A quality product by
Brainheaters™ LLC

Brainheaters Notes

DM | Computer Sem-6

SERIES 313 - 2018 (A.Y 2020 - 21)

BH.Index

(Learn as per the Priority to prepare smartly)

MODULE-1

Q1. Define noise data. Enlist the reasons for the presence of noise in data collection. Explain the methods to deal with noise. (P2-Appeared 3 times) (5-10M)

Ans: Noisy data is meaningless data.

- It includes any data that cannot be understood and interpreted correctly by machines, such as unstructured text.
- Noisy data unnecessarily increases the amount of storage space required and can also adversely affect the results of any data mining analysis.
- Noisy data can be caused by faulty data collection instruments, human or computer errors occurring at data entry, data transmission errors, the limited buffer size for coordinating synchronized data transfer, inconsistencies in naming conventions or data codes used and inconsistent formats for input fields(eg:date).

Noisy data can be handled by following the given procedures:

1. **Binning:**
    - Binning methods smooth a sorted data value by consulting the values around it.
    - The sorted values are distributed into a number of "buckets," or bins.
    - Because binning methods consult the values around it, they perform local smoothing.

- 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.
- Binning is also used as a discretization technique.

2. **Regression:**
   - Here data can be smoothed by fitting the data to a function.
   - Linear regression involves finding the "best" line to fit two attributes 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.
3. **Clustering:**
   - Outliers may be detected by clustering, where similar values are organized into groups, or "clusters."
   - Similarly, values that fall outside of the set of clusters may also be considered outliers.

Q2. Discuss the major issues/challenges in data mining.

(P2-Appeared 3 times)(5-10M)

Ans: **There are many major issues in data mining:**

1.Mining methodology and user interaction:

- Mining different kinds of knowledge in databases.
- Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks.
- Eg: data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.
- Interactive mining of knowledge at multiple levels of abstraction.
- Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
- Knowledge should be mined by drilling down, rolling up, and pivoting through the data space and knowledge space interactively.

2.Performance and scalability:

- Efficiency and scalability of data mining algorithms.
- To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.
- In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases. • Parallel, distributed and incremental mining methods.

- The huge size of many databases, the wide distribution of data, the high cost of some data mining processes and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms.
- Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.
- They incorporate database updates without having to mine the entire data again "from scratch."

3.Issues relating to the diversity of data types:

- Handling relational and complex types of data.
- It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data.
- Therefore, one may expect to have different data mining systems for different kinds of data.
- Mining information from heterogeneous databases and global information systems (WWW).
- Data mining may help disclose high-level data regularities in multiple heterogeneous databases that are unlikely to be discovered by simple query systems and may improve information exchange and interoperability in heterogeneous databases.Web mining, uncovers interesting knowledge about Web contents, Web structures, Web usage, and Web dynamics.

4.Issues related to applications and social impacts:

- Application of discovered knowledge
- Domain specific data mining tools.
- Intelligent query answering.
- Process control and decision making.
- Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem.
- Protection of data security, integrity, and privacy.

Q3. Define data mining and list its features.(P3-Appeared 2 times)(5-10M)

Ans: **Data Mining**

- Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and to predict the likelihood of future events based on past events. Data mining is also known as Knowledge Discovery in Data (KDD).

The key features of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

Q4. Discuss possible ways for integration of a Data Mining system with a Database or DataWarehouse system (P4- Appeared 1 times)(5-10M)

Ans: The possible ways for integration of a Data Mining system with a Database or DataWarehouse system are:

I) Data integration:

Data Integration is one of the steps of data pre-processing that involves combining data residing in different sources and providing users with a unified view of these data.

- It merges the data from multiple data stores (data sources)
- It includes multiple databases, data cubes or flat files.
- Metadata, Correlation analysis, data conflict detection, and resolution of semantic heterogeneity contribute towards smooth data integration.
- There are mainly 2 major approaches for data integration - commonly known as "tight coupling approach" and "loose coupling approach".

Tight Coupling:

- Here data is pulled over from different sources into a single physical location through the process of ETL - Extraction, Transformation and Loading.
- The single physical location provides an uniform interface for querying the data.
- ETL layer helps to map the data from the sources so as to provide a uniform data warehouse.

- This approach is called tight coupling since in this approach the data is tightly coupled with the physical repository at the time of query.

ADVANTAGES:

1. Independence (Lesser dependency to source systems since data is physically copied over)
2. Faster query processing
3. Complex query processing
4. Advanced data summarization and storage possible
5. High Volume data processing

DISADVANTAGES:

1. Latency (since data needs to be loaded using ETL)
2. Costlier (data localization, infrastructure, security)

Loose Coupling:

- Here a virtual mediated schema provides an interface that takes the query from the user, transforms it in a way the source database can understand and then sends the query directly to the source database to obtain the result.
- In this approach, the data only remains in the actual source databases.
- However, the mediated schema contains several "adapters" or "wrappers" that can connect back to the source systems in order to bring the data to the front end.

ADVANTAGES:

- Data Freshness (low latency - almost real time)
- Higher Agility (when a new source system comes or existing source system changes - only the corresponding adapter is created or changed - largely not affecting the other parts of the system)

- Less costlier (Lot of infrastructure cost can be saved since data localization not required)

DISADVANTAGES:

1. Semantic conflicts
2. Slower query response
3. High order dependency to the data sources

For example, let's imagine that an electronics company is preparing to roll out a new mobile device.

- The marketing department might want to retrieve customer information from a sales department database and compare it to information from the product department to create a targeted sales list.
- A good data integration system would let the marketing department view information from both sources in a unified way, leaving out any information that didn't apply to the search.

2) DATA TRANSFORMATION

- In data mining pre-processes and especially in metadata and data warehouses, we use data transformation in order to convert data from a source data format into destination data.

Data transformation - 2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

We can divide data transformation into 2 steps:

- Data Mapping: It maps the data elements from the source to the destination and captures any transformation that must occur.
- Code Generation: It creates the actual transformation program.

Data transformation:

- Here the data are transformed or consolidated into forms appropriate for mining.
- Data transformation can involve the following:

Smoothing:

- It works to remove noise from the data.
- It is a form of data cleaning where users specify transformations to correct data inconsistencies.
- Such techniques include binning, regression, and clustering.

Aggregation:

- Here summary or aggregation operations are applied to the data.
- This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
- Aggregation is a form of data reduction.

Generalization :

- Here low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
- For example, attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.
- Generalization is a form of data reduction.

Normalization:

- Here the attribute data are scaled so as to fall within a small specified range, such as 1:0 to 1:0, or 0:0 to 1:0.
- Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest-neighbor classification and clustering
- For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., income).

Q5. Discuss fraud detection and click-stream analysis using data mining (P4- Appeared 1 times)(5-10M)

Ans: **Fraud detection for Telecommunication Industry**

- The telecommunications industry has expanded dramatically in the last few years with the development of affordable mobile phone technology.
- Fraud is an adaptive crime, so it needs a special method of intelligent data analysis to detect and prevent it.
- This method exists in the areas of Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning and Statistics.
- They offer applicable and successful solutions in different areas of fraud crimes.
- At a low level, simple rule-based detection systems use rules such as the apparent use of the same phone in two very distant geographical locations in quick succession, calls which appear to overlap in time and very high value and very long calls.
- At a higher level, statistical summaries of call distributions (often called profiles or signatures at the user level) are compared with thresholds determined either by experts or by application of supervised learning methods to known fraud/nonfraud cases.
- Some forensic accountants specialize in forensic analytics which is the procurement and analysis of electronic data to

reconstruct, detect, and otherwise support a claim of financial fraud.

- The main steps in forensic analytics are (a)data collection,(b)data preparation,(c)data analysis, and(d)reporting.
- For example, forensic analytics may be used to review an employees' purchasing card activity to assess whether any of the purchases were diverted or divertible for personal use.

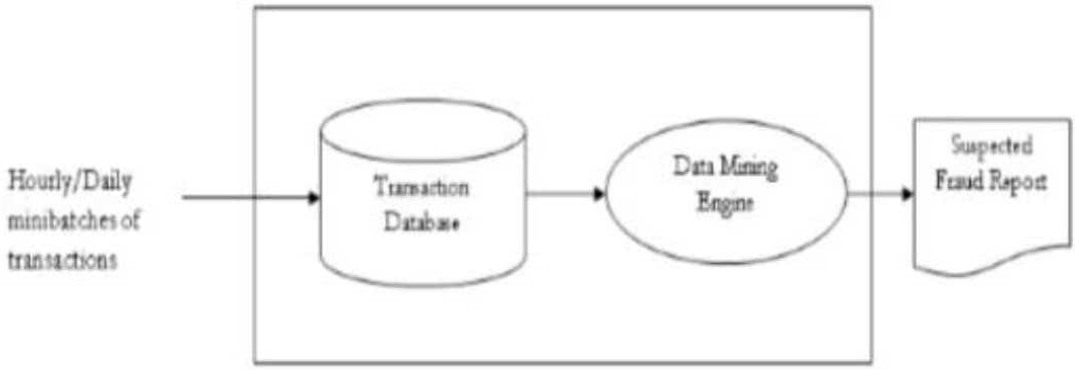


Fig: Fraud Detection

- Techniques used for fraud detection fall into two primary classes: Statistical techniques and Artificial intelligence.

Q6. What is Data Mining? Why is it called data mining rather than knowledge mining? (P4- Appeared 1 times)(5-10M)

Ans : **Data Mining**

- Data mining refers to extracting or "mining" knowledge from large amounts of data
- It is the computational process of discovering patterns in large data sets involving methods at the intersection of

artificial intelligence, machine learning, statistics, and database systems.

Why is it called data mining rather than knowledge mining?

- Data mining means extracting facts from the available data. While Knowledge means a deep study of those facts.
- We do not collect knowledge but facts.
- Hence it is called Data Mining rather than knowledge mining

Q7. What is the difference between KDD and Data Mining?(P4- Appeared 1 times)(5-10M)

Ans: KDD is a non-trivial process for identifying valid, potentially useful and ultimately understandable patterns in data.

- It consists of nine steps that begin with the development and understanding of the application domain to the action on the knowledge discovered.
- Data mining is one of the steps (seventh) and the KDD process is basically the search for patterns of interest in a particular representational form or a set of these representations.

Q8. Explain mining in the following Databases with examples. 1. Temporal Databases 2. Sequence Databases 3. Spatial Databases 4. Spatiotemporal Databases. (P4- Appeared 1 times)(5-10M)

Ans: **Temporal data mining**

- It can be defined as "process of knowledge discovery in temporal databases that enumerates structures (temporal

patterns or models) over the temporal data, and any algorithm that enumerates temporal patterns from, or fits models to, temporal data is a temporal data mining algorithm" (Lin et al., 2002).

- The aim of temporal data mining is to discover temporal patterns, unexpected trends, or other hidden relations in the larger sequential data, which is composed of a sequence of nominal symbols from the alphabet known as a temporal sequence and a sequence of continuous real-valued elements known as a time series, by using a combination of techniques from machine learning, statistics, and database technologies.
- In fact, temporal data mining is composed of three major works including representation of temporal data, definition of similarity measures and mining tasks.

Spatial data mining

- It refers to the process of the retrieval of information or patterns that are not explicitly stored in the spatial databases.
- Spatial data mining methods are used for the better understanding of spatial data, identifying the relationships between spatial data and non- spatial data, query optimization in spatial databases etc.
- Statistical Spatial analysis is the most commonly and widely used data mining technique.
- It assumes that the spatial data are independent which in fact is not true as the spatial data are interrelated with their neighboring objects.
- Statistical methods cannot handle symbolic values and non linear rules and are also very costly in the result computation.

- Several Machine learning techniques like learning from examples and generalization and specialization are used in spatial data mining.

Q9. Discuss the application of data warehousing and data mining (P4- Appeared 1 times)(5-10M)

Ans: **Applications of a Data Warehouse**

1. **Banking**
   - Identify the potential risk of default and manage and control collections
   - Performance analysis of each product, service, interchange, and exchange rates
   - Track performance of accounts and user data
   - Provide feedback to bankers regarding customer relationships and profitability
2. **Finance**
   - Evaluation of customer expenses trends
   - Maintain transparency in transactions
   - Predict/spot defaulters and act accordingly
   - Analyze and forecast different aspects of business, stock, and bond performance
3. **Government**
   - Maintain and analyze tax records, health policy records, and their respective providers
   - Prediction of criminal activities from patterns and trends
   - Searching terrorist profile

- Threat assessment and fraud detection

4. **Education**
    - Store and analyze information about faculty and students
    - Maintain student portals to facilitate student activities
    - Extract information for research grants and assess student demographics
    - Integrate information from different sources into a single repository for analysis and strategic decision-making
5. **Healthcare**
    - Generate patient, employee, and financial records
    - Share data with other entities, like insurance companies, NGOs, and medical aid services
    - Use data mining to identify patient trends
    - Provide feedback to physicians on procedures and tests
6. **Insurance**
    - Analyze data patterns and customer trends – Maintain records of all internal and external sources, including existing participants
    - Design customized offers and promotions for customers
    - Predict and analyze changes in the industry
7. **Manufacturing**
    - Predict market changes and analyze current business trends
    - Analyze previous and current market data

- Track customer feedback and identify opportunities for improvement
- Gather, standardize, and store data from various internal and external sources
- Identify profitable product lines and required product features

8. **Retail**
    - Maintain records of producers and consumers
    - Track items, their promotion strategies, and consumer buying trends (trend analysis)
    - Analyze sales to determine shelf space
    - Understanding the patterns of complaints, claims, and returns
9. **Services**
    - Maintenance of financial and employee records
    - Customer profiling and screening
    - Resource allocation and management
    - Revenue patterns and profitability

Data Mining Applications

- Here is the list of areas where data mining is widely used -
    - Financial Data Analysis
    - Retail Industry
    - Telecommunication Industry
    - Biological Data Analysis
    - Other Scientific Applications
    - Intrusion Detection

Financial Data Analysis

- The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining.

