Data Mining and Business Intelligence

(22MCA252)

Q.1a. Define Data warehouse. Explain the key features of data warehouse.

- Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions.
- A data warehouse refers to a database that is maintained separately from an organization's operational databases.
- Data warehouse systems allow for the integration of a variety of application systems.
- They support information processing by providing a solid platform of consolidated historical data for analysis.
- According to William H. Inmon, a leading architect in the construction of data warehouse systems, "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process"
- The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems.

  - Subject-oriented:
    - A data warehouse is organized around major subjects, such as customer, supplier, product, and sales.
    - Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers.
    - Data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
  - Integrated:
    - A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.
    - Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.
  - Time-variant:

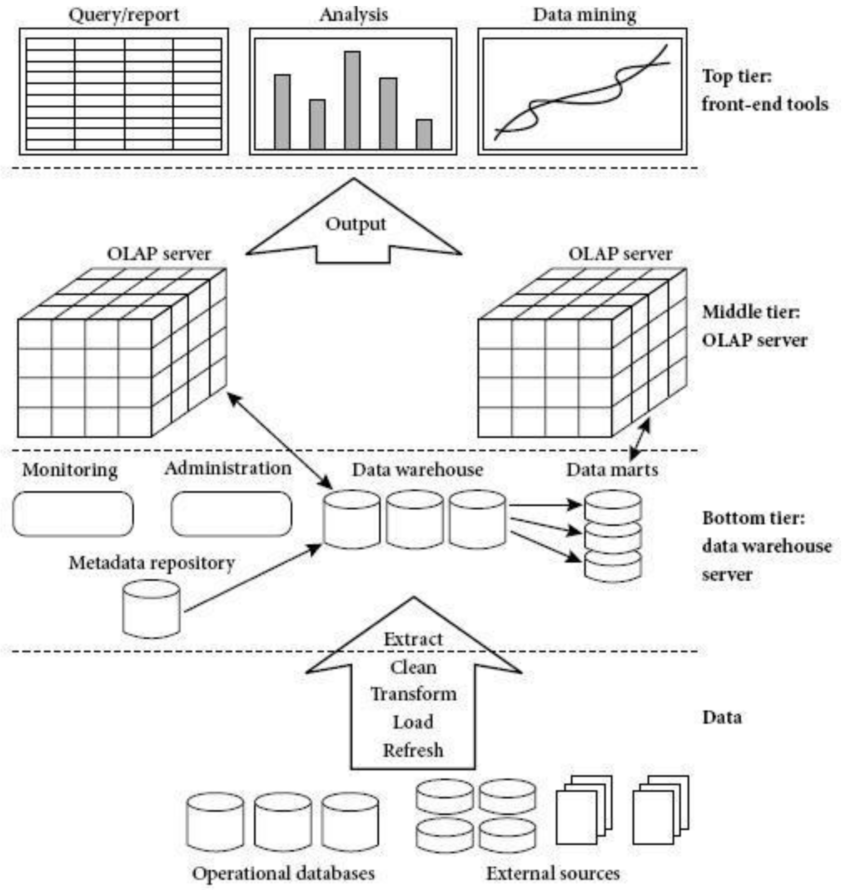
    - Data are stored to provide information from a historical perspective (e.g., the past 5–10 years).
    - Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
- Nonvolatile:
    - A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.
    - Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms.
    - It usually requires only two operations in data accessing: initial loading of data and access of data.

b. With a neat diagram explain the three tier data ware house architecture.

Bottom tier:

- The bottom tier is a warehouse database server that is almost always a relational database system.
- Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources.
- These tools and utilities perform data extraction, cleaning, and transformation, as well as load and refresh functions to update the data warehouse.
- The data are extracted using application program interfaces known as gateways.
- A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.
- Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection).

- This tier also contains a metadata repository, which stores information about the data warehouse and its contents.



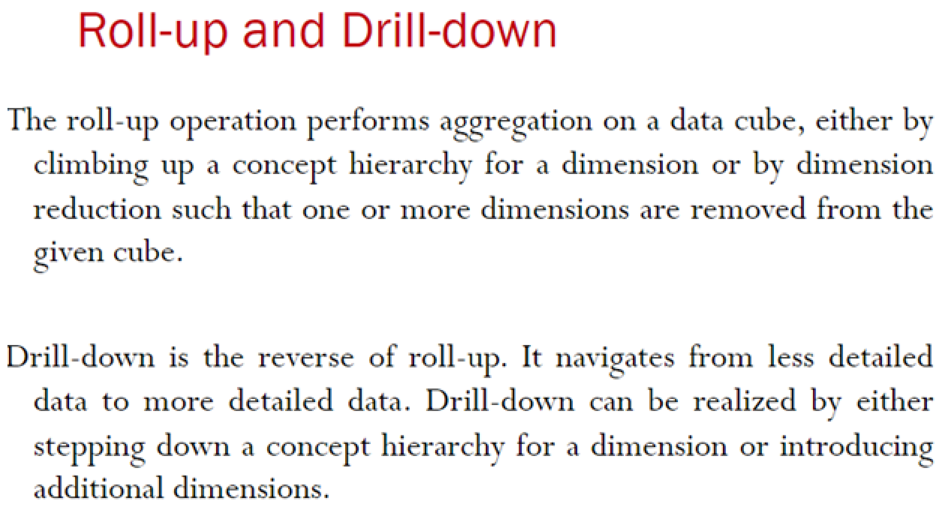
Middle tier:

- The middle tier is an OLAP server that is typically implemented using either.
- A relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations or,
- A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

Top tier:

- The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools.

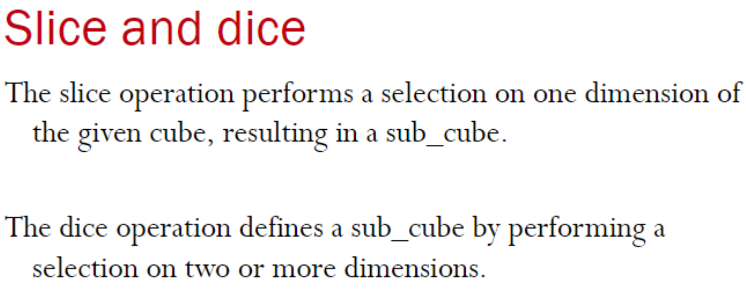
c. Give differences between i) Rollup and Drill down ii) Slice and dice



Roll-up and Drill-down

The roll-up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction such that one or more dimensions are removed from the given cube.

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.



Slice and dice

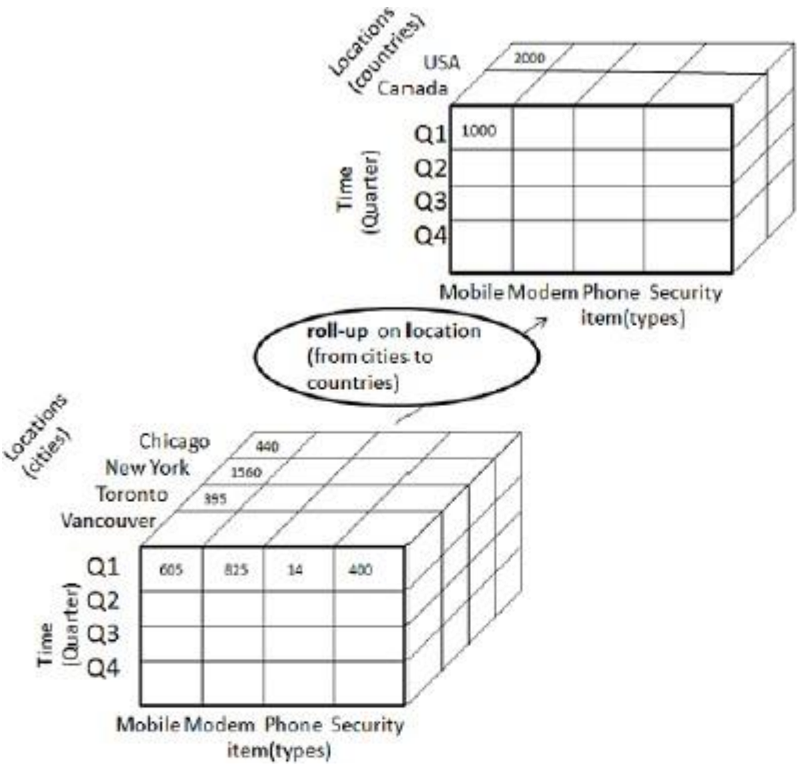
The slice operation performs a selection on one dimension of the given cube, resulting in a sub_cube.

The dice operation defines a sub_cube by performing a selection on two or more dimensions.

Q2. A. Discuss typical OLAP operations with an example.

