- A database or data warehouse server which fetches the relevant data based on users' data

  A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple Heterogeneous sources.

- data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.

- A pattern evaluation module that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns. A graphical user interface that allows the user an interactive approach to the data miningsystem.
- A graphical user interface that allows the user an interactive approach to the data mining system.



Architecture of a typical data mining system.

3. How is a data warehouse different from a database? How are theysimilar?

Differences between a data warehouse and a database: A data warehouse is a repository of information collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a database, is a collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another.

4. List out Data miningtasks?

The two "high-level" primary goals of data mining, in practice, are *prediction* and *description*.

1. **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables ofinterest.
2. **Description** focuses on finding human-interpretable patterns describing thedata.

The relative importance of prediction and description for particular data mining applications can vary considerably. However, in the context of KDD, description tends to be more important than prediction. This is in contrast to pattern recognition and machine learning applications (such as speech recognition) where prediction is often the primary goal of the KDD process.

The goals of prediction and description are achieved by using the following primary **data miningtasks**:

1. **Classification** is learning a function that maps (classifies) a data item into one of several predefinedclasses.
2. **Regression** is learning a function which maps a data item to a real-valued prediction variable.
3. **Clustering** is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe thedata.
    - Closely related to clustering is the task of *probability density estimation* which consists of techniques for estimating, from data, the joint multi-variate probability density function of all of the variables/fields in the database.
4. **Summarization** involves methods for finding a compact description for a subset ofdata.
5. **Dependency Modeling** consists of finding a model which describes significant dependencies between variables. Dependency models exist at twolevels:
    1. The *structural* level of the model specifies (often graphically) which variables are locally dependent on each other,and
    2. The *quantitative* level of the model specifies the strengths of the dependencies using some numericalscale.

    Change and Deviation Detection focuses on discovering the most significant changes in the data from previously measured or normative values.

5. What do you mean by Attribute sub selection / Feature selection?

Feature selection is a must for any data mining product. That is because, when you build a data mining model, the dataset frequently contains more information than is needed to build the model. For example, a dataset may contain 500 columns that describe characteristics of customers, but perhaps only 50 of those columns are used to build a particular model. If you keep the unneeded columns while building the model, more CPU and memory are required during the training process, and more storage space is required for the completed model.

In which select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features

Basic heuristic methods of attribute subset selection include the following techniques, some of which are illustrated below:

Step-wise forward selection: The procedure starts with an empty set of **attributes.** Thebest of the original attributes is determined and added to the set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.

Step-wise backward elimination: The procedure starts with the full set of **attributes.** At each step, it removes the worst attribute remaining in the set.
Combination forward selection and backward elimination: The step-wise **forward** selection and backward elimination methods can be combined, where at each step one selects the best attribute and removes the worst from among the remaining attributes.

Decision tree induction: Decision tree induction constructs a flow-chart-like structure where

each internal (non-leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction. At each node, the algorithm chooses the “best" attribute to partition the data into individual classes. When decision tree induction is used for attribute subset selection, a tree is constructed from the given data. All attributes that do not appear in the tree are assumed to be irrelevant. The set of attributes appearing in the tree form the reduced subset of attributes.

Wrapper approach/Filter approach:

The mining algorithm itself is used to determine the attribute sub set, then it is called wrapper approach or filter approach. Wrapper approach leads to greater accuracy since it optimizes the evaluation measure of the algorithm while removing attributes.

Data compression

In data compression, data encoding or transformations are applied so as to obtain a reduced or “compressed" representation of the original data. If the original data can be reconstructed from the compressed data without any loss of information, the data compression technique used is called lossless. If, instead, we can reconstruct only an approximation of the original data, then the data compression technique is called lossy. Effective methods of lossy datacompression:

UNIT-3- Association rule mining

1. TWO MARKS QUESTION WITHANSWERS:

1. Define frequent set and border set.

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set Of computer, antivirus software is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of the itemset.Where each variation involves “playing” with the support thresholdin slightly different way.The variations, where nodes indicate an item or itemset that has been examined, and nodes with thick borders indicate that an examined item or itemset isfrequent.

2. How is association rule mined from large databases?

Suppose, however, that rather than using a transactional database, sales and related information are stored in a relational database or data warehouse. Such data stores are multidimensional, by definition. For instance, in addition to keeping track of the items purchased in sales transactions, a relational database may record other attributes associated with the items, such as the quantity purchased or the price, or the branch location of the sale. Additional relational information regarding the customers who purchased the items, such as customer age, occupation, credit rating, income, and address, may also be stored.

3. List two interesting measures for association rules. (OR)Rulesupport and confidence are two measures of ruleinterestingness.

They respectively reflect the usefulness and certainty of discovered rules. A support

of 2% for Association Rule (5.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlationsbetween associated items.

4. What is over fitting and what can you do to prevent it?

Tree pruning methods address this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable branches. An unpruned tree and a pruned version of it. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data (i.e., of previously unseen tuples) than unpruned trees.

2. THREE MARKS QUESTION WITH ANSWERS:

***1.* in classification trees, what are surrogate splits, and how are theyused?**

Decision trees can suffer from repetition and replication, making them overwhelming to interpret. Repetition occurs when an attribute is repeatedly tested along a given branch of the tree (such as "age < 60?" followed by "age < 45"? and so on).In replication, duplicate sub trees exist within the tree. These situations can impede the accuracy and comprehensibility of a decision tree. The use of Tree pruning methods addresses this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable branches. An unpruned tree and a pruned version of it. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data (i.e., of previously unseen tuples) than unprunedtrees.

***2.* Explain the market basket analysis problem.**

Market basket analysis, which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence). This process analyzes customerbuyinghabitsbyfindingassociationsbetweenthedifferentitemsthatcustomersplace in their "shopping baskets". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelfspace.

***3.* Give the difference between Boolean association rule and quantitative Associationrule.**

Based on the types of values handled in the rule: If a rule involves associations between the presence or absence of items, it is a Boolean association rule. For example, the following three rules are Boolean association rules obtained from market basketanalysis.