Retail Industry

- Data Mining has its great application in the Retail Industry because it collects large amounts of data from sales, customer purchasing history, goods transportation, consumption and services.
- It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.
- Data mining in the retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction.

Telecommunication Industry

- Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc.
- Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding.
- This is the reason why data mining is becoming very important to help and understand the business.
- Data mining in the telecommunication industry helps in identifying the telecommunication patterns, catching

fraudulent activities, making better use of resources, and improving the quality of service.

Biological Data Analysis

- In recent times, we have seen tremendous growth in the field of biologies such as genomics, proteomics, functional genomics and biomedical research.
- Biological data mining is a very important part of Bioinformatics.

Other Scientific Applications

- The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate.
- Huge amounts of data have been collected from scientific domains such as geosciences, astronomy, etc.
- A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modelling, chemical engineering, fluid dynamics, etc.

Intrusion Detection

- Intrusion refers to any kind of action that threatens the integrity, confidentiality, or availability of network resources.
- In this world of connectivity, security has become a major issue.
- The increased usage of the internet and availability of the tools and tricks for intruding and attacking networks prompted intrusion detection to become a critical component of network administration.

Q10. Explain Various Data Mining Functionalities with an example (P4- Appeared 1 time)(5-10M)

Ans: **Data Mining Functionality:**

1. Class/Concept Descriptions:

- Classes or definitions can be correlated with results.
- In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts.
- These class or concept definitions are referred to as class/concept descriptions.
    - **Data Characterization:**
      This refers to the summary of general characteristics or features of the class that is under the study. For example, To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these types of data related to such products by running SQL queries.
    - **Data Discrimination:**
      It compares common features of the class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts.

2. Mining Frequent Patterns, Associations, and Correlations:

Frequent patterns are nothing but things that are found to be most common in the data.

There are different kinds of frequencies that can be observed in the dataset.

- **Frequent item set:**
  This applies to a number of items that can be seen together regularly for eg: milk and sugar.
- **Frequent Subsequence:**
  This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.
- **Frequent Substructure:**
  It refers to the different kinds of data structures such as trees and graphs that may be combined with the itemset or subsequence.

Association Analysis:

- The process involves uncovering the relationship between data and deciding the rules of the association.
- It is a way of discovering the relationship between various items.
- For example, it can be used to determine the sales of items that are frequently purchased together.

Correlation Analysis:

- Correlation is a mathematical technique that can show whether and how strongly the pairs of attributes are related to each other.
- For example, Highted people tend to have more weight.

Q11. How is data mining used in the 1) retail industry? 2) telecommunication industry? (P4- Appeared 1 time)(5-10M)

Ans: **Retail Industry**

- Data Mining has its great application in the Retail Industry because it collects large amounts of data from sales, customer purchasing history, goods transportation, consumption and services.
- It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.
- Data mining in the retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction.
- Here is the list of examples of data mining in the retail industry-
    - Design and Construction of data warehouses based on the benefits of data mining.
    - Multidimensional analysis of sales, customers, products, time and region.
    - Analysis of the effectiveness of sales campaigns.
    - Customer Retention.
    - Product recommendation and cross-referencing of items.

Telecommunication Industry

- Today the telecommunication industry is one of the most emerging industries providing various services

such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc.

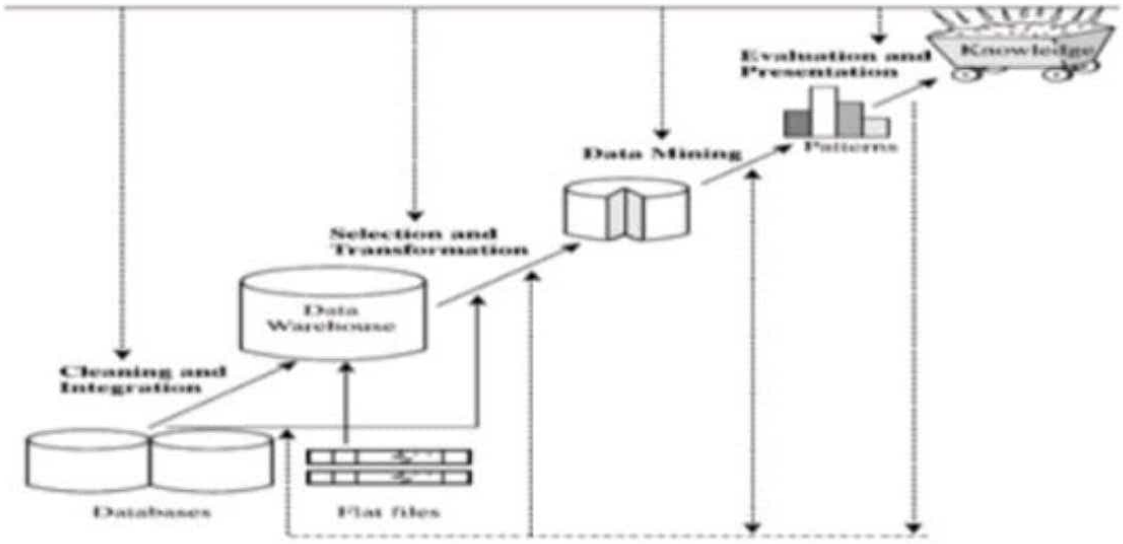
- Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding.
- This is the reason why data mining is becoming very important to help and understand the business.
- Data mining in the telecommunication industry helps in identifying the telecommunication patterns, catching fraudulent activities, making better use of resources, and improving the quality of service.
- Here is the list of examples for which data mining improves telecommunication services -
    - Multidimensional Analysis of Telecommunication data.
    - Fraudulent pattern analysis.
    - Identification of unusual patterns.
    - Multidimensional association and sequential patterns analysis.
    - Mobile Telecommunication services.
    - Use of visualization tools in telecommunication data analysis.

Q12. Describe the steps involved in data mining when viewed as a process of knowledge discovery. (P4- Appeared 1 time)(5-10M)

Ans: Knowledge discovery as a process consists of an iterative sequence of the following steps:

- Data cleaning: It can be applied to remove noise and correct inconsistencies in the data.
- Data integration: Data integration merges data from multiple sources into a coherent data store, such as a data warehouse.
- 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 example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.
- 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.

DIAGRAM:



Q13. Define following terms: Data Mart, Meta Data, Enterprise Warehouse & Virtual Warehouse (P4- Appeared 1 time)(5-10M)

Ans: **Data Mart**

- A data mart is a simple form of data warehouse focused on a single subject or line of business. With a data mart, teams can access data and gain insights faster, because they don't have to spend time searching within a more complex data warehouse or manually aggregating data from different sources.

Metadata: can be defined as follows.

- Metadata is the roadmap to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.

- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Enterprise Warehouse

- An enterprise data warehouse (EDW) is a relational data warehouse containing a company's business data, including information about its customers. An EDW enables data analytics, which can inform actionable insights. Like all data warehouses, EDWs collect and aggregate data from multiple sources, acting as a repository for most or all organizational data to facilitate broad access and analysis.

Virtual Warehouse

- A virtual warehouse is another term for a data warehouse. A data warehouse is a computing tool designed to simplify decision-making in business management. It collects and displays business data relating to a specific moment in time, creating a snapshot of the condition of the business at that moment. Virtual warehouses often collect data from a wide variety of sources.

Q14. A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data – Justify. (P4-Appeared 1 time)(5-10M)

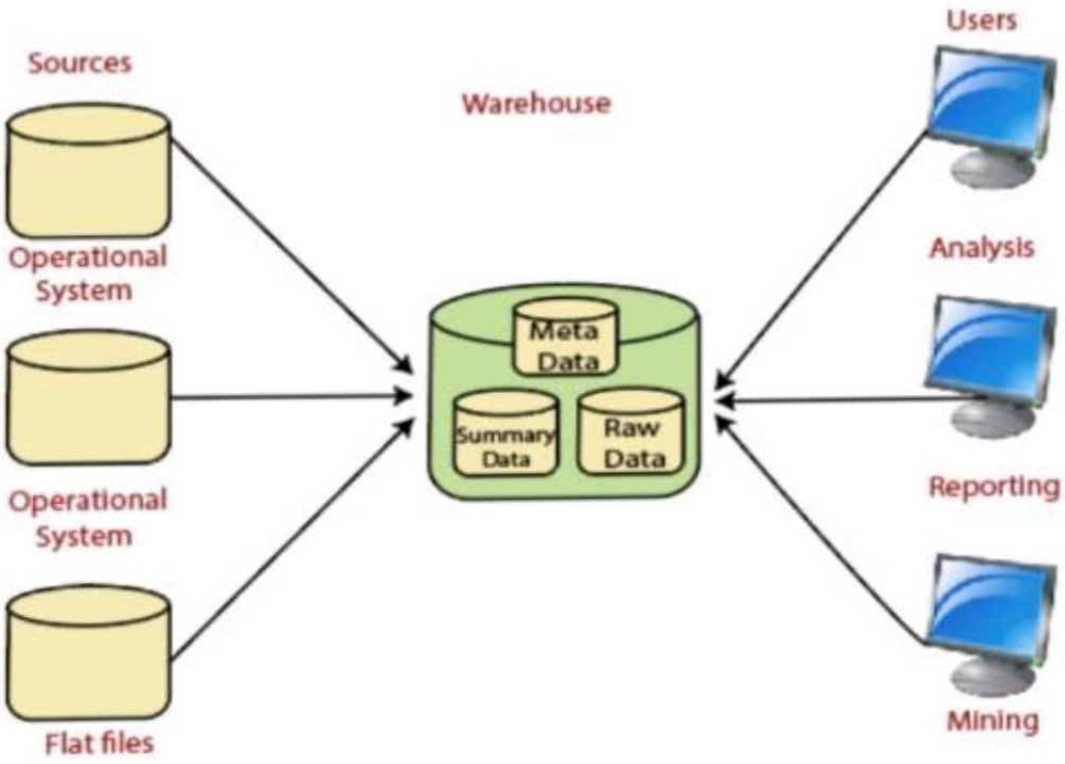
Ans: A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

- Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.
- Integrated: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.
- Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.
- Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

Q15. Discuss data warehouse architecture in detail.(P4-Appeared 1 time)(5-10M)

Ans: Data Warehouse Architecture: Basic

Architecture of a Data Warehouse



Operational System

- An operational system is a method used in data warehousing to refer to a system that is used to process the day-to-day transactions of an organization.

Flat Files

- A Flat file system is a system of files in which transactional data is stored, and every file in the system must have a different name.

Meta Data

- A set of data that defines and gives information about other data.
- Metadata used in Data Warehouse for a variety of purpose, including:
    - Meta Data summarizes necessary information about data, which can make finding and working with particular instances of data more accessible.
    - For example, author, data build, and data changed, and file size are examples of very basic document metadata.
    - Metadata is used to direct a query to the most appropriate data source.
- Lightly and highly summarized data
    - The area of the data warehouse saves all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.
    - The goals of the summarized information are to speed up query performance. The summarized record is updated continuously as new information is loaded into the warehouse.
- End-User access Tools
    - The principal purpose of a data warehouse is to provide information to the business managers for strategic decision-making.
    - These customers interact with the warehouse using end-client access tools.

  The examples of some of the end-user access tools can be:

- Reporting and Query Tools
- Application Development Tools
- Executive Information Systems Tools
- Online Analytical Processing Tools
- Data Mining Tools

Q16. What is an apex cuboid? Discuss drill down and roll up operation with diagram(P4- Appeared 1 time)(5-10M)

Ans: The apex cuboid, or 0-D cuboid, refers to the case where the group-by is empty.

- It contains the total sum of all sales.
- The base cuboid is the least generalized (most specific) of the cuboids.
- The apex cuboid is the most generalized (least specific) of the cuboids, and is often denoted as all.

