DATA MINING AND DATA WAREHOUSING

UNIT –I:

Introduction: Why Data Mining? What Is Data Mining? 1.3 What Kinds of Data Can Be Mined? 1.4 What Kinds of Patterns Can Be Mined? Which Technologies Are Used? Which Kinds of Applications Are Targeted? Major Issues in Data Mining. Data Objects and Attribute Types, Basic Statistical Descriptions of Data, Data Visualization, Measuring Data Similarity and Dissimilarity.

INTRODUCTION: Data mining is nothing but discovery of *knowledged data* from large database. Generally the term mining refers to mining of gold from rocks or sand is called gold mining.

1.1. Why Data Mining?

- The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and 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.
- Data mining can be viewed as a result of the natural evolution of information technology. It means, providing a path to extract the required data of an industry from warehousing machine. This is the witness of developing knowledge of an industry.
- It includes *data collection, database creation, data management (i.e* data storage and retrieval, and database transaction processing) and *data analysis and understanding(*involving data warehousing and data mining).

1.1.1. **Evolution of data mining and data warehousing:** In the development of data mining, we should know the evolution of database. This includes,

Data collection and Database creation*:* In the 1960's, database and information technology began with file processing system. It is powerful database system. But it is providing inconsistency of data. It means, a user needs to maintain duplicate data of an industry.

Database Management System*:* In b/w 1970 – 1980, the progress of database is
→ Hierarchical and network database systems were developed.
→ Relational database systems were developed
→ Data modeling tools were developed in early 1980s (such as E-R model etc.
→ Indexing and data organization techniques were developed. (such as B+ tree, hashing etc).
→ Query languages were developed. (such as SQL, PL/SQL)
→ User interfaces, forms and reports, query processing.
→ On-line transaction processing (OLTP)

Advanced Database Systems: In mid 1980s to till date,
→ Advanced data models were developed. (such as extended relational, object-oriented, object-relational, spatial, temporal, multimedia, scientific databases etc.

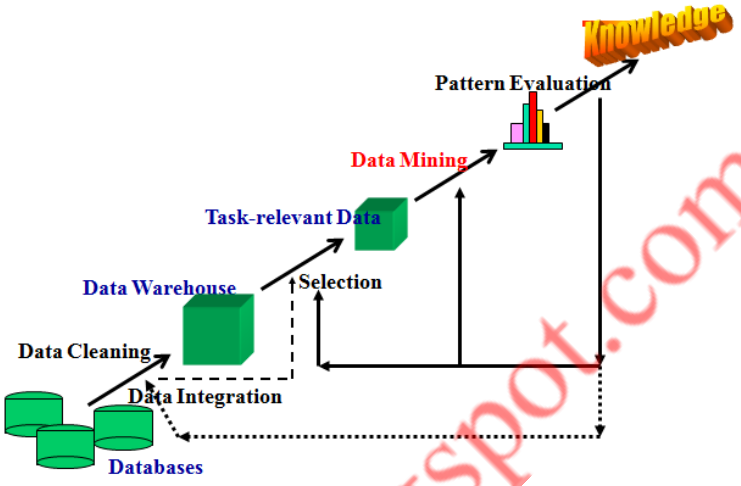
Data Warehousing and Data mining*:* In late 1980 to till date
→ Developed Data warehouse and OLAP technology
→ Data mining and knowledge discovery were introduced.

Web-based Databases Systems: In 1990 – till date
→ XML based database systems and web mining were developed.

***New Generation of Integrated Information Systems*:** From 2000 onwards developed an integrated information system.

1.2. What is Data Mining: The term Data Mining refers to *extracting or "mining" knowledge from large amounts of data*. The term mining is actually a misnomer (i.e. unstructured data). For example, mining of gold from rocks or sand is referred to as gold mining.

Data mining is the process of discovering meaningful new trends by storing the large amount of data in repository of database. It also uses pattern recognition techniques as well as statistical techniques.



1.2.1. Data mining steps in the knowledge discovery process (KDD): The Data mining is a step in the *Knowledge Discovery in Databases (KDD)*. It has different stages, such as

→***Data Cleaning:*** It is the process of removing noise and inconsistent data.

→***Data Integrating:*** It is the process of combining data from multiple sources.

→***Data Selection:*** It is the process of retrieving relevant data from database.