1. Roll-up

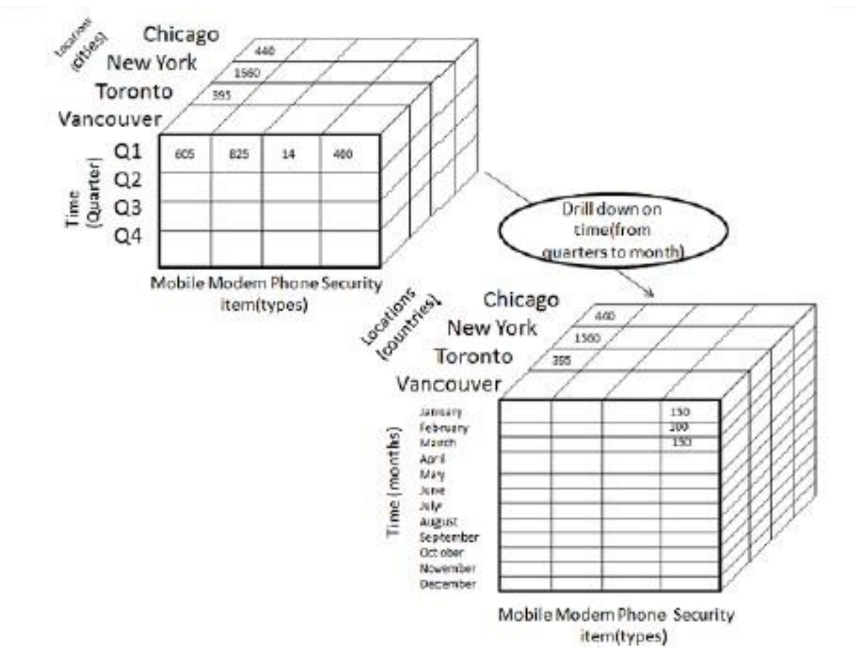
- Roll-up performs aggregation on a data cube in any of the following ways:
  - By climbing up a concept hierarchy for a dimension
  - By dimension reduction
- The following diagram illustrates how roll-up works.



- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

2. Drill-down

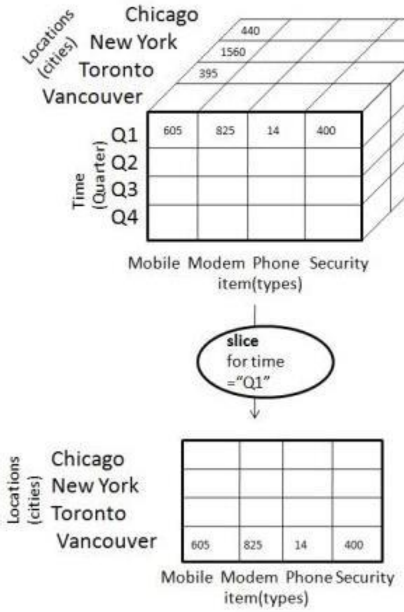
- Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:
  - By stepping down a concept hierarchy for a dimension
  - By introducing a new dimension.
- The following diagram illustrates how drill-down works:



- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

3. Slice

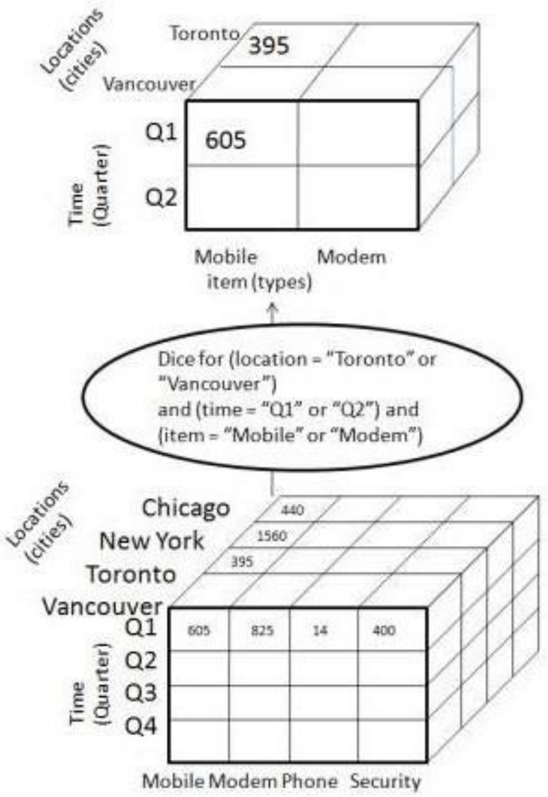
- The slice operation selects one particular dimension from a given cube and provides a new subcube.
- Consider the following diagram that shows how slice works.



- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

4. Dice

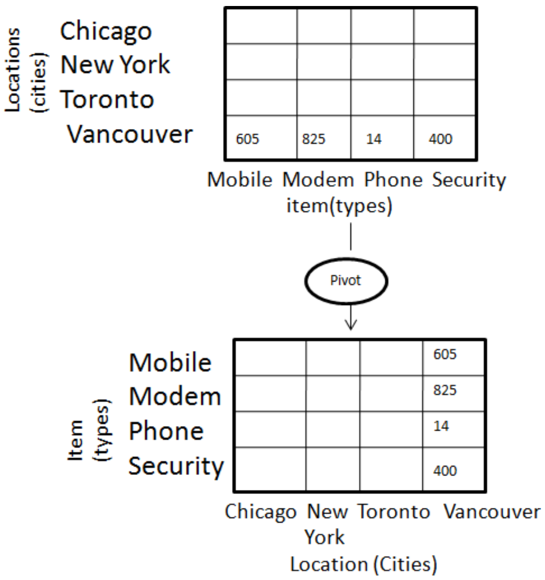
- Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- Consider the following diagram that shows the dice operation.



- The dice operation on the cube based on the following selection criteria involves three dimensions.
  - (location = "Toronto" or "Vancouver")
  - (time = "Q1" or "Q2")
  - (item =" Mobile" or "Modem")

5. Pivot

- The pivot operation is also known as rotation.
- It rotates the data axes in view in order to provide an alternative presentation of data.
- Consider the following diagram that shows the pivot operation.
- In this the item and location axes in 2-D slice are rotated.



b. Explain the following terms.

i)Data Mart

□Data marts contain a subset of organization-wide data that is valuable to specific groups of people in an organization.

□A data mart contains only those data that is specific to a particular group.

□Data marts improve end-user response time by allowing users to have access to the specific type of data they need to view most often by providing the data in a way that supports the collective view of a group of users.

□A data mart is basically a condensed and more focused version of a data warehouse that reflects the regulations and process specifications of each business unit within an organization.

□Each data mart is dedicated to a specific business function or region.

□For example, the marketing data mart may contain only data related to items, customers, and sales. Data marts are confined to subjects.

ii)Virtual Warehouse

- A virtual warehouse is a set of views over operational databases.

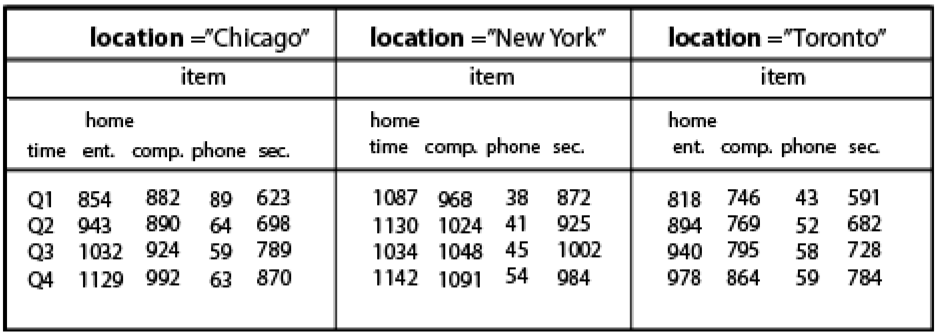
For efficient query processing, only some of the possible summary views may be materialized.

iii)Data Cube

When data is grouped or combined in multidimensional matrices called Data Cubes. The data cube method has a few alternative names or a few variants, such as "Multidimensional databases," "materialized views," and "OLAP (On-Line Analytical Processing)." The general idea of this approach is to materialize certain expensive computations that are frequently inquired.

Let suppose we would like to view the sales data with a third dimension. For example, suppose we would like to view the data according to time, item as well as the location for the cities Chicago, New York, Toronto, and Vancouver. The measured display in dollars sold (in thousands). These 3-D data are shown in the table. The 3-D data of the table are represented as a series of 2-D tables.

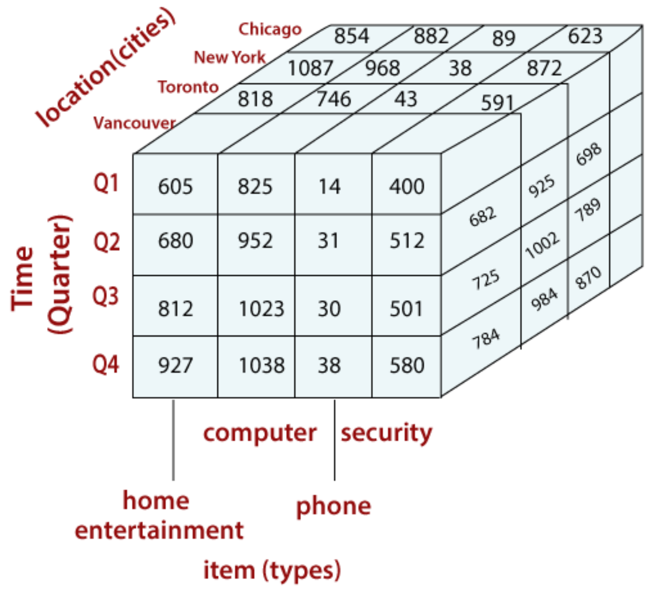
3-D view of Sales Data



location ="Chicago"					location ="New York"				location ="Toronto"			
item					item				item			
time	home ent.	comp.	phone	sec.	home time	comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784

Conceptually, we may represent the same data in the form of 3-D data cubes, as shown in fig:

3-D Data Cube



iv) ROLAP and MOLAP

Q3. A. What is data mining? Explain the process of knowledge discovery in Database (KDD) with a neat diagram.