Computer => antivirus software [support = 2%; confidence =

60%] buys(X, “computer”) => buys(X, “HP printer”)
buys(X, “laptop computer”) => buys(X, “HP printer”)

Quantitative association rules involve numeric attributes that have an implicit ordering among values (e.g., age). If a rule describes associations between quantitative items or attributes, then it is a quantitative association rule. In these rules, quantitative values for items or attributes are partitioned into intervals. Following rule is considered a quantitative association rule. Note that the quantitative attributes, age and income, have been discretized.

age(X, “30: : :39”)^income(X, “42K....48K”) => buys(X, “high resolution TV”)

4. List the techniques to improve the efficiency of Apriorialgorithm.

- Hash based technique
- Transaction Reduction
- Portioning
- Sampling
- Dynamic item counting

5. What is FP growth?

FP-growth, which adopts a divide-and-conquer strategy as follows. First, it compresses the database representing frequent items into a frequent-pattern tree, or FP-tree,which retains the itemset association information.

2. FIVE MARKS QUESTION WITHANSWERS:

“pattern fragment,” and mines each such database separately.

1. Explain Association rule?

It is an important data mining model studied extensively by the database and data mining community.

Assume all data are categorical. No good algorithm for numeric data. Initially used for Market Basket Analysis to find how items purchased by customers are related.

Bread $\rightarrow$ Milk [sup = 5%, conf =100%]

$I = \{i1, i2, \ldots, im\}$: a set of *items*.
Transaction t :t a set of items, and $t \subseteq I$.
TransactionDatabase T:asetoftransactions$T=\{t1,t2, \ldots,tn\}$. A transaction t contains X, a set of items (itemset) in I, if $X \subseteq t$. An association rule is an implication of theform:

$X \rightarrow Y$, where $X, Y \subset I, and X \cap Y = \emptyset$

An itemset is a set of items.

- E.g., X = {milk, bread, cereal} is an itemset.

A *k*-itemset is an itemset with *k* items.

- E.g., {milk, bread, cereal} is a 3-itemset

Rule strength measures:

Support: The rule holds with support *sup* in *T* (the transaction data set) if sup% of transactions contain $X \cup Y$.

- *sup* = Pr($X \cup Y$).

Confidence: The rule holds in *T* with confidence *conf* if *conf*% of tranactions that contain *X* also contain *Y*.

- *conf* = Pr($Y \mid X$)

An association rule is a pattern that states when *X* occurs, *Y* occurs with certain probability.

Support count: The support count of an itemset*X*, denoted by *X.count*, in a data set *T* is the number of transactions in *T* that contain *X*. Assume *T* has *n* transactions.

Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

2). Describe Apriorialgorithm?

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant and came to be known as Apriori. This algorithm uses two steps "join" and "prune" to reduce the search space. It is an iterative approach to discover the most frequent itemsets.

Apriori says:

The probability that item I is not frequent is if:

- P(I) < minimum support threshold, then I is not frequent.
- P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.
- If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus can be ignored. This property is called the Antimonotone property.

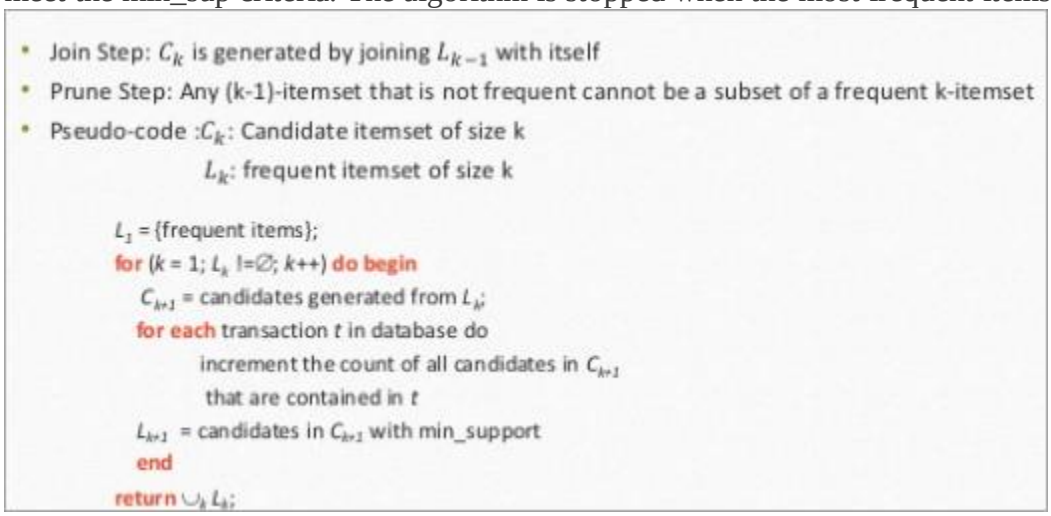
The steps followed in the Apriori Algorithm of data mining are:

1. **Join Step**: This step generates (K+1) itemset from K-itemsets by joining each item with itself.
2. **Prune Step**: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

Steps In Apriori

Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

#1) In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.
#2) Let there be some minimum support, min_sup (eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min_sup, are taken ahead for the next iteration and the others are pruned.
#3) Next, 2-itemset frequent items with min_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.
#4) The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.
#5) The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.
#6) Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min_sup criteria. The algorithm is stopped when the most frequent itemset is achieved.



Source:https://www.softwaretestinghelp.com/apriori-algorithm/

3). What is an item set & frequent item set?

A set of items together is called an itemset. If any itemset has k-items it is called a k-itemset. An itemset consists of two or more items. An itemset that occurs frequently is called a frequent itemset. **Thus frequent itemset mining is a data mining technique to identify the items that often occur together.For Example**, Bread and butter, Laptop and Antivirus software, etc.
A set of items is called frequent if it satisfies a minimum threshold value for support and confidence. Support shows transactions with items purchased together in a single transaction. Confidence shows transactions where

the items are purchased one after the other.For frequent itemset mining method, we consider only those transactions which meet minimum threshold support and confidence requirements. Insights from these mining algorithms offer a lot of benefits, cost-cutting and improved competitive advantage.There is a tradeoff time taken to mine data and the volume of data for frequent mining. The frequent mining algorithm is an efficient algorithm to mine the hidden patterns of itemsets within a short time and less memory consumption.

4.Write The MSapriori algorithm?