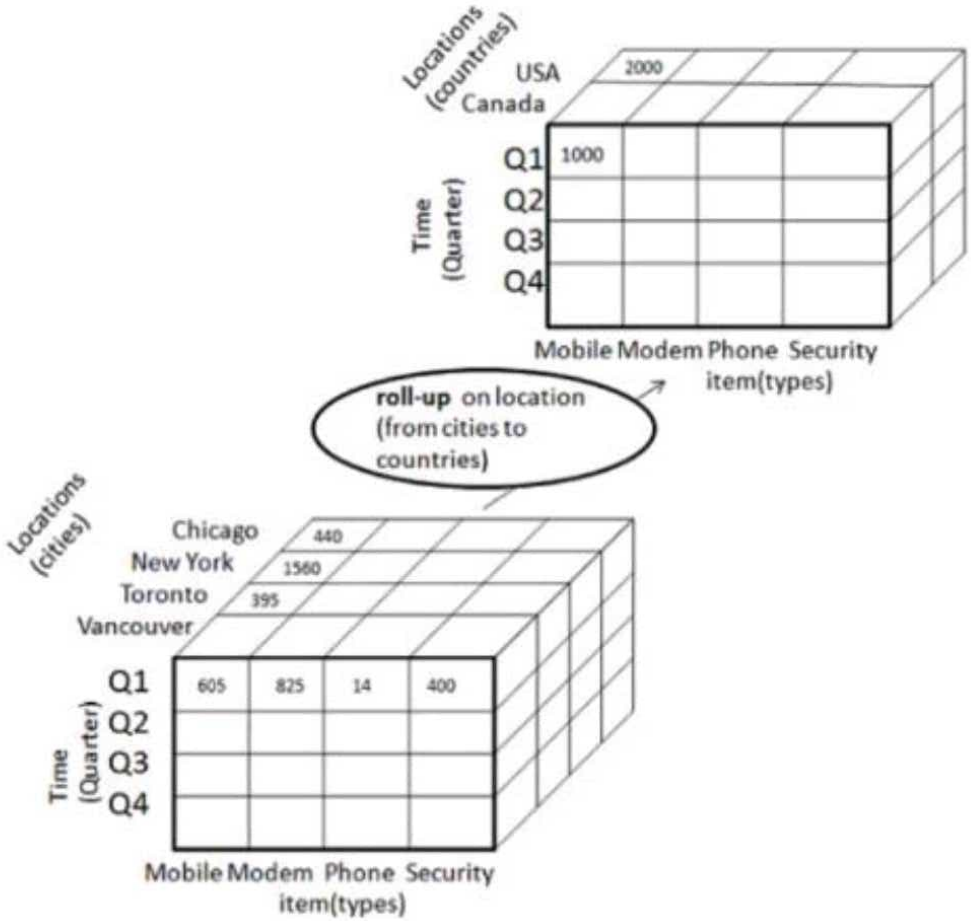
Roll-up

- 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 the city to the level of the 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.

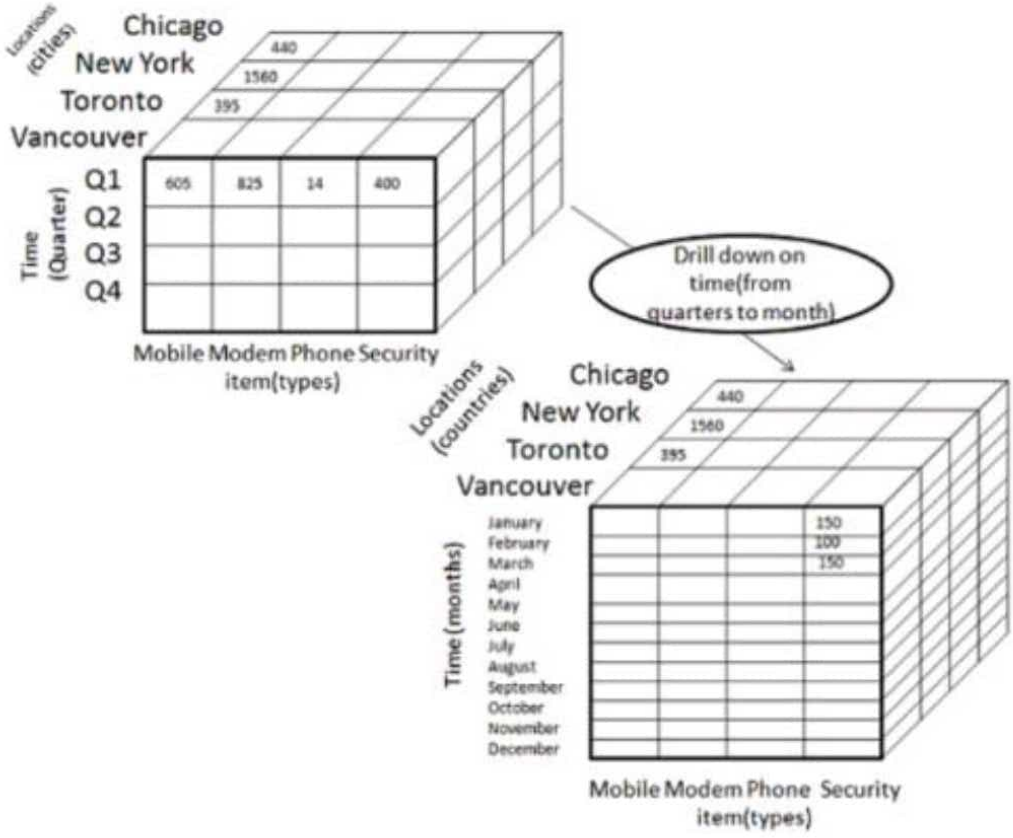


Drill-down

- 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.



Q17. Explain why data warehouses are needed for developing business solutions from today's perspective. Discuss the role of data marts. (P4- Appeared 1 time)(5-10M)

Ans: **Benefits of data warehouses**

- A goal common to all businesses is to make better business decisions than their competitors.

Once a data warehouse is implemented into your business intelligence plans, your company can benefit from it in many ways.

- **Better decision-making –** Corporate decision makers will no longer have to make important business decisions based on limited data and hunches. Data warehouses store credible facts and statistics, and decision makers will be able to retrieve that information from the data warehouse based on their personal needs. In addition to making strategic decisions, a data warehouse can also assist in marketing segmentation, inventory management, financial management, and sales.
- **Quick and easy access to data –** Speed is an important factor that sets you above your competitors. Business users can quickly access data from multiple sources from a data warehouse, meaning that precious time won't be wasted on retrieving data from multiple sources. This allows you to make quick and accurate decisions, with little or no support from your IT department.
- **Data quality and consistency –** Since data warehouses gather information from different sources and convert it into a single and widely used format, departments will produce

results that are in line and consistent with each other. When data is standardized, you can have confidence in its accuracy, and accurate data is what makes for strong business decisions.

- A data warehouse is essential for any business that wants to profit from sound business decisions.
- A data mart is the access layer of a data warehouse that is used to provide users with data.
- Data marts are often seen as small slices of the data warehouse.
- Data warehouses typically house enterprise-wide data, and information stored in a data mart usually belongs to a specific department or team.

Q18. Can BI be used for DM? Or vice versa? Justify -

Ans : Business intelligence encompasses data analysis with the intent of uncovering trends, patterns and insights. Findings based on data provide accurate, astute views of your company's processes and the results those processes are yielding. Beyond standard metrics such as financial measures, in-depth business intelligence reveals the impact of current practices on employee performance, overall company satisfaction, conversions, media reach and a number of other factors.

BI vs Data Mining

- As previously stated, business intelligence is defined as the methods and tools used by organizations to glean analytical findings from data. It also consists of how companies can gain information from big data and data mining. This means business intelligence is not confined to technology — it includes the business processes and data analysis procedures that facilitate the collection of big data.
- Data mining falls under the umbrella term of "business intelligence," and can be considered a form of BI. Data mining can be considered a function of BI, used to collect relevant information and gain insights. Moreover, business intelligence could also be thought of as the result of data mining. As stated, business intelligence involves using data to acquire insights. Data mining business intelligence is the collection of necessary data, which will eventually lead to answers through in-depth analysis.
- The link between data mining and business intelligence can be thought of as a cause-and-effect relationship. Data mining searches for the "what" (relevant data sets) and business intelligence processes uncover the "how" and "why" (insights). Analysts utilize data mining to find the information they need and use business intelligence to determine why it is important.
- Business intelligence and data mining differ in core aspects, including purpose, volume and results. The purpose of BI is to convert raw data into useful information for executives and stakeholders. It tracks and presents key performance

matrices on reports and dashboards to facilitate robust, data-driven decisions.

- On the other hand, the primary objective of data mining is to explore and analyze data to uncover solutions to specific business problems. It leverages complex algorithms and computational intelligence to detect trends and patterns.
- Data mining algorithms process datasets from specific departments, customer segments or competitors. Digging deeper into datasets lets them find answers to specific business problems. On the contrary, business intelligence processes enterprise-wide data to deduce business performance.
- Since data mining focuses on resolving complex business problems, the end result is a statistical data model that looks for patterns and relationships within similar datasets. However, business intelligence produces charts, graphs, dashboards and reports.

MODULE-2

Q1. Describe Concept Hierarchy? List and briefly explain types of Concept Hierarchy (P3-Appeared 2 times)

Ans: **Concept Hierarchy**

- A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts

Types of concept hierarchy

1. **Binning**
   - In binning, first sort data and partition into (equi-depth) bins then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
2. **Histogram analysis**
   - Histogram is a popular data reduction technique
   - Divide data into buckets and store average (sum) for each bucket
   - Can be constructed optimally in one dimension using dynamic programming
   - Related to quantization problems.
3. **Clustering analysis**
   - Partition data set into clusters, and one can store cluster representation only
   - Can be very effective if data is clustered but not if data is "smeared"

- Can have hierarchical clustering and be stored in multi-dimensional index tree structures

4. **Entropy-based discretization**
5. **Segmentation by natural partitioning**
    - 3-4-5 rules can be used to segment numeric data into relatively uniform, "natural" intervals.
    - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
    - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
    - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Q2. What is the need for preprocessing the data? (P4- Appeared 1 times)(5-10M)

Ans: **Data Preprocessing is required because:**

Real world data are generally:

- Incomplete: Missing attribute values, missing certain attributes of importance, or having only aggregate data
- Noisy: Containing errors or outliers
- Inconsistent: Containing discrepancies in codes

Q3. Explain about Data Transformation methods with suitable examples. (P4- Appeared 1 times)(5-10M)

Ans: **1. Smoothing:**

- It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.
- The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

2. Aggregation:

- Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.
-
- The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies.

- For example, Sales data may be aggregated to compute monthly & annual total amounts.

3. Discretization:

- It is a process of transforming continuous data into a set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.
- Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.
- For example, (1-10, 11-20) (age:- young, middle age, senior).

4. Attribute Construction:

- Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.

5. Generalization:

- It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).
- For example, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

6. Normalization:

- Data normalization involves converting all data variables into a given range.

Q4. Differentiate Fact table vs. Dimension table (P4- Appeared 1 times)(5-10M)

Ans: Difference between Dimension table vs. Fact table

Parameters	**Fact Table**	**Dimension Table**
Definition	Measurements, metrics or facts about a business process.	Companion table to the fact table contains descriptive attributes to be used as query constraining.
Characteristic	Located at the center of a star or snowflake schema and surrounded by dimensions.	Connected to the fact table and located at the edges of the star or snowflake schema
Design	Defined by their grain or its most atomic level.	Should be wordy, descriptive, complete, and quality assured.
Task	Fact table is a measurable event for which dimension table data is collected and is used for analysis and reporting.	Collection of reference information about a business.

Type of Data	Facts tables could contain information like sales against a set of dimensions like Product and Date.	Every dimension table contains attributes which describe the details of the dimension. E.g., Product dimensions can contain Product ID, Product Category, etc.
Key	Primary Key in fact table is mapped as foreign keys to Dimensions.	Dimension table has a primary key column that uniquely identifies each dimension.
Storage	Helps to store report labels and filter domain values in dimension tables.	Load detailed atomic data into dimensional structures.
Hierarchy	Does not contain Hierarchy	Contains Hierarchies. For example Location could contain, country, pin code, state, city, etc.

Q5. List and describe methods for handling missing values in data cleaning. (P4- Appeared 1 times)(5-10M)

Ans: **Ways to handle missing values :**

1. **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
2. **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
3. **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.

4. **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
5. **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
6. 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

Q6. Enlist data reduction strategies and explain any two (P4-Appeared 1 time)(5-10M)

Ans: **Methods of data reduction:**

1. Data Cube Aggregation
2. Dimension reduction
3. Data Compression
4. Numerosity Reduction
5. Discretization & Concept Hierarchy Operation

Numerosity Reduction:

- In this reduction technique the actual data is replaced with mathematical models or smaller representations of the data instead of actual data, it is important to only store the model parameter.
- Or non-parametric methods such as clustering, histogram, sampling. For More Information on Numerosity Reduction Visit the link below:

Discretization & Concept Hierarchy Operation:

- Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals.
- We replace many constant values of the attributes with labels of small intervals.
- This means that mining results are shown in a concise, and easily understandable way.
    - Top-down discretization: If you first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat this

method up to the end, then the process is known as top-down discretization also known as splitting.

- Bottom-up discretization: If you first consider all the constant values as split-points, some are discarded through a combination of the neighbourhood values in the interval, that process is called bottom-up discretization.

Q7. Discuss attribute subset selection (P4- Appeared 1 time)(5-10M)

Ans: Attribute subset Selection is a technique that is used for data reduction in the data mining process. Data reduction reduces the size of data so that it can be used for analysis purposes more efficiently.

Need of Attribute Subset Selection-

- The data set may have a large number of attributes.
- But some of those attributes can be irrelevant or redundant.
- The goal of attribute subset selection is to find a minimum set of attributes such that dropping of those irrelevant attributes does not much affect the utility of data and the cost of data analysis could be reduced.
- Mining on a reduced data set also makes the discovered pattern easier to understand.

Process of Attribute Subset Selection-

- The brute force approach can be very expensive in which each subset (2^n possible subsets) of the data having n attributes can be analyzed.

- The best way to do the task is to use statistical significance tests such that the best (or worst) attributes can be recognized.
- Statistical significance test assumes that attributes are independent of one another.
- This is a kind of greedy approach in which a significance level is decided (statistically ideal value of significance level is 5%) and the models are tested again and again until the p-value (probability value) of all attributes is less than or equal to the selected significance level.
- The attributes having a p-value higher than the significance level are discarded.
- This procedure is repeated again and again until all the attributes in the data set have a p-value less than or equal to the significance level.
- This gives us a reduced data set having no irrelevant attributes.