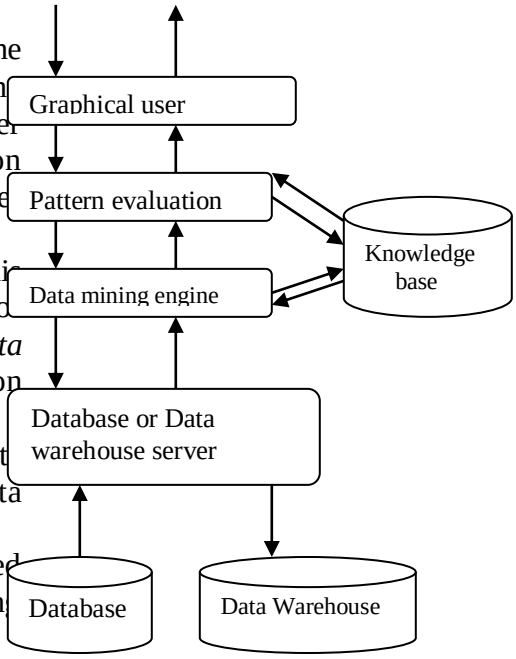
→***Data Transformation:*** In this process, data are transformed or consolidated into forms or reports by performing summary or aggregation operation.

→***Data Mining:*** It is an essential process to extracting data from raw data by using intelligent methods.

→***Pattern Evaluation:*** to identify the discovered data is in the knowledge based on some interestingness measures.(i.e identify the mined data is in the required format or not.).

→***Knowledge presentation:*** Visualization and knowledge representation techniques are used to present the mined data to the user.

1.2.2 Architecture of Data Mining System: The architecture of data mining is the process of discovering th interesting knowledge from large amounts of data stored either in databases or in the data warehouse or information repositories. It has various stages to extract the data into use view from unstructured sources.



→***Database, data warehouse, or information repositories****:* Thi is single or set of databases, data warehouses, spreadsheets, d other kinds of information repositories. In this step, the *Data Cleaning and Data Integration techniques* may be performed on the data.

→***Database or data warehouse server:*** The database or dat warehouse server is responsible for fetching the relevant data based on the user's data mining request.

→***Knowledge base:*** This is the domain knowledge that is used to guide the search or evaluate the interestingness of resultin patterns.

→***Data mining engine:*** This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis, and evolution and deviation analysis.

→***pattern evaluation module:*** This step providing measures, constraints(rules), methods etc to filter out the discovered patterns or data. This is most useful for efficient data mining.

→***Graphical user interface:*** This step provides the communication b/w user and data mining system. It allows the user to interact with the system by specifying a data mining query or task.

1.3. What Kind of Data Can Be Mined?

Data mining can be applied to any kind of information repositories such as Databases data, data warehouse, transactional data bases, advanced systems, flat files and the *World Wide Web*. Advanced databases systems include object-oriented, object-relation databases, time series databases, text databases and multimedia databases.

→**1.3.1.** ***Databases Data:*** A database system is also called a ***database management system* (DBMS).** It consists of a collection of interrelated data, known as a database, and set of software programs to manage and access the data. The software programs provide mechanisms for defining database structures and data storage. These also provide data consistency and security, concurrency, shared or distributed data access etc.

A ***relational database*** is a collection of tables, each of which is assigned a unique name. Each table consists of a set of **attributes** (columns or fields) and a set of **tupples** (records or rows). Each **tupple** is identified by a unique key and is described by a set of attribute values. For this, ER models are constructed for relational databases. For example, *AllElectronics* Industry illustrated with following information. *custer, item, employee, branch.*

customer table

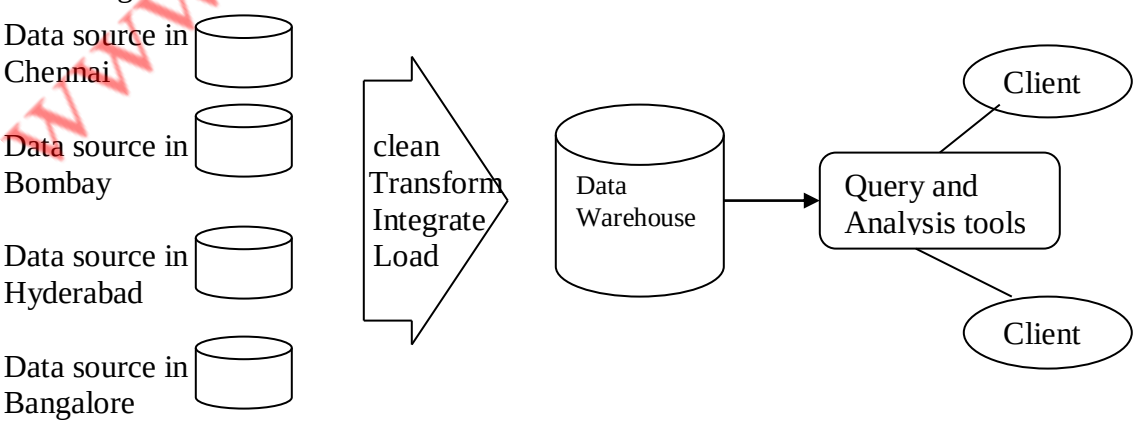
Cust-id	cust_name	Gender	Address_Place	item-id

