

## Unit-2

### Introduction to Data Mining

#### Data Mining

Data mining is the process of discovering interesting patterns and knowledge from the huge amount of data. Data Mining is one of the essential step in the process of KDD (Knowledge Discovery in Database).

#### *Why Data Mining? (Motivation)*

- Data mining helps to turn the huge amount of data into useful information and knowledge that can have different applications.
- Data mining helps in
  - a. Automatic discovery of patterns
  - b. Prediction of likely outcomes
  - c. Creation of actionable information
- Data mining can answer questions that cannot be addressed through simple query and reporting techniques.

#### Types of Data that can be mined on Data Mining

Different kinds of data can be mined. Some of the examples are mentioned below:

- ***Flat Files:*** Flat files are in the binary form or text form and having a structure that can be easily extracted by data mining algorithms. The data stored in the flat file has no relationship or path to each other. Flat files are represented by data dictionary. E.g. CSV file. It is often used in data warehousing to store data, in carrying data to and from servers, etc.
- ***Relational Databases:*** A relational database is a data collection organized into tables with rows and columns. The physical schema of a relational database is the schema that defines the structure of the table. A relational database logical schema is a schema that defines the relationships between tables.
- ***Data Warehouses:*** A data warehouse is defined as the collection of data integrated from multiple sources (often heterogeneous) that will queries and decision making. Data warehouses consist of three types, enterprise data warehouses, data marts, and virtual warehouses. It is widely used in everyday business decision-making.
- ***Transaction Databases:*** A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed. Object databases, ATM machine, Banking, and Distributed systems are very famous applications of a transactional database.
- ***Multimedia Databases:*** Multimedia databases include video, images, audio and text media. They can be stored on Object-Oriented Databases. E-book databases, video website databases, news website databases, etc. are famous applications of multimedia databases.
- ***Spatial Databases:*** Spatial databases are databases that store geographical information like maps and global or regional positioning. It stores data in the form of coordinates, topology, lines, polygons, etc.

## Data Mining Architecture

The major components of a *data mining system architecture* are as follows:

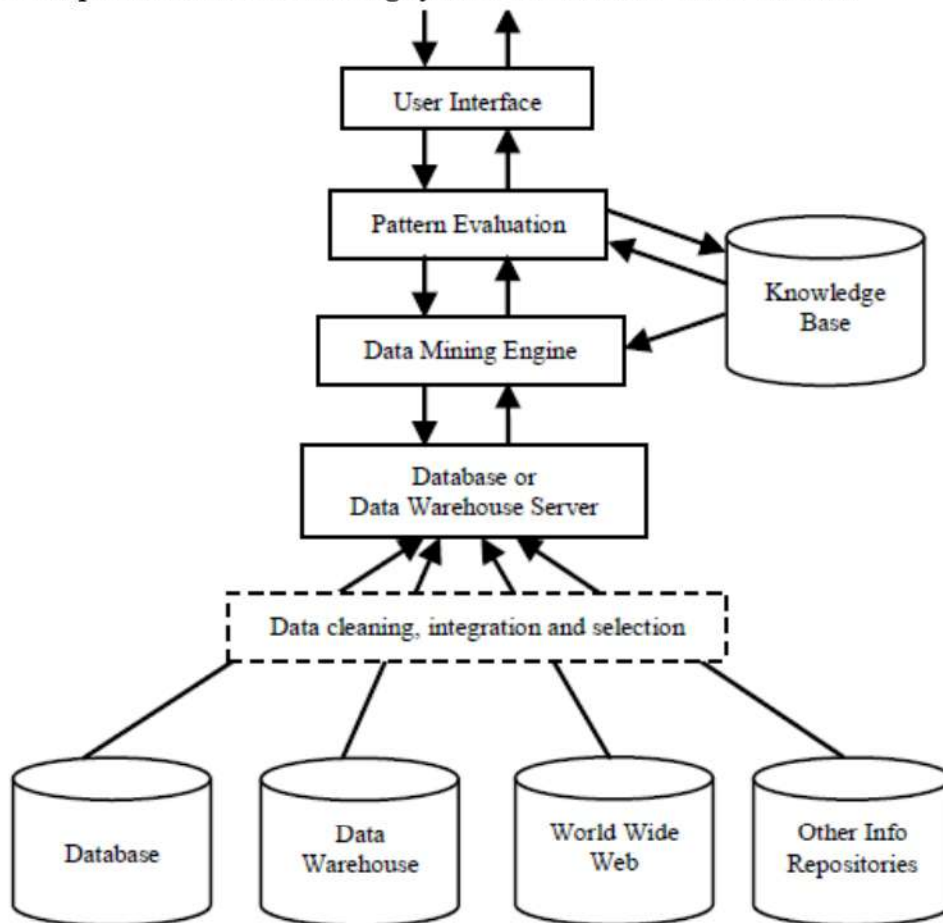


Fig: Architecture of typical data mining system

- **Database, Data Warehouse or Other Information Repository:** This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.
- **Database or Data Warehouse Server:** It fetches the data as per the users' requirement which one need for data mining task.
- **Knowledge Base:** This is the domain knowledge that is used to – guide the search or evaluate the interestingness of resulting patterns. It is simply stored in the form of set of rules.
- **Data Mining Engine:** It performs the data mining task such as characterization, association, classification, prediction, cluster analysis etc.
- **Pattern Evaluation Module:** They are responsible for finding interesting patterns in the data using a threshold value. It interacts with the data mining engine to focus the search on interesting patterns.
- **Graphical User Interface:** This module is used to communicate between user and the data mining system and allow users to browse databases or data warehouse schemas by specifying a data mining query or task.

## Data Mining Functionalities – What kinds of Patterns Can Be Mined?

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: *descriptive* and *predictive*.

- **Descriptive** mining tasks characterize the general properties of the data in the database.
- **Predictive** mining tasks perform inference on the current data in order to make predictions.

Data mining functionalities or the kinds of patterns that can be mined are as follows:

1. **Class/Concept Description:** Data can be associated with classes or concepts that can be described in summarized, concise and yet precise, terms. Such descriptions of a concept or class are called class/concept descriptions. These descriptions can be derived via:
  - **Data Characterization:** Characterization is a summarization of the general characteristics or features of a target class of data which creates what is called a characteristic rule.
  - **Data Discrimination:** Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.
2. **Association analysis on frequent patterns:** Frequent patterns are patterns that occur frequently in data. Association analysis aims to discover associations between items occurring together frequently.  
 E.g. buys(X,“computer”) => buys(X,“software”) [support=1%,confidence=50%]  
 where X is a variable representing a customer. Confidence=50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.
3. **Classification and Prediction:** Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. This model is derived based on the analysis of a set of training data and used to predict the class label of objects for which the class label is unknown.  
 Prediction is used to predict missing or unavailable numeric data values rather than class labels. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.
4. **Cluster Analysis / Clustering:** Clustering analyzes data objects without consulting class labels. It can be used to generate class labels for a group of data which did not exist at the beginning. The objects are clustered or grouped based on the principle of *maximizing the intra-class similarity and minimizing the interclass similarity*. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.
5. **Outlier Analysis:** Outliers are objects that do not comply with the general behavior or model of the data. Most data mining methods discard outliers as noise or exceptions. However, in some events these kind of events are more interesting. This analysis of outlier data is referred to as outlier analysis. E.g. Fraud detection
6. **Evolution Analysis:** Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. This may include characterization, discrimination, association and correlation analysis, classification, prediction or clustering of time related data. Distinct features of such data include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

## Knowledge Discovery in Database (KDD)

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data.