Methods of Attribute Subset Selection-

1. Stepwise Forward Selection.
2. Stepwise Backward Elimination.
3. Combination of forwarding Selection and Backward Elimination.
4. Decision Tree Induction.

All the above methods are greedy approaches for attribute subset selection.

1. Stepwise Forward Selection: This procedure starts with an empty set of attributes as the minimal set. The most relevant attributes are chosen(having minimum p-value) and are

added to the minimal set. In each iteration, one attribute is added to a reduced set.

2. Stepwise Backward Elimination: Here all the attributes are considered in the initial set of attributes. In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than the significance level.
3. Combination of forwarding Selection and Backward Elimination: The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently. This is the most common technique which is generally used for attribute selection.
4. Decision Tree Induction: This approach uses decision trees for attribute selection. It constructs a flow chart like structure having nodes denoting a test on an attribute. Each branch corresponds to the outcome of the test and leaf nodes are a class prediction. The attribute that is not part of the tree is considered irrelevant and hence discarded.

Q8. In data preprocessing, why do we need data smoothing? Discuss data smoothing by Binning. (P4- Appeared 1 time)(5-10M)

Ans: Data binning, bucketing is a data pre-processing method used to minimize the effects of small observation errors.

- The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin.

- This has a soothing effect on the input data and may also reduce the chances of overfitting in case of small datasets
- There are 2 methods of dividing data into bins "
    1. Equal Frequency Binning: bins have an equal frequency.
    2. Equal Width Binning : bins have equal width with a range of each bin are defined as [min + w], [min + 2w] .... [min + nw] where w = (max-min) / (no of bins).

Equal frequency
Input :[5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]
Output :
[5, 10, 11, 13]
[15, 35, 50, 55]
[72, 92, 204, 215]

Equal Width
Input :[5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]
Output :
[10, 11, 13, 15, 35, 50, 55, 72]
[92]
[204]

Q9. What is dimensionality reduction?(P4- Appeared 1 times)(5-10M)

Ans: **Dimensionality Reduction**

- Dimensionality reduction is the process of reducing the number of random variables or attributes under

consideration. High-dimensional data reduction, as part of a data pre-processing-step, is extremely important in many real-world applications.

- High-dimensionality reduction has emerged as one of the significant tasks in data mining applications. For an example you may have a dataset with hundreds of features (columns in your database).
- Then dimensionality reduction is that you reduce those features of attributes of data by combining or merging them in such a way that it will not lose much of the significant characteristics of the original dataset. One of the major problems that occurs with high dimensional data is widely known as the "Curse of Dimensionality".

Q11. Enlist the preprocessing steps with example. Explain procedure of any technique of preprocessing. - (7 Marks) (P4 - Appeared 1 Time)

Ans: The knowledge discovery process is an iterative sequence of the following steps:
1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the
database)
4. Data transformation (where data are transformed and consolidated into forms

appropriate for mining by performing summary or aggregation operations)4
5. Data mining (an essential process where intelligent methods are applied to extract
data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge
based on interestingness measures—see Section 1.4.6)
7. Knowledge presentation (where visualization and knowledge representation tech-
niques are used to present mined knowledge to users)
Steps 1 through 4 are different forms of data preprocessing, where data are prepared
for mining. The data mining step may interact with the user or a knowledge base. The
interesting patterns are presented to the user and may be stored as new knowledge in the
knowledge base.
The preceding view shows data mining as one step in the knowledge discovery pro-
cess, albeit an essential one because it uncovers hidden patterns for evaluation. However,
in industry, in media, and in the research milieu, the term data mining is often used to
refer to the entire knowledge discovery process (perhaps because the term is shorter
than knowledge discovery from data). Therefore, we adopt a broad view of data min-

ing functionality: Data mining is the process of discovering interesting patterns and
knowledge from large amounts of data. The data sources can include databases, data
warehouses, the Web, other information repositories, or data that are streamed into the
system dynamically.

MODULE-3

Q1. Explain various data normalization techniques. (P4- Appeared 1 time)(5-10M)

Ans: **There are three methods for data normalization:**

1. **min-max normalization :**
   - performs a linear transformation on the original data
   - Suppose that minAand maxA are the minimum and maximum values of an attribute, A.
   - Min-max normalization maps a value, v, of A to v0 in the range [new minA;newmaxA] by computing

$$V_0 = \frac{v - minA}{maxA - minA}(new_{maxA} - new_{minA}) + newminA$$

   - Min-max normalization preserves the relationships among the original data values.
2. **z-score normalization**
   - Here the values for an attribute, A, are normalized based on the mean and standard deviation of A.
   - Value, v of A is normalized to v0 by computing$V_0 \frac{v-\bar{A}}{\sigma A}$, where A and σA are the mean and standard deviation, respectively.
   - This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

3. **normalization by decimal scaling:**
    - Here the normalization is done by moving the decimal point of values of attribute A.
    - The number of decimal points moved depends on the maximum absolute value of A.
    - Value, v of A is normalized to v0 by computing $v_0 = \frac{v}{10^j}$ where j is the smallest integer such that Max $(|v_0|)<1$
    - Attribute construction:
    - Here new attributes are constructed and added from the given set of attributes to help the mining process.
    - Attribute construction helps to improve the accuracy and understanding of structure in high-dimensional data.
    - By combining attributes, attribute construction can discover missing information about the relationships between data attributes that can be useful for knowledge discovery.
    - EG:The structure of stored data may vary between applications, requiring semantic mapping prior to the transformation process. For instance, two applications might store the same customer credit card information using slightly different structures:

APPLICATION A	EXAMPLE	APPLICATION B	EXAMPLE
Cardholder First Name	JOHN	Cardholder Name	JOHN DOE
Cardholder Last Name	DOE	Card Type	VISA
Card Type and Card Number	VISA 0123 4567 8910 1112	Card Number	0123 4567 8910 1112
Expiration Date	05/2012	Expiration Date	05/2012

Q2. What is meant by multidimensional association rules? (P4-Appeared 1 time)(5-10M)

Ans: MULTIDIMENSIONAL ASSOCIATION RULES:

1) In Multi dimensional association:

- Attributes can be categorical or quantitative.
- Quantitative attributes are numeric and incorporate hierarchy.
- Numeric attributes must be discretized.
- Multi dimensional association rule consists of more than one dimension:

Eg: buys(X,"IBM Laptop computer")buys(X,"HP Inkjet Printer")

2) Three approaches in mining multidimensional association rules:

1. Using static discretization of quantitative attributes.
    - Discretization is static and occurs prior to mining.
    - Discretized attributes are treated as categorical.

- Use apriori algorithm to find all k-frequent predicate sets(this requires k or k+1 table scans).
- Every subset of a frequent predicate set must be frequent.
- Eg: If in a data cube the 3D cuboid (age, income, buys) is frequent implies (age, income), (age, buys), (income, buys) are also frequent.
- Data cubes are well suited for mining since they make mining faster.
- The cells of an n-dimensional data cuboid correspond to the predicate cells.

2. Using dynamic discretization of quantitative attributes:
   - Known as mining Quantitative Association Rules.
   - Numeric attributes are dynamically discretized.
   - Eg: age(X,"20..25") ∧ income(X,"30K..41K")buys (X,"Laptop Computer")

	Age=20	Age=21	Age=22	Age=23	Age=24	Age=25
Income,38 to 41						
Income,34 to 37						
Income,30 to 33						

GRID FOR TUPLES

3. Using distance based discretization with clustering.
   - This id dynamic discretization process that considers the distance between data points.

1. It involves a two step mining process:
   a. Perform clustering to find the interval of attributes involved.

b. Obtain association rules by searching for groups of clusters that occur together.

2. The resultant rules may satisfy:
   a. Clusters in the rule antecedent are strongly associated with clusters of rules in the consequent.
   b. Clusters in the antecedent occur together.
   c. Clusters in the consequent occur together.

Q3. Discuss the following terms. 1) Supervised learning 2) Correlation analysis 3) Tree pruning (P4- Appeared 1 times)(5-10M)

Ans: **Supervised learning**

- Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

Supervised learning is classified into two categories of algorithms:

- **Classification:** A classification problem is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease".
- **Regression:** A regression problem is when the output variable is a real value, such as "dollars" or "weight".

Correlation analysis

- Correlation Analysis is a statistical method that is used to discover if there is a relationship between two variables/datasets, and how strong that relationship may be.
- In terms of market research this means that correlation analysis is used to analyze quantitative data gathered from research methods such as surveys and polls, to identify whether there are any significant connections, patterns, or trends between the two.

Tree pruning

- Pruning a decision tree helps to prevent overfitting the training data so that our model generalizes well to unseen data. Pruning a decision tree means to remove a subtree that is redundant and not a useful split and replace it with a leaf node. Decision tree pruning can be divided into two types:

1. **Pre-pruning**
    - Pre-pruning, also known as Early Stopping Rule, is the method where the subtree construction is halted at a particular node after evaluation of some measure. These measures can be the Gini Impurity or the Information Gain. In pre-pruning, we evaluate the pruning condition based on the above measures at each node.
    - Examples of pruning conditions include informationGain(Attr)> minGain or treeDepth == MaxDepth. If the condition is satisfied, we prune the subtree. That means we replace the decision node

with a leaf node. Otherwise, we continue building the tree using our decision tree algorithm.

2. **Post-pruning**
    - As the name suggests, post-pruning means to prune after the tree is built. You grow the tree entirely using your decision tree algorithm and then you prune the subtrees in the tree in a bottom-up fashion. You start from the bottom decision node and, based on measures such as Gini Impurity or Information Gain, you decide whether to keep this decision node or replace it with a leaf node.
    - For example, say we want to prune out subtrees that result in least information gain. When deciding the leaf node, we want to know what leaf our decision tree algorithm would have created if it didn't create this decision node.

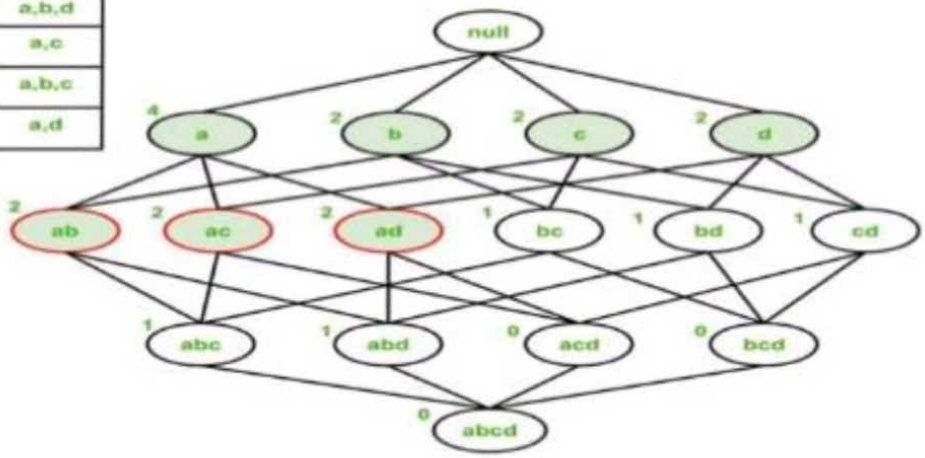
Q4. What is meant by Maximal Frequent Item Set (P4- Appeared 1 times)(5-10M)

Ans: A maximal frequent itemset is a frequent itemset for which none of its immediate supersets are frequent.

- To illustrate this concept, consider the example given below:
- The support counts are shown on the top left of each node. Assume support count threshold = 50%, that is, each item must occur in 2 or more transactions.

- Based on that threshold, the frequent itemsets are: a, b, c, d, ab, ac and ad (shaded nodes).

TID	Items
1	a,b,d
2	a,c
3	a,b,c
4	a,d



Out of these 7 frequent itemsets, 3 are identified as maximal frequent (having red outline):

- ab: Immediate supersets abc and abd are infrequent.
- ac: Immediate supersets abc and acd are infrequent.
- ad: Immediate supersets abd and acd are infrequent.

- The remaining 4 frequent nodes (a, b, c and d) cannot be maximally frequent because they all have at least 1 immediate superset that is frequent.

Advantage:

- Maximal frequent itemsets provide a compact representation of all the frequent itemsets for a particular dataset.
- In the above example, all frequent itemsets are subsets of the maximal frequent itemsets, since we can obtain sets a, b, c, d

by enumerating subsets of ab, ac and ad (including the maximal frequent itemsets themselves).

Disadvantage:

- The support count of maximal frequent itemsets does not provide any information about the support count of their subsets.
- This means that an additional traversal of data is needed to determine support count for non-maximal frequent itemsets, which may be undesirable in certain cases.

Q5. Define time-series database. Explain how to characterize time series data using trend analysis. (P4- Appeared 1 time)(5-10M)

Ans: A time-series database (TSDB) is a software system that is optimized for storing and serving time series through associated pairs of time(s) and value(s).

- Gamma is used when a series has a trend in data.
- Delta is used when seasonality cycles are present in data.
- A model is applied according to the pattern of the data.
- Curve fitting in time series analysis: Curve fitting regression is used when data is in a non-linear relationship.

MODULE-4

Q1. Briefly explain linear and non-linear regression (P2-Appeared 3 times)(5-10M)

Ans: **Linear Regression**

- Linear regression involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other.

