Data Warehousing and Data Mining
Important topics from UNIT 3

- OLAP guidelines -refer class notes/text book
- Categories of tools –refer class notes /text book
- Essential step in the process of knowledge discovery in databases
- Data mining: on what kind of data? / Describe the following advanced database systems and applications: object-relational databases, spatial databases, text databases, multimedia databases, the World Wide Web.
- Data mining functionalities
- Five primitives for specifying a data mining task
- Data cleaning
- Data smoothing techniques
- Data Transformation
- Numerosity reduction techniques:
- Measure the Central Tendency
- Measure the Dispersion of Data

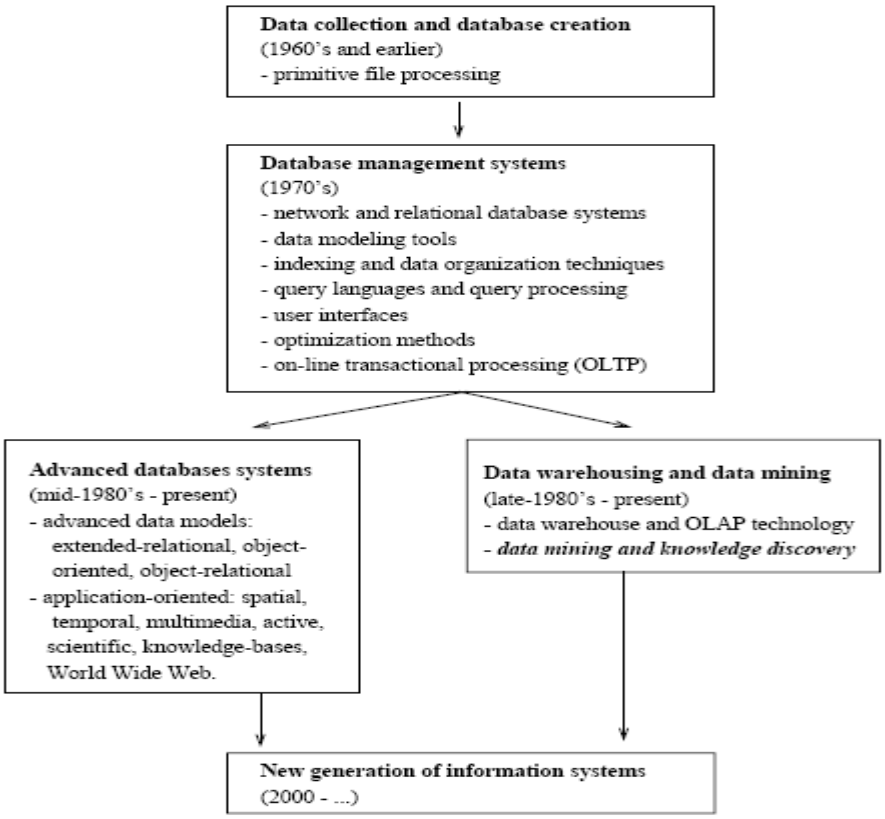
UNIT III DATA MINING
8

Introduction – Data – Types of Data – Data Mining Functionalities – Interestingness of Patterns – Classification of Data Mining Systems – Data Mining Task Primitives –Integration of a Data Mining System with a Data Warehouse – Issues –Data Preprocessing.

What motivated data mining? Why is it important?

The major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

The evolution of database technology



What is data mining?

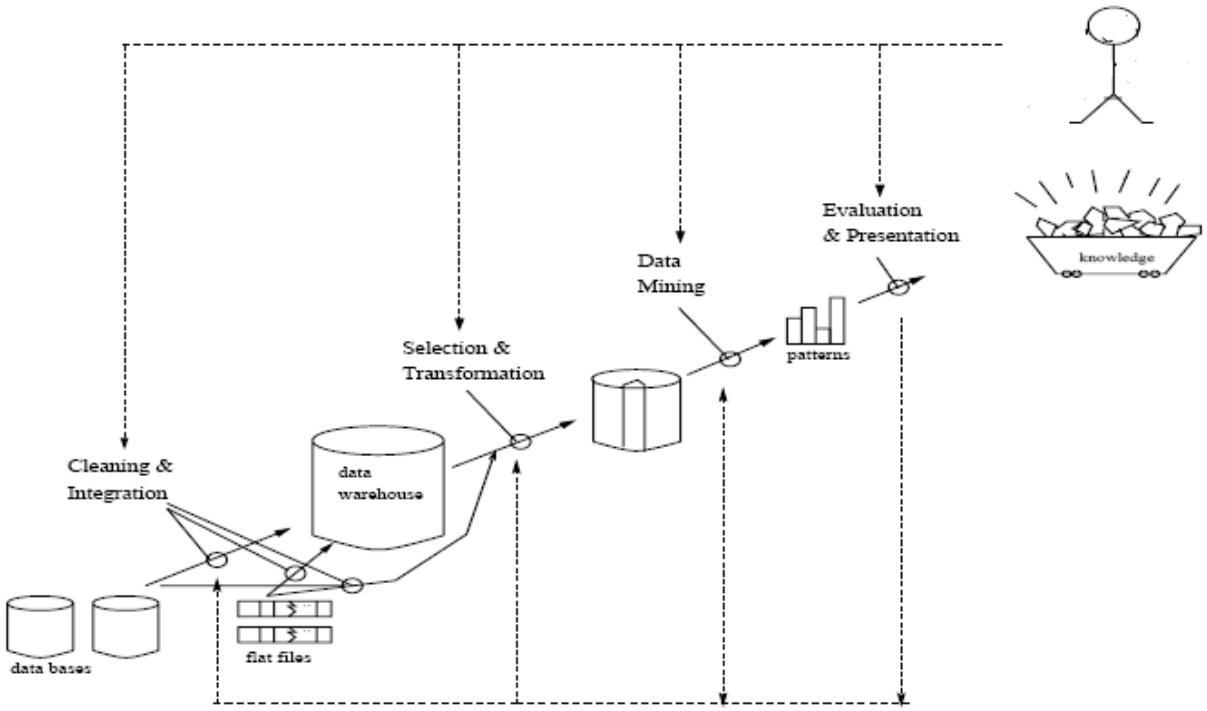
Data mining refers to extracting or mining" knowledge from large amounts of data. There are many other terms related to data mining, such as knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly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

Essential step in the process of knowledge discovery in databases

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

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

- knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.



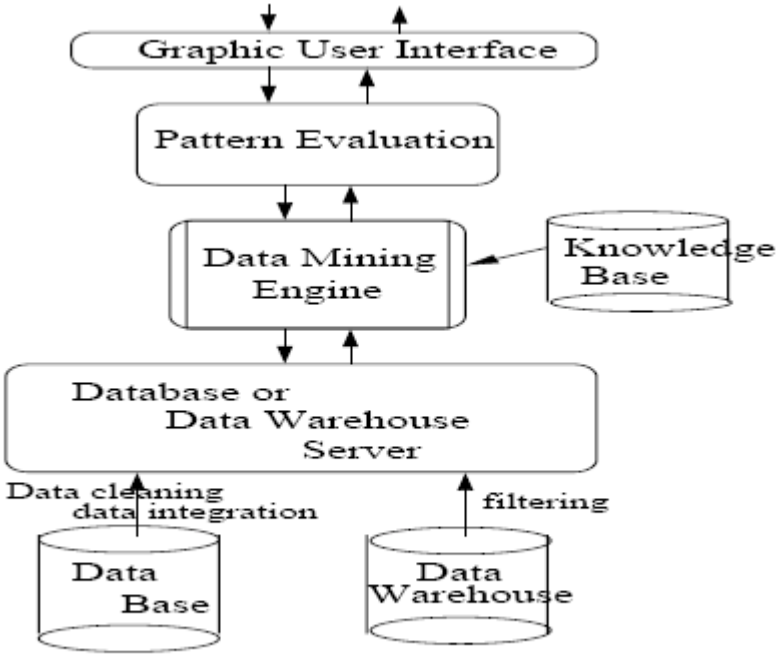
Data mining as a process of knowledge discovery.

Architecture of a typical data mining system/Major Components

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

1. A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.
2. A database or data warehouse server which fetches the relevant data based on users' data mining requests.
3. A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.
4. A data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.

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



Architecture of a typical data mining system.

How is a data warehouse different from a database? How are they similar?

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

• Similarities between a data warehouse and a database: Both are repositories of information, storing huge amounts of persistent data.

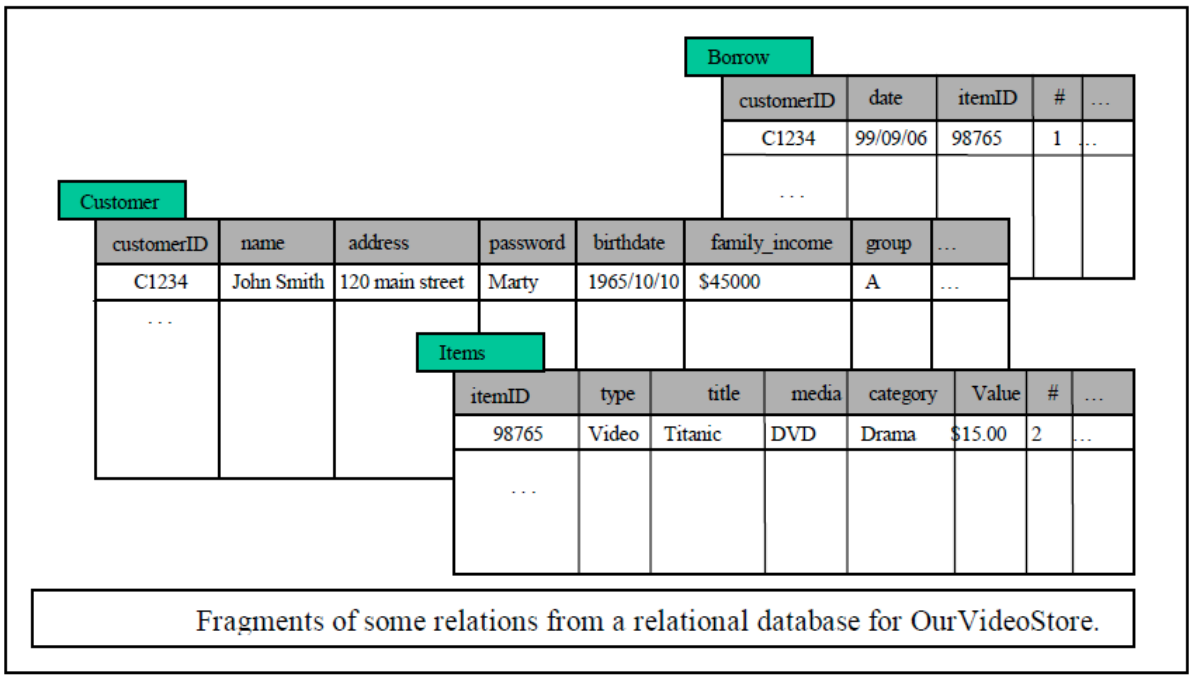
Data mining: on what kind of data? / Describe the following advanced database systems and applications: object-relational databases, spatial databases, text databases, multimedia databases, the World Wide Web.

In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database

systems,
flat files, and the World-Wide Web. Advanced database systems include object-oriented and object-relational databases, and special c application-oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases.

Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.

Relational Databases: a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. In following figure it presents some relations Customer, Items, and Borrow representing business activity in a video store. These relations are just a subset of what could be a database for the video store and is given as an example.



Fragments of some relations from a relational database for OurVideoStore.