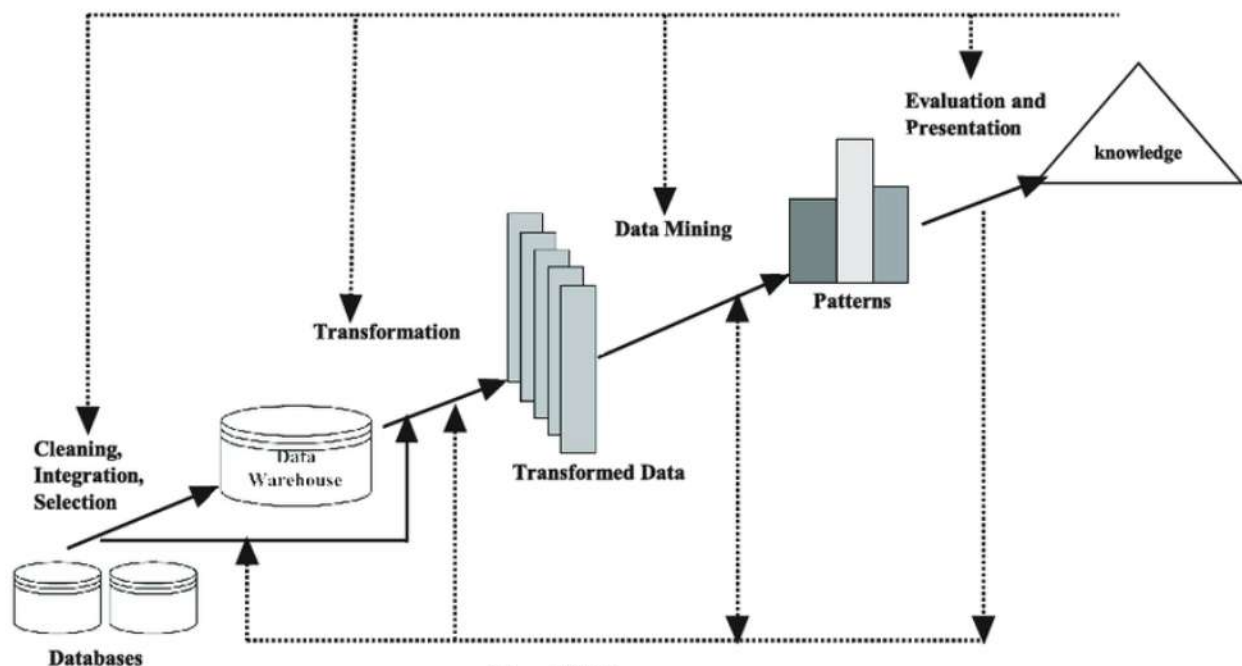


Fig: KDD process

The steps involved in knowledge discovery process:

1. **Data Cleaning**: Data cleaning is a process of removing unnecessary and inconsistent data from the databases. The main purpose of cleaning is to improve the quality of the data by filling the missing values, configuring the data to make sure that it is in consistent format.
2. **Data Integration**: In this step data from various sources such as database, data warehouse and transactional data are combined.
3. **Data Selection**: Data which is required for data mining process can be extracted from multiple and heterogeneous data sources such as databases, files etc. Data selection is a process where the appropriate data required for analysis is fetched from the databases.
4. **Data Transformation**: In the transformation stage data extracted from multiple data sources are converted into an appropriate format for data mining process. Data reduction or summarization is used to decrease the number of possible values of data without affecting the integrity of data.
5. **Data Mining**: It is the most essential step of KDD process where intelligent methods are applied in order to extract hidden patterns from data stored in databases.
6. **Pattern Evaluation**: This step identifies the truly interesting patterns representing knowledge on the basis of some interestingness measures. Support and confidence are two widely used interestingness measures. These patterns are helpful for decision support systems.
7. **Knowledge Presentation**: In this step, visualization and knowledge representation techniques are used to present mined knowledge to users. Visualizations can be in form of graphs, charts or table.