item table

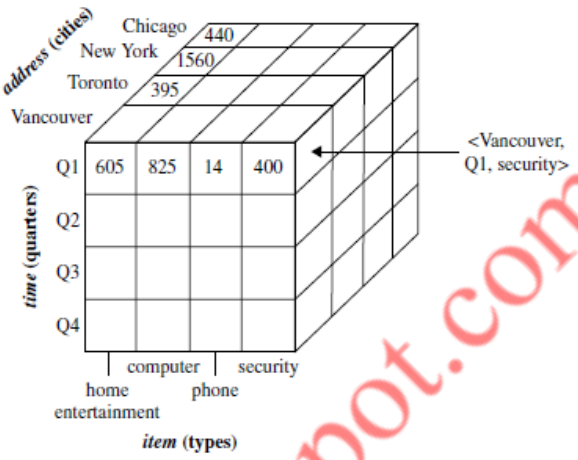
item-id	item_name	Price	Manufacturing

AllElectronics company sales his products (such as computers and printers) to the customers. Here providing the relation b/w **custormer** table (file) and **product** table. By this relation can identify what types of products are taken the customer.

→**1.3.2.** ***Data Warehouses*:** A data warehouse is a repository of information collected from multiple sources, stored under a schema and resides at a single site. The data warehouses are constructed by a process of data cleaning, data transformation, data integration, data loading and periodic data refreshing.



A data warehouse is mainly 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 sales amount. The 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 preprocess and fast accessing of summarized data.



A data cube for summarized sales data of AllElectronics is presented in fig. The cube has three dimensions such as address (Chennai, Bombay, Hyd, Bang), time with Q1,Q2,Q3,Q4 and item with home needs, computer, phone and security. In this, aggregate value stored in each cell of the cube.

By providing multidimensional data views, performed the OLAP operations. Such as **drill-down,** and **roll-up.**

→**1.3.3.** ***Transactional Databases*****:** A transactional database consists of a file where each record represents a transaction. A transaction includes a unique transaction such as data of the transaction, the customer id number, the ID number of the sales person and so on.

AllElectronics transactions can be stored in a table with one record per transaction. This is shown in fig.

Transaction_id	List of items	Transaction dates
T100	I1, I3, I8, I16	18-12-2018
T200	I2, I8	18-12-2018

1.4 What Kinds of Patterns Can Be Mined? (or) Data Mining Functionalities:

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks are classified into two categories ***descriptive*** and ***predictive***.

→ ***Descriptive*** mining tasks characterize the general properties of the data in the database.

→***Predictive*** mining tasks perform inference on the current data in order to make predictions.

1.4.1. Concept/Class Description: Descriptions of a individual classes or a concepts in summarized, concise and precise terms called class or concept descriptions. These descriptions can be divided into 1. Data Characterization 2. Data Discrimination.

Data Characterization:

- It is summarization of the general characteristics of a target class of data (forms).
- The data corresponding to the user specified class are collected by a database query.

The output of data characterization can be presented in various forms like *pie charts, bar charts curves, multidimensional cubes, multidimensional tables* etc. The resulting descriptions can be presented as generalized relations are called *characteristic rules*.

Data Discriminations: Comparison of two target class data objects from one or set of contrasting (distinct) classes. The target and contrasting classes can be specified by the user, and the corresponding data objects are retrieved through database queries.

For example, comparison of products whose sales increased by 10% in the last year with those whose sales decreased by 30% during the same period. This is called data discrimination.

1.4.2. Mining Frequent Patterns, Associations and Correlations:

1.4.2.1. Frequent Patterns: A *frequent itemset* typically refers to a set of items that often appear in a transactional data. For example, milk, and bread are frequently purchased by many customers. *AllElectronics industry occurring the products which are frequently purchased by the customers.* Generally, home needs are frequently used by the more customers.

1.4.2.2. Association Analysis: "*What is association analysis ?"*

Association analysis is the discovery of *association rules* showing attribute with value conditions that occur frequently together the given set of data. It is used for transaction data analysis. The Association rule of the form X ==> Y.

For example, In AllElectronics relational database, data mining system may find association rules like

buys(X, "computer") ==> buys(X, "software")

Here, who buys "computer", they buys "software".

age (X, "20 .. 29") & income (X, "20k .. 29k") ==> buys(X,

In this, the Association rule indicate that that indicates who employee of AllElectronics have the age b/w 20 to 29 and earning income b/w 20000 to 29000 are purchased CD player at AllElectronics Company.

1.4.2.3. Classification and Regressive prediction:

Classification is the process of finding a set of models that describes and distinguishes data classes or concepts.