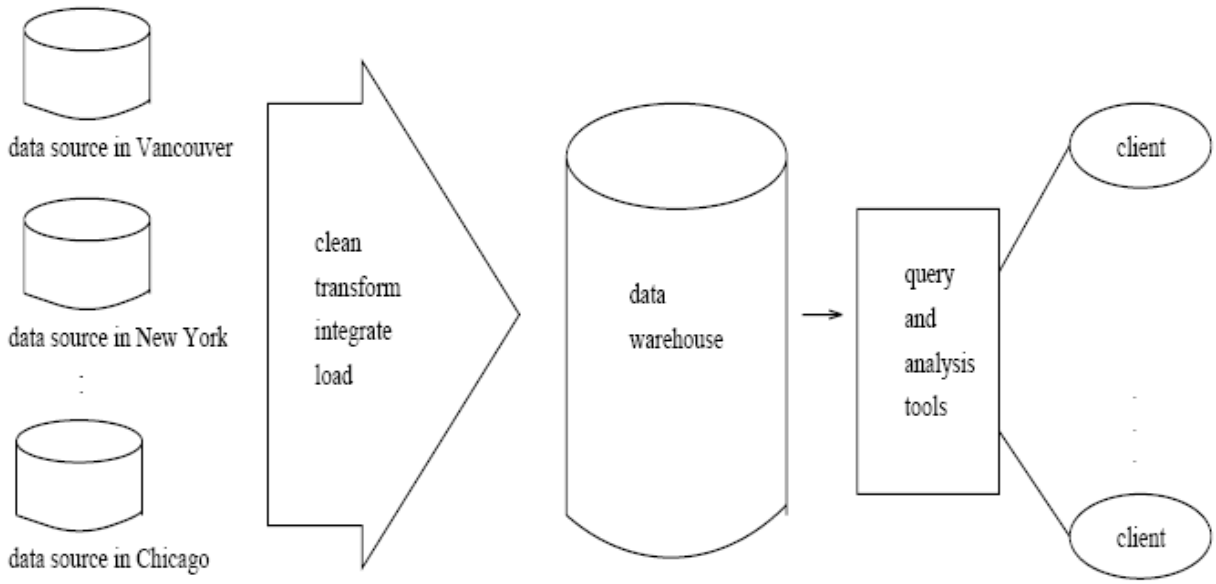
The most commonly used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count. For instance, an SQL query to select the videos grouped by category would be:

SELECT count(*) FROM Items WHERE type=video GROUP BY category.

Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files, since they can take advantage of the structure inherent to relational databases. While data mining can benefit from SQL for data selection, transformation and consolidation, it goes beyond what SQL could provide, such as predicting, comparing, detecting deviations, etc.

Data warehouses

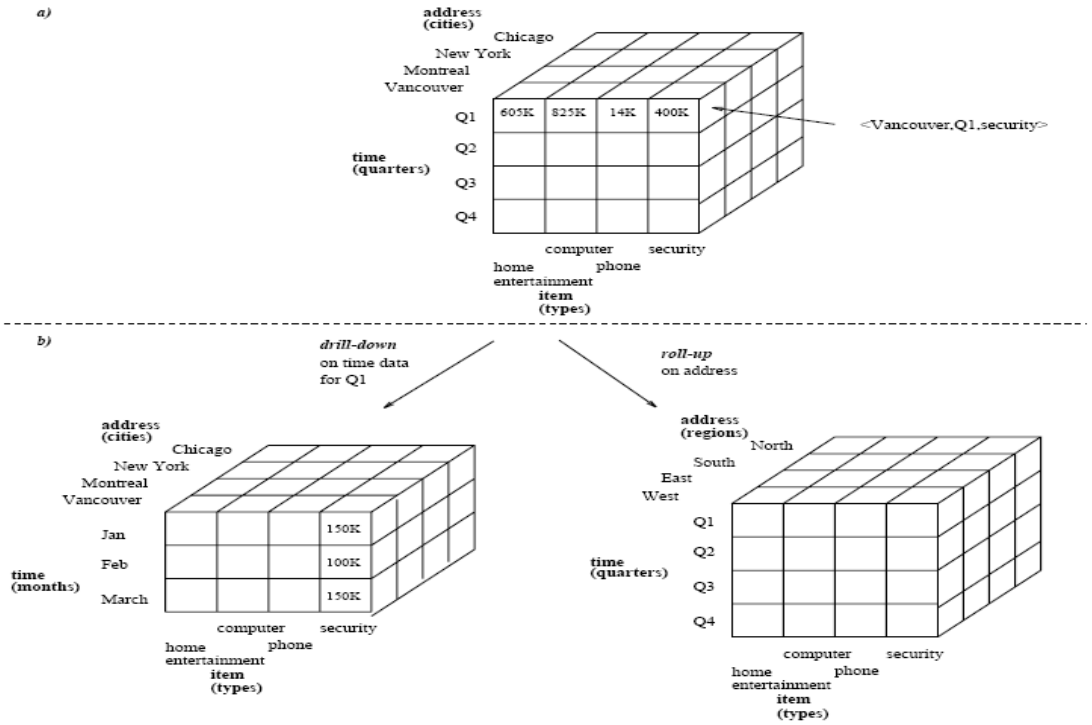
A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing. The figure shows the basic architecture of a data warehouse



Architecture of a typical data warehouse.

In order to facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity. The data are stored to provide information from a historical perspective and are typically summarized.

A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the pre-computation and fast accessing of summarized data.



A multidimensional data cube, commonly used for data warehousing, *a)* showing summarized data for and *b)* showing summarized data resulting from drill-down and roll-up operations on the cube in *a*

The data cube structure that stores the primitive or lowest level of information is called a base cuboid. Its corresponding higher level multidimensional (cube) structures are called (non-base) cuboids. A base cuboid together with all of its corresponding higher level cuboids form a data cube. By providing multidimensional data views and the precomputation of summarized data, data warehouse systems are well suited for On-Line Analytical Processing, or OLAP. OLAP operations make use of background knowledge regarding the domain of the data being studied in order to allow the presentation of data at different levels of abstraction. Such operations accommodate different user viewpoints. Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization, as illustrated in above figure.

Transactional databases

In general, a transactional database consists of a flat file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction (such as items purchased in a store) as shown below:

sales

trans_ID	list of item_ID's
T100	I1, I3, I8, I16
...	...

Advanced database systems and advanced database applications

• **An objected-oriented database** is designed based on the object-oriented programming paradigm where data are a large number of objects organized into classes and class hierarchies. Each entity in the database is considered as an object. The object contains a set of variables that describe the object, a set of messages that the object can use to communicate with other objects or with the rest of the database system and a set of methods where each method holds the code to implement a message.

• **A spatial database** contains spatial-related data, which may be represented in the form of raster or vector data. Raster data consists of n-dimensional bit maps or pixel maps, and vector data are represented by lines, points, polygons or other kinds of processed primitives, Some examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.

• **Time-Series Databases:** Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

• **A text database** is a database that contains text documents or other word descriptions in the form of long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

• **A multimedia database** stores images, audio, and video data, and is used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces.

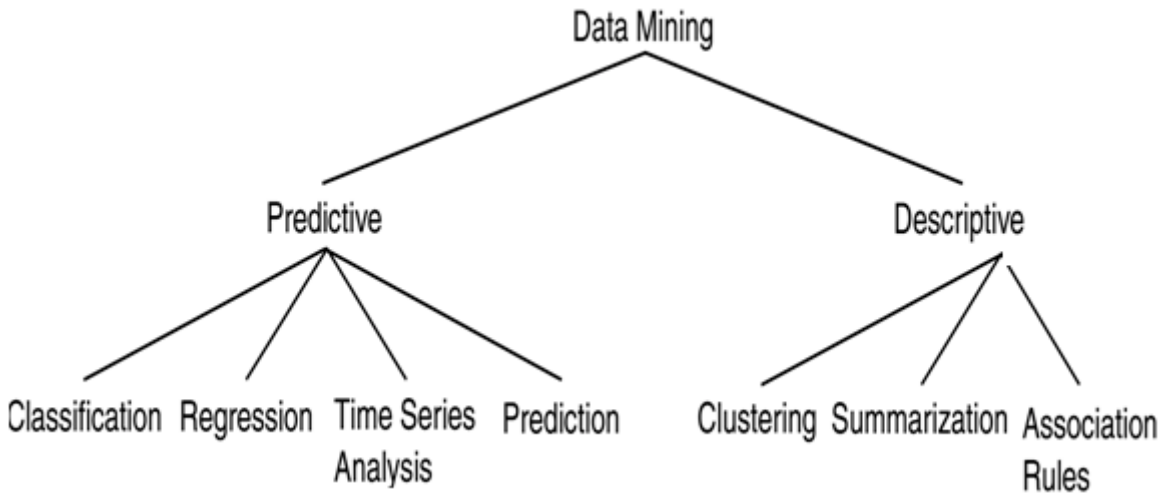
• **The World-Wide Web** provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access. Some examples of distributed information services associated with the World-Wide Web include America Online, Yahoo!, AltaVista, and Prodigy.

Data mining functionalities/Data mining tasks: what kinds of patterns can be mined?

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories:

- Descriptive
- 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.



Describe data mining functionalities, and the kinds of patterns they can discover

(or)

Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, prediction, clustering, and evolution analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.

1 Concept/class description: characterization and discrimination

Data can be associated with classes or concepts. It describes a given set of data in a concise and summarative manner, presenting interesting general properties of the data. These descriptions can be derived via

1. data characterization, by summarizing the data of the class under study (often called the target class)
2. data discrimination, by comparison of the target class with one or a set of comparative classes
3. both data characterization and discrimination

Data characterization

It is a summarization of the general characteristics or features of a target class of data.

Example:

A data mining system should be able to produce a description summarizing the characteristics of a student who has obtained more than 75% in every semester; the result could be a general profile of the student.

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.

Example

The general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.

The output of data characterization can be presented in various forms. Examples include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs. The resulting descriptions can also be presented as generalized relations, or in rule form called characteristic rules.

Discrimination descriptions expressed in rule form are referred to as discriminant rules.

2 Association

Mining Frequent Patterns, Associations, and Correlations

Frequent patterns, are patterns that occur frequently in data. There many kinds of frequent patterns, including **itemsets**, **subsequences**, and **substructures**.

A **frequent itemset** typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.

A frequently occurring **subsequence**, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.

A **substructure** can refer to different structural forms, such as graphs, trees, or lattices, which may be combined with itemsets ro subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns lead to discovery of interesting associations and correlations within data.

It is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like

major(X, "computing science") ⇒ owns(X, "personal computer")

[support = 12%, confidence = 98%]

where X is a variable representing a student. The rule indicates that of the students under study, 12% (support) major in computing science and own a personal computer. There is a 98% probability (confidence, or certainty) that a student in this group owns a personal computer.

Example:

A grocery store retailer to decide whether to but bread on sale. To help determine the impact of this decision, the retailer generates association rules that show what other products are frequently purchased with bread. He finds 60% of the times that bread is sold so are pretzels and that 70% of the time jelly is also sold. Based on these facts, he tries to capitalize on the association between bread, pretzels, and jelly by placing some pretzels and jelly at the end of the aisle where the bread is placed. In addition, he decides not to place either of these items on sale at the same time.

3 Classification and prediction

Classification:

Classification:

- It predicts categorical class labels
- It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

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

Example:

An airport security screening station is used to deter mine if passengers are potential terrorist or criminals. To do this, the face of each passenger is scanned and its basic pattern(distance

between eyes, size, and shape of mouth, head etc) is identified. This pattern is compared to entries in a database to see if it matches any patterns that are associated with known offenders

A classification model can be represented in various forms, such as

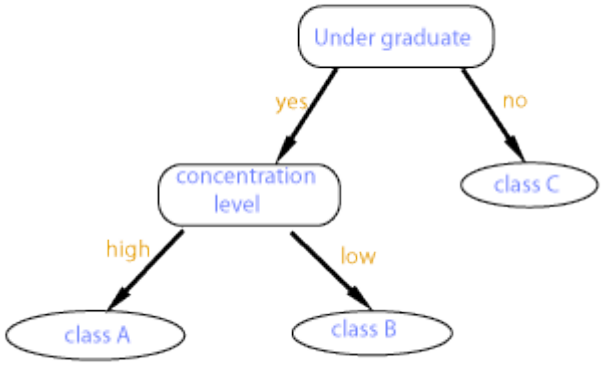
1) IF-THEN rules,

student (class , "undergraduate") AND concentration (level, "high") ==> class A

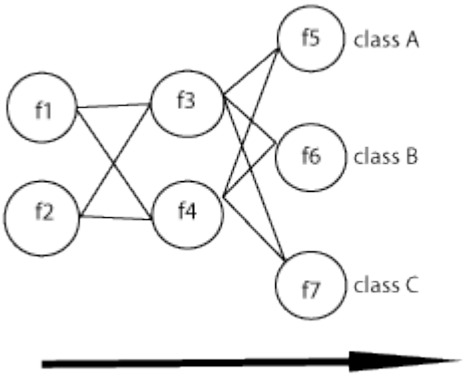
student (class ,"undergraduate") AND concentrtion (level,"low") ==> class B

student (class , "post graduate") ==> class C

2) Decision tree



3) Neural network.