## Classification of Data Mining System

The data mining system can be classified according to the following criteria:

### 1. *Classification according to kind of databases mined*

We can classify the data mining system according to kind of databases mined. Database system can be classified according to different criteria such as data models, types of data etc. And the data mining system can be classified accordingly. For example if we classify the database according to data model then we may have a relational, transactional, object-relational, or data warehouse mining system.

### 2. *Classification according to kind of knowledge mined*

We can classify the data mining system according to kind of knowledge mined. It means data mining system are classified on the basis of functionalities such as: Characterization, Discrimination, Association and Correlation Analysis, Classification, Prediction, Clustering, Outlier Analysis, Evolution Analysis

### 3. *Classification according to kinds of techniques utilized*

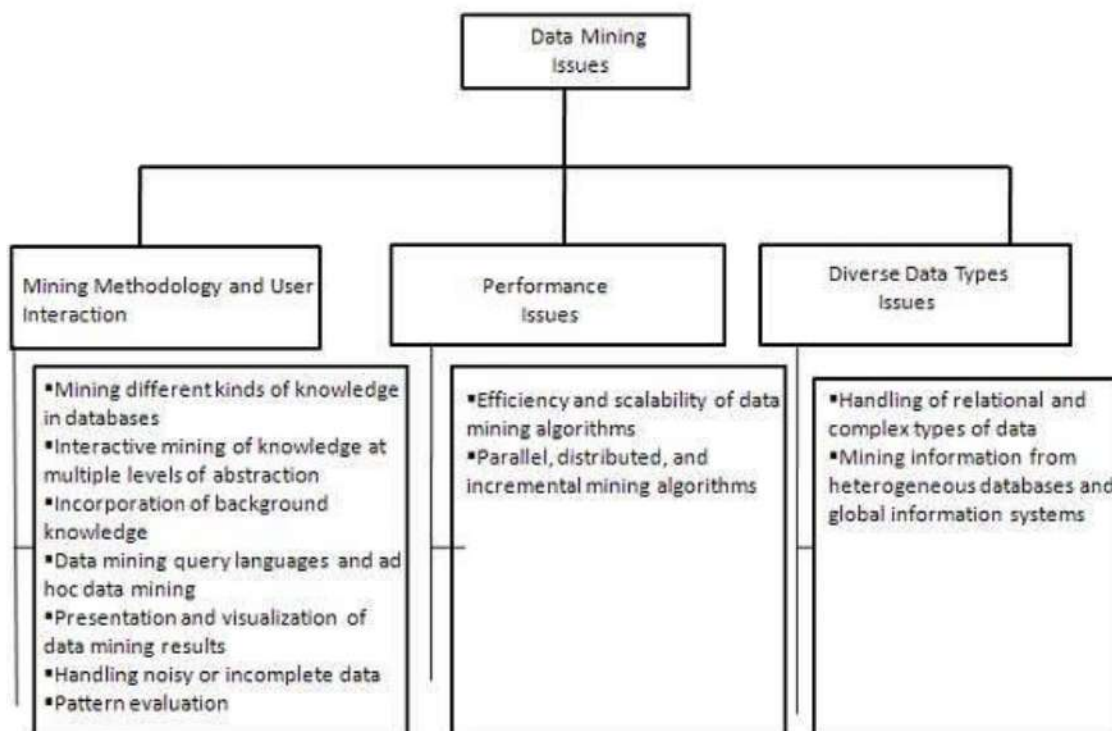
We can classify the data mining system according to kind of techniques used. We can describe these techniques according to degree of user interaction involved or the methods of analysis employed.