- The derived model may be represented in various forms such as *classification (IF-THEN) rules, decision trees, mathematical formulae or neural networks.*
- A *decision tree* is a **flow-chart** like tree structure. The decision trees can easily converted to classification rule. The neural networks are used for classification to provide connection b/w computers.

age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*
age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*
age(X, "middle_aged") ⟶ *class(X, "C")*
age(X, "senior") ⟶ *class(X, "C")*

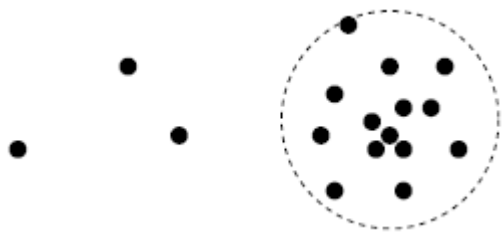
Regression for Predication is used to predict missing or unavailable data values rather than class labels. Prediction refers to both *data value prediction* and class *label prediction.* The predicted values are numerical data and are often referred to as ***prediction.***

1.4.2.4. Cluster Analysis: ("*What is cluster analysis?")*

Clustering is a method of grouping data into different groups, so that in each group share similar trends and patterns. The objectives of clustering are

- To uncover natural groupings
- To initiate hypothesis about the data
- To find consistent and valid organization of data.

For example, Cluster analysis can be performed on AllElectronics customers. It means, to identify homogeneous (same group) customers. By this cluster may represent target groups for marketing to increase the sales.

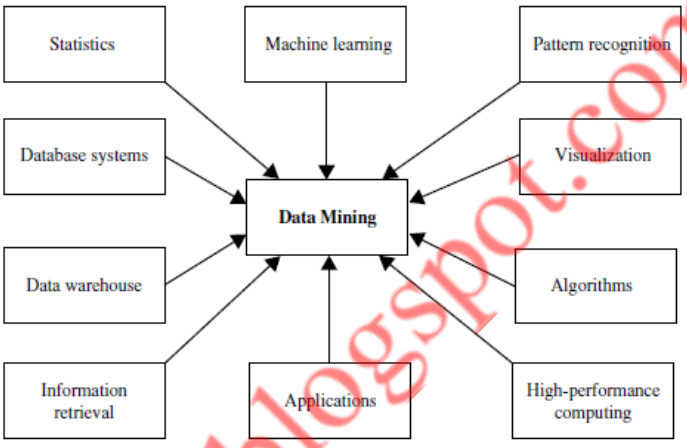


1.4.2.5.Outlier Analysis: In this analysis, a database may contain data objects that do not do what someone wants. Most data mining methods discard outliers as noise or exceptions. Finding such type of applications are fraud detection is referred as **outlier mining.**

For example, Outlier analysis may uncover usage of credit cards by detecting purchases of large amount of products when comparing with regular purchase of large product customers.

1.5. Which Technologies Are Used? (or) Classification of Data Mining Systems:

Data mining is classified with many techniques. Such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high performance computing, and many application domains (Shown in Figure).Data mining system can be categorized according to various criteria.



Statistics :A ***statistical model*** is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes. For example, in data mining tasks like data characterization and classification, statistical models of target classes can be built.

Machine Learning: *Machine learning* investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to *automatically* learn to recognize complex patterns and make intelligent decisions based on data. Machine learning is a fast-growing discipline.

- **Supervised learning** is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.
- **Unsupervised learning** is essentially a synonym for clustering. The learning processis unsupervised since the input examples are not class labeled. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits. Suppose that it finds 10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to 9, respectively.
- **Semi-supervised** learning is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model. For a two-class problem, one class as the *positive examples* and the other class as the *negative examples.*
- **Active learning** is a machine learning approach that lets users play an active role in the learning process.

Database Systems and Data Warehouses:

- **Database systems** can focuses on the creation, maintenance, and use of databases for organizations and end-users. Particularly, database systems principles in data models, query languages, query processing and optimization methods, data storage, and indexing and accessing methods. Many data mining tasks need to handle large data sets or even real-time,

fast streaming data. Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities.

- **A data warehouse** integrates data from multiple sources and various timeframes. It provides OLAP facilities in multidimensional databases to promote multidimensional data mining. It maintain recent data, previous data and historical data in database.

Information Retrieval:

- **Information retrieval** (IR) is the science of searching for documents or information in documents. The typical approaches in information retrieval adopt probabilistic models. For example, a text document can be observing as a container of words, that is, a multi set of words appearing in the document.

Pattern recognition is the process of recognizing patterns by using machine learning algorithm. Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation. One of the important aspects of the pattern recognition is its application potential. **Examples:** Speech recognition, speaker identification, multimedia document recognition (MDR), automatic medical diagnosis.

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

An ***algorithm*** in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends.

High Performance Computing (HPC) framework which can abstract the increased complexity in current computing systems and at the same time provide performance benefits by exploiting multiple forms of parallelism in Data Mining algorithms.