Prediction:

Find some missing or unavailable data values rather than class labels referred to as prediction. Although prediction may refer to both data value prediction and class label prediction, it is

usually confined to data value prediction and thus is distinct from classification. Prediction also encompasses the identification of distribution trends based on the available data.

Example:

Predicting flooding is difficult problem. One approach is uses monitors placed at various points in the river. These monitors collect data relevant to flood prediction: water level, rain amount, time, humidity etc. These water levels at a potential flooding point in the river can be predicted based on the data collected by the sensors upriver from this point. The prediction must be made with respect to the time the data were collected.

Classification vs. Prediction

Classification differs from prediction in that the former is to construct a set of models (or functions) that describe and distinguish data class or concepts, whereas the latter is to predict some missing or unavailable, and often numerical, data values. Their similarity is that they are both tools for prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

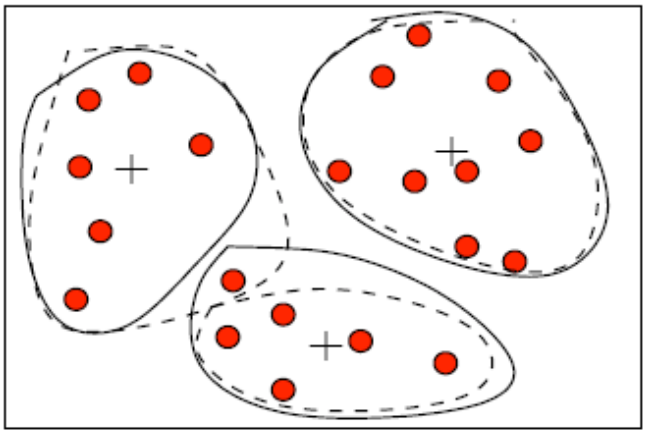
4 Clustering analysis

Clustering analyzes data objects without consulting a known class label.

The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

Each cluster that is formed can be viewed as a class of objects.

Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together as shown below:



customer data with respect to customer locations in a city, showing three data clusters. Each cluster 'center' is marked with a '+'.

Example:

A certain national department store chain creates special catalogs targeted to various demographic groups based on attributes such as income, location and physical characteristics of potential customers (age, height, weight, etc). To determine the target mailings of the various catalogs and to assist in the creation of new, more specific catalogs, the company performs a clustering of potential customers based on the determined attribute values. The results of the clustering exercise are the used by management to create special catalogs and distribute them to the correct target population based on the cluster for that catalog.

Classification vs. Clustering

- In general, in classification you have a set of predefined classes and want to know which class a new object belongs to.
- Clustering tries to group a set of objects and find whether there is *some* relationship between the objects.
- In the context of machine learning, classification is *supervised learning* and clustering is *unsupervised learning*.

5 Outlier analysis: A database may contain data objects that do not comply with general model of data. These data objects are outliers. In other words, the data objects which do not fall within the cluster will be called as outlier data objects. Noisy data or exceptional data are also called as outlier data. The analysis of outlier data is referred to as outlier mining.

Example
Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred

by the same account. Outlier values may also be detected with respect to the location and type of purchase, or the purchase frequency.

6 Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

Example:
The data of result the last several years of a college would give an idea if quality of graduated produced by it

7 Correlation analysis

Correlation analysis is a technique use to measure the association between two variables. A **correlation coefficient** (**r**) is a statistic used for measuring the strength of a supposed linear association between two variables. Correlations range from -1.0 to +1.0 in value.

A correlation coefficient of 1.0 indicates a perfect positive relationship in which high values of one variable are related perfectly to high values in the other variable, and conversely, low values on one variable are perfectly related to low values on the other variable.

A correlation coefficient of 0.0 indicates no relationship between the two variables. That is, one cannot use the scores on one variable to tell anything about the scores on the second variable.

A correlation coefficient of -1.0 indicates a perfect negative relationship in which high values of one variable are related perfectly to low values in the other variables, and conversely, low values in one variable are perfectly related to high values on the other variable.

What is the difference between discrimination and classification? Between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?

Answer:

• Discrimination differs from classification in that the former refers to 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, while the latter is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Discrimination and classification are similar in that they both deal with the analysis of class data objects.

• Characterization differs from clustering in that the former refers to a summarization of the general characteristics or features of a target class of data while the latter deals with the analysis of data objects without consulting a known class label. This pair of tasks is similar in that they both deal with grouping together objects or data that are related or have high similarity in comparison to one another.

• Classification differs from prediction in that the former is the process of finding a set of models (or functions) that describe and distinguish data class or concepts while the latter predicts missing or unavailable, and often numerical, data values. This pair of tasks is similar in that they both are tools for

Prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

Are all of the patterns interesting? / What makes a pattern interesting?

A pattern is interesting if,

(1) It is easily understood by humans,

(2) Valid on new or test data with some degree of certainty,

(3) Potentially useful, and

(4) Novel.

A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge.

- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Several objective measures of pattern interestingness exist.

An objective measure for association rules of the form S ==> Y is rule support

Another objective measure of association rules is confidence

support(X=> Y) = P(XUY)

confidence(X=> Y) = P(Y/X)

support (X=> Y) = No. of tuples containing both X and Y / total number of tuples

$$\text{confidence } (X \Rightarrow Y) = \frac{\text{No. of tuples\_ containing both X and Y}}{\text{Number of tuples containing X}}$$

Classification of data mining systems

There are many data mining systems available or being developed. Some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities, other are more versatile and comprehensive. Data mining systems can be categorized according to various criteria among other classification are the following:

· **Classification according to the type of data source mined:** this classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

· **Classification according to the data model drawn on:** this classification categorizes data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional, etc.

· **Classification according to the kind of knowledge discovered:** this classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

· **Classification according to mining techniques used:** Data mining systems employ and provide different techniques. This classification categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

Primitives for specifying a data mining task

• **Task-relevant data:** This primitive specifies the data upon which mining is to be performed. It involves specifying the database and tables or data warehouse containing the relevant data, conditions for selecting the relevant data, the relevant attributes or dimensions for exploration, and instructions regarding the ordering or grouping of the data retrieved.

• **Knowledge type to be mined:** This primitive specifies the specific data mining function to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. As well, the user can be more specific and provide pattern templates that all discovered patterns must match. These templates or meta patterns (also called meta rules or meta queries), can be used to guide the discovery process.

• **Background knowledge:** This primitive allows users to specify knowledge they have about the domain to be mined. Such knowledge can be used to guide the knowledge discovery process and evaluate the patterns that are found.

• **Pattern interestingness measure:** This primitive allows users to specify functions that are used to separate uninteresting patterns from knowledge and may be used to guide the mining process, as well as to evaluate the discovered patterns. This allows the user to confine the number of uninteresting patterns returned by the process, as a data mining process may generate a large number of patterns. Interestingness measures can be specified for such pattern characteristics as simplicity, certainty, utility and novelty.

• **Visualization of discovered patterns:** This primitive refers to the form in which discovered patterns are to be displayed. In order for data mining to be effective in conveying knowledge to users, data mining systems should be able to display the discovered patterns in multiple forms such as rules, tables, cross tabs (cross-tabulations), pie or bar charts, decision trees, cubes or other visual representations.

Major issues in data mining

Major issues in data mining is regarding mining methodology, user interaction, performance, and diverse data types

1 Mining methodology and user-interaction issues:

_ Mining different kinds of knowledge in databases: Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques.

_ **Interactive mining of knowledge at multiple levels of abstraction:** Since it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive.

_ **Incorporation of background knowledge:** Background knowledge, or information regarding the domain under study, may be used to guide the discovery patterns. Domain knowledge related

to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.

_ **Data mining query languages and ad-hoc data mining:** Knowledge in Relational query languages (such as SQL) required since it allow users to pose ad-hoc queries for data retrieval.

_ **Presentation and visualization of data mining results:** Discovered knowledge should be expressed in high-level languages, visual representations, so that the knowledge can be easily understood and directly usable by humans

_ **Handling outlier or incomplete data:** The data stored in a database may reflect outliers: noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods which can handle outliers are required.

_ **Pattern evaluation: refers to interestingness of pattern**: A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns,

2. Performance issues. These include efficiency, scalability, and parallelization of data mining algorithms.

_ **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.

_ **Parallel, distributed, and incremental updating algorithms:** Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

3. Issues relating to the diversity of database types

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

_ **Mining information from heterogeneous databases and global information systems:** Local and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining.

Data preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user

Why Data Preprocessing?

Data in the real world is dirty. It can be in incomplete, noisy and inconsistent from. These data needs to be preprocessed in order to help improve the quality of the data, and quality of the mining results.

- If no quality data, then no quality mining results. The quality decision is always based on the quality data.
- If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult

Incomplete data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”.

Noisy data: containing errors or outliers data. e.g., Salary=“-10”

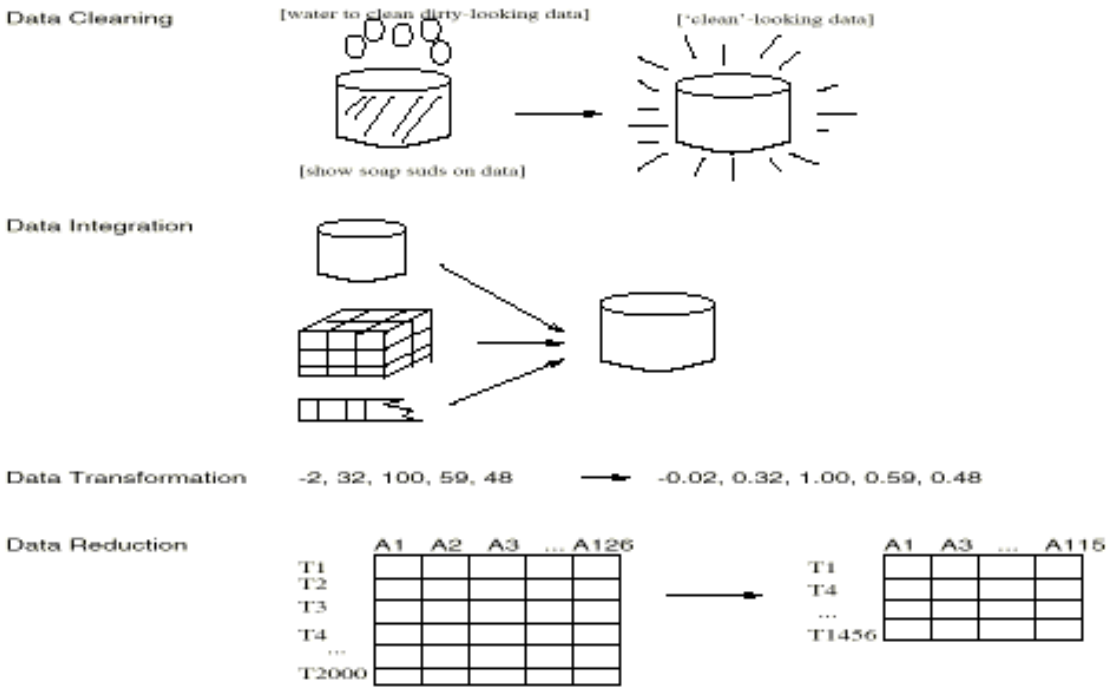
Inconsistent data: containing discrepancies in codes or names. e.g., Age=“42” Birthday=“03/07/1997”

- Incomplete data may come from
  - “Not applicable” data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection by instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Forms of Data Preprocessing



Data cleaning:

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Various methods for handling this problem:

The various methods for handling the problem of missing values in data tuples include:

(a) Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). 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.