1. **Straight-line regression:**
    - Straight-line regression analysis involves a response variable, y, and a single predictor variable, x.
    - It is the simplest form of regression and models y as a linear function of x.
    - That is, y = b+wx;
    - where the variance of y is assumed to be constant, and there are regression coefficients specifying the Y-intercept and slope of the line, respectively.

2. **Multiple linear regression:**
- Multiple linear regression is an extension of straight-line regression so as to involve more than one predictor variable.
- It allows response variable y to be modelled as a linear function of n predictor variables or attributes.
- The equations(obtained from the method of least squares), become long and are tedious to solve by hand.

- Multiple regression problems are instead commonly solved with the use of statistical software packages, such as SAS, SPSS, and S-Plus .

Non-Linear regression

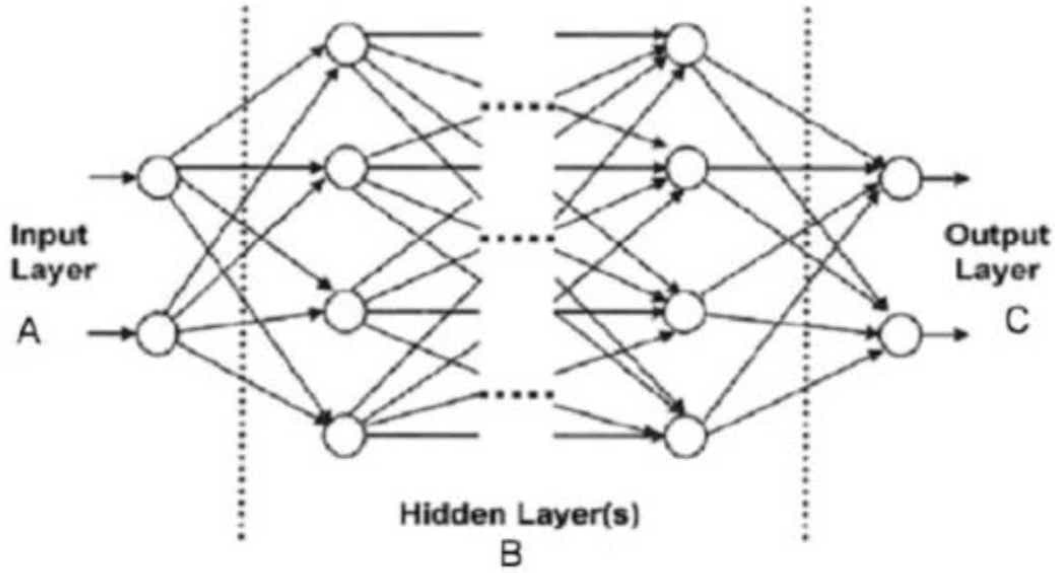
- Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function. Simple linear regression relates two variables (X and Y) with a straight line (y = mx + b), while nonlinear regression relates the two variables in a nonlinear (curved) relationship.
- Nonlinear regression modeling is similar to linear regression modeling in that both seek to track a particular response from a set of variables graphically.
- Nonlinear models are more complicated than linear models to develop because the function is created through a series of approximations (iterations) that may stem from trial and error.

Q2. Draw and explain the topology of a multilayer feed-forward Neural Network with diagrams. (P2-Appeared 3 times)

Ans: Schematic diagram of a multilayer feedforward neural network Following some suitable operation, it results in the desired output.

- Knowledge is usually stored as a set of connecting weights (presumably corresponding to synapse efficiency in the biological neural system) (Santosh et al., 1993).

- A neural network is a massively parallel-distributed processor that has a natural propensity for storing experiential knowledge and making it available for use.
- It resembles the human brain in two respects; the knowledge is acquired by the network through a learning process, and interneuron connection strengths known as synaptic weights are used to store the knowledge (Haykin, 1994).



- Training is the process of modifying the connection weights in an orderly fashion using a suitable learning method.
- The network uses a learning mode, in which an input is presented to the network along with the desired output and the weights are adjusted so that the network attempts to produce the desired output.
- Weights after training contain meaningful information whereas before training they are random and have no meaning (Kalogirou, 2001).

- Two different types of learning can be distinguished: supervised and unsupervised learning, in supervised learning it is assumed that at each instant of time when the input is applied, the desired response d of the system is provided by the teacher.
- This is illustrated in Figure 5-a. The distance ρ [d,o] between the actual and the desired response serves as an error measure and is used to correct network parameters externally.
- Since adjustable weights are assumed, the teacher may implement a reward-and-punishment scheme to adopt the network's weight.
- For instance, in learning classifications of input patterns or situations with known responses, the error can be used to modify weights so that the error decreases.
- This mode of learning is very pervasive.
- Also, it is used in many situations of learning. A set of input and output patterns called a training set is required for this learning mode.
- Figure 5-b shows the block diagram of unsupervised learning. In unsupervised learning, the desired response is not known; thus, explicit error information cannot be used to improve the network's behaviour.
- Since no information is available as to correctness or incorrectness of responses, learning must somehow be accomplished based on observations of responses to inputs that we have mar- ginal or no knowledge about (Zurada, 1992).

Q3. Differentiate classification and prediction (P3-Appeared 2 times) (3-7 M)

Ans: **DIFFERENCE BETWEEN CLASSIFICATION & PREDICTION.**

CLASSIFICATION	PREDICTION
Prediction is about predicting a missing/unknown element(continuous value) of a dataset	Classification is about determining a (categorial) class (or label) for an element in a dataset
Eg. We can think of prediction as predicting the correct treatment for a particular disease for an individual person.	Eg. Whereas the grouping of patients based on their medical records can be considered classification.
The model used to predict the unknown value is called a predictor.	The model used to classify the unknown value is called a classifier.
The predictor is constructed from a training set and its accuracy refers to how well it can estimate the value of new data.	A classifier is also constructed from a training set composed of the records of databases and their corresponding class names.

Q4. State the Apriori Property. Generate large itemsets and association rules using Apriori algorithm on the following data set with minimum support value and minimum confidence value set as 50% and 75% respectively. Given Table: (P4- Appeared 1 times)(5-10M)

Ans: Consider the following transaction database.

TID	Items
01	A,B,C,D
02	A,B,C,D,E,G
03	A,C,G,H,K
04	B,C,D,E,K
05	D,E,F,H,L
06	A,B,C,D,L
07	B,I,E,K,L
08	A,B,D,E,K
09	A,E,F,H,L
10	B,C,D,F

Q5. What is classification? Explain classification as a two step process with diagrams. (P4- Appeared 1 times)(5-10M)

Ans: **Classification**

- Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.
- A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on.
- Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.

It is a two-step process such as :

1. **Learning Step (Training Phase):**
   Construction ofClassification Model Different Algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.
2. **Classification Step:**

Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.Test data are used to estimate the accuracy of the classification rule.

Q6. **Consider a transactional database where 1, 2, 3, 4, 5, 6, 7 are items. Given Table: Suppose the minimum support is 60%. Find all frequent itemsets using Apriori algorithm. (P4- Appeared 1 times)(5-10M)**

Ans: Use Apriori algorithm to find all frequent itemsets.

- Itemsets shaded in gray are removed because they fail the minimum support constraint.
- Those shaded in light yellow are removed because there exists a subset of itemsets that is not frequent. Minimum support = 5 × 60% = 3

C_1

Itemset	Support
{1}	5
{2}	4
{3}	4
{4}	2
{5}	3
{6}	2
{7}	1

→

L_1

Itemset	Support
{1}	5
{2}	4
{3}	4
{5}	3

C_2

Itemset	Support
{1, 2}	4
{1, 3}	4
{1, 5}	3
{2, 3}	3
{2, 5}	3
{3, 5}	2

→

L_2

Itemset	Support
{1, 2}	4
{1, 3}	4
{1, 5}	3
{2, 3}	3
{2, 5}	3

C_3

Itemset	Support
{1, 2, 3}	3
{1, 2, 5}	3
{1, 3, 5}	
{2, 3, 5}	

→

L_3

Itemset	Support
{1, 2, 3}	3
{1, 2, 5}	3

C_4

Itemset	Support
{1, 2, 3, 5}	

Q7. Briefly outline the major steps of decision tree classification.

Why is tree pruning useful in decision tree induction?(P4- Appeared 1 times)(5-10M)

Ans: Step 1: Determine the Root of the Tree.

Step 2: Calculate Entropy for The Classes.

Step 3: Calculate Entropy After Split for Each Attribute.

Step 4: Calculate Information Gain for each split.

Step 5: Perform the Split.
Step 6: Perform Further Splits.
Step 7: Complete the Decision Tree.

- The decision tree built may overfit the training data. There could be too many branches, some of which may reflect anomalies in the training data due to noise or outliers.
- Tree pruning addresses this issue of overfitting the data by removing the least reliable branches (using statistical measures).
- This generally results in a more compact and reliable decision tree that is faster and more accurate in its classification of data.
- The drawback of using a separate set of tuples to evaluate pruning is that it may not be representative of the training tuples used to create the original decision tree.
- If the separate set of tuples are skewed, then using them to evaluate the pruned tree would not be a good indicator of the pruned tree's classification accuracy.
- Furthermore, using a separate set of tuples to evaluate pruning means there are less tuples to use for creation and testing of the tree.
- While this is considered a drawback in machine learning, it may not be so in data mining due to the availability of larger data sets.

Q8. Explain the following as attribute selection measure: (i) Information Gain (ii) Gain Ratio. (P4- Appeared 1 times)(5-10M)

Ans : **Information Gain :**

- Information gain is the amount of information that's gained by knowing the value of the attribute, which is the entropy of the distribution before the split minus the entropy of the distribution after it. The largest information gain is equivalent to the smallest entropy.
- Information gain is the amount of information gained by knowing the value of the attribute
- Information gain=(Entropy of distribution before the split)−(entropy of distribution after it)

Gain Ratio :

- Gain Ratio is modification of information gain that reduces its bias. Gain ratio overcomes the problem with information gain by taking into account the number of branches that would result before making the split.It corrects information gain by taking the intrinsic information of a split into account.We can also say Gain Ratio will add penalty to information gain.
- We already know how to calculate Gain or Information Gain Where S1,S2.....Sn are the subset resulting partitioning S for Attribute A

For example:

From the below image I can say A is sunny where S1,S2...Sn are the subsets like {Yes,No}

Q9. Discuss the following terms: 1) Sequence Databases. 2) Correlation Databases. (P4- Appeared 1 times)(5-10M)

Ans: In the field of bioinformatics, a sequence database is a type of biological database that is composed of a large collection of computerized nucleic acid sequences, protein sequences, or other polymer sequences stored on a computer.

- The UniProt database is an example of a protein sequence database.
- A correlational analysis is a statistical technique employed to investigate the magnitude and significance of such relationships.
- This paper presents commonly used techniques to examine bivariate relationships of interval/ratio, ordinal and nominal variables.

Q10. Discuss the following terms: 1) Spatiotemporal Databases 2) Tree Pruning (P4- Appeared 1 times)(5-10M)

Ans: A spatiotemporal database is a database that manages both space and time information. Common examples include:

- Tracking of moving objects, which typically can occupy only a single position at a given time.
- A database of wireless communication networks, which may exist only for a short timespan within a geographic region.

- An index of species in a given geographic region, where over time additional species may be introduced or existing species migrate or die out.
- Historical tracking of plate tectonic activity.
- Spatiotemporal databases are an extension of spatial databases and temporal databases.
- A spatiotemporal database embodies spatial, temporal, and spatiotemporal database concepts, captures spatial and temporal aspects of data and deals with:
  - geometry changing over time and/or
  - location of objects moving over invariant geometry
  - Pruning a decision tree helps to prevent overfitting the training data so that our model generalizes well to unseen data. Pruning a decision tree means removing a subtree that is redundant and not a useful split and replacing it with a leaf node.

Q11. Enlist the steps of the ID3 decision tree generation algorithm.

Explain it with suitable examples and generate the tree. (P4-Appeared 1 times)(5-10M)

Ans: **ID3 Steps**

1. Calculate the Information Gain of each feature.
2. Considering that all rows don't belong to the same class, split the dataset S into subsets using the feature for which the Information Gain is maximum.
3. Make a decision tree node using the feature with the maximum Information gain.

4. If all rows belong to the same class, make the current node as a leaf node with the class as its label.
5. Repeat for the remaining features until we run out of all features, or the decision tree has all leaf nodes.

Q12. Explain Bayes Theorem and Naïve Bayesian Classification with suitable example (P4- Appeared 1 time)(5-10M)

Ans: Bayes' theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability.

- Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring.
- Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.
- In finance, Bayes' theorem can be used to rate the risk of lending money to potential borrowers.
- Bayes' theorem is also called Bayes' Rule or Bayes' Law and is the foundation of the field of Bayesian statistics.

Numerical Example Of Bayes' Theorem

- As a numerical example, imagine there is a drug test that is 98% accurate, meaning 98% of the time it shows a true positive result for someone using the drug and 98% of the time it shows a true negative result for nonusers of the drug.
- Next, assume 0.5% of people use the drug.

- If a person selected at random tests positive for the drug, the following calculation can be made to see whether the person is actually a user of the drug.
  $(0.98 \times 0.005) / [(0.98 \times 0.005) + ((1 - 0.98) \times (1 - 0.005))] = 0.0049 / (0.0049 + 0.0199) = 19.76\%$
- Bayes' theorem shows that even if a person tested positive in this scenario, it is actually much more likely the person is not a user of the drug.
- Bayes' theorem allows you to update the predicted probabilities of an event by incorporating new information.
- Bayes' theorem was named after 18th-century mathematician Thomas Bayes.
- It is often employed in finance in updating risk evaluation.
- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.
- It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
- The fundamental Naive Bayes assumption is that each feature makes an:
  - independent
  - Equal

contribution to the outcome.