□ KDD stands for knowledge discoveries from database. There are some pre-processing operations which are required to make pure data in data warehouse before use that data for Data Mining processes.

□ A view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in Figure 2 and consists of an iterative sequence of the following steps:

✓ Data cleaning: To remove noise and inconsistent 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, for instance.

✓ 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.

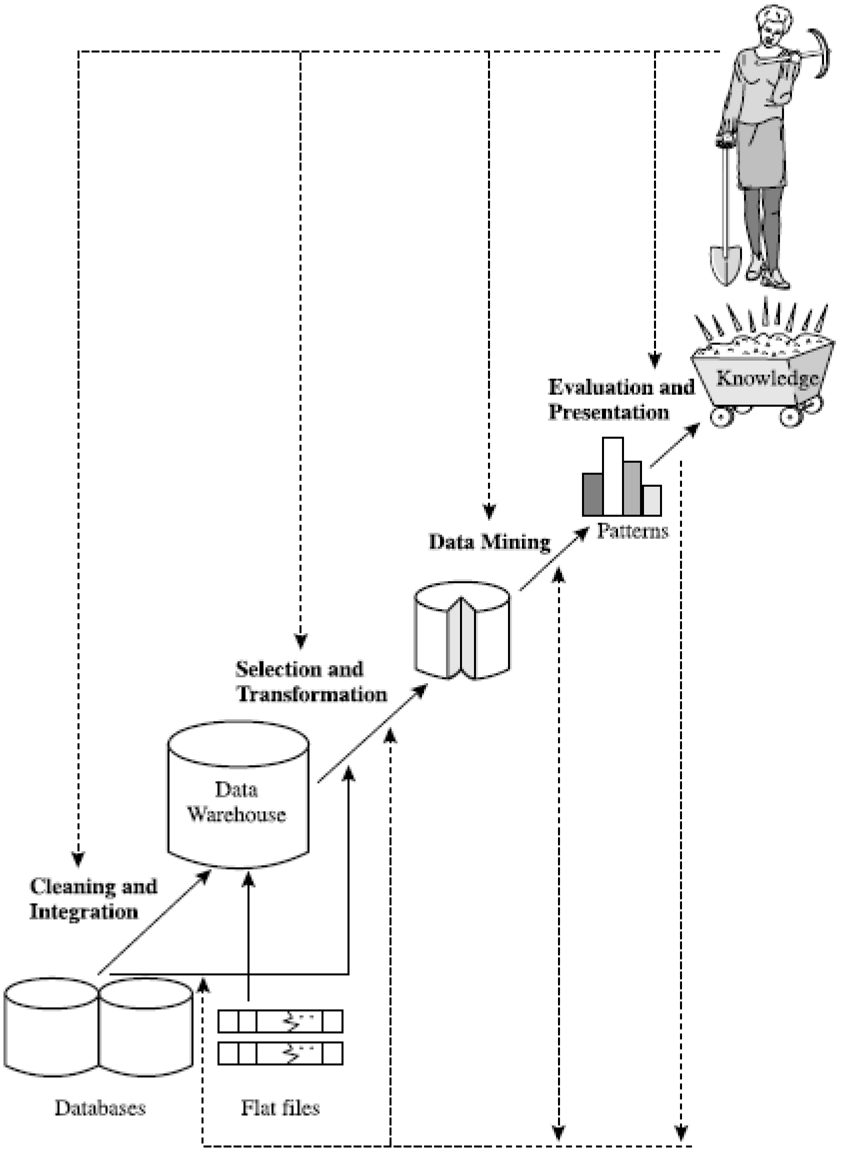


Fig. 2 Data mining as a step in the process of knowledge discovery

- KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.
- Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.
- Objective of Pre-processing on data is to remove noise from data or to remove redundant data.
- There are mainly 4 types of Pre-processing Activities included in KDD Process that is shown in fig. as Data cleaning, Data integration, Data transformation, Data reduction.

b. Explain the classification of data mining systems.

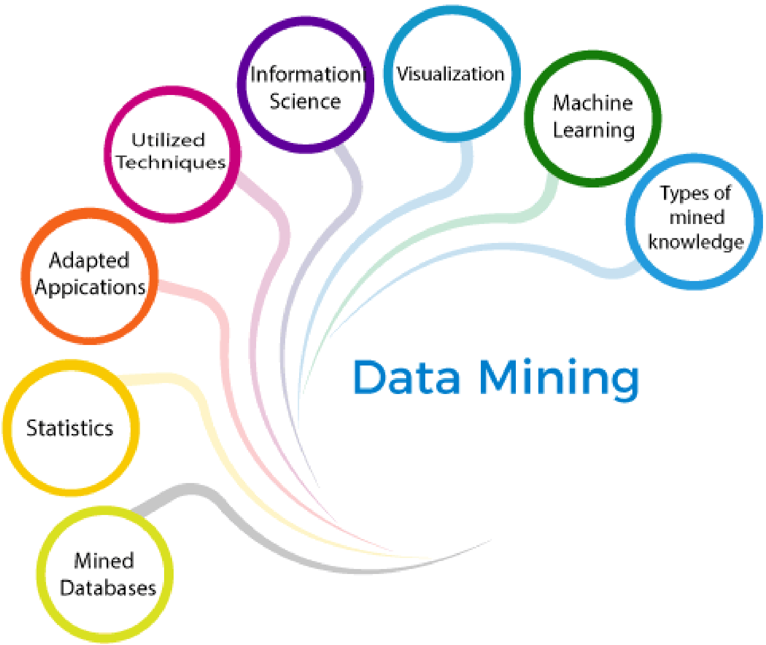
Classification of Data Mining Systems

Data mining refers to the process of extracting important data from raw data. It analyses the data patterns in huge sets of data with the help of several software. Ever since the development of data mining, it is being incorporated by researchers in the research and development field.

With Data mining, businesses are found to gain more profit. It has not only helped in understanding customer demand but also in developing effective strategies to enforce overall business turnover. It has helped in determining business objectives for making clear decisions.

Data collection and data warehousing, and computer processing are some of the strongest pillars of data mining. Data mining utilizes the concept of mathematical algorithms to segment the data and assess the possibility of occurrence of future events.

To understand the system and meet the desired requirements, data mining can be classified into the following systems:



- Classification based on the mined Databases
- Classification based on the type of mined knowledge
- Classification based on statistics
- Classification based on Machine Learning

- Classification based on visualization
- Classification based on Information Science
- Classification based on utilized techniques
- Classification based on adapted applications

Classification Based on the mined Databases

A data mining system can be classified based on the types of databases that have been mined. A database system can be further segmented based on distinct principles, such as data models, types of data, etc., which further assist in classifying a data mining system.

For example, if we want to classify a database based on the data model, we need to select either relational, transactional, object-relational or data warehouse mining systems.

Classification Based on the type of Knowledge Mined

A data mining system categorized based on the kind of knowledge mind may have the following functionalities:

1. Characterization
2. Discrimination
3. Association and Correlation Analysis
4. Classification
5. Prediction
6. Outlier Analysis
7. Evolution Analysis

Classification Based on the Techniques Utilized

A data mining system can also be classified based on the type of techniques that are being incorporated. These techniques can be assessed based on the involvement of user interaction involved or the methods of analysis employed.

Classification Based on the Applications Adapted

Data mining systems classified based on adapted applications adapted are as follows:

1. Finance
2. Telecommunications
3. DNA
4. Stock Markets
5. E-mail

Examples of Classification Task

Following is some of the main examples of classification tasks:

- Classification helps in determining tumor cells as benign or malignant.
- Classification of credit card transactions as fraudulent or legitimate.
- Classification of secondary structures of protein as alpha-helix, beta-sheet, or random coil.
- Classification of news stories into distinct categories such as finance, weather, entertainment, sports, etc.

Q.4.a. What is data processing? What are the steps involved in it? Explain any 2 in detail.

- Today’s real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources.
- Low-quality data will lead to low-quality mining results. How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process?
- Data have quality if they satisfy the requirements of the intended use. There are many

factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.

- Inaccurate, incomplete, and inconsistent data are commonplace properties of large real-world databases and data warehouses.
- There are many possible reasons for inaccurate data (i.e., having incorrect attribute values). The data collection instruments used may be faulty.
- There may have been human or computer errors occurring at data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value "January 1" displayed for birthday). This is known as disguised missing data. Errors in data transmission can also occur.
- There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date).
- 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 they were 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 data history or modifications may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.
- Data Preprocessing Methods/Techniques:
  - Data Cleaning routines work to "clean" the data by filling in missing values, smoothing noisy

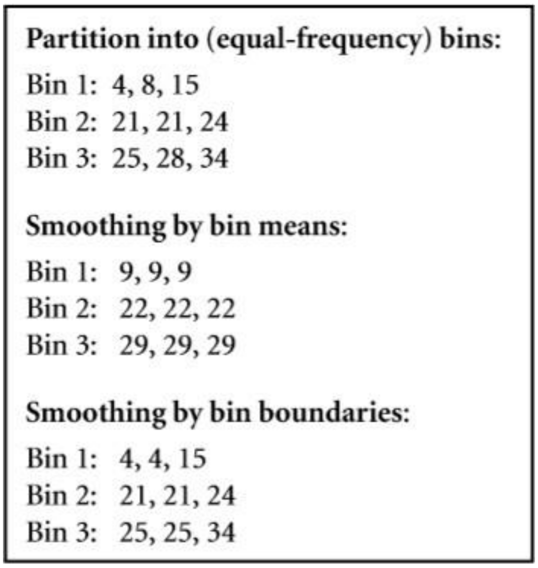
    data, identifying or removing outliers, and resolving inconsistencies.
  - Data Integration which combines data from multiple sources into a coherent data store, as in data warehousing.
  - Data Transformation, the data are transformed or consolidated into forms appropriate for mining

    - Data Reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.
- Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.
- Missing Values: Imagine that you need to analyze AllElectronics sales and customer data. You note that many tuples have no recorded value for several attributes such as customer income. How can you go about filling in the missing values for this attribute? Let's look at the following methods.
    - Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
    - By ignoring the tuple, we do not make use of the remaining attributes values in the tuple. Such data could have been useful to the task at hand.
    - Fill in the missing value manually: In general, this approach is time consuming and may not be feasible given a large data set with many missing values.
    - Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant such as a label like "Unknown" or 1. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not foolproof.
    - Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value: For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median.
    - For example, suppose that the data distribution regarding the income of AllElectronics customers is symmetric and that the mean income is $56,000. Use this value to replace the missing value for income.
    - Use the attribute mean or median for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit

risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.

  - Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

- Noisy Data: Noise is a random error or variance in a measured variable. Given a numeric attribute such as say, price, how can we "smooth" out the data to remove the noise? Let's look at the following data smoothing techniques.
  - Binning: Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighborhood of values, they perform local smoothing.
  - Figure 1 illustrates some binning techniques. In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values).
  - In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
  - For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be equal width, where the interval range of values in each bin is constant.
  - Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34



Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Fig. 1: Binning methods for data smoothing

- Regression: Data smoothing can also be done by regression, a technique that conforms data values to a function. Linear regression involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.

  Multiple linear regression is an extension of linear regression, where more than two attributes are

  involved and the data are fit to a multidimensional surface.

- Outlier analysis: Outliers may be detected by clustering, for example, where similar values are organized into groups, or “clusters.” Intuitively, values that fall outside of the set of clusters may be considered outliers.

- In data transformation, the data are transformed or consolidation to forms appropriate for mining. Strategies for data transformation include the following:

- Smoothing, which works to remove noise from the data. Techniques include binning, regression, and clustering.

- Attribute construction (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

- Aggregation, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for data analysis at multiple abstraction levels.

- Normalization, where the attribute data are scaled so as to fall within a smaller range, such as−1.0

to 1.0, or 0.0 to 1.0.

Example: Data Transformation -2, 32, 100, 59, 48⟶

- Discretization, where the raw values of a numeric attribute (e.g. Age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The labels, in turn, can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numeric attribute. Figure 2 shows a concept hierarchy for the attribute price. More than one concept hierarchy can be defined for the same attribute to accommodate the needs of various users.

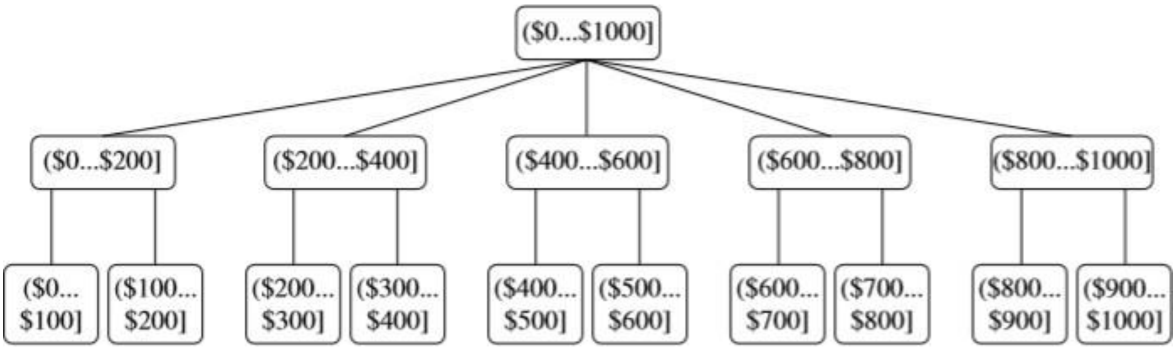


Fig. 2 A concept hierarchy for the attribute price, where an interval ($X... $Y] denotes the range from $X (exclusive) to $Y (inclusive).

- Concept hierarchy generation for nominal data, where attributes such as street can be generalized to higher-level concepts, like city or country. Many hierarchies for nominal attributes are implicit within the database schema and can be automatically defined at the schema definition level.

b. Explain any two data mining primitives in detail.

A data mining query is defined in terms of the following primitives, such as:

1. The set of task-relevant data to be mined

This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (the relevant attributes or dimensions).

In a relational database, the set of task-relevant data can be collected via a relational query involving operations like selection, projection, join, and aggregation.

The data collection process results in a new data relational called the initial data relation. The initial data relation can be ordered or grouped according to the conditions specified in the query. This data retrieval can be thought of as a subtask of the data mining task.

This initial relation may or may not correspond to physical relation in the database. Since virtual relations are called Views in the field of databases, the set of task-relevant data for data mining is called a minable view.

2. The kind of knowledge to be mined

This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

3. The background knowledge to be used in the discovery process

This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction.

Concept hierarchy defines a sequence of mappings from low-level concepts to higher-level, more general concepts.

- Rolling Up - Generalization of data: Allow to view data at more meaningful and explicit abstractions and makes it easier to understand. It compresses the data, and it would require fewer input/output operations.
- Drilling Down - Specialization of data: Concept values replaced by lower-level concepts. Based on different user viewpoints, there may be more than one concept hierarchy for a given attribute or dimension.

An example of a concept hierarchy for the attribute (or dimension) age is shown below. User beliefs regarding relationships in the data are another form of background knowledge.

Q5. A. What is the concept of description? Explain data generalization and summarization based characterization in detail.

From Data Analysis point of view, data mining can be classified into two categories: Descriptive mining and predictive mining.
Descriptive mining: It describes the data set in a concise and summative manner and presents interesting general properties of data.

Predictive mining: It analyzes the data to construct one or a set of models, and attempts to predict the behavior of new data sets.

Databases usually store a large amount of data in great detail. However, users often like to view sets of summarized data in concise, descriptive terms.

Such data descriptions may provide an overall picture of a class of data or distinguish it from a set of comparative classes.
Such descriptive data mining is called concept descriptions and forms an important component of data mining.

The simplest kind of descriptive data mining is called concept description. A concept usually refers to a collection of data such as frequent_buyers, graduate_students and so on.

As data mining task concept description is not a simple enumeration of the data. Instead, concept description generates descriptions for characterization and comparison of the data.

It is sometimes called class description when the concept to be described refers to a class of objects

- Characterization: It provides a concise and succinct summarization of the given collection of data.
- Comparison: It provides descriptions comparing two or more collections of data.

Data Generalization & Summarization

Data and objects in databases contain detailed information at the primitive concept level.
For example, the item relation in a sales database may contain attributes describing low-level item information such as item_ID, name, brand, category, supplier, place_made and price.
It is useful to be able to summarize a large set of data and present it at a high conceptual level.

For example, summarizing a large set of items relating to Christmas season sales provides a general description of such data, which can be very helpful for sales and marketing managers.
This requires an important functionality called data generalization.

Data Generalization

A process that abstracts a large set of task-relevant data in a database from a low conceptual level to higher ones.

Data Generalization is a summarization of general features of objects in a target class and produces what is called characteristic rules.

The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

For example, one may want to characterize the "OurVideoStore" customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the attribute-oriented induction method can be used, for example, to carry out data summarization.

Note that with a data cube containing a summarization of data, simple OLAP operations fit the purpose of data characterization.

Approaches:

- Data cube approach(OLAP approach).
- Attribute-oriented induction approach.

Presentation Of Generalized Results

Generalized Relation:

- Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.

Cross-Tabulation:

- Mapping results into cross-tabulation form (similar to contingency tables).

Visualization Techniques:

- Pie charts, bar charts, curves, cubes, and other visual forms.

Quantitative characteristic rules:

- Mapping generalized results in characteristic rules with quantitative information associated with it.

b. Explain Apriori algorithms for frequent item set generation in detail.

- **Purpose**: The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
- **Key Concepts**:
  - **Frequent Itemsets**: The sets of item which has minimum support (denoted by L_i for ith-Itemset).
  - **Apriori Property**: Any subset of frequent itemset must be frequent.
  - **Join Operation**: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} itself.
    - Find the frequent itemsets: the sets of items that have minimum support – A subset of a frequent itemset must also be a frequent itemset **(Apriori Property)**
    - i.e., if {AB} is a frequent itemset, both {A} and {B} should be a frequent itemset –

Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)

- Use the frequent itemsets to generate association rules.