(b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

(c) Using 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 $-\infty$. 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 recommended.

(d) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

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

Noisy data:

Noise is a random error or variance in a measured variable. Data smoothing tech is used for removing such noisy data.

Several Data smoothing techniques:

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

In this technique,

1. The data for first sorted
2. Then the sorted list partitioned into equi-depth of bins.
3. Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
   a. **Smoothing by bin means**: Each value in the bin is replaced by the mean value of the bin.
   b. **Smoothing by bin medians**: Each value in the bin is replaced by the bin median.
   c. **Smoothing by boundaries**: The min and max values of a bin are identified as the bin boundaries. Each bin value is replaced by the closest boundary value.

- Example: Binning Methods for Data Smoothing

  - Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  - Partition into (equi-depth) bins(equi depth of 3 since each bin contains three values):
    - **Bin 1**: 4, 8, 9, 15
    - **Bin 2**: 21, 21, 24, 25
    - **Bin 3**: 26, 28, 29, 34
  - Smoothing by bin means:
    - **Bin 1**: 9, 9, 9, 9
    - **Bin 2**: 23, 23, 23, 23
    - **Bin 3**: 29, 29, 29, 29
  - Smoothing by bin boundaries:
    - **Bin 1**: 4, 4, 4, 15
    - **Bin 2**: 21, 21, 25, 25
    - **Bin 3**: 26, 26, 26, 34

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in

increasing order): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
(a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps.
Comment on the effect of this technique for the given data.

The following steps are required to smooth the above data using smoothing by bin means with a bin
depth of 3.

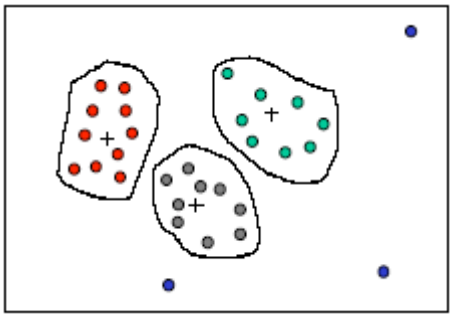
• Step 1: Sort the data. (This step is not required here as the data are already sorted.)

• Step 2: Partition the data into equidepth bins of depth 3.
Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22
Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35
Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70

• Step 3: Calculate the arithmetic mean of each bin.

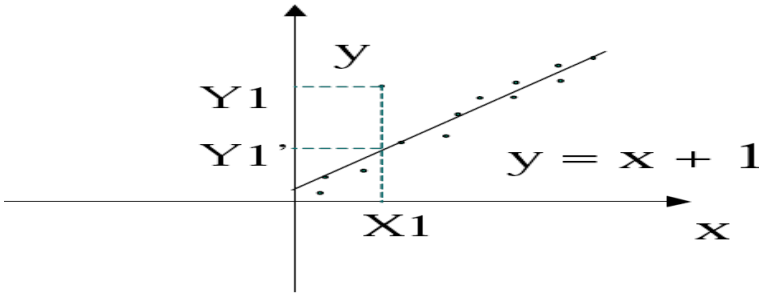
• Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin.
Bin 1: 14, 14, 14 Bin 2: 18, 18, 18 Bin 3: 21, 21, 21
Bin 4: 24, 24, 24 Bin 5: 26, 26, 26 Bin 6: 33, 33, 33
Bin 7: 35, 35, 35 Bin 8: 40, 40, 40 Bin 9: 56, 56, 56

2 Clustering: Outliers in the data may be detected by clustering, where similar values are organized into groups, or 'clusters'. Values that fall outside of the set of clusters may be considered outliers.



3 Regression : smooth by fitting the data into regression functions.

- Linear regression involves finding the best of line to fit two variables, so that one variable can be used to predict the other.



- Multiple linear regression is an extension of linear regression, where more than two variables are involved and the data are fit to a multidimensional surface.

Using regression to find a mathematical equation to fit the data helps smooth out the noise.

Field overloading: is a kind of source of errors that typically occurs when developers compress new attribute definitions into unused portions of already defined attributes.

Unique rule is a rule says that each value of the given attribute must be different from all other values of that attribute

Consecutive rule is a rule says that there can be no missing values between the lowest and highest values of the attribute and that all values must also be unique.

Null rule specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition and how such values should be handled.

Data Integration

It combines data from multiple sources into a coherent store. There are number of issues to consider during data integration.

Issues:

- **Schema integration:** refers integration of metadata from different sources.
- **Entity identification problem:** Identifying entity in one data source similar to entity in another table. For example, customer_id in one db and customer_no in another db refer to the same entity
- **Detecting and resolving data value conflicts:** Attribute values from different sources can be different due to different representations, different scales. E.g. metric vs. British units
- **Redundancy:** is another issue while performing data integration. Redundancy can occur due to the following reasons:
    - Object identification: The same attribute may have different names in different db
    - Derived Data: one attribute may be derived from another attribute.

Handling redundant data in data integration

1. **Correlation analysis**

For numeric data

Some redundancy can be identified by correlation analysis. The correlation between two variables A and B can be measured by

$$r_{A,B} = \frac{\Sigma(A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B}$$

$\overline{A}$, $\overline{B}$ are respective mean values of A and B
σ_A, σ_B are respective standard deviation of A and B
n is the number of tubles

- The result of the equation is > 0, then A and B are positively correlated, which means the value of A increases as the values of B increases. The higher value may indicate redundancy that may be removed.
- The result of the equation is = 0, then A and B are independent and there is no correlation between them.
- If the resulting value is < 0, then A and B are negatively correlated where the values of one attribute increase as the value of one attribute decrease which means each attribute may discourages each other.

  -also called Pearson's product moment coefficient

For categorical data

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Example:

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Data Transformation

Data transformation can involve the following:

- **Smoothing:** which works to remove noise from the data
- **Aggregation**: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute weekly and annual total scores.
- **Generalization of the data:** where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country.
- **Normalization:** where 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.
- **Attribute construction (feature construction)**: this is where new attributes are constructed and added from the given set of attributes to help the mining process.

Normalization

In which data are scaled to fall within a small, specified range, useful for classification algorithms involving neural networks, distance measurements such as nearest neighbor classification and clustering. There are 3 methods for data normalization. They are:

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Min-max normalization: performs linear transformation on the original data values. It can be defined as,

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

v is the value to be normalized
min_A,max_A are minimum and maximum values of an attribute A
new_ max_A, new_ min_A are the normalization range.

Z-score normalization / zero-mean normalization: In which values of an attribute A are normalized based on the mean and standard deviation of A. It can be defined as,

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

This method is useful when min and max value of attribute A are unknown or when outliers that are dominate min-max normalization.

Normalization by decimal scaling: normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value v of A is normalized to v' by computing,

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that Max}(|v'|)<1$$

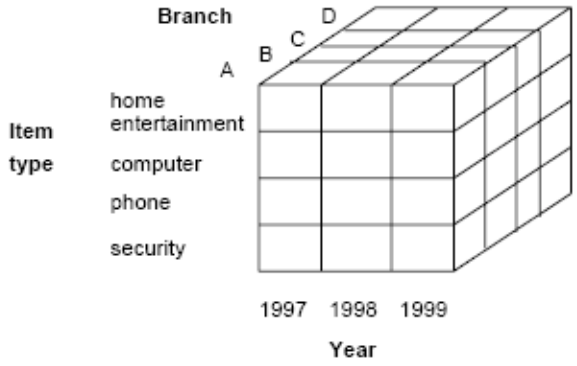
Data Reduction techniques

These techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Data reduction includes,

- Data cube aggregation
- Dimension reduction
- Data compression
- Discretization
- Numerosity reduction
  - Regression
  - Histograms
  - Clustering
  - Sampling

1. **Data cube aggregation**, where aggregation operations are applied to the data in the construction of a data cube.
2. **Attribute subset selection**, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
3. Dimensionality reduction, where encoding mechanisms are used to reduce the data set size. Examples: Wavelet Transforms Principal Components Analysis
4. **Numerosity reduction**, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.
5. **Discretization and concept hierarchy generation**, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies.

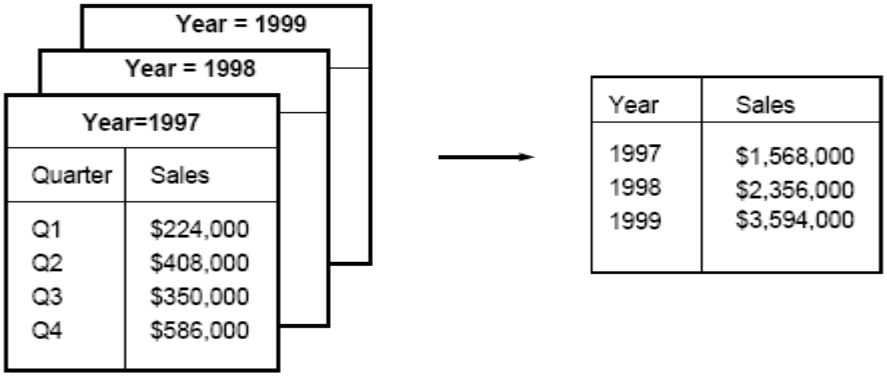
Data cube aggregation: Reduce the data to the concept level needed in the analysis. Queries regarding aggregated information should be answered using data cube when possible. Data cubes store multidimensional aggregated information. The following figure shows a data cube for multidimensional analysis of sales data with respect to annual sales per item type for each branch.



Each cells holds an aggregate data value, corresponding to the data point in multidimensional space.
Data cubes provide fast access to precomputed, summarized data, thereby benefiting on-line analytical processing as well as data mining.

The cube created at the lowest level of abstraction is referred to as the base cuboid. A cube for the highest level of abstraction is the apex cuboid. The lowest level of a data cube (base cuboid). Data cubes created for varying levels of abstraction are sometimes referred to as cuboids, so that a "data cube" may instead refer to a lattice of cuboids. Each higher level of abstraction further reduces the resulting data size.

The following database consists of sales per quarter for the years 1997-1999.



Suppose, the annalyser interested in the annual sales rather than sales per quarter, the above data can be aggregated so that the resulting data summarizes the total sales per year instead of per quarter. The resulting data in smaller in volume, without loss of information necessary for the analysis task

Dimensionality Reduction

It reduces the data set size by removing irrelevant attributes. This is a method of attribute subset selection are applied. A heuristic method of attribute of sub set selection is explained here:

Attribute sub selection / Feature selection

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

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

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

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

2. Step-wise backward elimination: The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.

3. Combination forward selection and backward elimination: The step-wise forward selection and backward elimination methods can be combined, where at each step one selects the best attribute and removes the worst from among the remaining attributes.

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

Forward Selection

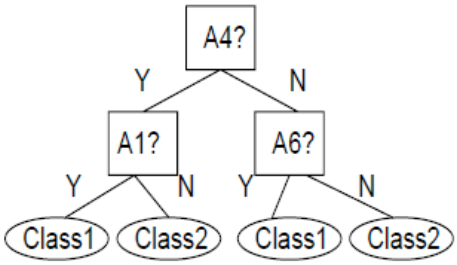
Initial attribute set:
{A1, A2, A3, A4, A5, A6}
Initial reduced set:
{}
-> {A1}
--> {A1, A4}
---> Reduced attribute set:
{A1, A4, A6}

Backward Elimination

Initial attribute set:
{A1, A2, A3, A4, A5, A6}
-> {A1, A3, A4, A5, A6}
--> {A1, A4, A5, A6}
---> Reduced attribute set:
{A1, A4, A6}

Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



---> Reduced attribute set:
{A1, A4, A6}

Greedy (heuristic) methods for attribute subset selection.