Algorithm MSapriori(*T*,*MS*)

$M \leftarrow$ sort(I,MS);

$L \leftarrow$ init-pass(M,T);

$F_1 \leftarrow$ {{i} | $i \in L$, $i.count/n \geq$ MIS(i)};

for (k =2;$F_{k-1} \neq \emptyset$; k++) **do**
if k=2 **then**
$C_k \leftarrow$ level2-candidate-gen(L)
else $C_k \leftarrow$ MScandidate-gen(F_{k-1}); **end;**
for each transaction $t \in T$ **do** **for** each candidate $c \in C_k$ **do** **if** c is contained in t **then**
$c.count$++;
if c − {c[1]} is contained in t **then** $c.tailCount$++
end
end
$F_k \leftarrow$ {$c \in C_k$ | $c.count/n \geq MIS(c[1])$}
end
return $F \leftarrow \bigcup_k F_k$;

Candidate itemset generation

Special treatmentsneeded:

- Sorting the items according to their MIS values
- First pass over data (the first three lines)
  - Let us look at this in detail.
- Candidate generation at level-2
  - Read it in the handout.
- Pruning step in level-k (k> 2) candidate generation.
  - Read it in the handout.

First pass over data

It makes a pass over the data to record the support count of each item.

It then follows the sorted order to find the first item i in M that meets MIS(i).

- *i* is inserted into *L*.
- For each subsequent item *j* in *M* after *i*, if *j.count*/*n* ≥ MIS(*i*) then *j* is also inserted into *L*, where *j.count* is the support count of *j* and *n* is the total number of transactions in *T*. Why?

L is used by function level2-candidate-gen

3. Explain PartitionAlgorithm?

The pseudocode of PAM algorithm is shown below:
In the R programming language, the PAM algorithm is available in the cluster package and can be called by the following command:pam(x, k, diss, metric, medoids, stand, cluster.only, do.swap, keep.diss, keep.data,trace.lev) Where the parameters are:
x: numerical data matrix representing the dataset entities, or can be the dissimilarity matrix, it depends on the value of the diss parameter. In case x is a data matrix each row is an entity and each column is an variable, and in this case missing values are allowed as long as every pair of entities has at least one case not missing. In case x is a dissimilarity matrix it is not allowed to have missing values.**k:** number of clusters that the dataset will be partitioned where $0 < k < n$, where n is the number of entities.**diss:** logical flag, if it is TRUE x is used as the dissimilarity matrix, if it is FALSE, then x will be considered as a data matrix.
metric: an string specifying each of the two metrics will be used to calculate the dissimilarity matrix, the metric variable can be "euclidean" to use the Euclidean distance, or can be "manhattan" to use the Manhattan distance.

stand: logical flag, if it is TRUE then the measurements in x will be standardized before calculating the dissimilarities. Measurements are standardized for each column, by subtracting the column's mean value and dividing by the variable's mean absolute deviation. If x is a dissimilarity matrix then this parameter isignored.

cluster.only: logical flag, if it is TRUE, only the clustering will be computed and returned.

do.swap: logical flag, indicates if the swap phase should happen (TRUE) or not (FALSE).

keep.diss: logical flag indicating if the dissimilarities should (TRUE) or not (FALSE) be kept in the result.

keep.data: logical flag indicating if the input data x should (TRUE) or not (FALSE) be kept in the result.

trace.lev: an numeric parameters specifying a trace level for printing diagnostics during the build and swap phase of the algorithm. Default 0 does not printanything.

The PAM algorithm returns a pam object that contains the information about the result of the execution of the algorithm.

4. Illustrate FP-Growthalgorithm?

The FP-Growth Algorithm is an alternative way to find frequent itemsets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the itemset associationinformation.

In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step it divides the compressed database into a set of conditional databases, each one associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity. Inlarge databases, it's not possible to hold the FP-tree in the main memory. A strategy to cope with this problem is to firstly partition the database into a set of smaller databases (called projected databases), and then construct an FP-tree from each of these smaller databases.

The next subsections describe the FP-tree structure and FP-Growth Algorithm, finally an example is presented to make it easier to understand theseconcepts.

FP-Tree structure

The frequent-pattern tree (FP-tree) is a compact structure that stores quantitative information about frequent patterns in a database

Han defines the FP-tree as the tree structure defined below

1. One root labeled as "null" with a set of item-prefix sub trees as children, and a frequent-item-header table (presented in the left side of Figure1);
2. Each node in the item-prefix sub tree consists of threefields:

    1. Item-name: registers which item is represented by thenode;
    2. Count: the number of transactions represented by the portion of the path reaching thenode;
    3. Node-link: links to the next node in the FP-tree carrying the same item-name, or null if there isnone.

    1. Each entry in the frequent-item-header table consists of twofields:

        1. Item-name: as the same to thenode;
        2. Head of node-link: a pointer to the first node in the FP-tree carrying the item-name.

The original algorithm to construct the FP-Tree defined by Han in [1] is presented below in Algorithm 1.

Algorithm 1: FP-tree construction

Input: A transaction database DB and a minimum support threshold ?.

Output: FP-tree, the frequent-pattern tree of DB.

Method: The FP-tree is constructed as follows.

1. Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequentitems.
2. Create the root of anFP-tree,T, and label it as "null". For each transaction Trans in DB do thefollowing:

- Select the frequent items in Trans and sort them according to the order of FList. Let the sorted frequent-item list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree([p | P], T).
- The function insert tree([p | P], T) is performed as follows. If T has a child N such that N.item-name=p.item-name, then increment N 's count by 1; else create a new node N , with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree(P, N)recursively.

By using this algorithm, the FP-tree is constructed in two scans of the database. The first scan collects and sort the set of frequent items, and the second constructs the FP-Tree.

5.What are the methods to improve apriori algorithm efficiency?

Methods To Improve Apriori Efficiency

Many methods are available for improving the efficiency of the algorithm.

1. **Hash-Based Technique:** This method uses a hash-based structure called a hash table for generating the k-itemsets and its corresponding count. It uses a hash function for generating the table.
2. **Transaction Reduction:** This method reduces the number of transactions scanning in iterations. The transactions which do not contain frequent items are marked or removed.
3. **Partitioning:** This method requires only two database scans to mine the frequent itemsets. It says that for any itemset to be potentially frequent in the database, it should be frequent in at least one of the partitions of the database.
4. **Sampling:** This method picks a random sample S from Database D and then searches for frequent itemset in S. It may be possible to lose a global frequent itemset. This can be reduced by lowering the min_sup.
5. **Dynamic Itemset Counting:** This technique can add new candidate itemsets at any marked start point of the database during the scanning of the database.

6.What are the application ofapriorialgorithm?

Applications of Apriori Algorithm

Some fields where Apriori is used:

1. **In Education Field:** Extracting association rules in data mining of admitted students through characteristics and specialties.
2. **In the Medical field:** For example, Analysis of the patient's database.
3. **In Forestry:** Analysis of probability and intensity of forest fire with the forest fire data.
4. Apriori is used by many companies like Amazon in the **Recommender System** and by Google for the auto-complete feature.