Data Mining Applications: The list of areas where data mining is widely used − *Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Other Scientific Applications, Intrusion Detection.*

1.6. Which Kinds of Applications Are Targeted?

Data mining has seen great successes in many applications. Presentations of data mining in knowledge-intensive application domains, such as bioinformatics and software engineering,

- **Business intelligence** (**BI**) technologies provide historical, current, and predictive views of business operations. Examples include reporting, online analytical processing, business performance management, competitive intelligence, benchmarking, and predictive analytics.
    - Data mining is the core of business intelligence. Online analytical processing tools in business intelligence depend on data warehousing and multidimensional data mining. Classification and prediction techniques are the core of predictive analytics in business intelligence, for which there are many applications in analyzing markets, supplies, and sales.
- A **Web search engine** is a specialized computer server that searches for information on the Web. The search results of a user query are often returned as a list (sometimes called hits). The hits may consist of web pages, images, and other types of files.

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

(e.g., deciding which pages should be crawled and the crawling frequencies), indexing (e.g., selecting pages to be indexed and deciding to which extent the index should be constructed), and searching (e.g., deciding how pages should be ranked, which advertisements should be added, and how the search results can be personalized or made "context aware").

1.7. Major issues in Data Mining: Data mining is a dynamic and fast-expanding field with great strengths. Major issues in data mining research, partitioning them into five groups: *mining methodology, user interaction, efficiency and scalability, diversity of data types*, and *data mining and society*.

→ **Mining methodology**: In this methodology the user interaction on different issues such as

- *Mining various and new kinds of knowledge.*
- *Mining knowledge in multidimensional space.*
- *Data mining—an interdisciplinary effort.*
- *Boosting the power of discovery in a networked environment.*
- *Handling uncertainty, noise, or incompleteness of data.*
- *Pattern evaluation and pattern- or constraint-guided mining.*

→**User Interaction:** Interesting areas of research include *how to interact with a data mining system, how to incorporate a user's background knowledge in mining*, and *how to visualize and comprehend data mining results.*

- *Interactive mining.*
- *Incorporation of background knowledge.*
- *Ad hoc data mining and data mining query language.*
- *Presentation and visualization of data mining results.*

→**Efficiency and Scalability:**

- *Efficiency and scalability of data mining algorithms.*
- *Parallel, distributed, and incremental mining algorithms.*
- *Cloud computing* and *cluster computing.*

→**Diversity of Database Types:**

- *Handling complex types of data*
- *Mining dynamic, networked, and global data repositories*

→**Data Mining and Society:**

- *Social impacts of data mining.*
- *Privacy-preserving data mining.*
- *Invisible data mining.*

1.8. Data Objects and Attribute Types:

A **data object** represents an entity.

- In a sales database, the objects may be customers, store items, and sales;
- in a medical database, the objects may be patients;
- in a university database, the objects may be students, professors, and courses.
- Data objects are typically described by attributes. Data objects can also be referred to as *samples, examples, instances, data points*, or *objects*.
- The data objects are stored in a database, they are *data tuples*. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.

1.8.1. What Is an Attribute?

- An **attribute** is a data field, representing a characteristic or feature of a data object.
- The attribute may also called, dimension, feature, and variable. The term *dimension* is commonly used in data warehousing. The term *feature* is commonly used in Machine learning, while statisticians prefer the term *variable*. Data mining and database professionals commonly use the term *attribute*.
- For example, Attributes is described for a customer object is as *customer ID*, *name*, and *address*.

Types of Attribute: The type of an attribute is determined by the set of possible values. They are —nominal, binary, ordinal, or numeric.

- **Nominal Attributes:** Nominal means "relating to names." The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values are also known as enumerations.
  - Example: *hair color* and *marital status* are two attributes describing person objects. In our application, possible values for *hair color* are black, brown, blond, red, auburn, gray, and white. The attribute *marital status* can take on the values single, married, divorced, and widowed.
- **Ordinal Attributes:** An **ordinal attribute** is an attribute with possible values that have a meaningful order or *ranking* among them, but the magnitude between successive values is not known.
  - **Example:** *drink size* corresponds to the size of drinks available at a fast-food restaurant. This ordinal attribute has three possible values: *small*, *medium*, and *large*. The values have a meaningful sequence (which corresponds to increasing drink size.
  - Other examples of ordinal attributes include *grade* (e.g., *A+*, *A*, *A-*, *B+* and so on).
  - Professional ranks can be enumerated in a sequential order: for example, *assistant*, *associate*, and *professors*.
- **Binary:** Nominal attribute with only 2 states (0 and 1). Eg: true or false, yes or no.
  - Symmetric binary: both outcomes equally important e.g., gender
  - Asymmetric binary: outcomes not equally important. e.g., medical test (positive vs. negative).
- **Numeric Attributes:** A **numeric attribute** is *quantitative*; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be *interval-scaled* or *ratio-scaled*.
  - ***Interval-Scaled Attributes****:* Interval-scaled attributes are measured on a scale of equal-size units.
    - **Example**: A temperature attribute is interval-scaled. Suppose that we have the outdoor temperature value for a number of different days, where each day is an object. For example, a temperature of 20_C is five degrees higher than a temperature of 15_C. Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart.
- **Ratio-Scaled Attributes:**
  - A **ratio-scaled attribute** is a numeric attribute with an inherent zero-point. e.g., *temperature in Kelvin, length, counts.*