Wrapper approach/Filter approach:

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

Data compression

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

- **Wavelet transforms**
- **Principal components analysis**.

Wavelet compression is a form of data compression well suited for image compression. The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector D, transforms it to a numerically different vector, D0, of wavelet coefficients.

The general algorithm for a discrete wavelet transform is as follows.

1. The length, L, of the input data vector must be an integer power of two. This condition can be met by padding the data vector with zeros, as necessary.

2. Each transform involves applying two functions:

- data smoothing

- calculating weighted difference

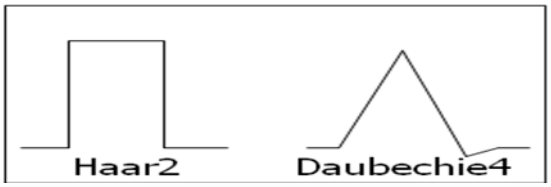
3. The two functions are applied to pairs of the input data, resulting in two sets of data of length L/2.

4. The two functions are recursively applied to the sets of data obtained in the previous loop, until the resulting data sets obtained are of desired length.

5. A selection of values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data.

If wavelet coefficients are larger than some user-specified threshold then it can be retained. The remaining coefficients are set to 0.

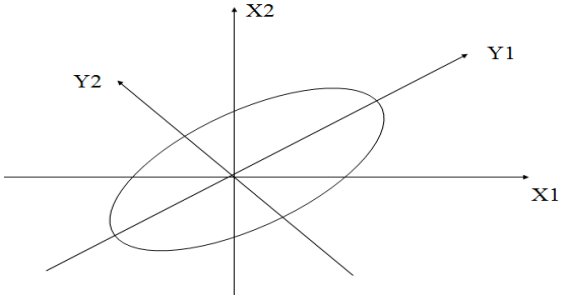
Haar2 and Daubechie4 are two popular wavelet transforms.



Principal Component Analysis (PCA)
-also called as Karhunen-Loeve (K-L) method
Procedure

- Given *N* data vectors from *k*-dimensions, find $c <= k$ orthogonal vectors that can be best used to represent data
  - The original data set is reduced (projected) to one consisting of N data vectors on c principal components (reduced dimensions)
- Each data vector is a linear combination of the *c* principal component vectors
- Works for ordered and unordered attributes
- Used when the number of dimensions is large

The principal components (new set of axes) give important information about variance. Using the strongest components one can reconstruct a good approximation of the original signal.



Numerosity Reduction

Data volume can be reduced by choosing alternative smaller forms of data. This tech. can be

- Parametric method
- Non parametric method

Parametric: Assume the data fits some model, then estimate model parameters, and store only the parameters, instead of actual data.
Non parametric: In which histogram, clustering and sampling is used to store reduced form of data.

Numerosity reduction techniques:

1 Regression and log linear models:

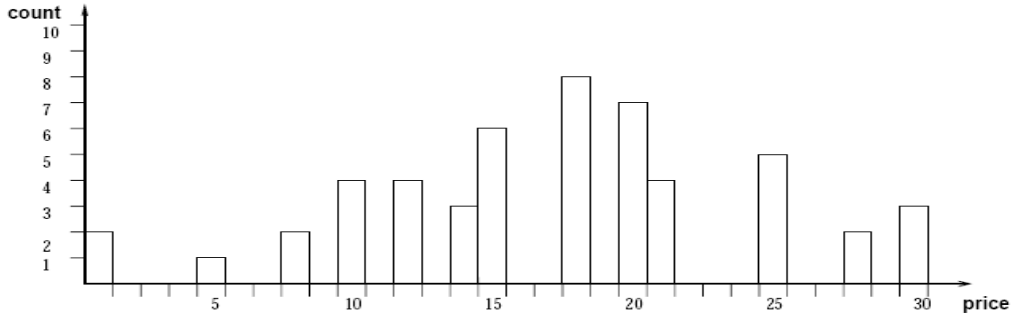
- Can be used to approximate the given data
- In linear regression, the data are modeled to fit a straight line using
  Y = α + β X, where α, β are coefficients
- Multiple regression: $Y = b0 + b1\ X1 + b2\ X2$.
  - Many nonlinear functions can be transformed into the above.

Log-linear model: The multi-way table of joint probabilities is approximated by a product of lower-order tables.

Probability: $p(a, b, c, d) = \alpha ab\ \beta ac \chi ad\ \delta bcd$

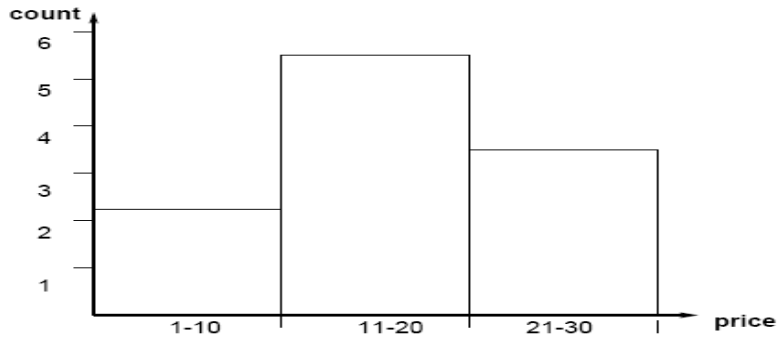
2 Histogram

- Divide data into buckets and store average (sum) for each bucket
- A bucket represents an attribute-value/frequency pair
- It can be constructed optimally in one dimension using dynamic programming
- It divides up the range of possible values in a data set into classes or groups. For each group, a rectangle (bucket) is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group.
- The buckets are displayed in a horizontal axis while height of a bucket represents the average frequency of the values.



Example:
The following data are a list of prices of commonly sold items. The numbers have been sorted.
1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Draw histogram plot for price where each bucket should have equi width of 10

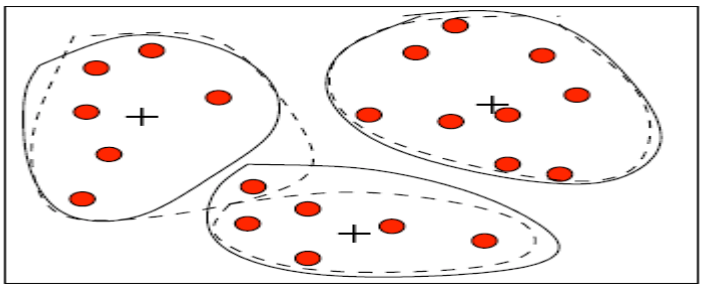


The buckets can be determined based on the following partitioning rules, including the following.

1. Equi-width: histogram with bars having the same width
2. Equi-depth: histogram with bars having the same height
3. V-Optimal: histogram with least variance å ($count_b$*$value_b$)
4. MaxDiff: bucket boundaries defined by user specified threshold

V-Optimal and MaxDiff histograms tend to be the most accurate and practical. Histograms are highly effective at approximating both sparse and dense data, as well as highly skewed, and uniform data.

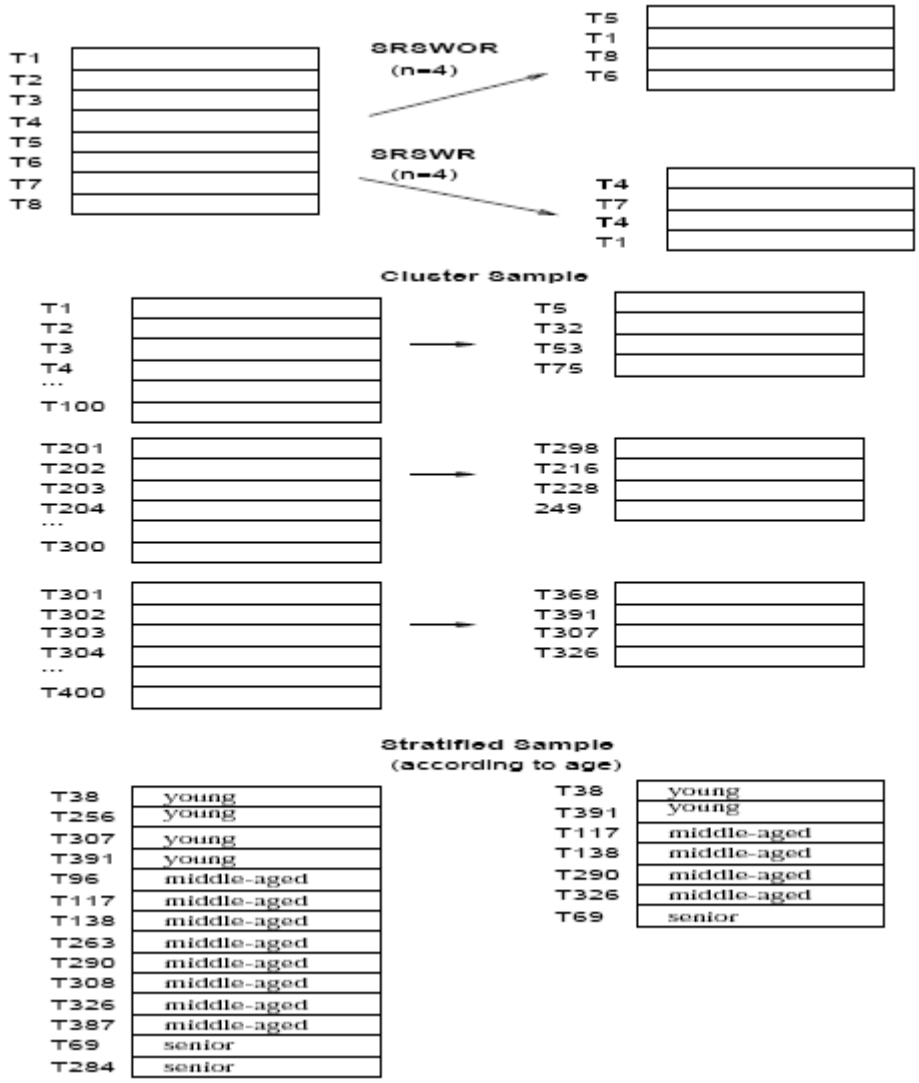
Clustering techniques consider data tuples as objects. They partition the objects into groups or clusters, so that objects within a cluster are “similar" to one another and “dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how “close" the objects are in space, based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



Quality of clusters measured by their diameter (max distance between any two objects in the cluster) or centroid distance (avg. distance of each cluster object from its centroid)

Sampling

Sampling can be used as a data reduction technique since it allows a large data set to be represented by a much smaller random sample (or subset) of the data. Suppose that a large data set, D, contains N tuples. Let's have a look at some possible samples for D.



1. Simple random sample without replacement (SRSWOR) of size n: This is created by drawing n of the
N tuples from D (n < N), where the probably of drawing any tuple in D is 1=N, i.e., all tuples are qually
likely.

2. Simple random sample with replacement (SRSWR) of size n: This is similar to SRSWOR, except
that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again.

3. Cluster sample: If the tuples in D are grouped into M mutually disjoint “clusters", then a SRS of m clusters can be obtained, where m < M. For example, tuples in a database are usually retrieved a page at a time, so that each page can be considered a cluster. A reduced data

representation can be obtained by applying, say, SRSWOR to the pages, resulting in a cluster sample of the tuples.
4. Stratified sample: If D is divided into mutually disjoint parts called “strata", a stratified sample of D is generated by obtaining a SRS at each stratum. This helps to ensure a representative sample, especially when the data are skewed. For example, a stratified sample may be obtained from customer data, where stratum is created for each customer age group. In this way, the age group having the smallest number of customers will be sure to be represented.

Advantges of sampling

1. An advantage of sampling for data reduction is that the cost of obtaining a sample is proportional to the size of the sample, n, as opposed to N, the data set size. Hence, sampling complexity is potentially sub-linear to the size of the data.