With relation to our dataset, this concept can be understood as:

- We assume that no pair of features are dependent. For example, the temperature being 'Hot' has nothing to do with the humidity or the outlook being 'Rainy' has no effect on the winds. Hence, the features are assumed to be independent.

- Secondly, each feature is given the same weight(or importance). For example, knowing the only temperature and humidity alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing equally to the outcome.

Q13. Discuss the Gain ratio as an attribute selection measure.

Ans : Prediction of precipitation is a necessary tool in meteorology. To date, it is technologically and scientifically a challenging task for scientists and researchers around the globe.

- Rainfall is a liquid form of precipitation that depends primarily on humidity, temperature, pressure, wind speed, dew point, and so on. Because rainfall depends on several parameters, its prediction becomes very complex.
- Approaches such as the backpropagation, linear regression, support vector machine, Bayesian networks, and fuzzy logic can be applied, but their rate of prediction is very low, which leads to unpredictable results.
- This paper aims at improving the prediction of precipitation compared to Supervised Learning in Quest (SLIQ) decision trees, especially when datasets are large.
- Because SLIQ decision trees take more computational steps to find split points, they consume more time and thus cannot be applied to huge datasets.
- An elegant decision tree using the gain ratio as an attribute selection measure is adopted, which increases the accuracy rate and decreases the computation time.

- This approach provides an average accuracy of 76.93% with a reduction of 63% in computational time over SLIQ decision trees.

Q14. Discuss Hash-based technique to improve the efficiency of the Apriori algorithm.

Ans : Apriori with hashing Algorithm As we know the apriori algorithm has some weaknesses so to reduce the span of the hopeful k-item sets, Ck hashing technique is used. Our hash-based Apriori execution utilizes the data structure that specifically speaks to a hash table.

- Specifically, the 2-itemsets, since that is the way to enhance execution. This calculation utilizes a hash-based procedure to minimize the quantity of applicant item sets created in the 1st pass.
- It is guaranteed that the number of item sets in C2 produced utilizing hashing can be smaller so that the output required to decide L2 is more efficient.
- For instance, while scanning every transaction in the database to create the Frequent-itemsets, L1, from the candidate 1-itemsets in C1, we can produce the majority of the 2-itemsets for every transaction, hash(i.e) map into the diverse bucket of a hash table structure, and increment the complementary bucket count.
- A 2-itemset whose complementary bucket count in the hash table is below the min_sup threshold cannot be frequent and thus should be reduced from the candidate set. So

hash-based apriori reduces the number of candidate k-item sets.

Steps:

1. Scan database transaction. generate frequent-1 item set. Then after generating a frequent-2 item set.
2. Let's take a hash table of size 7.
3. For each bucket appoint a candidate set utilizing the ASCII estimations of the itemsets.
4. Each bucket in the hash table has a count, which is expanded by 1 every item an item set is hashed to that bucket.
5. If the bucket count satisfies the min_sup threshold value then the bit vector is set to 1, otherwise is set to 0.
6. The candidate pairs that bit vector bits are not sets that are removed.

Comparisons of algorithms:

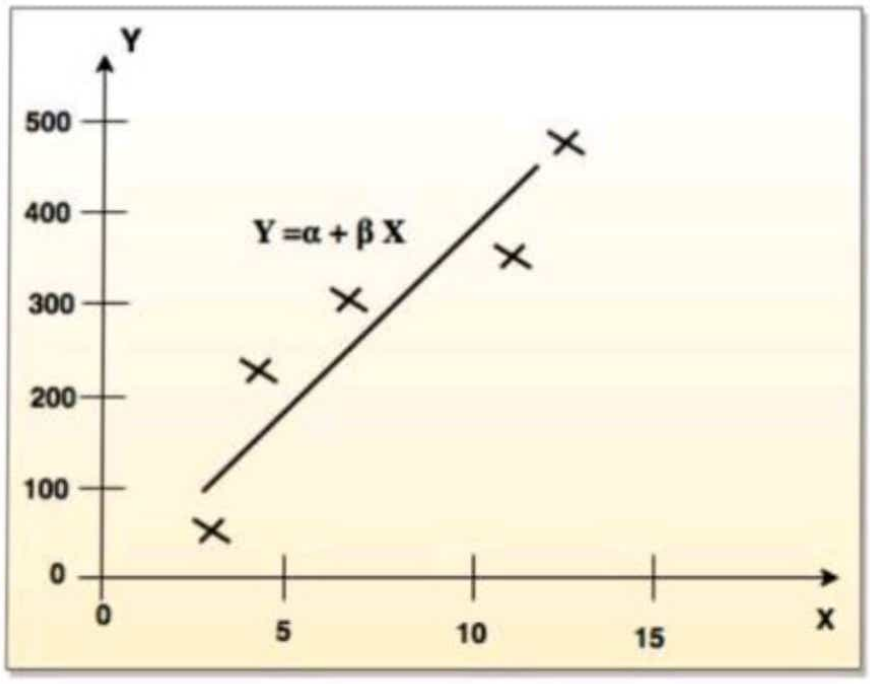
Apriori	**Apriori with hashing**
Pros: 1) Algo uses info from previous steps to produce the frequent itemsets. 2) Easy to implement	**Pros :** 1)Reduce the number of scans. 2)Remove the large candidates that cause high Input/Output cost.
Cons: 1) Databases need to scan at every level. 2) Uses more space and	**Cons:** 1) As DB size increases the size of the bucket also increases.

memory time. 3) In the case of a large database it is not efficient.	2) For large DBs it is difficult to handle hash tables and candidate sets. 3) Execution time is small for a small DB.

Q15. Explain Linear regression with an example.

Ans : **Linear regression**

- It is the simplest form of regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe the data.
- Linear regression attempts to find the mathematical relationship between variables.
- If the outcome is a straight line then it is considered a linear model and if it is a curved line, then it is a non linear model.
- The relationship between dependent variables is given by a straight line and it has only one independent variable.
  $Y = \alpha + B X$
- Model 'Y', is a linear function of 'X'.
- The value of 'Y' increases or decreases in a linear manner according to which the value of 'X' also changes.



Linear Regression

Q16. Explain various conflict resolution strategies in rule based classification.

Ans :IF-THEN Rules

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

IF condition THEN conclusion

Let us consider a rule R1,

R1: IF age = youth AND student = yes

THEN buy_computer = yes

Points to remember -

- The IF part of the rule is called rule antecedent or precondition.
- The THEN part of the rule is called rule consequent.
- The antecedent part of the condition consists of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.

R1: (age = youth) ^ (student = yes))(buys computer = yes)

If the condition holds true for a given tuple, then the antecedent is satisfied.

Rule Extraction

Here we will learn how to build a rule-based classifier by extracting IF-THEN rules from a decision tree.

Points to remember -

To extract a rule from a decision tree -

- One rule is created for each path from the root to the leaf node.
- To form a rule antecedent, each splitting criterion is logically ANDed.
- The leaf node holds the class prediction, forming the rule consequent.

Rule Induction Using Sequential Covering Algorithm

- Sequential Covering Algorithm can be used to extract IF-THEN rules from the training data. We do not require to generate a decision tree first. In this algorithm, each rule for a given class covers many of the tuples of that class.
- Some of the sequential Covering Algorithms are AQ, CN2, and RIPPER. As per the general strategy the rules are learned one at a time. For each time rules are learned, a tuple covered by the rule is removed and the process continues for the rest of

the tuples. This is because the path to each leaf in a decision tree corresponds to a rule.

- The Following is the sequential learning Algorithm where rules are learned for one class at a time. When learning a rule from a class Ci, we want the rule to cover all the tuples from class C only and no tuples from any other class.

Algorithm: Sequential Covering

Input:
D, a data set class-labeled tuples,
Att_vals, the set of all attributes and their possible values.

Output: A Set of IF-THEN rules.
Method:
Rule_set={ }; // initial set of rules learned is empty

for each class c do

  repeat
    Rule = Learn_One_Rule(D, Att_valls, c);
    remove tuples covered by Rule form D;
  until termination condition;

  Rule_set=Rule_set+Rule; // add a new rule to rule-set
end for
return Rule_Set;

Rule Pruning
The rule is pruned is due to the following reason -

- The Assessment of quality is made on the original set of training data. The rule may perform well on training data but less well on subsequent data. That's why the rule pruning is required.
- The rule is pruned by removing conjunctions. The rule R is pruned if the pruned version of R has greater quality than what was assessed on an independent set of tuples.

FOIL is one of the simple and effective methods for rule pruning. For a given rule R,

FOIL_Prune = pos - neg / pos + neg

where pos and neg is the number of positive tuples covered by R, respectively.

Q17. Discuss the variations of the Apriori algorithm to improve the efficiency.

Ans :There are some variations of the Apriori algorithm that have been projected that target developing the efficiency of the original algorithm which are as follows -

The hash-based technique (hashing itemsets into corresponding buckets) -

- A hash-based technique can be used to decrease the size of the candidate k-itemsets, Ck, for k > 1. For instance, when scanning each transaction in the database to create the frequent 1-itemsets,L1, from the candidate 1-itemsets in C1, it can make some 2-itemsets for each transaction, hash (i.e., map) them into the several buckets of a hash table structure, and increase the equivalent bucket counts.

Transaction reduction -

- A transaction that does not include some frequent k-itemsets cannot include some frequent (k + 1)-itemsets. Thus, such a transaction can be marked or deleted from further consideration because subsequent scans of the database for j-itemsets, where j > k, will not need it.

Partitioning -

- A partitioning technique can be used that needs two database scans to mine the frequent itemsets. It includes two phases involving In Phase I, the algorithm subdivides the transactions of D into n non-overlapping partitions. If the minimum support threshold for transactions in D is min_sup, therefore the minimum support count for a partition is min_sup × the number of transactions in that partition.

Sampling -

- The fundamental idea of the sampling approach is to select a random sample S of the given data D, and then search for frequent itemsets in S rather than D. In this method, it can trade off some degree of accuracy against efficiency. The sample size of S is such that the search for frequent itemsets in S can be completed in main memory, and therefore only one scan of the transactions in S is needed overall.

MODULE-5

Q1. Define "clustering"? Mention any two applications of clustering (P3-Appeared 2 times)(5-10M)

Ans: **Clustering**:

- Clustering is the process of making a group of abstract objects into classes of similar objects.
- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over-classification is that it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

Q2. Explain Mean, Median, Mode, Variance, Standard Deviation & five-number summary with suitable database example.
(P3-Appeared 2 times)

Ans: **The Mean and Mode:**

- The *sample mean* is the average and is computed as the sum
- of all the observed outcomes from the sample divided by the total number of events.
- We use x as the symbol for the sample mean. In math terms,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

where n is the sample size and the x corresponds to the observed value.

Example:

Suppose you randomly sampled six acres in the Desolation Wilderness for a non-indigenous weed and came up with the following counts of this weed in this region:

34, 43, 81, 106, 106 and 115

We compute the sample mean by adding and dividing by the number of samples, 6.

$$\frac{34 + 43 + 81 + 106 + 106 + 115}{6} = 80.83$$

We can say that the sample means of non-indigenous weed is 80.83.

- The *mode* of a set of data is the number with the highest frequency. In the above example, 106 is the mode, since it occurs twice and the rest of the outcomes occur only once.
- The *population mean* is the average of the entire population and is usually impossible to compute. We use the Greek letter m for the population mean.

Median:

- One problem with using the mean, is that it often does not depict the typical outcome.
- If there is one outcome that is very far from the rest of the data, then the mean will be strongly affected by this outcome.
- Such an outcome is called an *outlier*.
- An alternative measure is the median.
- The *median* is the middle score.
- If we have an even number of events we take the average of the two middles.
- The median is better for describing the typical value. It is often used for income and home prices.

Example:

Suppose you randomly selected 10 house prices in the South Lake Tahoe area.

You are interested in the typical house price. In $100,000 the prices were

2.7, 2.9, 3.1, 3.4, 3.7, 4.1, 4.3, 4.7, 4.7, 40.8

- If we computed the mean, we would say that the average house price is 744,000.
- Although this number is true, it does not reflect the price for available housing in South Lake Tahoe.

- A closer look at the data shows that the house valued at 40.8 x $100,000 = $4.08 million skews the data.
- Instead, we use the median. Since there is an even number of outcomes, we take the average of the middle two

$$\frac{3.7 + 4.1}{2} = 3.9$$

- The median house price is $390,000.
- This better reflects what house shoppers should expect to spend.

Variance and Standard Deviation:

- The mean, mode, median, and trimmed mean do a nice job in telling where the center of the data set is, but often we are interested in more.
- For example, a pharmaceutical engineer develops a new drug that regulates iron in the blood.
- Suppose she finds out that the average sugar content after taking the medication is the optimal level.
- This does not mean that the drug is effective.
- There is a possibility that half of the patients have dangerously low sugar content while the other half have dangerously high content.
- Instead of the drug being an effective regulator, it is a deadly poison.
- What the pharmacist needs is a measure of how far the data is spread apart.
- This is what the variance and standard deviation do.
- First we show the formulas for these measurements.
- Then we will go through the steps on how to use the formulas.