1.9. Basic Statistical Descriptions of Data: Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

1.9.1. Measures of central tendency: It means, measure the location of the middle or center of a data distribution. It includes mean, median, mode, and midrange.

- **Mean**: measure of the "center" of a set of data is the *(arithmetic) mean*.
  - Let $x1, x2, : : : , xN$ be a set of N values or *observations*, such as for some numeric attribute X, like *salary*. The **mean** of this set of values is

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

Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58.$$

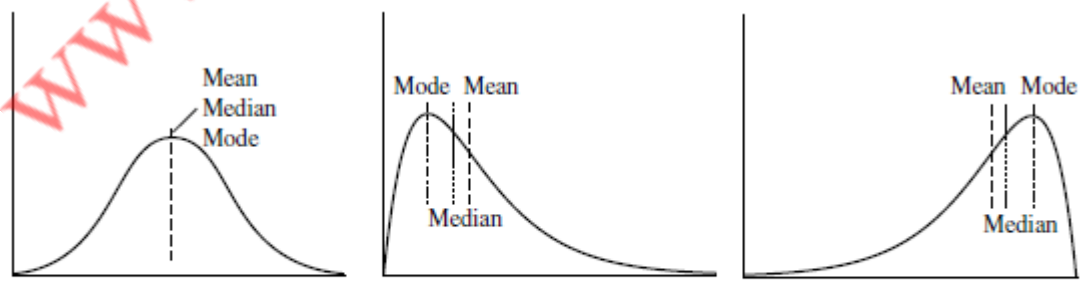
- **Median.** Let's find the median of the data from the above example. The data are already sorted in increasing order. The median can be any value within the two middlemost values of 52 and 56. $\frac{52+56}{2} = \frac{108}{2} = 54.$ Thus, the median is \$54,000.

- **Mode:** In the example, the data are bimodal. The two modes are \$52,000 and \$70,000.

Step 1: The number that occurs most frequently in a data set is called the mode.
Step 2: 30, 36, 47,50, 52, 52, 56, 60, 63, 70, 70, 110.
Step 3: Since the number 52 and 70 appears two times. So, the mode of the data set are 52 and 70.

- **Midrange.** The midrange of the data of above Example is $\frac{30,000+110,000}{2} = \$70,000.$

symmetric data distribution: In a unimodal frequency curve with perfect **symmetric** data distribution, the mean, median, and mode are all at the same center value, as shown in Figure (a).
Data in most real applications are not symmetric. They may instead be either **positively skewed**, where the mode occurs at a value that is smaller than the median (Figure b), or **negatively skewed**, where the mode occurs at a value greater than the median (Figure c).



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

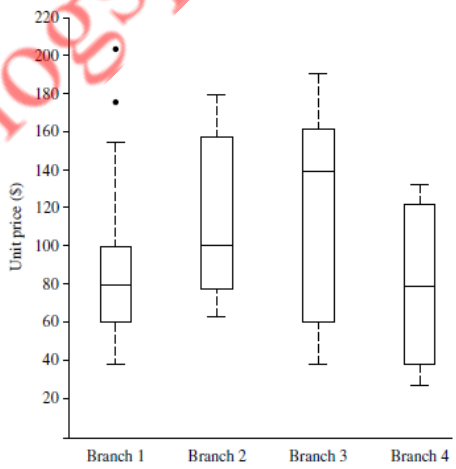
Mean, median, and mode of symmetric versus positively and negatively skewed data.

1.9.2. Measuring the Dispersion of Data: The measures include range, quartiles, quartiles, percentiles, and the inter-quartile range. The five-number summary can be displayed as a boxplot, outliers, variance and standard deviation. also indicate the spread of a data distribution.

- The **range** of the set is the difference between the largest (max()) and smallest (min()) values.
- **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially *equal size* consecutive sets.
- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **inter-quartile range** (**IQR**) and is defined as $IQR = Q3 - Q1$.
    - For example, the quartiles are the three values that split the sorted data set into four equal parts. The data of above example contain 12 values, already sorted in increasing order. Thus, the qartiles for this data are the third, sixth, and ninth values, respectively, in the sorted list. Therefore, $Q1$ is \$47,000 and $Q3$ is \$63,000. Thus, the inter-quartile range is $IQR = 63 - 47 = \$16{,}000$.