2. When applied to data reduction, sampling is most commonly used to estimate the answer to an aggregate query.

Discretization and concept hierarchies

Discretization:

Discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

Concept Hierarchy

A concept hierarchy for a given numeric 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 numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

Discretization and Concept hierarchy for numerical data:

Three types of attributes:
Nominal — values from an unordered set, e.g., color, profession
Ordinal — values from an ordered set, e.g., military or academic rank
Continuous — real numbers, e.g., integer or real numbers

There are five methods for numeric concept hierarchy generation. These include:

1. binning,
2. histogram analysis,
3. clustering analysis,
4. entropy-based discretization, and
5. data segmentation by “natural partitioning".

1 Bining: refer previous tpoic
2 Histogram: refer previous tpoic
3Clustering: refer previous tpoic
4 Entroby based discretization

An information-based measure called “entropy" can be used to recursively partition the values of a numeric attribute A, resulting in a hierarchical discretization.

Procedure:

- Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the information gain after partitioning is

$$I(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

- Entropy is calculated based on class distribution of the samples in the set. Given m classes, the entropy of S_1 is

$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

  where p_i is the probability of class i in S_1

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

5 Segmentation by Natural Partitioning

- A simply 3-4-5 rule 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

Example:
Suppose that profits at different branches of a company for the year 1997 cover a wide range, from -$351,976.00 to $4,700,896.50. A user wishes to have a concept hierarchy for profit automatically generated

Suppose that the data within the 5%-tile and 95%-tile are between -$159,876 and $1,838,761. The results of applying the 3-4-5 rule are shown in following figure

Step 1: Based on the above information, the minimum and maximum values are: MIN = -$351, 976.00, and MAX = $4, 700, 896.50. The low (5%-tile) and high (95%-tile) values to be considered for the top or first level of segmentation are: LOW = -$159, 876, and HIGH = $1, 838,761.

Step 2: Given LOW and HIGH, the most significant digit is at the million dollar digit position (i.e., msd =
1,000,000). Rounding LOW down to the million dollar digit, we get LOW' = -$1; 000; 000; and rounding
HIGH up to the million dollar digit, we get HIGH' = +$2; 000; 000.

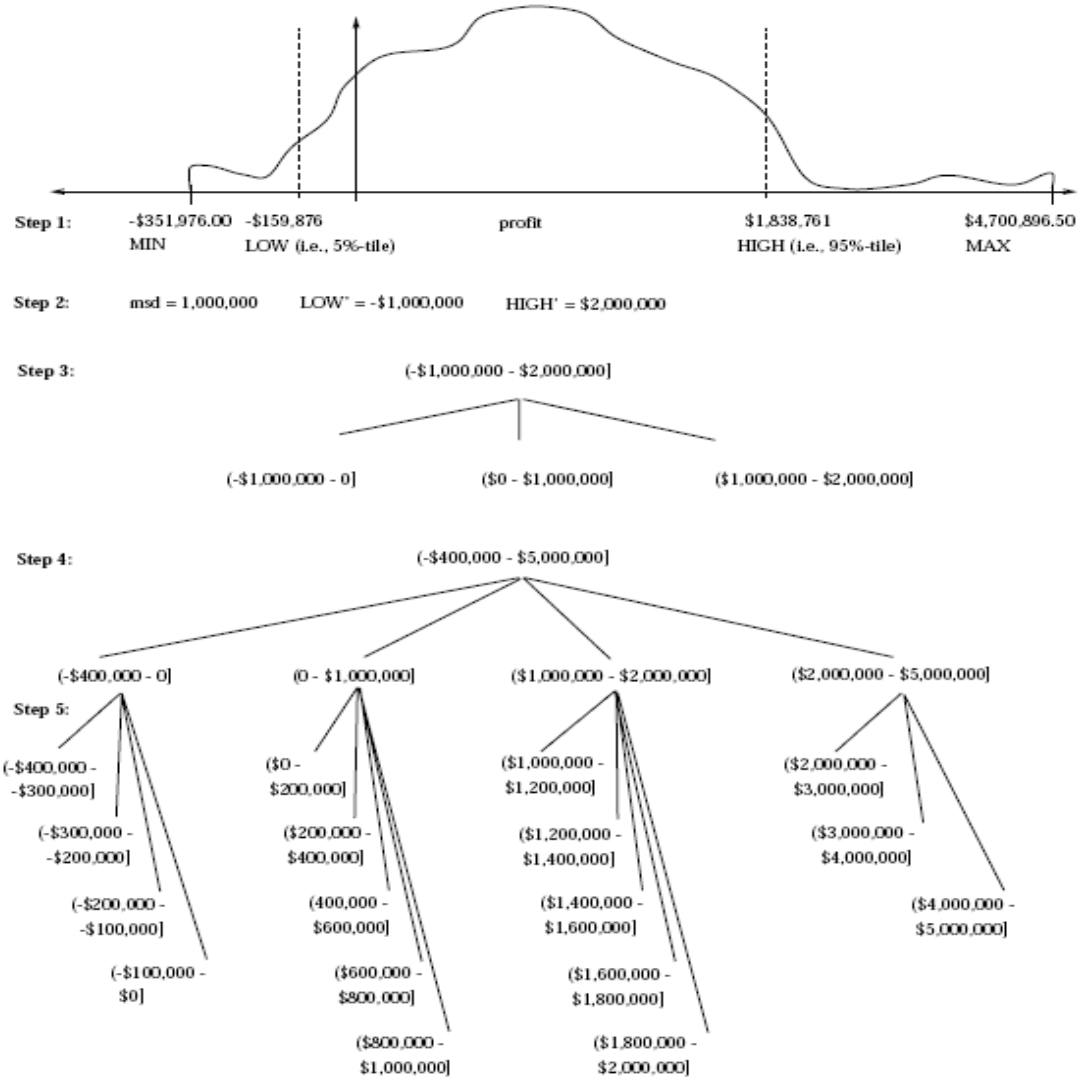
Step 3: Since this interval ranges over 3 distinct values at the most significant digit, i.e., (2; 000; 000-(-1, 000; 000))/1, 000, 000 = 3, the segment is partitioned into 3 equi-width sub segments according to the 3-4-5 rule: (-$1,000,000 - $0], ($0 - $1,000,000], and ($1,000,000 - $2,000,000]. This represents the top tier of the hierarchy.

Step 4: We now examine the MIN and MAX values to see how they "fit" into the first level partitions. Since the first interval, (-$1, 000, 000 - $0] covers the MIN value, i.e., LOW' < MIN, we can adjust the left boundary of this interval to make the interval smaller. The most significant

digit of MIN is the hundred thousand digit position. Rounding MIN down to this position, we get MIN0' = -$400, 000.
Therefore, the first interval is redefined as (-$400,000 - 0]. Since the last interval, ($1,000,000-$2,000,000] does not cover the MAX value, i.e., MAX > HIGH', we need to create a new interval to cover it. Rounding up MAX at its most significant digit position, the new interval is ($2,000,000 - $5,000,000]. Hence, the top most level of the hierarchy contains four partitions, (-$400,000 - $0], ($0 - $1,000,000], ($1,000,000 - $2,000,000], and ($2,000,000 - $5,000,000].

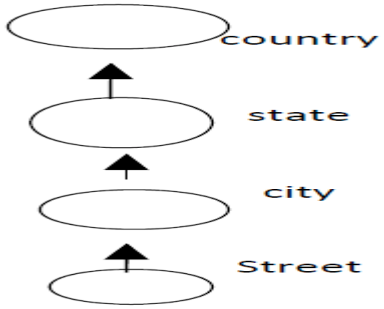
Step 5: Recursively, each interval can be further partitioned according to the 3-4-5 rule to form the next lower level of the hierarchy:

- The first interval (-$400,000 - $0] is partitioned into 4 sub-intervals: (-$400,000 - -$300,000], (-$300,000 - -$200,000], (-$200,000 - -$100,000], and (-$100,000 - $0].
- The second interval, ($0- $1,000,000], is partitioned into 5 sub-intervals: ($0 - $200,000], ($200,000 - $400,000], ($400,000 - $600,000], ($600,000 - $800,000], and ($800,000 - $1,000,000].
- The third interval, ($1,000,000 - $2,000,000], is partitioned into 5 sub-intervals: ($1,000,000 - $1,200,000], ($1,200,000 - $1,400,000], ($1,400,000 - $1,600,000], ($1,600,000 - $1,800,000], and ($1,800,000 - $2,000,000].
- The last interval, ($2,000,000 - $5,000,000], is partitioned into 3 sub-intervals: ($2,000,000 - $3,000,000], ($3,000,000 - $4,000,000], and ($4,000,000 - $5,000,000].



Concept hierarchy generation for category data

A concept hierarchy defines a sequence of mappings from set of low-level concepts to higher-level, more general concepts.
It organizes the values of attributes or dimension into gradual levels of abstraction. They are useful in mining at multiple levels of abstraction



Commercial tools available to find data discrepancy detection:

Data scrubbing tools use simple domain knowledge to detect errors and make corrections in the data

Data auditing tools find discrepancies by analyzing the data to discover rules and relationships and detecting data that violate such conditions

Data migration tools allow simple transformation to be specified such as replace the string "gender" by "sex"

ETL(Extraction/Transformation/Loading tools: allow users to specify transforms through a graphical user interface.

Integration of a data mining system with a database or data warehouse system

Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semitight coupling, and tight coupling. State which approach you think is the most popular, and why.

The differences between the following architectures for the integration of a data mining system with a database or data warehouse system are as follows.

Integration of a Data Mining System with a Database or Data Warehouse System

• No coupling:

The data mining system uses sources such as flat files to obtain the initial data set to be mined since no database system or data warehouse system functions are implemented as part of the process. Thus, this architecture represents a poor design choice.

• Loose coupling:

The data mining system is not integrated with the database or data warehouse system beyond their use as the source of the initial data set to be mined, and possible use in storage of the results. Thus, this architecture can take advantage of the flexibility, efficiency and features such as indexing that the database and data warehousing systems may provide. However, it is difficult for loose coupling to achieve high scalability and good performance with large data sets as many such systems are memory-based.

• Semitight coupling:

Some of the data mining primitives such as aggregation, sorting or pre computation of statistical functions are efficiently implemented in the database or data warehouse system, for use by the

data mining system during mining-query processing. Also, some frequently used inter mediate mining results can be pre computed and stored in the database or data warehouse system, thereby enhancing the performance of the data mining system.

• Tight coupling:

The database or data warehouse system is fully integrated as part of the data mining system and thereby provides optimized data mining query processing. Thus, the data mining sub system is treated as one functional component of an information system. This is a highly desirable architecture as it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment

From the descriptions of the architectures provided above, it can be seen that tight coupling is the best alternative without respect to technical or implementation issues. However, as much of the technical infrastructure needed in a tightly coupled system is still evolving, implementation of such a system is non-trivial. Therefore, the most popular architecture is currently semi tight coupling as it provides a compromise between loose and tight coupling.

Categorize the measures

- A measure is distributive, if we can partition the dataset into smaller subsets, compute the measure on the individual subsets, and then combine the partial results in order to arrive at the measure's value on the entire (original) dataset
- A measure is algebraic if it can be computed by applying an algebraic function to one or more distributive measures
- A measure is holistic if it must be computed on the entire dataset as a whole

Measure the Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location.

In other words, in many real-life situations, it is helpful to describe data by a single number that is most representative of the entire collection of numbers. Such a number is called a measure of central tendency. The most commonly used measures are as follows. **Mean, Median, and Mode**

Mean: mean, or average, of numbers is the sum of the numbers divided by n. That is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n} \quad \text{i.e.,} \quad \text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

shortly,

$$\bar{x} = \frac{\sum x}{n}$$

where $\bar{x}$ (read as 'x bar') is the mean of the set of x values,
$\sum x$ is the sum of all the x values, and
n is the number of x values.

Example 1

The marks of seven students in a mathematics test with a maximum possible mark of 20 are given below:

15 13 18 16 14 17 12

Find the mean of this set of data values.

Solution:

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$
$$= \frac{15+13+18+16+14+17+12}{7}$$
$$= \frac{105}{7}$$
$$= 15$$

So, the mean mark is 15.

Midrange

The midrange of a data set is the average of the minimum and maximum values.