### 4. *Classification according to applications adapted*

We can classify the data mining system according to application adapted. These applications are as follows: Finance, Telecommunications, DNA, Stock Markets, E-mail

## Issues in Data Mining

In data mining, the algorithm used is complex and data is not available from single sources so these factors also create some issues. The major issues are:



### **1. Mining Methodology and User Interaction Issues**

- a) ***Mining different kinds of knowledge in databases:*** Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- b) ***Interactive mining of knowledge at multiple levels of abstraction:*** The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- c) ***Incorporation of background knowledge:*** To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- d) ***Data mining query languages and ad hoc data mining:*** Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- e) ***Presentation and visualization of data mining results:*** Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- f) ***Handling noisy or incomplete data:*** The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- g) ***Pattern evaluation:*** The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

### **2. Performance Issues**

- a) ***Efficiency and scalability of data mining algorithms:*** In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- b) ***Parallel, distributed, and incremental mining algorithms:*** The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

### **3. Diverse Data Types Issues**

- a) ***Handling of relational and complex types of data:*** The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- b) ***Mining information from heterogeneous databases and global information systems:*** The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

## Data Object and Attribute Types

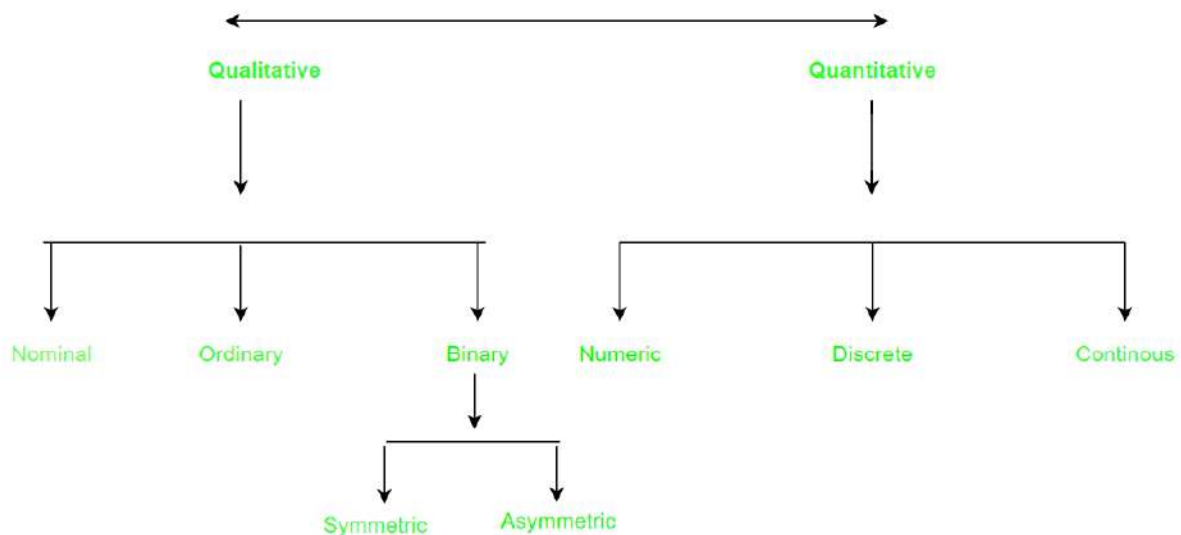
### Data Objects

Data sets are made up of data objects. A data object represents an entity - in a sales database, the objects may be customers, store items, and sales. Data objects are typically described by attributes. If the data objects are stored in a database, they are data tuples.

### Attribute

An attribute is a data field, representing a characteristic or feature of a data object. Attributes describing a customer object can include, for example, customer\_ID, name, and address.

On the basis of set of possible values attributes can be divided into following types:



1) **Nominal Attributes**: Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful order. E.g.

- Hair\_color: possible values are: {black, brown, red, grey, white}
- Marital\_status: possible values are: {Married, Single, Divorced, Widowed}

2) **Binary Attributes**: A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. E.g. Given the attribute *smoker* describing a *patient* object, 1 indicates that the patient smokes, while 0 indicates that the patient does not.

- A binary attribute is ***symmetric*** if both of its states are equally valuable. E.g. attribute *gender* having the states *male* and *female*.
- A binary attribute is ***asymmetric*** if the outcomes of the states are not equally important, such as the *positive* (1) and *negative* (0) outcomes of a medical test for HIV.