- **The Apriori Algorithm : Pseudo code**
  - **Join Step**: C k is generated by joining Lk-1with itself
  - **Prune Step**: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

- Pseudo-code:

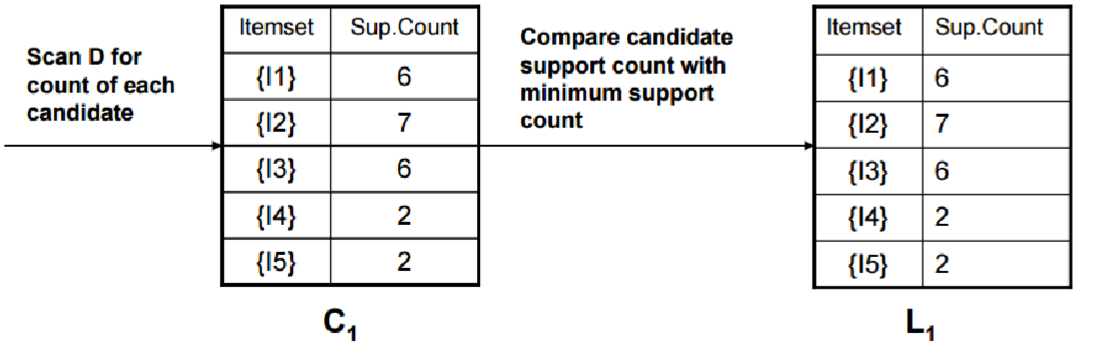
```
$C_k$: Candidate itemset of size
k $L_k$: frequent itemset of
size k $L_1$= {frequent
items};
for (k = 1; $L_k$ != ∅ ; k++) do begin
        $C_{k+1}$ = candidates generated from $L_k$;
for each transaction t in database do
        Increment the count of all candidates in
        $C_{k+1}$ That are contained in t
        $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
        end
        return $\cup_k L_k$;
```

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

Example

- Consider a database, **D**, consisting of 9 transactions.
- Suppose min. support count required is **2** (i.e. min_sup = 2/9 = 22 %)
- Let minimum confidence required is **70%**.
- We have to first find out the frequent itemset using Apriori algorithm.
- Then, Association rules will be generated using min. support & min. confidence.

Step 1: Generating 1-itemset Frequent Pattern



Scan D for count of each candidate

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

C_1

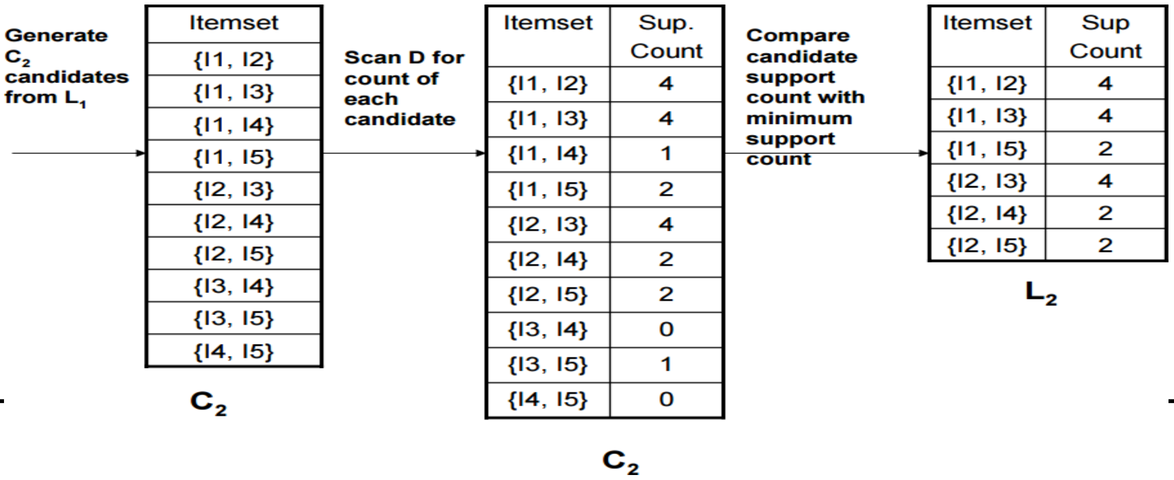
Compare candidate support count with minimum support count

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

L_1

- **The set of frequent 1-itemsets, L1**, consists of the candidate 1- itemsets satisfying minimum support.
- In the first iteration of the algorithm, each item is a member of the set of candidate.

Step 2: Generating 2-itemset Frequent Pattern



Generate C_2 candidates from L_1

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

C_2

Scan D for count of each candidate

Itemset	Sup. Count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

C_2

Compare candidate support count with minimum support count

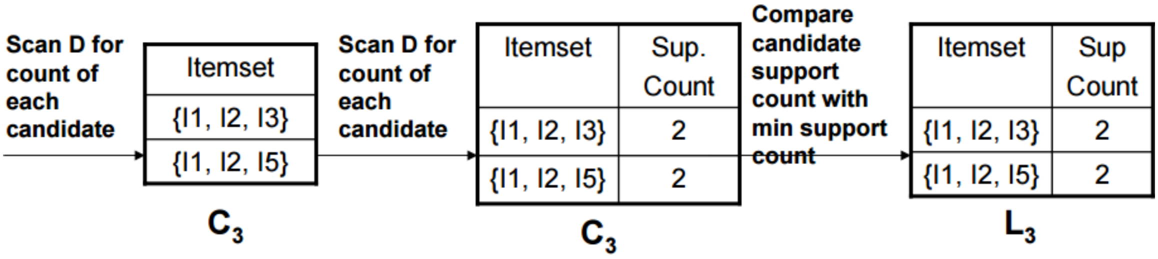
Itemset	Sup Count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

L_2

- To discover the set of frequent 2-itemsets, L2, the algorithm uses L1 Join L1 to generate a candidate set of 2-itemsets, C2.
- Next, the transactions in D are scanned and the support count for each candidate itemset in C2 is accumulated (as shown in the middle table).
- The set of frequent 2-itemsets, L2, is then determined, consisting of those candidate 2-itemsets in C2 having minimum support.

- **Note**: We haven't used Apriori Property yet.

Step 3: Generating 3-itemset Frequent Pattern



Scan D for count of each candidate

Itemset
{I1, I2, I3}
{I1, I2, I5}

C_3

Scan D for count of each candidate

Itemset	Sup. Count
{I1, I2, I3}	2
{I1, I2, I5}	2

C_3

Compare candidate support count with min support count

Itemset	Sup Count
{I1, I2, I3}	2
{I1, I2, I5}	2

L_3

- The generation of the set of candidate 3-itemsets, C3 , involves use of the Apriori Property.
- In order to find C3, we compute L2 Join L2.
- C3 = L2 join L2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}.
- Now, Join step is complete and Prune step will be used to reduce the size of C3. Prune step helps to avoid heavy computation due to large Ck.
- Based on the **Apriori property** that all subsets of a frequent itemset must also be frequent, we can determine that four latter candidates cannot possibly be frequent. How ?
- For example, lets take **{I1, I2, I3}**. The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2- item subsets of {I1, I2, I3} are members of L2, We will keep {I1, I2, I3} in C3.
- Lets take another example of **{I2, I3, I5}** which shows how the pruning is performed. The 2-item subsets are {I2, I3}, {I2, I5} & {I3,I5}.
- But, {I3, I5} is not a member of L2 and hence it is not frequent **violating Apriori Property**. Thus We will have to remove {I2, I3, I5} from C3.

- Therefore, C3 = {{I1, I2, I3}, {I1, I2, I5}} after checking for all members of result of Join operation for Pruning.
- Now, the transactions in D are scanned in order to determine **L3, consisting of those candidates 3- itemsets in C3 having minimum support.**

- **Step 4: Generating 4-itemset Frequent Pattern**
  - The algorithm uses L3 Join L3 to generate a candidate set of 4-itemsets, **C4**. Although the join results in {{I1, I2, I3, I5}}, this itemset is pruned since its subset {{I2, I3, I5}} is not frequent.
  - Thus, **C4 = φ**, and algorithm terminates, **having found all of the frequent items. This completes our Apriori Algorithm.** What's Next?
  - These frequent itemsets will be used to generate **strong association rules** (where strong association rules satisfy both minimum support & minimum confidence).
- **Step 5: Generating Association Rules from Frequent Itemsets**

Procedure:

- For each frequent itemset "l", generate all nonempty subsets of l.
- For every nonempty subset s of l, output the rule "s -> (l-s)" if support_count(l) /

  support_count(s) >= min_conf where min_conf is minimum confidence threshold.

Back to Example:

- We had L = {{I1}, {I2}, {I3}, {I4}, {I5}, {I1, I2}, {I1, I3}, {I1, I5}, {I2, I3}, {I2, I4}, {I2, I5}, {I1, I2, I3}, {I1, I2, I5}}.
- Let's take l = {I1, I2, I5}. – It's all nonempty subsets are {I1, I2}, {I1, I5}, {I2, I5}, {I1}, {I2}, {I5}.
- Let **minimum confidence threshold** is, say 70%.
- The resulting association rules are shown below, each listed with its confidence.
- R1: I1 ^ I2 -> I5 Confidence = sc{I1, I2, I5}/sc{I1,I2} = 2/4 = 50% (R1 is Rejected)
- R2: I1 ^ I5 -> I2 Confidence = sc{I1, I2, I5}/sc{I1,I5} = 2/2 = 100% (**R2 is Selected**)
- R3: I2 ^ I5 -> I1 Confidence = sc{I1, I2, I5}/sc{I2,I5} = 2/2 = 100% (**R3 is Selected**)
- R4: I1 -> I2 ^ I5 Confidence = sc{I1, I2, I5}/sc{I1} = 2/6 = 33% (R4 is Rejected)

- R5: I2 -> I1 ^ I5 Confidence = sc{I1, I2, I5}/{I2} = 2/7 = 29% (R5 is Rejected)
- R6: I5 -> I1 ^ I2 Confidence = sc{I1, I2, I5}/ {I5} = 2/2 = 100% (**R6 is Selected**)
- In this way, we have found **three strong association rules**.