1.9.2.1. Boxplot: Boxplot incorporates the five-number summary as follows:

- *Data is represented with a box*
- *The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR*
- *The median is marked by a line within the box*
- *Two lines(whiskers) outside the box extended to Minimum and Maximu.*

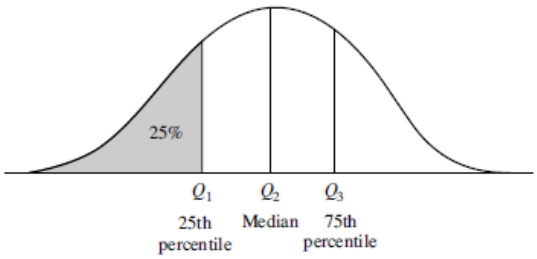


The figure shows Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period. For branch 1, the median price of items sold is \$80, $Q1$ is \$60, and $Q3$ is \$100. Notice that two outlying observations for this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR here of 40.

Note: Boxplots can be computed in $O.n\log n/$ time.

- **Outliers:** points beyond a specified outlier threshold, plotted individually.

- **Percentile:** The three Quartiles are shown as : Q_1 (25th percentile), Q_3 (75th percentile)



1.9.3. Variance and Standard Deviation: Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

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

where x^- is the mean value of the observations, as defined in Eq. mean formula. The **standard deviation**,σ, of the observations is the square root of the variance, σ^2 .
In the example 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. we found $\bar{x} = \$58{,}000$
Using mean value i.e.

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12} = \frac{696}{12} = 58.$$

To determine the variance and standard deviation of the data from that example, we set $N = 12$ and use Eq. **variance** to obtain

$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + 47^2 \ldots + 110^2) - 58^2$$
$$\approx 379.17$$
$$\sigma \approx \sqrt{379.17} \approx 19.47.$$

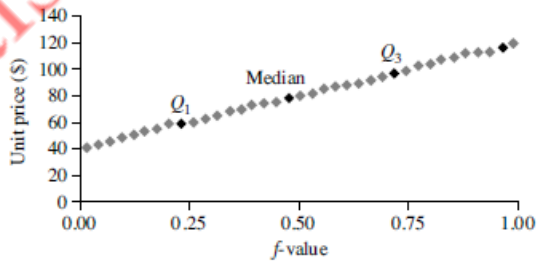
The basic properties of the standard deviation, σ, as a measure of spread are as follows:

- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

1.9.4. Graphic Displays of Basic Statistical Descriptions of Data

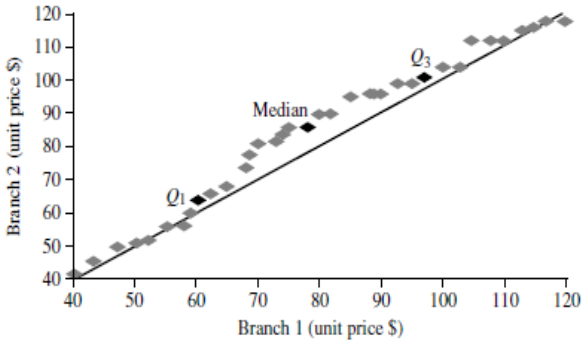
Graphic displays of basic statistical descriptions of data include *quantile plots*, *quantile–quantile plots*, *histograms*, and *scatter plots*. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing.

1. Quantile Plot: A **quantile plot** is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, it plots quantile information. Each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$.



For example, given the quantile plots of sales data for two different time periods, we can compare their *Q*1, median, *Q*3, and other *fi* values at a glance. This is shown in fig.

2. Quantile–Quantile Plot: A**quantile–quantile plot**, or **q-q plot**, graphs 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 fromone distribution to another.



- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.
- A q-q plot for unit price data from two *AllElectronics* branches.

3. Histograms: "Histos" means pole or mast, and "gram" means chart, so a histogram is a chart of poles. Below figure shows a histogram for the data set of Table 2.1, where buckets (or bins) are defined by equal-width ranges representing $20 increments and the frequency is the count of items sold.