7.What are the advantages and disadvantages of apriori algorithm?

Advantages

1. Easy to understand algorithm
2. Join and Prune steps are easy to implement on large itemsets in large databases

Disadvantages

1. It requires high computation if the itemsets are very large and the minimum support is kept very low.
2. The entire database needs to be scanned.

UNIT-4 Clustering techniques

2. TWO MARKS QUESTION WITH ANSWERS:

1. **What is treepruning?**

   Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

2. **List the requirements of clustering in datamining.**

   Mining data streams involves the efficient discovery of general patterns and dynamic changes within stream data. For example, we may like to detect intrusions of a computer network based on the anomaly of message flow, which may be discovered by clustering data streams, dynamic construction of stream models, or comparing the current frequent patterns with that at a certain previous time.

3. **What is classification?**

   Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

4. What is the difference between classification and decision tree:

Classification	Clustering
It is supervised learning.	It is unsupervised learning.
Classification contains previously categorized training set.	In clustering, the characteristics of similarity of data is not known.

Decision tree is used to partition and segment record.	There are a variety of algorithms for clustering, which generally share the same property of interactively assigning records to a cluster.

5.What is the objective function of the K-meansalgorithm?

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low.Clustersimilarityismeasuredinregardtothemeanvalueoftheobjectsinacluster,

which can be viewed as the cluster's centroid or center of gravity.
First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes thenewmeanforeachcluster.Thisprocessiteratesuntilthecriterionfunctionconverges.

Typically, the square-error criterion is used, defined as where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and mi is the mean of cluster Ci (both p and mi are multidimensional).

6.The naïve Bayes classifier makes what assumption that motivates itsname?

Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers.

Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered

"naïve."

2. THREE MARKS QUESTION WITH ANSWERS:

1. **What is an outlier? (OR)**
Define outliers. List various outlier detection approaches.

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. These can be categorized into four approaches: the statistical approach, the distance-based approach, the density-based local outlier approach, and the deviation-based approach.

2. **Compare clustering and classification.**

Clustering techniques consider data tuples as objects. They partition the objects into groups or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how "close" the objects are in space, based on a distance function. The "quality" of a cluster may

be represented by its diameter, the maximum distance between any two objects in the cluster. Outliers may be detected by clustering, where similar values are organized intogroups, or "clusters." Intuitively, values that fall outside of the set of clusters may be consideredoutliers.

***3.* What is meant by hierarchical clustering?**

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

***4.* What is Bayesian theorem?**

Let X be a data tuple. In Bayesian terms, X is considered "evidence." As usual, it is

described by measurements made on a set of n attributes. Let H be some hypothesis, such as that the data tuple X belongs to a specified class C. For classification problems, we want to determine P(HjX), the probability that the hypothesis H holds given the "evidence" or observed data tuple X. In other words, we are looking for the probability that tuple X belongs to class C, given that we know the attribute description of X.

***5.* What is Association based classification?**

Association-based classification, which classifies documents based on a set of associated, frequently occurring text patterns. Notice that very frequent terms are likely poor discriminators. Thus only those terms that are not very frequent and that have good discriminative power will be used in document classification. Such an association-based classification method proceeds as follows: First, keywords and terms can be extracted by information retrieval and simple association analysis techniques. Second, concept hierarchies of keywords and terms can be obtained using available term classes, such as WordNet, or relying on expert knowledge, or some keyword classificationsystems.

***6.* Compare the advantages of and disadvantages of eager classification (e.g., decision tree) versus lazy classification (k-nearest neighbor)**

Eager learners, when given a set of training tuples, will construct a generalization (i.e., classification) model before receiving new (e.g., test) tuples to classify. We can think of the learned model as being ready and eager to classify previously unseen tuples. Imagine a contrasting lazy approach, in which the learner instead waits until the last minute before doing

any model construction in order to classify a given test tuple. That is, when given a training tuple, a lazy learner simply stores it (or does only a little minor processing) and waits until it is given a test tuple.

4. FIVE MARKS QUESTION AND ANSWERS

1) What isclassification?

Classification:

- Used for prediction (future analysis) to know the unknown attributes with theirvalues. By using classifier algorithms and decision tree.(in data mining)
- Which constructs some models (like decision trees) then which classifies the attributes.
- Already we know the types of attributes are
  1. Categorical attributeand
  2. Numericalattribute
- These classifications can work on both the above mentioned attributes.

Prediction: prediction also used for to know the unknown or missing values.

which also uses some models in order to predict the attributes
models like neural networks, if else rules and other mechanisms