Median: median of numbers is the middle number when the numbers are written in order. If is even, the median is the average of the two middle numbers.

Example 2

The marks of nine students in a geography test that had a maximum possible mark of 50 are given below:

47 35 37 32 38 39 36 34 35

Find the median of this set of data values.

Solution:

Arrange the data values in order from the lowest value to the highest value:

32 34 35 35 36 37 38 39 47

The fifth data value, 36, is the middle value in this arrangement.

$\therefore$ Median = 36

Note:

The number of values, n, in the data set = 9

$$\text{Median} = \frac{1}{2}(9+1)\text{ th value}$$
$$= 5\text{th value}$$
$$= 36$$

In general:

$$\text{Median} = \frac{1}{2}(n+1)\text{ th value, where } n \text{ is the number of data values in the sample}$$

If the number of values in the data set is even, then the **median** is the average of the two middle values.

Example 3

Find the median of the following data set:

12 18 16 21 10 13 17 19

Solution:

Arrange the data values in order from the lowest value to the highest value:

10 12 13 16 17 18 19 21

The number of values in the data set is 8, which is even. So, the median is the average of the two middle values.

$$\therefore \text{Median} = \frac{4\text{th data value} + 5\text{th data value}}{2}$$
$$= \frac{16+17}{2}$$
$$= \frac{33}{2}$$
$$= 16.5$$

Trimmed mean

A trimming mean eliminates the extreme observations by removing observations from each end of the ordered sample. It is calculated by discarding a certain percentage of the lowest and the highest scores and then computing the mean of the remaining scores.

Mode of numbers is the number that occurs most frequently. If two numbers tie for most frequent occurrence, the collection has two modes and is called bimodal.

The mode has applications in printing. For example, it is important to print more of the most popular books; because printing different books in equal numbers would cause a shortage of some books and an oversupply of others.

Likewise, the mode has applications in manufacturing. For example, it is important to manufacture more of the most popular shoes; because manufacturing different shoes in equal numbers would cause a shortage of some shoes and an oversupply of others.

Example 4

Find the mode of the following data set:

48 44 48 45 42 49 48

Solution:

The mode is 48 since it occurs most often.

- It is possible for a set of data values to have more than one mode.
- If there are two data values that occur most frequently, we say that the set of data values is **bimodal**.
- If there is three data values that occur most frequently, we say that the set of data values is **trimodal**

- If two or more data values that occur most frequently, we say that the set of data values is **multimodal**
- If there is no data value or data values that occur most frequently, we say that the set of data values has no mode.

The mean, median and mode of a data set are collectively known as measures of **central tendency** as these three measures focus on where the data is centered or clustered. To analyze data using the mean, median and mode, we need to use the most appropriate measure of central tendency. The following points should be remembered:

- The mean is useful for predicting future results when there are no extreme values in the data set. However, the impact of extreme values on the mean may be important and should be considered. E.g. The impact of a stock market crash on average investment returns.
- The median may be more useful than the mean when there are extreme values in the data set as it is not affected by the extreme values.
- The mode is useful when the most common item, characteristic or value of a data set is required.

Measures of Dispersion

Measures of dispersion measure how spread out a set of data is. The two most commonly used measures of dispersion are the variance and the standard deviation. Rather than showing how data are similar, they show how data differs from its variation, spread, or dispersion.
Other measures of dispersion that may be encountered include the Quartiles, Interquartile range (IQR), Five number summary, range and box plots

Variance and Standard Deviation

Very different sets of numbers can have the same mean. You will now study two measures of dispersion, which give you an idea of how much the numbers in a set differ from the mean of the set. These two measures are called the variance of the set and the standard deviation of the set

Consider a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with a mean of $\bar{x}$. The variance of the set is

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

and the standard deviation of the set is $\sigma = \sqrt{v}$ (σ is the lowercase Greek letter *sigma*).

The standard deviation of a set is a measure of how much a typical number in the set differs from the mean. The greater the standard deviation, the more the numbers in the set *vary* from the mean. For instance, each of the following sets has a mean of 5.

{5, 5, 5, 5}, {4, 4, 6, 6}, and {3, 3, 7, 7}

The standard deviations of the sets are 0, 1, and 2.

$$\sigma_1 = \sqrt{\frac{(5-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2}{4}}$$
$$= 0$$
$$\sigma_2 = \sqrt{\frac{(4-5)^2 + (4-5)^2 + (6-5)^2 + (6-5)^2}{4}}$$
$$= 1$$
$$\sigma_3 = \sqrt{\frac{(3-5)^2 + (3-5)^2 + (7-5)^2 + (7-5)^2}{4}}$$
$$= 2$$

Percentile

Percentiles are values that divide a sample of data into one hundred groups containing (as far as possible) equal numbers of observations.

The pth percentile of a distribution is the value such that p percent of the observations fall at or below it.

The most commonly used percentiles other than the median are the 25th percentile and the 75th percentile.

The 25th percentile demarcates the first quartile, the median or 50th percentile demarcates the second quartile, the 75th percentile demarcates the third quartile, and the 100^{th} percentile demarcates the fourth quartile.

Quartiles

Quartiles are numbers that divide an ordered data set into four portions, each containing approximately one-fourth of the data. Twenty-five percent of the data values come before the first quartile (Q1). The median is the second quartile (Q2); 50% of the data values come before the median. Seventy-five percent of the data values come before the third quartile (Q3).

Q1=25^{th} percentile=(n*25/100), where n is total number of data in the given data set

Q2=median=50^{th} percentile=(n*50/100)

Q3=75^{th} percentile=(n*75/100)

Inter quartile range (IQR)

The inter quartile range is the length of the interval between the lower quartile (Q1) and the upper quartile (Q3). This interval indicates the central, or middle, 50% of a data set.

$$IQR = Q3 - Q1$$

Range

The range of a set of data is the difference between its largest (maximum) and smallest (minimum) values. In the statistical world, the range is reported as a single number, the difference between maximum and minimum. Sometimes, the range is often reported as "from (the minimum) to (the maximum)," i.e., two numbers.

Example1:

Given data set: 3, 4, 4, 5, 6, 8

The range of data set is 3–8. The range gives only minimal information about the spread of the data, by defining the two extremes. It says nothing about how the data are distributed between those two endpoints.

Example2:

In this example we demonstrate how to find the minimum value, maximum value, and range of the following data: 29, 31, 24, 29, 30, 25

1. Arrange the data from smallest to largest.

24, 25, 29, 29, 30, 31

2. Identify the minimum and maximum values:

Minimum = 24, Maximum = 31

3. Calculate the range:

Range = Maximum-Minimum = 31–24 = 7.

Thus the range is 7.

Five-Number Summary

The Five-Number Summary of a data set is a five-item list comprising the minimum value, first quartile, median, third quartile, and maximum value of the set.

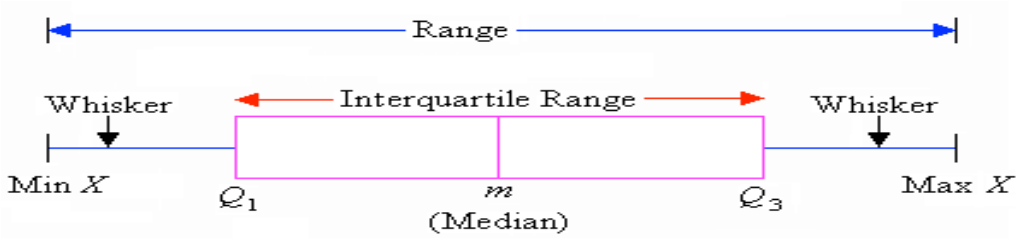
{MIN, Q1, MEDIAN (Q2), Q3, MAX}

Box plots

A box plot is a graph used to represent the range, median, quartiles and inter quartile range of a set of data values.

Constructing a Box plot: To construct a box plot:

(i) Draw a box to represent the middle 50% of the observations of the data set.
(ii) Show the median by drawing a vertical line within the box.
(iii) Draw the lines (called **whiskers**) from the lower and upper ends of the box to the minimum and maximum values of the data set respectively, as shown in the following diagram.



- X is the set of data values.
- Min X is the minimum value in the data set.
- Max X is the maximum value in the data set.

Example: Draw a boxplot for the following data set of scores:

76 79 76 74 75 71 85 82 82 79 81

Step 1: Arrange the score values in ascending order of magnitude:

71 74 75 76 76 79 79 81 82 82 85

There are 11 values in the data set.

Step 2: Q1=25th percentile value in the given data set

Q1=11*(25/100) th value

=2.75 =>3rd value

=75

Step 3: Q2=median=50th percentile value

=11 * (50/100) th value

=5.5th value => 6th value

=79

Step 4: Q3=75th percentile value

=11*(75/100)th value

=8.25th value=>9th value

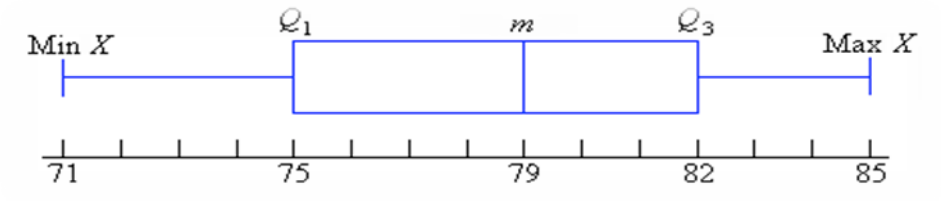
= 82

Step 5: Min X= 71

Step 6: Max X=85

Step 7: Range= 85-71 = 14

Step 5: IQR=height of the box=Q3-Q1=9-3=6th value=79



Since the medians represent the middle points, they split the data into four equal parts. In other words:

- one quarter of the data numbers are less than 75
- one quarter of the data numbers are between 75 and 79
- one quarter of the data numbers are between 79 and 82
- one quarter of the data numbers are greater than 82

Outliers

Outlier data is a data that falls outside the range. Outliers will be any points below $Q_1 - 1.5\times IQR$ or above $Q_3 + 1.5\times IQR$.

Example:

Find the outliers, if any, for the following data set:

10.2, 14.1, 14.4, **14.4**, 14.4, 14.5, 14.5, **14.6,** 14.7, 14.7, 14.7, **14.9,** 15.1, 15.9, 16.4

To find out if there are any outliers, I first have to find the IQR. There are fifteen data points, so the median will be at position (15/2) = 7.5=8th value=14.6. That is, $Q_2 = 14.6$.

Q_1 is the fourth value in the list and Q_3 is the twelfth: $Q_1 = 14.4$ and $Q_3 = 14.9$.

Then IQR = 14.9 – 14.4 = 0.5.

Outliers will be any points below:

$Q_1 - 1.5 \times IQR = 14.4 - 0.75 = 13.65$ or above $Q_3 + 1.5 \times IQR = 14.9 + 0.75 = 15.65$.

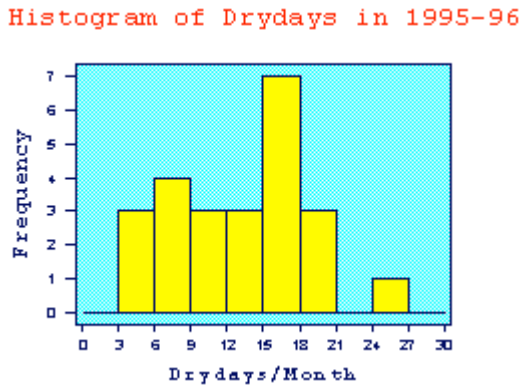
Then the outliers are at 10.2, 15.9, and 16.4.

The values for $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ are the "fences" that mark off the "reasonable" values from the outlier values. Outliers lie outside the fences.

Graphic Displays of Basic Descriptive Data Summaries

1 Histogram

A histogram is a way of summarizing data that are measured on an interval scale (either discrete or continuous). It is often used in exploratory data analysis to illustrate the major features of the distribution of the data in a convenient form. It divides up the range of possible values in a data set into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group. This means that the rectangles might be drawn of non-uniform height.