We define the *variance* to be

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

and the *standard deviation* to be

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

Five-number summary:

- The five-number summary of a data set consists of the five numbers determined by computing the minimum, Q_1, median, Q_3, and maximum of the data set.

Example 1: Find the five-number summary for the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.

From our Example 1's on the previous pages, we see that the five-number summary is:

Minimum: 3 Q_1: 6 Median: 12 Q_3: 16

Maximum: 21

Q3. Explain cluster analysis and outlier analysis with examples.

(P4- Appeared 1 times)(5-10M)

Ans: **Cluster Analysis**

- Cluster Analysis is the process to find similar groups of objects in order to form clusters.It is an unsupervised machine

learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

Properties of Clustering :

1. **Clustering Scalability:** Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable if it is not scalable, then we can't get the appropriate result and would lead to wrong results.
2. **High Dimensionality:** The algorithm should be able to handle high dimensional space along with the data of small size.
3. **Algorithm Usability with multiple data kinds:** Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.
4. **Dealing with unstructured data:** These would be some databases that contain missing values, noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data and give it some structure to the data by organizing it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.
5. **Interpretability:** The outcomes of clustering should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Outlier Analysis :

- Outliers are mostly discarded when methods of data mining are applied. But, it's still used in certain applications like fraud detection. This is mainly because the events that rarely occur can store much more interesting facts than the events that occur more regularly.

Other applications where outlier detection plays a major role are:

- Detection of frauds in the insurance sector, credit cards, and the healthcare sector.
- Fraud detection in telecom.
- In cybersecurity for detecting any form of intrusion.
- In the field of medical analysis.
- Detection of faults in the safety-critical systems.
- In marketing, outlier analysis helps in identifying the customer's nature of spending.
- Any sort of unusual responses that occur due to certain medical treatments can be analyzed through outlier analysis in data mining.

The process where the anomalous behavior of the outliers is identified in a dataset is known as outlier analysis. Also, known as "outlier mining", the process is defined to be an important task of data mining.

Q4. What do you mean by learning-by-observation?

Ans : Observational learning is the process of learning by watching the behaviors of others.

- The targeted behavior is watched, memorized, and then mimicked.
- Also known as shaping and modeling, observational learning is most common in children as they imitate behaviors of adults.

Four Processes of Observational Learning

1. Attention
2. Retention
3. Reproduction
4. Motivation

1. Attention

- To learn, an observer must pay attention to something in the environment. They must notice the model and the behavior occurring. Attention levels can vary based on the characteristics of the model and environment – including the model's degree of likeness, or the observer's current mood.
- In humans, it is likely the observer will pay attention to behaviors of models that are high-status, talented, intelligent, or similar to the observer in any wa
- For example, if you want to become a VP at your company, it makes sense that you'd observe the current VP's (or other renowned VP's in your industry) and try to mimic their behavior

2. Retention

- Simple attention is not enough to learn a new behavior. An observer must also retain, or remember, the behavior at a later time
- To increase chances of retention, the observer must structure the information in an easy-to-remember format. Maybe they use a mnemonic device. Or form a daily learning habit
- The behavior must be easily remembered so the action can be performed with little or no effort
- Using our VP example above, let's say the current VP is giving a company-wide presentation. You notice that they are calm, confident, engaging, and use eye contact. You make a list of these attributes and remember them for the next time you give a presentation

3. Reproduction

- The behavior is remembered. But can it be performed in real-life? Reproduction is the process where the observer must be able to physically perform the behavior in the
- real-world. Easier said than done
- Often, producing a new behavior requires hours of practice to obtain the skills. You can't just watch your VP give a brilliant company-wide presentation, then use only the observed tactics in your own presentation 20-minutes later. Those skills take years to craft and perfect
- Using our VP example again, you've observed and identified four skills that the current VP uses during presentations. To be able to perform these skills yourself, you need to deliberately practice these skills. Maybe you hold small team meetings to

test your skills. Or you ask team members for feedback on your presentation skills. In a few months, you will have sharpened your presenting skills and may be ready to produce a behavior similar to the current VP.

4. Motivation

- All learning requires some degree of personal motivation. For observational learning, the observer must be motivated to produce the desired behavior
- Sometimes this motivation is intrinsic to the observer. Other times, motivation can come in the form of external reinforcement – rewards and punishments.
- Using our VP example again, the motivation is intrinsic. You understand that the path to becoming a VP at your company requires a certain skill set.

Q5. Explain k-Means clustering algorithm with suitable examples.

Enlist the steps of the K-Mean clustering algorithm. (P4- Appeared 1 times)(5-10M)

Ans: K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

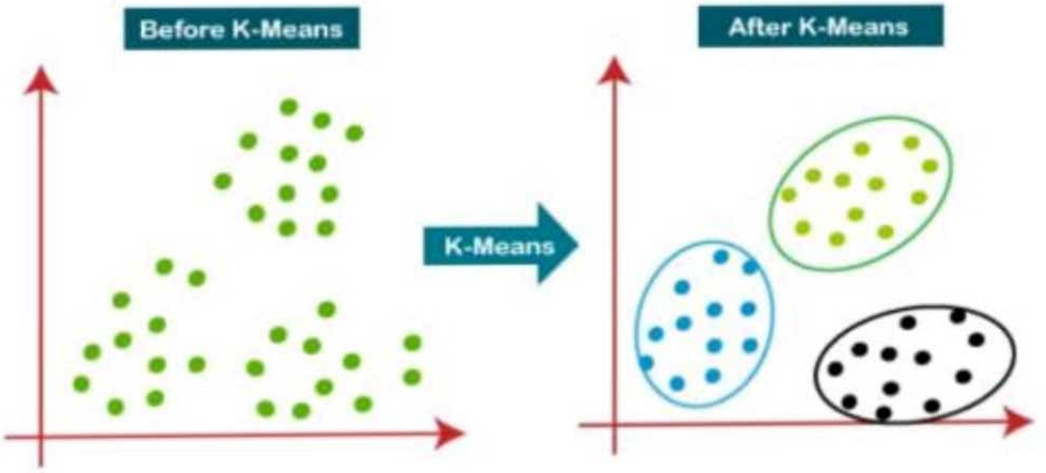
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters.

- Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

- It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties.
- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.
- Hence each cluster has data points with some commonalities, and it is away from other clusters.
- The below diagram explains the working of the K-means Clustering Algorithm:



How does the K-Means Algorithm Work?

- The working of the K-Means algorithm is explained in the below steps:
- Step-1: Select the number K to decide the number of clusters.
- Step-2: Select random K points or centroids. (It can be different from the input dataset).
- Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
- Step-4: Calculate the variance and place a new centroid of each cluster.
- Step-5: Repeat the third steps, which means assign each datapoint to the new closest centroid of each cluster.
- Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
- Step-7: The model is ready.

K-means Clustering – Example 1:

- A pizza chain wants to open its delivery centers across a city. What do you think would be the possible challenges?

- They need to analyze the areas from where the pizza is being ordered frequently.
- They need to understand how many pizza stores have to be opened to cover delivery in the area.
- They need to figure out the locations for the pizza stores within all these areas in order to keep the distance between the store and delivery points minimum.
- Resolving these challenges includes a lot of analysis and mathematics.

Q6. Define outlier analysis? Why is outlier mining important? (P4-Appeared 1 time)(5-10M)

Ans: **Outlier Analysis :**

- Outliers are mostly discarded when methods of data mining are applied. But, it's still used in certain applications like fraud detection. This is mainly because the events that rarely occur can store much more interesting facts than the events that occur more regularly.

Other applications where outlier detection plays a major role are:

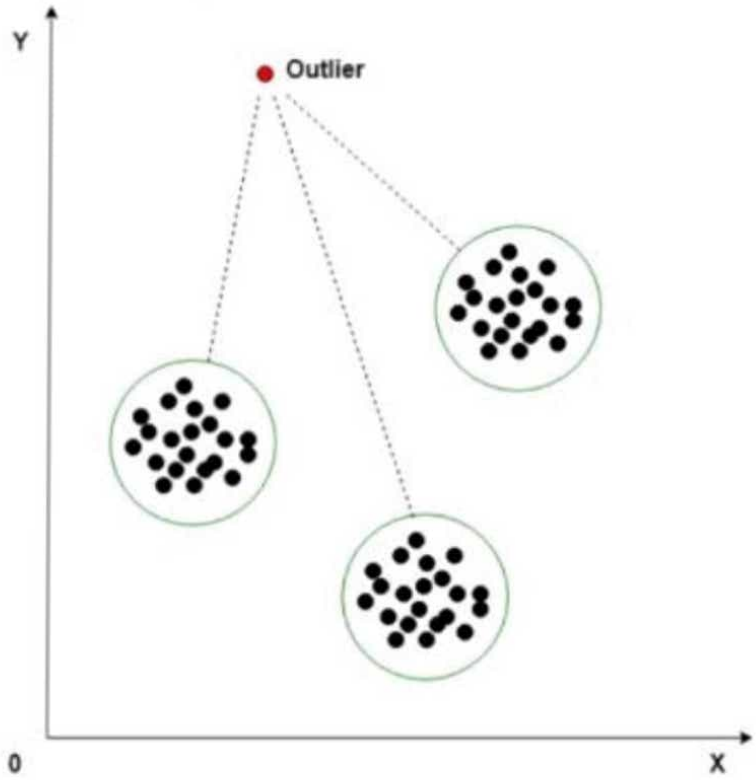
- Detection of frauds in the insurance sector, credit cards, and the healthcare sector.
- Fraud detection in telecom.
- In cybersecurity for detecting any form of intrusion.
- In the field of medical analysis.
- Detection of faults in the safety-critical systems.
- In marketing, outlier analysis helps in identifying the customer's nature of spending.

- Any sort of unusual responses that occur due to certain medical treatments can be analyzed through outlier analysis in data mining.

The process where the anomalous behavior of the outliers is identified in a dataset is known as outlier analysis. Also, known as "outlier mining", the process is defined to be an important task of data mining.

Why outlier analysis?

- Most data mining methods discard outliers noise or exceptions, however, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring one and hence, the outlier analysis becomes important in such cases.



Q7. How are concept hierarchies useful in data mining? (P4-Appeared 1 time)(5-10M)

Ans: 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.
- Transforming nominal data with the use of concept hierarchies allows higher-level knowledge patterns to be found.
- It allows mining at multiple levels of abstraction, which is a common requirement for data mining applications.

Q8. List out the General applications of Clustering

Ans : **Applications of Clustering:**

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Q9. If Mean salary is 54,000Rs and standard deviation is 16,000 Rs then find z score value of 73,600 Rs salary.

Ans : The formula for z score is $\frac{(x1-mean)}{StandardDeviation}$;

z score= $\frac{73600-540000}{16000}$

z-score= $\frac{19600}{16000}$

z-score=1.225

MODULE-6

Q1. Explain Web structure and Web usage mining (P2-Appeared 3 times)(5-10M)

Ans: **Web Structure Mining:**

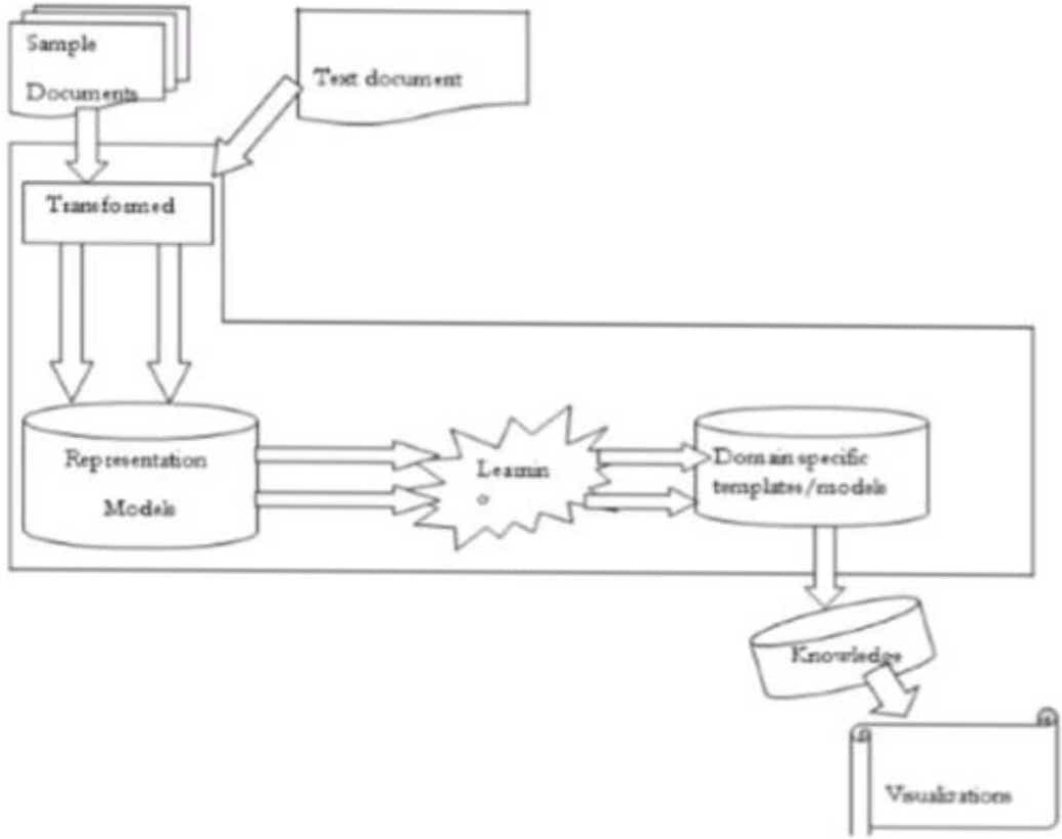
- Web structure mining is the application of discovering structure information from the web.
- The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages.
- Structure mining basically shows the structured summary of a particular website.
- It identifies relationships between web pages linked by information or direct link connection.
- To determine the connection between two commercial websites, Web structure mining can be very useful.