Classification—A Two-StepProcess

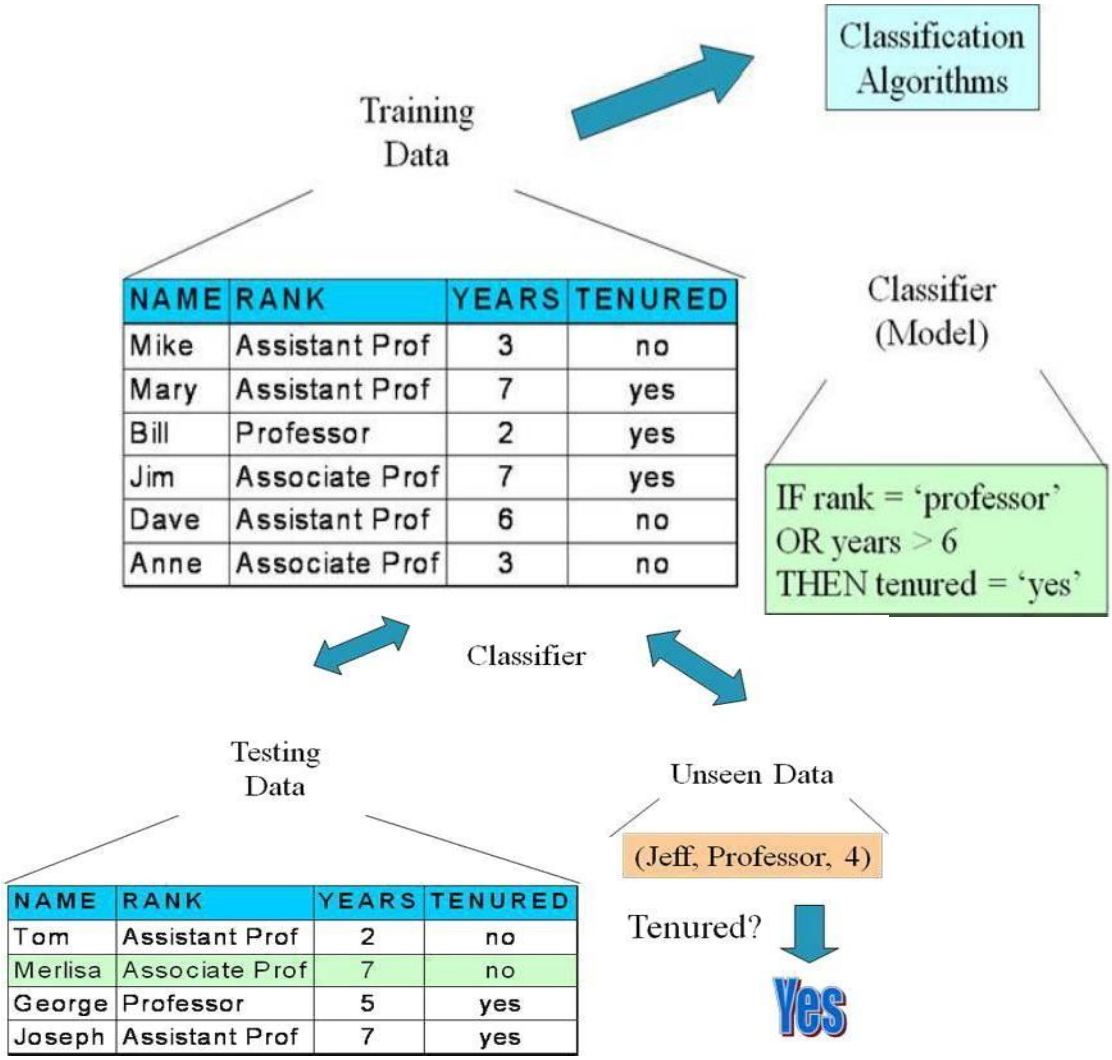
Model construction: describing a set of predetermined classes

- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
- The set of tuples used for model construction: training set
- The model is represented as classification rules, decision trees, or mathematical formulae

Model usage: for classifying future or unknown objects

- Estimate accuracy of the mode

The known label of test sample is compared with the classified result from the model



Accuracy rate is the percentage of test set samples that are correctly classified by the model
Test set is independent of training set, otherwise over-fitting will occurProcess (1): Model Construction

Process (2): Using the Model in Prediction

Supervised vs. Unsupervised Learning

Supervised learning (classification)

Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations.

New data is classified based on the training set

Unsupervised learning (clustering)

The class labels of training data is unknown

Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

2) What are the Issues regarding inclassification?

There are two issues regarding classification and prediction they are Issues (1): Data Preparation

Issues (2): Evaluating Classification Methods

Issues (1): Data Preparation: Issues of data preparation includes the following1) Data cleaning

Preprocess data in order to reduce noise and handle missing values (refer preprocessing techniques i.e. data cleaning notes)

2) Relevance analysis (feature selection)

Remove the irrelevant or redundant attributes (refer unit-iv AOI Relevance analysis)Data transformation (refer preprocessing techniques i.e data cleaning notes) Generalize and/or normalize data

Issues (2): Evaluating Classification Methods: considering classification methodsshould satisfy the followingproperties

Predictive accuracy
Speed and scalability
*time to construct the model *time to use the model

3. Robustness

Handling noise and missing values

4. Scalability

Efficiency in disk-resident databases

5. Interpretability:

Understanding and insight provided by the model

6. Goodness ofrules

Decision tree size

Compactness of classification rules

4). what is Decision Tree?

Decisiontree

- A flow-chart-like tree structure

- Internal node denotes a test on an attribute

  Branch represents an outcome of the test

  Leaf nodes represent class labels or class distribution

Decision tree generation consists of two phases

  Tree construction

At start, all the training examples are at the root
Partition examples recursively based on selected attributes

  Tree pruning

Identify and remove branches that reflect noise or outliers

Use of decision tree: Classifying an unknown sample

  Test the attribute values of the sample against the decision tree

4) Write the Algorithm for DecisionTree?

Basic algorithm (a greedy algorithm)

  Tree is constructed in a top-down recursive divide-and-conquer manner

  At start, all the training examples are at the root

  Attributes are categorical (if continuous-valued, they are discretized in advance)

- Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

Conditions for stopping partitioning

  - All samples for a given node belong to the same class.
  - There are no remaining attributes forfurtherpartitioning – majority voting is employed for classifying theleaf.
  - There are no samples left.

1. TWO MARKS QUESTION AND ANSWERS.

1. What do you go for clustering analysis?

Clustering can be used to generate a concept hierarchy for A by following either a top down splitting strategy or a bottom- up merging strategy, where each cluster forms a node of the concept hierarchy. In the former, each initial cluster or partition may be further decomposed into several sub clusters, forming a lower level of the hierarchy. In the latter, clusters are formed by repeatedly grouping neighboring clusters in order to form higher- level concepts.

2. What are the requirements of cluster analysis?

- Scalability
- Ability to deal with different types of attributes Discovery of clusters with arbitraryshape
- Minimal requirements for domain knowledge to determine input parameters Ability to deal with noisy data
- Incremental clustering and insensitivity to the order of input records High dimensionality Constraint-based clustering
-

3. What is mean by cluster analysis?

A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive object.

4.Mention the advantages of hierarchical clustering.