Q6.a. Explain market basket analysis in detail.

- Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. For example, if you are in a store and you buy a milk and don't buy a bread, you are more likely to buy eggs at the same time than somebody who didn't buy bread.
- The set of items a customer buys is referred to as an itemset, and market basket analysis seeks to find relationships between purchases.
- Typically, the relationship will be in the form of a rule:

  e.g IF {milk, eggs} THEN {bread}.
- The probability that a customer will buy milk without an eggs (i.e. that the antecedent is true) is referred to as the support for the rule. The conditional probability that a customer will purchase bread is referred to as the confidence.
- The algorithms for performing market basket analysis are fairly straightforward. The complexities mainly arise in exploiting taxonomies, avoiding combinatorial explosions (a supermarket may stock 10,000 or more line items), and dealing with the large amounts of transaction data that may be available.
- A major difficulty is that a large number of the rules found may be trivial for anyone familiar with the business. Although the volume of data has been reduced, we are still asking the user to find a needle in a haystack.

- Requiring rules to have a high minimum support level and a high confidence level risks missing any exploitable result we might have found. One partial solution to this problem is differential market basket analysis, as described below.

- ❖ How is it used?
    - In retailing, most purchases are bought on impulse. Market basket analysis gives clues as to what a customer might have bought if the idea had occurred to them.
    - As a first step, therefore, market basket analysis can be used in deciding the location and promotion of goods inside a store. If, as has been observed, purchasers of Barbie dolls have are more likely to buy candy, then high-margin candy can be placed near to the Barbie doll display. Customers who would have bought candy with their Barbie dolls had they thought of it will now be suitably tempted.
    - But this is only the first level of analysis. Differential market basket analysis can find interesting results and can also eliminate the problem of a potentially high volume of trivial results.
    - In differential analysis, we compare results between different stores, between customers in different demographic groups, between different days of the week, different seasons of the year, etc.
    - If we observe that a rule holds in one store, but not in any other (or does not hold in one store, but holds in all others), then we know that there is something interesting about that store. Perhaps its clientele is different, or perhaps it has organized its displays in a novel and more lucrative way. Investigating such differences may yield useful insights which will improve company sales.
- ❖ Application Areas
    - Although Market Basket Analysis conjures up pictures of shopping carts and supermarket shoppers, it is important to realize that there are many other areas in which it can be applied. These include:
        - Analysis of credit card purchases.
        - Analysis of telephone calling patterns.
        - Identification of fraudulent medical insurance claims. (Consider cases where common rules are broken).
        - Analysis of telecom service purchases.

3) Association Rule

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

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

Before we start defining the rule, let us first see the basic definitions.

Support Count( ) – Frequency of occurrence of a itemset.

Here ({Milk, Bread, Diaper})=2

Frequent Itemset – An itemset whose support is greater than or equal to minsup threshold.

Association Rule – An implication expression of the form X -> Y, where X and Y are any 2 itemsets.

Example: {Milk, Diaper}->{Beer}

Rule Evaluation Metrics –

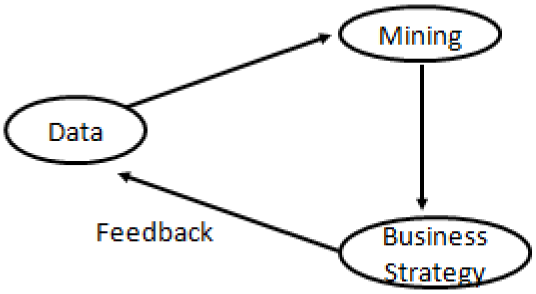
- Support(s) –
  The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction.It is a measure of how

frequently the collection of items occur together as a percentage of all transactions.

- Support = (X+Y) total –
  It is interpreted as fraction of transactions that contain both X and Y.
- Confidence(c) –
  It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
- Conf(X=>Y) = Supp(X Y) Supp(X) –
  It measures how often each item in Y appears in transactions that contains items in X also.
- Lift(l) –
  The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the itemsets X and Y are independent of each other.The expected confidence is the confidence divided by the frequency of {Y}.
- Lift(X=>Y) = Conf(X=>Y) Supp(Y) –
  Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected.Greater lift values indicate stronger association.

b. Describe briefly incremental associative role mining.

- It is noted that analysis of past transaction data can provide very valuable information on customer buying behavior, and thus improve the quality of business decisions.
- With the increasing use of the record-based databases whose data is being continuously added, updated, deleted etc.
- Examples of such applications include Web log records, stock market data, grocery sales data, transactions in e-commerce, and daily weather/traffic records etc.
- In many applications, we would like to mine the transaction database for a fixed amount of most recent data (say, data in the last 12 months).
- Mining is not a one-time operation, a naive approach to solve the incremental mining problem is to re-run the mining algorithm on the updated database.



Q7. A. What is classification and prediction? Explain the issues regarding classification and prediction.

- There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends.
- These two forms are as follows
  - Classification
  - Prediction
- Classification models predict categorical class labels.
- Prediction models predict

continuous valued functions. For example,

- We can build a classification model to categorize bank loan applications as either safe or risky.
- Prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

What is classification?

- 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 customer (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.
- In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

KNN Solved Example to predict Sugar of Diabetic Patient given BMI and Age

Apply K nearest neighbor classifier to predict the diabetic patient with the given features BMI, Age. If the training examples are,
Assume K=3,

Test Example BMI=43.6, Age=40, Sugar=?
Solution:
The given training dataset has 10 instances with two features BMI (Body Mass Index) and Age. Sugar is the target label. The target label has two possibilities 0 and 1. 0 means the diabetic patient has no sugar and 1 means the diabetic patient has sugar.

Given the dataset and new test instance, we need to find the distance from the new test instance to every training example. Here we use the euclidean distance formula to find the distance.

$$Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

In the next table, you can see the calculated distance from text example to training instances.

See also How to find the Entropy - Decision Tree Learning

Once you calculate the distance, the next step is to find the nearest neighbors based on the value of k. In this case, the value of k is 3. Hence we need to find 3 nearest neighbors.

Now, we need to apply the majority voting technique to decide the resulting label fro the new example. Here the 1st and 2nd nearest neighbors have target label 1 and the 3rd nearest neighbor has target label 0. Target label 1 has the majority. Hence the new example is classified as 1, That is the diabetic patient has Sugar.

What is prediction?

- Following are the examples of cases where the data analysis task is Prediction −
- Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company.
- In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction.
- In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

Classification and Prediction Issues

- The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities −
- Data Cleaning
    - Data cleaning involves removing the noise and treatment of missing values.
    - The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.
- Relevance Analysis
    - Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
- Data Transformation and reduction
    - The data can be transformed by any of the following methods.
- Normalization
    - The data is transformed using normalization.
    - Normalization involves scaling all values for given attribute in order to make them fall within a small specified range.
    - Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.
- Generalization
    - The data can also be transformed by generalizing it to the higher concept.
    - For this purpose we can use the concept hierarchies.

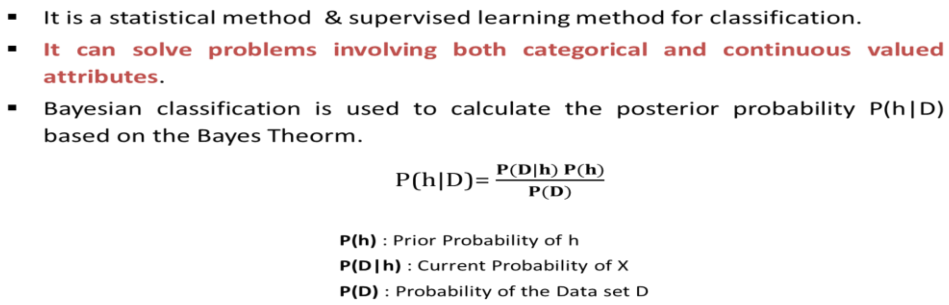
Comparison of Classification and Prediction Methods

- Here is the criteria for comparing the methods of Classification and Prediction −
- Accuracy − Accuracy of classifier refers to the ability of classifier. It predict the

class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

- Speed − this refers to the computational cost in generating and using the classifier or predictor.
- Robustness − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
- Scalability − Scalability refers to the ability to construct the classifier or predictor efficiently; given large
amount of data.
- Interpretability − It refers to what extent the classifier or predictor understands.

b. Explain Bayesian classification method in detail.



- It is a statistical method & supervised learning method for classification.
- **It can solve problems involving both categorical and continuous valued attributes**.
- Bayesian classification is used to calculate the posterior probability P(h|D) based on the Bayes Theorm.