Web Usage Mining:

- Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets.
- And these patterns enable you to understand the user behaviors or something like that.
- In web usage mining, users access data on the web and collect data in the form of logs. So, Web usage mining is also called log mining.

Q2. Explain text mining using examples.(P3-Appeared 2 times)(5-10M)

Ans: Text Mining is the procedure of synthesizing information, by analyzing relations, patterns, and rules among textual data-semi structured or unstructured text.



- This procedure contains text summarization, text categorization and text clustering.
- Text summarization is the procedure to extract its partial content reflection to its whole contents automatically.
- Text categorization is the procedure of assigning a category to the text among categories predefined by users.

- Text clustering is the procedure of segmenting texts into several clusters, depending on the substantial relevance.

Techniques:

- Data mining
- Machine learning
- Information retrieval
- Statistics
- Natural –language understanding
- Case-based reasoning

Text Mining Approaches:

1. Keyword based Association Analysis:
    a. Collect sets of keywords or terms that occur frequently together and then find the association or correlation relationship among them.
    b. First preprocess the text data by parsing, stemming, removing stop words, etc.
    c. Then evoke association mining algorithms -Consider each document as a transaction -View a set of keywords in the document as a set of items in the transaction.
    d. Term level association mining
    e. No need for human effort in tagging documents. -The number of meaningless results and the execution time is greatly reduced.
2. Document Classification Analysis:
   Automatic document classification:
    - Automatic classification for the tremendous number of on-line text documents(Web pages, emails, etc)

- Text document classification differs from the classification of relational data as document databases are not structured according to attribute-value pairs.

Association-Based Document Classification:

- Extract keywords and terms by information retrieval and simple association analysis techniques.
- Obtain concept hierarchies of keywords and terms using Available term classes such as WordNet, Expert knowledge.
- Classify documents in the training set into class hierarchies.
- Apply term association mining method to discover sets of associated terms.
- Use the term to maximally distinguish one class of documents from others.
- Derive a set of association rules associated with each document class.
- Order the classification rule based on their occurrence frequency and discriminative power.
- Used the rules to classify new documents.

3. Document Clustering Analysis:

1. Automatically group related documents based on their contents.
2. Require no training sets or predetermined taxonomies, generate a taxonomy at runtime,
3. Major steps:
4. Preprocessing: Remove stop words, stem, feature extraction.
5. Hierarchical clustering: Compute similarities applying clustering algorithms.

6. Slicing: Fan out controls; flatten the tree to a configurable number of levels.

Q3. Explain data mining application for fraud detection.

(P3-Appeared 2 times)(5-10M)

Ans : Fraud detection for Telecommunication Industry

- The telecommunications industry has expanded dramatically in the last few years with the development of affordable mobile phone technology.
- Fraud is an adaptive crime, so it needs a special method of intelligent data analysis to detect and prevent it.
- This method exists in the areas of Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning and Statistics. They offer applicable and successful solutions in different areas of fraud crimes.
- At a low level, simple rule-based detection systems use rules such as the apparent use of the same phone in two very distant geographical locations in quick succession, calls which appear to overlap in time and very high value and very long calls.
- At a higher level, statistical summaries of call distributions (often called profiles or signatures at the user level) are compared with thresholds determined either by experts or by application of supervised learning methods to known fraud/nonfraud cases.
- Some forensic accountants specialize in forensic analytics which is the procurement and analysis of electronic data to

reconstruct, detect, and otherwise support a claim of financial fraud.

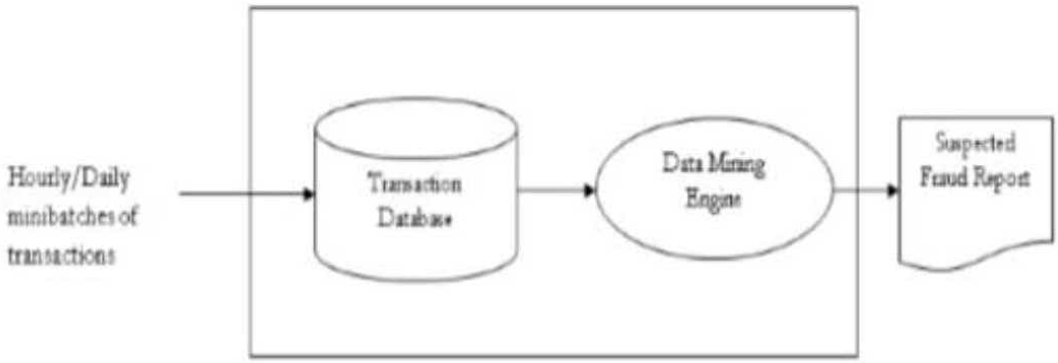


Fig: Fraud Detection

- The main steps in forensic analytics are (a) data collection, (b) data preparation, © data analysis, and (d) reporting.
- For example, forensic analytics may be used to review an employees' purchasing card activity to assess whether any of the purchases were diverted or divertible for personal use.
- Techniques used for fraud detection fall into two primary classes: Statistical techniques and Artificial intelligence.

Q4. Discuss the following terms. 1) DataNode 2) NameNode 3) Text mining (P3- Appeared 2 times)(5-10M)

Ans: **DataNode**

- DataNodes stores data in a Hadoop cluster and is the name of the daemon that manages the data.
- File data is replicated on multiple DataNodes for reliability and so that localized computation can be executed near the data.

- Within a cluster, DataNodes should be uniform. If they are not uniform, issues can occur.
- For example, DataNodes with less memory fill up more quickly than DataNodes with more memory, which can result in job failures.

NameNode

- NameNodes maintain the namespace tree for HDFS and a mapping of file blocks to DataNodes where the data is stored.
- A simple HDFS cluster can have only one primary NameNode, supported by a secondary NameNode that periodically compresses the NameNode edits log file that contains a list of HDFS metadata modifications.
- This reduces the amount of disk space consumed by the log file on the NameNode, which also reduces the restart time for the primary NameNode.
- A high-availability cluster contains two NameNodes: active and standby.

Text Mining

- Text mining, also referred to as text data mining, similar to text analytics, is the process of deriving high-quality information from text.
- It involves "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources."

Q5. Is Graphical visualization better than text data? Justify your answer and explain different data visualization techniques. (P4-Appeared 1 times)(5-10M)

Ans: Data visualization is the graphical representation of information and data.

- By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.
- Visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers.
- If we can see something, we internalize it quickly. It's storytelling with a purpose.
- If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

The uses of Data Visualization as follows.

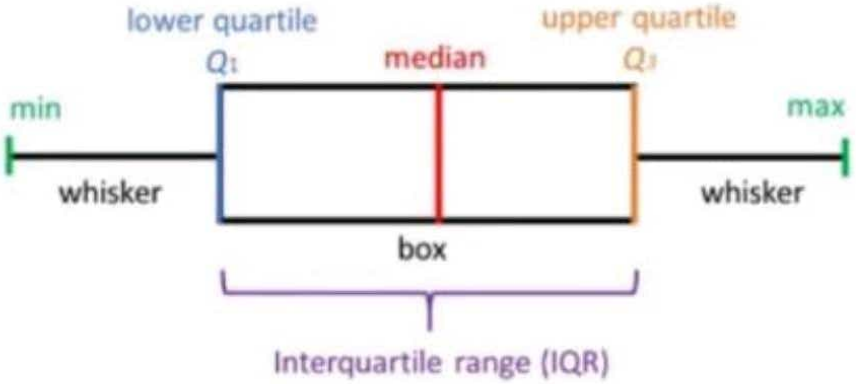
- Powerful way to explore data with presentable results.
- Primary use is the pre-processing portion of the data mining process.
- Supports the data cleaning process by finding incorrect and missing values.
- For variable derivation and selection means to determine which variable to include and discard in the analysis.

- Also play a role in combining categories as part of the data reduction process.

Data Visualization Techniques

- Box plots
- Histograms
- Heat maps
- Charts
- Tree maps
- Word Cloud/Network diagram

Box Plots:



- The image above is a box plot.
- A boxplot is a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").
- It can tell you about your outliers and what their values are.
- It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.
- A box plot is a graph that gives you a good indication of how the values in the data are spread out.

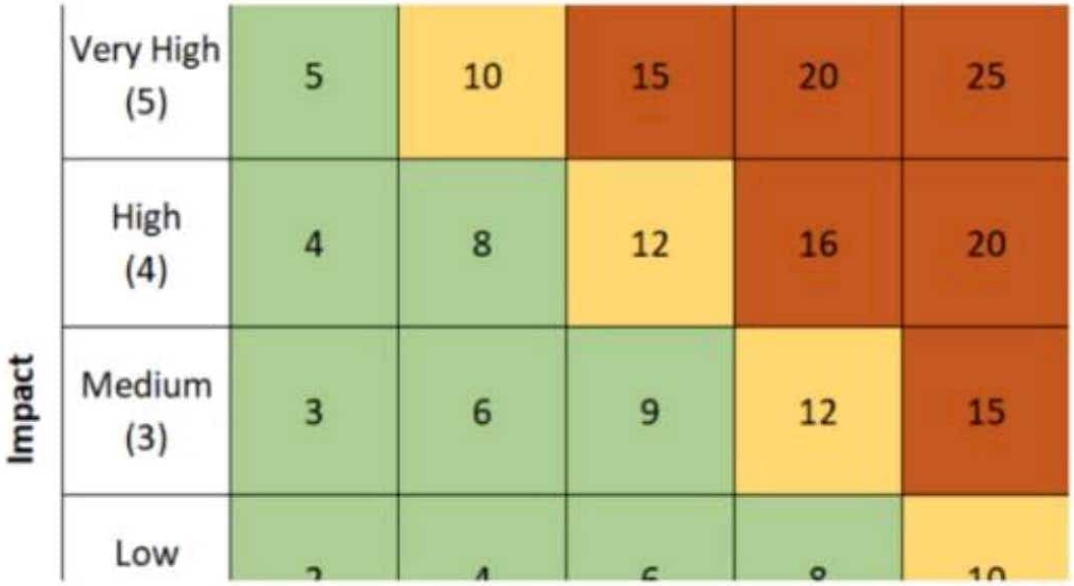
- Although box plots may seem primitive in comparison to a histogram or density plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets.
- For some distributions/datasets, you will find that you need more information than the measures of central tendency (median, mean, and mode).
- You need to have information on the variability or dispersion of the data.

Heat Maps:

- A heat map is data analysis software that uses color the way a bar graph uses height and width: as a data visualization tool.
- If you're looking at a web page and you want to know which areas get the most attention, a heat map shows you in a visual way that's easy to assimilate and make decisions from.

It is a graphical representation of data where the individual values contained in a matrix are represented as colors. Useful for two purposes: for visualizing correlation tables and for visualizing missing values in the data.

- In both cases, the information is conveyed in a two-dimensional table.



- Note that heat maps are useful when examining a large number of values, but they are not a replacement for more precise graphical displays, such as bar charts, because color differences cannot be perceived accurately.

Q5. Explain web mining using example. (3-7 marks) (P4- Appeared 1 Time)

Ans: The World Wide Web serves as a huge, widely distributed, global information center for news, advertisements, consumer information, financial management, education, government, and e-commerce.

- It contains a rich and dynamic collection of information about web page contents with hypertext structures and multimedia, hyperlink information, and access and usage information, providing fertile sources for data mining.

- Web mining is the application of data mining techniques to discover patterns, structures, and knowledge from the Web.
- According to analysis targets, web mining can be organized into three main areas: web content mining, web structure mining, and web usage mining.
- Web content mining analyzes web content such as text, multimedia data, and structured data (within web pages or linked across web pages).
- This is done to understand the content of web pages, provide scalable and informative keyword-based page indexing, entity/concept resolution, web page relevance and ranking, web page content summaries, and other valuable information related to web search and analysis.
- Web pages can reside either on the surface web or on the deep Web. The surface web is that portion of the Web that is indexed by typical search engines.
- The deep Web (or hidden Web) refers to web content that is not part of the surface web. Its contents are provided by underlying database engines.
- Web content mining has been studied extensively by researchers, search engines, and other web service companies. Web content mining can build links across multiple web pages for individuals; therefore, it has the potential to inappropriately disclose personal information. Studies on privacy-preserving data mining address this concern through the development of techniques to protect personal privacy on the Web.
- Web structure mining is the process of using graph and network mining theory and methods to analyze the nodes and

connection structures on the Web. It extracts patterns from hyperlinks, where a hyperlink is a structural component that connects a web page to another location. It can also mine the document structure within a page.

- Both kinds of web structure mining help us understand web contents and may also help transform web contents into relatively structured data sets.
- Web usage mining is the process of extracting useful information (e.g., user click streams) from server logs. It finds patterns related to general or particular groups of users; understands users' search patterns, trends, and associations; and predicts what users are looking for on the Internet.
- It helps improve search efficiency and effectiveness, as well as promotes products or related information to different groups of users at the right time. Web search companies routinely conduct web usage mining to improve their quality of service.