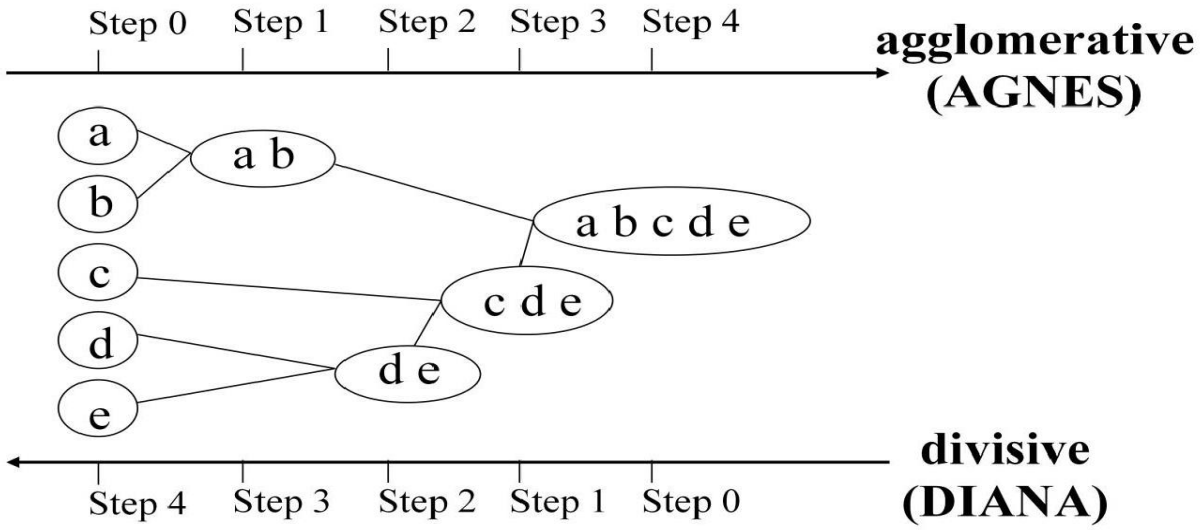
Hierarchical clustering (or hierarchic clustering) outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering. Hierarchical clustering does not require us to prespecify the number of clusters and most hierarchical algorithms that have been used in IR are deterministic. These advantages of hierarchical clustering come at the cost of lower efficiency.

1. Define time series analysis.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series are very frequently plotted via line charts.

2. Explain Hierarchical method clustering of classification with example?[Nov/Dec2014]

- Use distance matrix. This method does not require the number of clusters *k* as an input, but needs a termination condition



AGNES (Agglomerative Nesting):

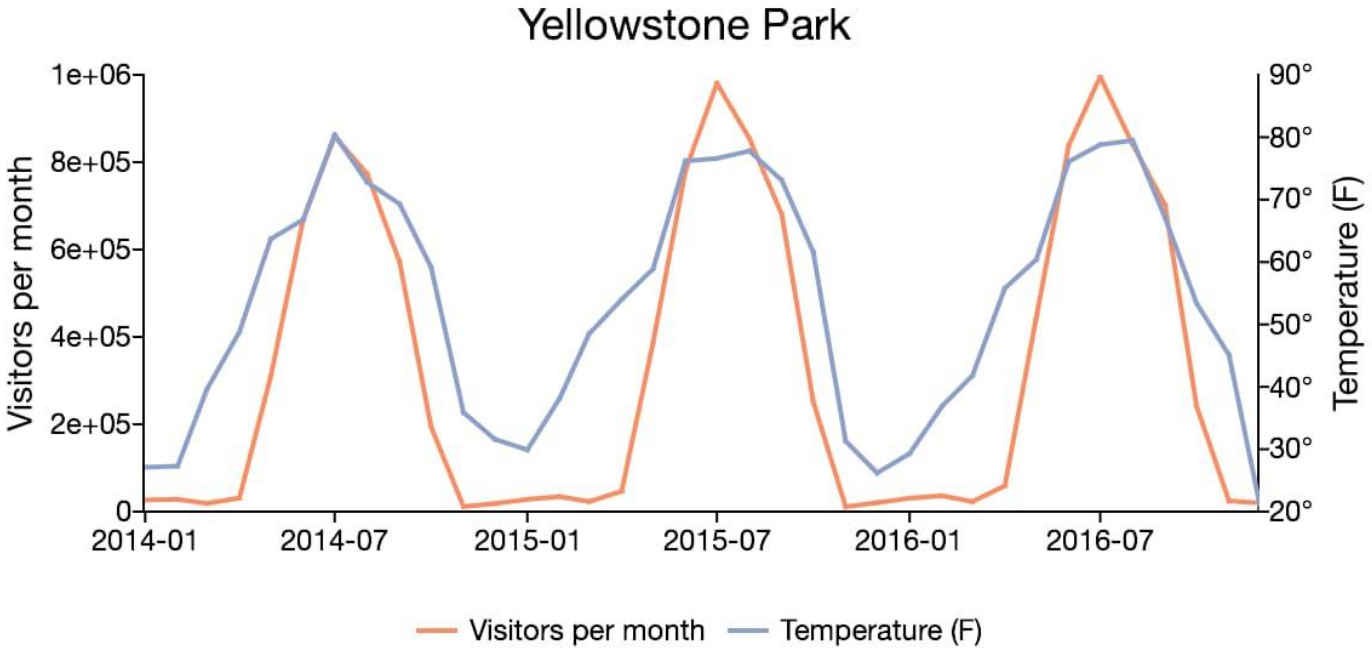
Dendrogram: **Shows How the Clusters are merged:**

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a*Dendrogram*.A clustering of the data objects is obtained by cutting the dendrogram at the desired level, and then each connected component forms a cluster.

Time Series analysis:

Define time series:

Time series data is a collection of quantities that are assembled over even intervals in time and ordered chronologically. The time interval at which data is collection is generally referred to as the time series frequency.



For example, the time series graph above plots the visitors per month to Yellowstone National Park with the

average monthly temperatures. The data ranges between January 2014 to December 2016 and is collected at a monthly frequency.

Define time series graph?

A time series graph plots observed values on the y-axis against an increment of time on the x-axis. These graphs visually highlight the behaviour and patterns of the data and can lay the foundation for building a reliable model.

More specifically, visualizing time series data provides a preliminary tool for detecting if data:

- Is mean-reverting or has explosive behavior;
- Has a time trend;
- Exhibits seasonality;
- Demonstrates structural breaks.

This, in turn, can help guide the testing, diagnostics, and estimation methods used during time series modelling and analysis.

Define seasonality:

Seasonality is another characteristic of time series data that can be visually identified in time series plots. Seasonality occurs when time series data exhibits regular and predictable patterns at time intervals that are smaller than a year.An example of a time series with seasonality is retail sales, which often increase between September to December and will decrease between January and February.

Difference between stationarity and time series:

A time series is stationary when all statistical characteristics of that series are unchanged by shifts in time. In technical terms, strict stationarity implies that the joint distribution of (yt,…,yt−h) depends only on the lag, h, and not on the time period, t. Strict stationarity is not widely necessary in time series analysis.This is not to imply that stationarity is not an important concept in time series analysis. Many time series models are valid only under the assumption of weak stationarity (also known as covariance stationarity).

Weak stationarity, henceforth stationarity, requires only that:

- A series has the same finite unconditional mean and finite unconditional variance at all time periods.
- That the series autocovariances are independent of time.

Nonstationary time series are any data series that do not meet the conditions of a weakly stationary time series.