- 3) **Ordinal Attributes:** An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known. E.g. Height: possible values are: {Tall, Medium, Short}. The values have a meaningful sequence (which corresponds to increasing height); however, we cannot tell from the values how much bigger, say, a medium is than a short. Other example of ordinal attributes include grade (e.g., A+, A, A-, B+, and so on).
- 4) **Numeric Attributes:** A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.
- **Interval-Scaled Attributes:** Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. E.g. Calendar Date (2002 and 2010 are 8 years apart)
  - **Ratio-Scaled Attributes:** If a measurement is ratio scaled means a value being multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode. E.g. Frequency of words in a document.
- 5) **Discrete versus Continuous Attributes:** A *discrete attribute* has a finite or countably infinite set of values, which may or may not be represented as integers. The attributes *hair\_color*, *smoker*, *medical\_test* each have a finite number of values, and so are discrete.
- A *continuous attribute* has an infinite no. of states. Continuous attributes are typically represented as floating-point variables. E.g. The attribute *Height* having the values 5.4,....., 6.5,.....etc.

## Statistical Description of Data

The basic statistical description of data can be used to identify properties of the data and highlight which data values should be treated as noise or outliers. Basic statistical descriptions include *Measure of Central Tendency* and *Measure of Dispersion*.

### Measure of Central Tendency

Measure of central tendency measures the location of the middle or center of a data distribution.

Measures of central tendency include the *mean*, *median*, *mode*, and *midrange*.

- **Mean:** Mean is the most common and effective numeric measure, which is used to measure the “center” of a set of data. Let  $x_1, x_2, \dots, x_n$  be the set of  $N$  observed values for  $X$ . The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

If each  $x_i$  is associated with a weight  $w_i$  for  $i = 1, \dots, N$  then the weighted mean is

$$\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_N x_N}{w_1 + w_2 + \dots + w_N}$$



- **Median:** A better measure of the center of data is the median, which is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half.

Suppose that a given data set of  $N$  values for an attribute  $X$  is sorted in increasing order. If  $N$  is odd, then the median is the middle value of the ordered set. If  $N$  is even, then the median is not unique; it is the two middlemost values and any value in between. If  $X$  is a numeric attribute in this case, by convention, the median is taken as the average of the two middlemost values.

- **Mode:** The mode for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes. Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal. In general, a data set with two or more modes is multimodal. At the other extreme, if each data value occurs only once, then there is no mode.

For unimodal numeric data, we have the following empirical relation:

$$\text{mean} - \text{mode} \approx 3 \times (\text{mean} - \text{median}).$$

- **Midrange:** The midrange can also be used to assess the central tendency of a numeric data set. It is the average of the largest and smallest values in the set.

**Example:**

Let 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110 are the values.

$$\text{➤ Mean}(\bar{x}) = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12} = 58$$

$$\text{➤ Median} = \frac{52+56}{2} = 54$$

➤ **Mode:** The given data are bimodal. Two modes are 52 and 70.

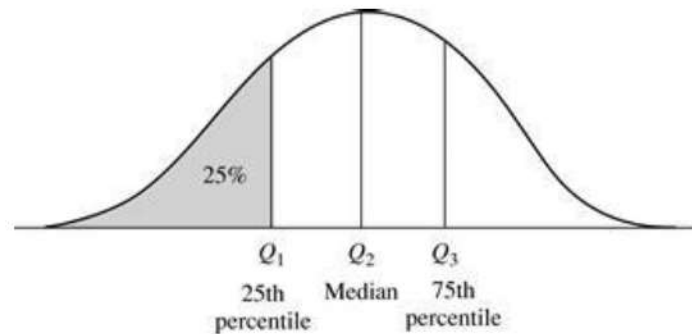
$$\text{➤ Midrange} = \frac{30+110}{2} = 70$$

### Measure of Dispersion

Measures of dispersion indicate how much the observed data is spread out around a measure of central tendency. The measures include **range**, **quantiles**, **quartiles**, **percentiles**, and the **interquartile range**. **Variance** and **standard deviation** also indicate the spread of a data distribution.