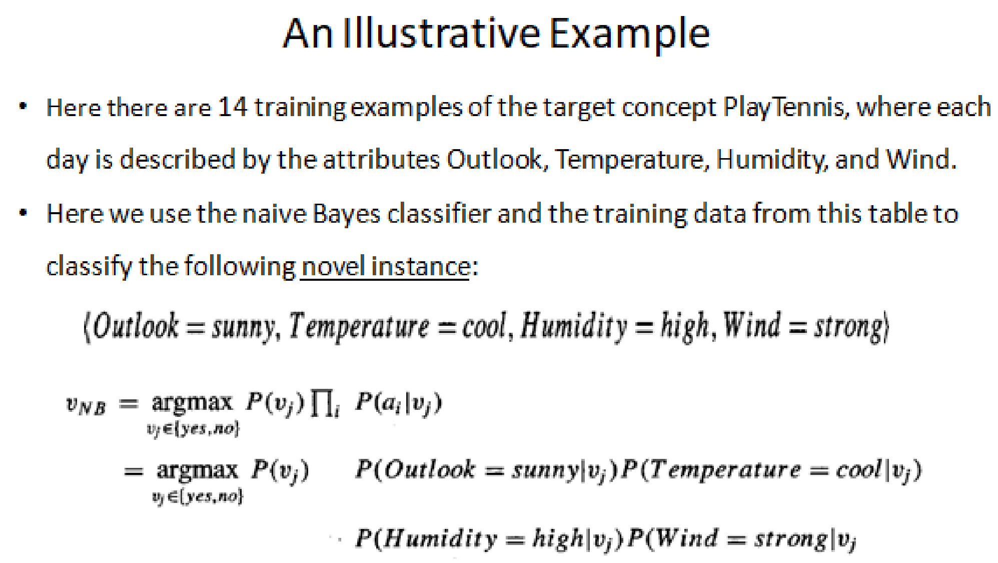
$$P(h|D)=\frac{P(D|h)\,P(h)}{P(D)}$$

P(h) : Prior Probability of h
P(D|h) : Current Probability of X
P(D) : Probability of the Data set D

Naive Bayes classifier:

$$v_{NB} = \operatorname*{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i|v_j)$$

$$\begin{aligned} v_{MAP} &= \operatorname*{argmax}_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n|v_j)P(v_j)}{P(a_1, a_2 \ldots a_n)} \\ &= \operatorname*{argmax}_{v_j \in V} P(a_1, a_2 \ldots a_n|v_j)P(v_j) \end{aligned}$$



An Illustrative Example

- Here there are 14 training examples of the target concept PlayTennis, where each day is described by the attributes Outlook, Temperature, Humidity, and Wind.
- Here we use the naive Bayes classifier and the training data from this table to classify the following novel instance:

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong\rangle$$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\text{argmax}}\ P(v_j)\prod_i P(a_i|v_j)$$

$$= \underset{v_j \in \{yes,no\}}{\text{argmax}}\ P(v_j)\quad P(Outlook = sunny|v_j)P(Temperature = cool|v_j)$$

$$\cdot\ P(Humidity = high|v_j)P(Wind = strong|v_j$$

Q8.a. Explain linear and nonlinear regression prediction methods.

- Regression is a data mining function that predicts a number.
- Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques.
- For example, a regression model could be used to predict children's height, given their age, weight, and other factors.
- A regression task begins with a data set in which the target values are known.
- For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time.
- The data might track age, height, weight, developmental milestones, family history, and so on.
- Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.
- Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values.
- It is required to understand the mathematics used in regression analysis to develop quality regression models for data mining.
- The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide.
- It shows that regression is the process of estimating the value of a continuous target (y) as a function (F)

of one or more predictors (x1 , x2 , ..., xn), a set of parameters (θ1 , θ2 , ..., θn), and a measure of error (e).

$$y = F(x,\theta) + e$$

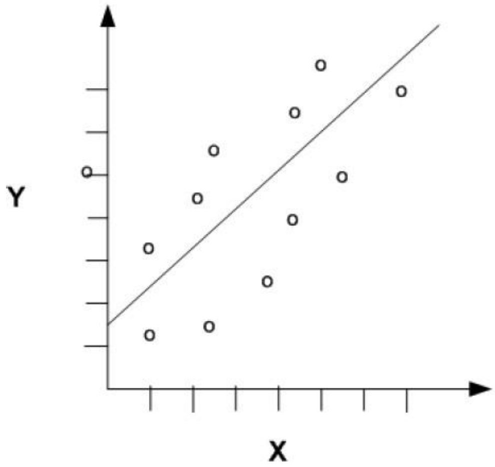
- The process of training a regression model involves finding the best parameter values for the function that minimize a measure of the error.

Linear Regression

- The simplest form of regression to visualize is linear regression with a single predictor.

- A linear regression technique can be used if the relationship between x and y can be approximated with a straight line.
- Linear regression with a single predictor can be expressed with the following equation.
- y = θ2x + θ1 + e
- The regression parameters in simple linear regression are:
- The slope of the line (θ) — the angle between a data point and the regression line
- The y intercept (θ) — the point where x crosses the y axis (x = 0)

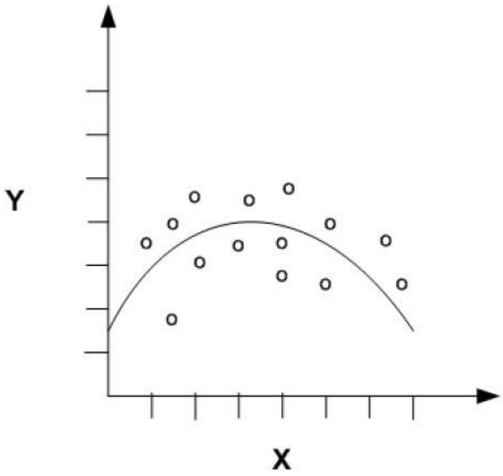
Figure 4-1 Linear Relationship Between x and y



Nonlinear Regression

- Often the relationship between x and y cannot be approximated with a straight line.
- In this case, a nonlinear regression technique may be used. Alternatively, the data could be preprocessed to make the relationship linear.

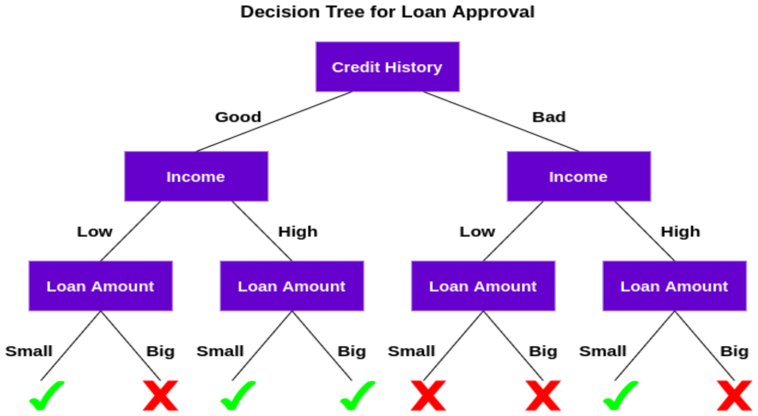
Figure 4-2 Nonlinear Relationship Between x and y



b. Explain decision tree classification method in detail.

Hunt’s Algorithm For Decision Tree

- **Algorithm:** Generate a decision tree. Generate a decision tree from the training tuples of data partition D.
- **Input:**
- **1)** Data partition, D, which is a set of training tuples and their associated class labels;
- **2)** Attribute list-the set of candidate attributes;
- **3)** Attribute selection method, a procedure to determine the splitting criterion that "*best*" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, a split point or splitting subset.
- **Output:** A decision tree.
- Decision tree is a classifier in the form of a tree structure
  - **Decision node**: Specifies a test on a single attribute
  - **Leaf node**: Indicates the value of the target attribute
  - **Arc/edge**: Split of one attribute
  - **Path**: A disjunction of test to make the final decision



Q9.a. Explain the following data mining business applications: i) Balance Score card ii)Fraud detection.

- The Balanced Scorecard (BSC) is a framework for managing business performance.
- Balanced scorecards provide concise, predictive and actionable information about how a company is performing and may perform in the future.
- BSC provides a framework for designing a set of measures for business activities as being the key drivers of the business or Key Performance Indicators (KPIs).
- KPIs are collected from CRM, ERP, Accounting, Personnel, Inventory, and so on.
- Good balanced scorecards might be said to have good representation on good quality business drivers or KPIs. Qualities of good KPIs include;

- Valid & agreed upon: drivers must be valid and agreed upon by stakeholders.
- Specific & measurable: drivers must be specific and measurable systematically.
- Reliable: information used as KPIs must be reliable.
- Relevant: drivers must be relevant to business.
- Achievable: targets assigned for drivers must be achievable. Otherwise drivers will be meaningless to include.
- Easily understood: drivers should be easily understood by users. Complex and obscure drivers may not be useful.
- Timely: drivers must use timely information obtained in a timely manner.

How Knowledge-Enhanced Predictive Balanced Scorecard improve business visibility.

- Predictive analytics can be used to detect patterns and trends in business drivers automatically from hidden numbers, and to predict future directions.
- It is known that leading predictive indicators are more useful than trailing indicators. Directions and projections can be very useful information to have.
- Rule-based expert systems can be used to leverage complexity of various business drivers and indicators.
- As the survey mentioned found, understanding too many drivers and complex numbers can be very daunting tasks for executives and business users.
- Expert systems based on business logic can take this task as an expert, making balanced scorecards friendlier and easier to understand.
- Web-based reporting & charting engines are essential in generating balanced scorecards in a timely real-time fashion so that executives and business users can recognize developing situation in real-time.

Fraud detection:

The term fraud here refers to the abuse of a profit organization's System without necessarily leading to direct legal consequences. In a competitive fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control automates and helps reduce the manual parts of a checking payee: This area has become one of the most established industry/government data mining applications, It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms.
Application of data mining techniques to fraud analysis. Present some classification and prediction data mining techniques which we consider important to handle fraud detection. There exist a number of data mining algorithms and we _ present statistics-based algorithm, decision tree based algorithm and rule-based algorithm. We present Bayesian classification. Model to detect fraud in automobile insurance. Naïve Bayesian visualisation is selected to analyze and interpret the classifier predictions.

b. Explain click stream mining in detail.