Durbin Watson test:

Durbin-Watson Test:

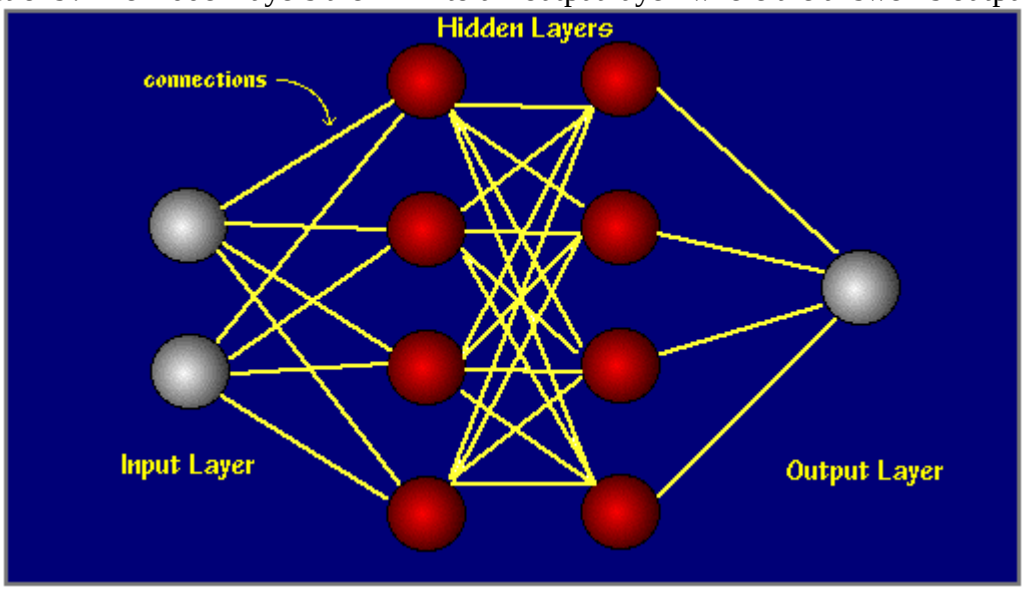
1. Estimate the model using ordinary least squares.
2. Predict the dependent variable using parameter estimates from Step One.
3. Compute the residuals by subtracted predicted dependent variables from the observed dependent variable.
4. Square and sum the residuals.
5. Compute the difference between the residual at each time period, t, and the previous time period, t-1. Then square the differences, and find the sum.
6. Compute the Durbin-Watson statistic by dividing the sum from Step Five by the sum in Step Four.

Neural network:

What Is A Neural Network?

The simplest definition of a neural network, more properly referred to as an 'artificial' neural network (ANN), is provided by the inventor of one of the first neurocomputers, Dr. Robert Hecht-Nielsen. He defines a neural network as:*"...a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.*In "Neural Network Primer: Part I" by Maureen Caudill, AI Expert, Feb. 1989
ANNs are processing devices (algorithms or actual hardware) that are loosely modeled after the neuronal structure of the mamalian cerebral cortex but on much smaller scales. A large ANN might have hundreds or thousands of processor units, whereas a mamalian brain has billions of neurons with a corresponding increase in magnitude of their overall interaction and emergent behavior. Although ANN researchers are generally not concerned with whether their networks accurately resemble biological systems, some have. For example, researchers have accurately simulated the function of the retina and modeled the eye rather well.Although the mathematics involved with neural networking is not a trivial matter, a user can rather easily gain at least an operational understanding of their structure and function.

Neural neworks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output as shown in the graphic below.



Most ANNs contain some form of 'learning rule' which modifies the weights of the connections according to the input patterns that it is presented with. In a sense, ANNs learn by example as do their biological counterparts; a child learns to recognize dogs from examples of dogs.

Although there are many different kinds of learning rules used by neural networks, this demonstration is concerned only with one; the delta rule. The delta rule is often utilized by the most common class of ANNs called 'backpropagational neural networks' (BPNNs). Backpropagation is an abbreviation for the backwards propagation of error.

With the delta rule, as with other types of backpropagation, 'learning' is a supervised process that occurs with each cycle or 'epoch' (i.e. each time the network is presented with a new input pattern) through a forward activation flow of outputs, and the backwards error propagation of weight adjustments.

Backpropagation performs a gradient descent within the solution's vector space towards a 'global minimum' along the steepest vector of the error surface. The global minimum is that theoretical solution with the lowest possible error. The error surface itself is a hyperparaboloid but is seldom 'smooth' as is depicted in the graphic

below. Indeed, in most problems, the solution space is quite irregular with numerous 'pits' and 'hills' which may cause the network to settle down in a 'local minum' which is not the best overall solution.

Since the nature of the error space can not be known aprioi, neural network analysis often requires a large number of individual runs to determine the best solution. Most learning rules have built-in mathematical terms to assist in this process which control the 'speed' (Beta-coefficient) and the 'momentum' of the learning. The speed of learning is actually the rate of convergence between the current solution and the global minimum. Momentum helps the network to overcome obstacles (local minima) in the error surface and settle down at or near the global miniumum.

Once a neural network is 'trained' to a satisfactory level it may be used as an analytical tool on other data. To do this, the user no longer specifies any training runs and instead allows the network to work in forward propagation mode only. New inputs are presented to the input pattern where they filter into and are processed by the middle layers as though training were taking place, however, at this point the output is retained and no backpropagation occurs. The output of a forward propagation run is the predicted model for the data which can then be used for further analysis and interpretation.

It is also possible to over-train a neural network, which means that the network has been trained exactly to respond to only one type of input; which is much like rote memorization. If this should happen then learning can no longer occur and the network is refered to as having been "grandmothered" in neural network jargon. In real-world applications this situation is not very useful since one would need a separate grandmothered network for each new kind of input.

How Do Neural Networks Differ From Conventional Computing?

To better understand artificial neural computing it is important to know first how a conventional 'serial' computer and it's software process information. A serial computer has a central processor that can address an array of memory locations where data and instructions are stored. Computations are made by the processor reading an instruction as well as any data the instruction requires from memory addresses, the instruction is then executed and the results are saved in a specified memory location as required. In a serial system (and a standard parallel one as well) the computational steps are deterministic, sequential and logical, and the state of a given variable can be tracked from one operation to another.