- **Range:** The range of the set is the difference between the largest ( $\text{max}()$ ) and smallest ( $\text{min}()$ ) values. **Example:** 1, 3, 5, 6, 7  $\Rightarrow$  Range = 7 - 1 = 6
- **Quantiles:** Suppose that the data for attribute  $X$  are sorted in increasing numeric order. Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.
  - The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the **median**.

- **Quartiles:** The 4-quartiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**.
- **Percentiles:** The 100-quartiles are more commonly referred to as **percentiles**; they divide the data distribution into 100 equal-sized consecutive sets.



- **Interquartile Range:** The distance between the first (25<sup>th</sup> percentile) and third (75<sup>th</sup> percentile) quartiles is called the interquartile range (IQR).

$$IQR = Q_3 - Q_1.$$

- **Variance:** The variance of  $N$  observations,  $x_1, x_2, \dots, x_n$ , for a numeric attribute  $X$  is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \left( \frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2,$$

where  $\bar{x}$  is the mean value of the observations.

- **Standard Deviation:** The standard deviation,  $\sigma$ , of the observations is the square root of the variance,  $\sigma^2$ . A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

**Example:**

Marks: 8, 10, 15, 20

Mean of marks  $\bar{x} = 13.25$

➤  $Variance(\sigma^2) = \frac{(8-13.25)^2 + (10-13.25)^2 + (15-13.25)^2 + (20-13.25)^2}{4} = 21.6$

➤  $Standard\ Deviation(\sigma) = \sqrt{21.6} = 4.6$

## **Applications of Data Mining**

Data mining can be applied in almost every field. Some of the major applications of data mining are briefly discussed below.

### **1. Market Analysis and Management**

Listed below are the various fields of market where data mining is used:

- **Customer Profiling:** Data mining helps determine what kind of people buy what kind of products.
- **Identifying Customer Requirements:** Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis:** Data mining performs association/correlations between product sales.
- **Target Marketing:** Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- **Determining Customer purchasing pattern:** Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information:** Data mining provides us various multidimensional summary reports.

### **2. Corporate Analysis and Risk Management**

Data mining is used in the following fields of the Corporate Sector:

- **Finance Planning and Asset Evaluation:** It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning:** It involves summarizing and comparing the resources and spending.
- **Competition:** It involves monitoring competitors and market directions.

### **3. Fraud Detection**

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

### **4. Intrusion Detection**

Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity.

### **5. Web Search Engines**

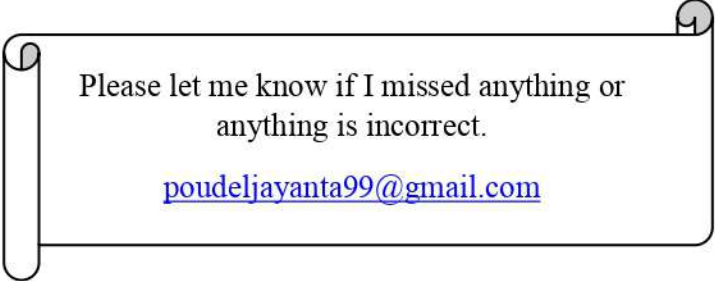
Web search engines are essentially very large data mining applications. Various data mining techniques are used in all aspects of search engines, ranging from crawling, indexing, and searching.

**6. Social Web and Networks**

There are a growing number of highly-popular user-centric applications such as blogs, wikis and Web communities that generate a lot of structured and semi-structured information. In these applications data mining can be used to explain and predict the evolution of social networks, personalized search for social interaction, user behavior prediction etc.

**7. Space Science**

Data mining can be used to automate the analysis image data collected from sky survey with better accuracy.



Please let me know if I missed anything or  
anything is incorrect.

[poudeljayanta99@gmail.com](mailto:poudeljayanta99@gmail.com)