Clickstream data and clickstream analytics are the processes involved in collecting, analyzing and reporting aggregate data about which pages a website visitor visits -- and in what order. The path the visitor takes through a website is called the clickstream.

Clickstreams are categorized into clickstream data and clickstream analytics, which is also referred to as clickstream analysis. The clickstream data is the information collected about a user while they browse through a website or use a web browser. Clickstream analytics is the process of tracking, analyzing and reporting data on the pages a user visits and user behavior while on a webpage.

Websites use clickstream data to show how a user progressed from an initial search or landing page to buying an item or service. Search engines use clickstream data sets to show where a user has searched for a term, when they have clicked on it and if they go back to searching after this. Internet service providers, advertising networks, and IT and telecom organizations also collect clickstream data.

Clickstream data collected from a single session of a user interacting with a website may not be useful. However, an organization can use aggregate data gathered from many visitors to improve its website or service.

For example, if a lot of visitors leave a site after landing on a page with too little information, the organization may need to enhance the page with more valuable information. Likewise, if visitors often land on a page that isn't the website's homepage, then the organization may want to redesign that page to be more inviting and informative to users.

Clickstream data does not include personal details about a user, and it is typically stored on the server that supports the website. Clickstream data is a useful addition to data from Google Analytics.

Organizations use clickstream analytics to uncover trends and draw conclusions from different metrics about their websites. This process typically uses a web server log file to monitor user activity on a website.

Using the clickstream analysis, an organization can collect data on the number of page visits, views, and unique and repeat visitors. This data provides an idea of how the organization's website performs and it can help approximate the typical user experience (UX). A website owner can then adjust the site to make it more user-friendly and increase the chance that visitors will stay longer, make a purchase or otherwise interact with the website and the organization behind it.

Because an extremely large volume of data can be gathered through clickstream analysis, many e-businesses rely on big data analytics and related tools such as Hadoop to interpret the data and generate reports on specific areas of interest.

Clickstream analysis is effective when used in conjunction with other, more traditional market research, evaluation resources, data sources and strategies.

There are two levels of clickstream analysis: traffic analytics and e-commerce analytics.

Traffic analytics

This analysis operates at the server level. It collects and analyzes the following data sets:

how many pages are served to a user;

how long it takes each page to load;

how often the user hits the browser's back button; and

how much data is transmitted before the user moves to a different webpage.

E-commerce analytics

This analysis uses clickstream data to determine the effectiveness of a website in terms of conversions and transactions. It is concerned with the following data points:

what pages the shopper lingers on;

what the shopper puts in or takes out of a shopping cart;

what items the shopper purchases;

whether the shopper belongs to a loyalty program;

whether the shopper uses a coupon code; and

the shopper's preferred method of payment.

Benefits of clickstream data analysis

There are a number of benefits organizations can get from clickstream data and clickstream analytics. Among them are the following:

User information. The data collected can include search terms used, pages landed on, webpage features used and the addition or removal of items from a cart, all of which can lead to more actionable insights.

User routes. Organizations can use data analysis to view the different routes their online visitors or customers take to reach a page or to make a purchase.

Customer trends and insights. Collecting and analyzing the clickstreams of a large number of visitors lets an organization identify trends in the following areas:

how visitors get to the website;

what they do once there;

how long they stay on a page;

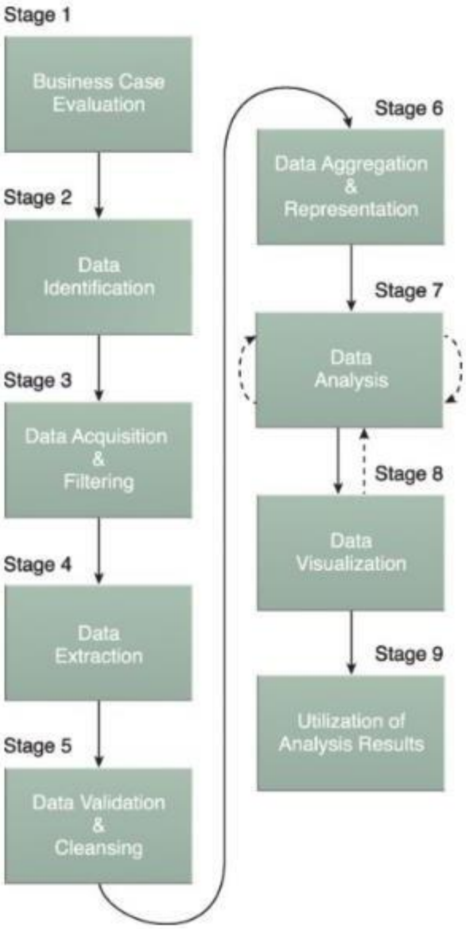
the number of page visits visitors make; and

the number of unique and repeat visitors.

UX. If a majority of users quickly leave a page or website, it could be a sign that the page is poorly optimized or doesn't contain enough information of value. Clickstream data enables an organization to recognize UX shortcomings, enabling them to make necessary changes.

Digital marketing. Clickstream data can be used to determine the amount of traffic coming from ad banners and campaigns. Such data provides insight as to which advertisements are most effective and lead to customer conversion rate optimization. Clickstream analysis can also derive what times of day, month or year a marketing strategy is most effective.

Q10.a. Explain data analytics life cycle in detail.



Data Analytics life cycle

- Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes.
- To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data.
- The upcoming sections explore a specific data analytics lifecycle that organizes and manages the tasks and activities associated with the analysis of Big Data.
- From a Big Data adoption and planning perspective, it is important that in addition to the lifecycle, consideration be made for issues of training, education, tooling and staffing of a data analytics team.
- The Big Data analytics lifecycle can be divided into the following nine stages,

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering

4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

Business Case Evaluation

- Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis.
- The Business Case Evaluation stage requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks.
- An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle.
- Based on business requirements that are documented in the business case, it can be determined whether the business problems being addressed are really Big Data problems.
- In order to qualify as a Big Data problem, a business problem needs to be directly related to one or more of the Big Data characteristics of volume, velocity, or variety.

Data Identification

- The Data Identification stage is dedicated to identifying the datasets required for the analysis project and their sources.
- Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for.
- Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.

Data Acquisition and Filtering

- During the Data Acquisition and Filtering stage, the data is gathered from all of the data sources that were identified during the previous stage.
- The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.
- Depending on the type of data source, data may come as a collection of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter.
- In many cases, especially where external, unstructured data is concerned, some or

most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process.

Data Extraction

- The Data Extraction lifecycle stage is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.
- The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution.
- For example, extracting the required fields from delimited textual data, such as with webserver log files, may not be necessary if the underlying Big Data solution can already directly process those files.

Data Validation and Cleansing

- The Data Validation and Cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data.
- Big Data solutions often receive redundant data across different datasets.
- This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data.

Data Aggregation and Representation

- The Data Aggregation and Representation stage is dedicated to integrating multiple datasets together to arrive at a unified view.
- Performing this stage can become complicated because of differences in:
    - Data Structure – Although the data format may be the same, the data model may be different.
    - Semantics – A value that is labeled differently in two different datasets may mean the same
      thing, for example “surname” and “last name.”
- The large volumes processed by Big Data solutions can make data aggregation a time and effort- intensive operation.
- Reconciling these differences can require complex logic that is executed automatically without the need for human intervention.

Data Analysis

- The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics.
- This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered.
- The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

Data Visualization

- The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

- The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated.

Utilization of Analysis Results

- The Utilization of Analysis Results stage is dedicated to determining how and where processed analysis data can be further leveraged.
- Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce “models” that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed.

b. What are core deliverables? Explain various functions involved in developing core deliverables for stack holders.

1. Stakeholders are the originator of the project management organization that is responsible for the delivery of stakeholders' expectation and satisfaction. The successful delivery of any project deliverables highly depend on stakeholder engagement and management.

2. The effective engagement and management of stakeholder relies on project manager's ability to identify stakeholders' expectations from the beginning to close-up.

3. Researchers described project stakeholders management as a process in which project team facilitates the needs of stakeholders to identify, discuss, agree and contribute to achieve their objectives, describes stakeholder relationship management through six continues processes, including identifying stakeholders, analyzing, engaging, identifying information flow, enforcing stakeholder agreement, and stakeholder debriefing.

4. 4. Developing core deliverables has drawn the key stakeholder management processes from the literature to construct its mediating factor.

5. 5. The mediating variable of manage-through-stakeholder consists of five main observed variables of stakeholder identification and classification, communication, engagement, empowerment, and risk control.

6. 6. The aim here is to investigate the mediating role of manage-through-stakeholder on the relationship between stakeholder influential variables and project success.

- ▪ Data products that result from developing a big data product are in most of the cases some of the following −

- Machine learning implementation − This could be a classification algorithm, a regression model or a segmentation model.
- Recommender system − The objective is to develop a system that recommends choices based on user behavior. Netflix is the characteristic example of this data product, where based on the ratings of users, other movies are recommended.
- Dashboard − Business normally needs tools to visualize aggregated data. A dashboard is a graphical mechanism to make this data accessible.
- Ad-Hoc analysis − Normally business areas have questions, hypotheses or myths that can be answered doing ad-hoc analysis with data.