In comparison, ANNs are not sequential or necessarily deterministic. There are no complex central processors, rather there are many simple ones which generally do nothing more than take the weighted sum of their inputs from other processors. ANNs do not execute programed instructions; they respond in parallel (either simulated or actual) to the pattern of inputs presented to it. There are also no separate memory addresses for storing data. Instead, information is contained in the overall activation 'state' of the network. 'Knowledge' is thus represented by the network itself, which is quite literally more than the sum of its individual components.

What Applications Should Neural Networks Be Used For?

Neural networks are universal approximators, and they work best if the system you are using them to model has a high tolerance to error. One would therefore not be advised to use a neural network to balance one's cheque book! However they work very well for:

- capturing associations or discovering regularities within a set of patterns;
- where the volume, number of variables or diversity of the data is very great;
- the relationships between variables are vaguely understood; or,
- the relationships are difficult to describe adequately with conventional approaches.

What Are Their Limitations?

There are many advantages and limitations to neural network analysis and to discuss this subject properly we would have to look at each individual type of network, which isn't necessary for this general discussion. In reference to backpropagational networks however, there are some specific issues potential users should be aware of.

- Backpropagational neural networks (and many other types of networks) are in a sense the ultimate 'black boxes'. Apart from defining the general archetecture of a network and perhaps initially seeding it with a random numbers, the user has no other role than to feed it input and watch it train and await the output. In fact, it has been said that with backpropagation, "you almost don't know what you're doing". Some software freely available software packages (NevProp, bp, Mactivation) do allow the user to sample the networks 'progress' at regular time intervals, but the learning itself progresses on its own. The final product of this activity is a trained network that provides no equations or coefficients defining a relationship (as in regression) beyond it's own internal mathematics. The network 'IS' the final equation of the relationship.
- Backpropagational networks also tend to be slower to train than other types of networks and sometimes require thousands of epochs. If run on a truly parallel computer system this issue is not really a problem, but if the BPNN is being simulated on a standard serial machine (i.e. a single SPARC, Mac or PC) training can take some time. This is because the machines CPU must compute the function of each node and connection separately, which can be problematic in very large networks with a large amount of data. However, the speed of most current machines is such that this is typically not much of an issue.

LOGISTIC REGRESSION

Introduction:

Socio-economic variables are very often categorical, rather than interval scale. In many cases research focuses on models where the dependent variable is categorical. For example, the dependent variable might be ‘unemployed’ / ‘employed’, and we could be interested in how this variable is related to sex, age, ethnic group, etc. In this case we could not carry out a multiple linear regression as many of the assumptions of this technique will not be met, as will be explained theoretically below. Instead we would carry out a logistic regression analysis. Hence, logistic regression may be thought of as an approach that is similar to that of multiple linear regression, but takes into account the fact that the dependent variable is categorical. Categorical data and 2 x 2 tables We can write categorical data in two forms: list form or table form. The important point to make about this is that whichever way we choose to think about this kind of data, the information is the same. For example, if we were interested in the association between unemployment and sex for a sample of 12 people (this is a smaller sample than we would tend to use in general but it illustrates the point)

ODDS and Relative ODDS

Odds and Relative Odds A useful way of using the information in cross tabulations where one dimension of the table is an outcome of interest (whether 2x2 tables or more complicated ones), is to calculate odds and relative odds (odds ratios). Odds In the above table, the odds of a white boy being seen to have a behaviour problem are 19/90 = 0.21 or 0.21 to 1. In betting terms that is about 5 : 1 against – much less than even money. For black boys, the corresponding odds are 33/30 = 1.1, or 1.1 to 1. Equivalent to 11 to 10 on, (or a little better than even money.). Note that odds are not the same as probabilities – they are not restricted to the range 0 to 1. Relative odds We can also think of the information in the table in terms of relative odds. The relative odds of a black boy compared with a white boy being seen as having a behaviour problem are 1.1 / 0.21 or 5.2 to 1. In other words a black boy is 5.2 times more likely than a white boy to be seen as having a behaviour problem. Equally, boys perceived to have behaviour problems are 5.2 times more likely to be black rather than white, compared with boys without perceived behaviour problems. Relative odds are symmetrical in that sense; like correlation, we do not think of this measure in terms of a dependent variable and an explanatory variable. We just think in terms of the association between two variables.

LOGISTIC REGRESSION THEORY:

When we want to look at a dependence structure, with a dependent variable and a set of explanatory variables (one or more), we can use the logistic regression framework. Multiple linear regression may be used to investigate the relationship between a continuous (interval scale) dependent variable, such as income, blood pressure or examination score. However, socio-economic variables are very often categorical, rather than interval scale. In many cases research focuses on models where the dependent variable is categorical.

For example, the dependent variable might be ‘unemployed’ or ‘not’ (as we saw in Exercise 1) , and we could be interested in how this variable is related to sex, age, ethnic group, etc. In this case we could not carry out a multiple linear regression as many of the assumptions of this technique will not be met, as will be explained theoretically below. Instead we would carry out a logistic regression.

The Theory If we wrote the ‘perceived behaviour problems’ table as data in list format, we would be interested in modelling the variation in the probability of being perceived to have behaviour problems, and for the table data we are interested in modelling the variations in the proportions with perceived behaviour problems amongst black boys compared with white boys. It is important to note that regardless of whether we consider the analysis in terms of data in a list or a table, the results will be exactly the same.

Proportions and probabilities are different from continuous variables in a number of ways. They are bounded by 0 and 1, whereas in theory continuous variables can take any value between plus or minus infinity. This means that we cannot assume normality for a proportion, and we must recognise that proportions have a binomial distribution. Unlike the normal distribution, the mean and variance of the Binomial distribution are not independent. The mean is denoted by P and the variance is denoted by P*(1-P)/n, where n is the number of observations, and P is the probability of the event occurring (e.g. the probability of being unemployed, or having ‘perceived behavioural problems’) in any one ‘trial’ (for any one individual in this example). If we were considering the data in ‘list’ rather than table form we would assume that the variable had a mean P and a variance P*(1-P) and this variable would have a Bernoulli distribution.

When we have a proportion as a response, we use a logistic or logit transformation to link the dependent variable to the set of explanatory variables. The logit link has the form: Logit (P) = Log [P / (1-P)] The term within the square brackets is the odds of an event occurring. In the example above this would be the odds of a person being perceived to have behaviour problems. Using the logit scale changes the scale of a proportion to plus and minus infinity, and also because Logit (P) = 0, when P=0.5. When we transform our results back from the logit (log odds) scale to the original probability scale, our predicted values will always be at least 0 and at most 1.

Video:

SAS codes and R programming videos will be shared with students. It is available in you tube also.