The histogram is only appropriate for variables whose values are numerical and measured on an interval scale. It is generally used when dealing with large data sets (>100 observations)

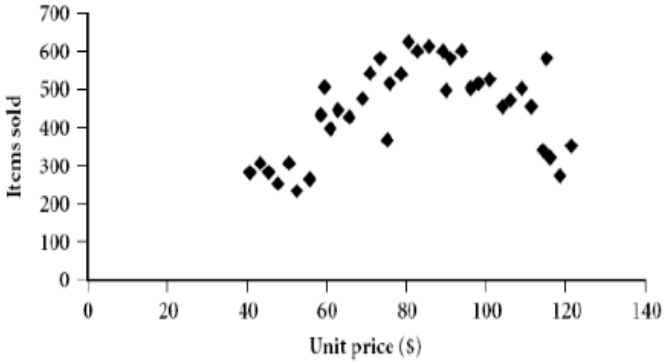
A histogram can also help detect any unusual observations (outliers), or any gaps in the data set.

2 Scatter Plot

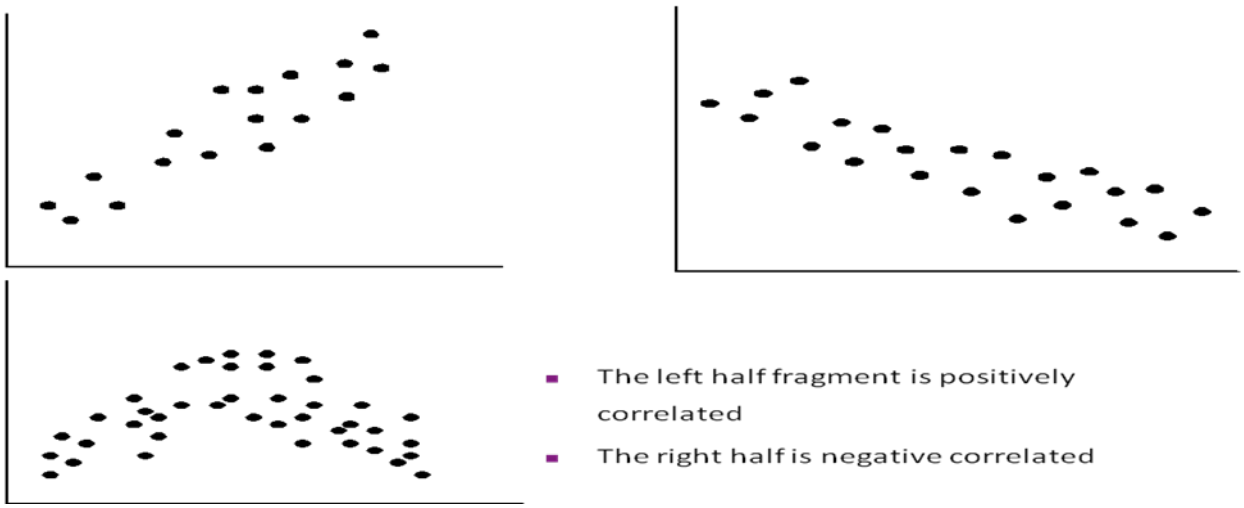
A scatter plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual

picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.

Each unit contributes one point to the scatter plot, on which points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between the two variables.



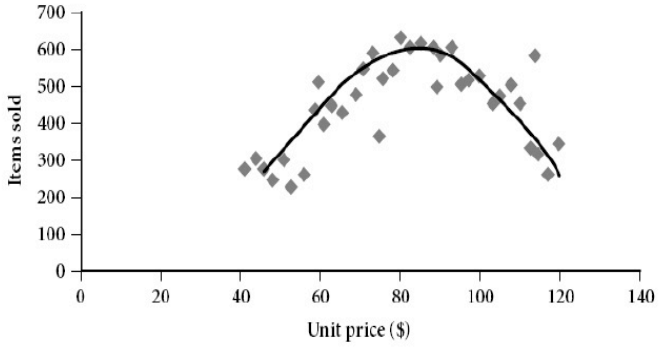
Positively and Negatively Correlated Data



A scatter plot will also show up a non-linear relationship between the two variables and whether or not there exist any outliers in the data.

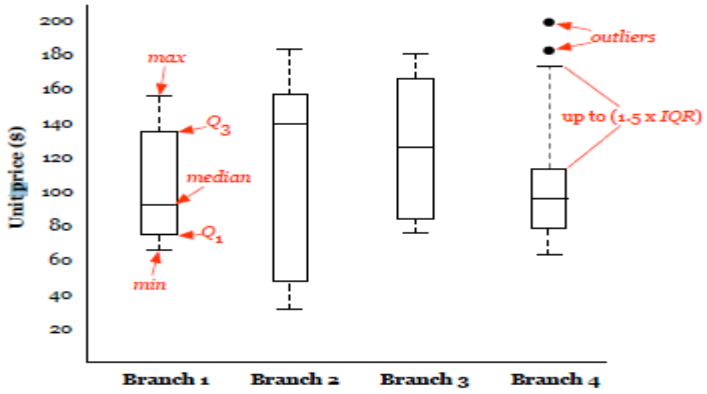
3 Loess curve

It is another important exploratory graphic aid that adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence. The word loess is short for "local regression."



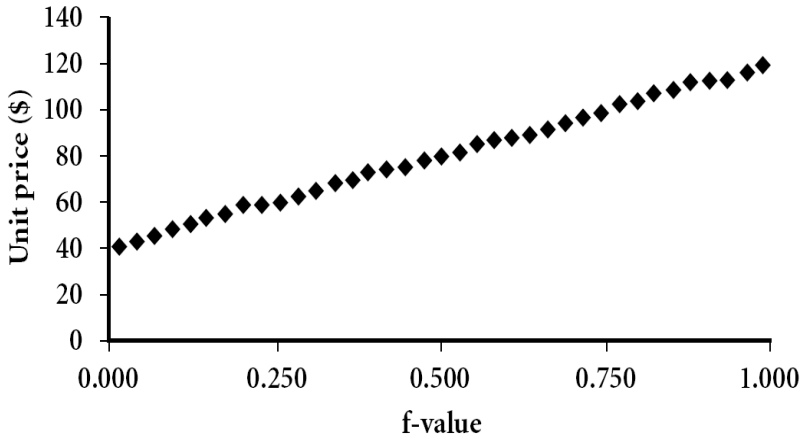
4 Box plot

The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median.



5 Quantile plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



The *f* quantile is the data value below which approximately a decimal fraction *f* of the data is found. That data value is denoted q(f). Each data point can be assigned an *f*-value. Let a time series *x* of length *n* be sorted from smallest to largest values, such that the sorted values have rank. The *f*-value for each observation is computed as . 1,2,..., n . The *f*-value for each observation is computed as, $f_i = \frac{i - 0.5}{n}$

6 Quantile-Quantile plots (Q-Q plot)

Quantile-quantile plots allow us to compare the quantiles of two sets of numbers.

This kind of comparison is much more detailed than a simple comparison of means or medians.

A normal distribution is often a reasonable model for the data. Without inspecting the data, however, it is risky to assume a normal distribution. There are a number of graphs that can be used to check the deviations of the data from the normal distribution. The most useful tool for assessing normality is a quantile quantile or QQ plot. This is a scatterplot with the quantiles of the scores on the horizontal axis and the expected normal scores on the vertical axis.

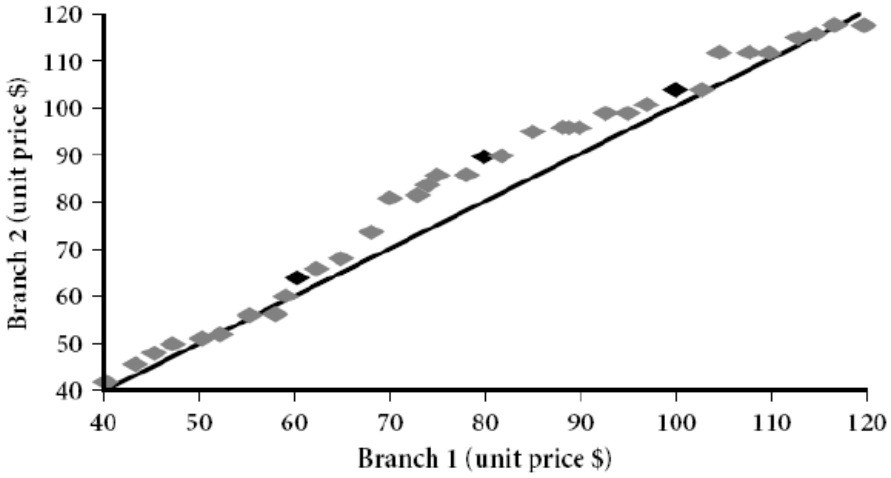
In other words, it is a graph that shows the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

The steps in constructing a QQ plot are as follows:

First, we sort the data from smallest to largest. A plot of these scores against the expected normal scores should reveal a straight line.

The expected normal scores are calculated by taking the z-scores of (I - ½)/n where I is the rank in increasing order.

Curvature of the points indicates departures of normality. This plot is also useful for detecting outliers. The outliers appear as points that are far away from the overall pattern op points



How is a quantile-quantile plot different from a quantile plot?

A quantile plot is a graphical method used to show the approximate percentage of values below or equal
to the independent variable in a univariate distribution. Thus, it displays quantile information for all the
data, where the values measured for the independent variable are plotted against their corresponding
quantile.

A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corre-
sponding quantiles of another univariate distribution. Both axes display the range of values measured
for their corresponding distribution, and points are plotted that correspond to the quantile values of the
two distributions. A line (y = x) can be added to the graph along with points representing where the
first, second and third quantiles lie, in order to increase the graph's informational value. Points that lie
above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than
for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points
lying below this line.

Example 1

- Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
- (a) What is the mean of the data? What is the median?
- (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
- (c) What is the midrange of the data?
- (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
- (e) Give the five-number summary of the data.
- (f) Show a boxplot of the data.

- (a) What is the mean of the data? What is the median?

The (arithmetic) mean of the data is: $\bar{x} = 809/27 = 30$. The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

- (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35.

- (c) What is the midrange of the data?

The midrange (average of the largest and smallest values in the data set) of the data is: $(70+13)/2 = 41.5$

- (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.

- (e) Give the five-number summary of the data.

The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and for this data is: 13, 20, 25, 35, 70.

- (f) Show a boxplot of the data.

Draw the figure.

On an interview for a job, the interviewer tells you that the average annual income of the company's 25 employees is $60,849. The actual annual incomes of the 25 employees are shown below. What are the mean, median, and mode of the incomes? Was the person telling you the truth?

$17,305,	$478,320,	$45,678,	$18,980,	$17,408,
$25,676,	$28,906,	$12,500,	$24,540,	$33,450,
$12,500,	$33,855,	$37,450,	$20,432,	$28,956,
$34,983,	$36,540,	$250,921,	$36,853,	$16,430,
$32,654,	$98,213,	$48,980,	$94,024,	$35,671

Solution

The mean of the incomes is

$$\text{Mean} = \frac{17{,}305 + 478{,}320 + 45{,}678 + 18{,}980 + \cdots + 35{,}671}{25}$$

$$= \frac{1{,}521{,}225}{25} = \$60{,}849.$$

To find the median, order the incomes as follows.

$12,500,	$12,500,	$16,430,	$17,305,	$17,408,
$18,980,	$20,432,	$24,540,	$25,676,	$28,906,
$28,956,	$32,654,	$33,450,	$33,855,	$34,983,
$35,671,	$36,540,	$36,853,	$37,450,	$45,678,
$48,980,	$94,024,	$98,213,	$250,921,	$478,320

From this list, you can see that the median (the middle number) is $33,450. From the same list, you can see that $12,500 is the only income that occurs more than once. So, the mode is $12,500. Technically, the person was telling the truth because the average is (generally) defined to be the mean. However, of the three measures of central tendency *Mean:* $60,849 *Median:* $33,450 *Mode:* $12,500 it seems clear that the median is most representative. The mean is inflated by the two highest salaries.