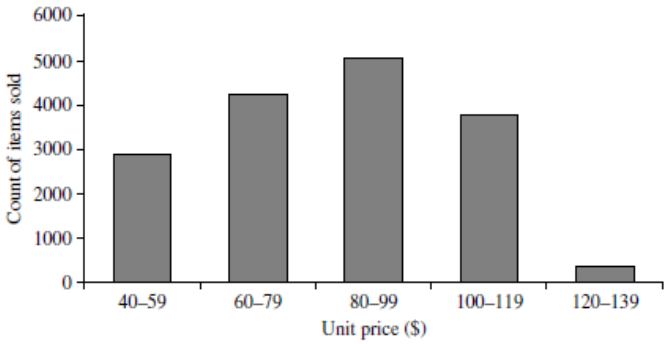
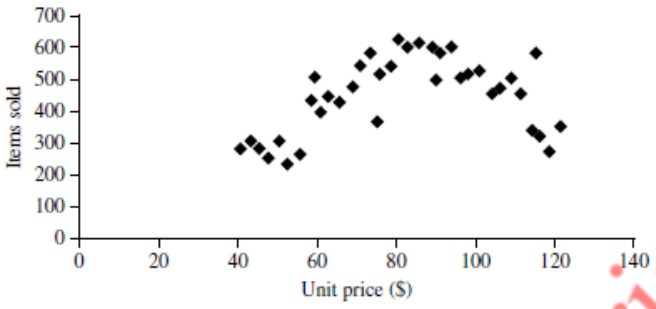


Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

4. Scatter Plots and Data Correlation: A **scatter plot** is one of the most effective graphical methods for determining a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane. Figure shows a scatter plot for the set of data in Table 2.1.



Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

Data Correlation Two attributes, *X*, and *Y*, are **correlated** if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated). Figure (a) shows examples of positive and negative correlations between two attributes. If the plotted point's pattern slopes from lower left to upper right, this means that the values of *X* increase as the values of *Y* increase, suggesting a *positive correlation* (Figure b). If the pattern of plotted points slopes from upper left to lower right, the values of *X* increase as the values of *Y* decrease, suggesting a *negative correlation* (Figure c).

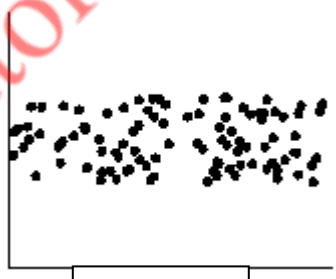


figure (a)

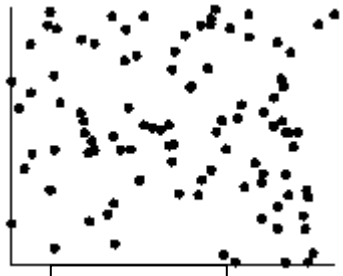


figure (b)

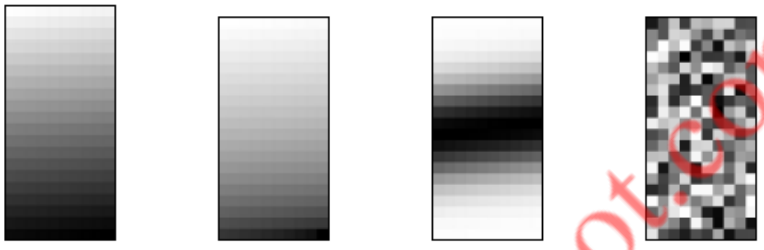
figure (c)

1.10. Data Visualization:

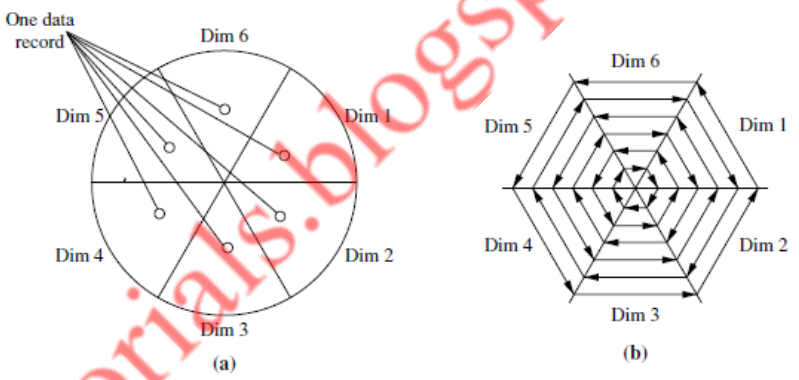
- **Data visualization** aims to communicate data clearly and effectively through graphical representation. Data visualization has been used extensively in many applications—for example, at work for reporting, managing business operations, and tracking progress of tasks.
- The basic concept of data visualization has several representative approaches, including pixel-oriented techniques, geometric projection techniques, icon-based techniques, and hierarchical and graph-based techniques.

1. Pixel-oriented visualization. *AllElectronics* maintains a customer information table, which consists of four dimensions: *income*, *credit limit*, *transaction volume*, and *age*. This shown in figure.

- For a data set of 'm' dimensions, create 'm' windows on the screen, one for each dimension. The 'm' dimension values of a record are mapped to 'm' pixels at the corresponding positions in the windows. The colors of the pixels reflect the corresponding values.

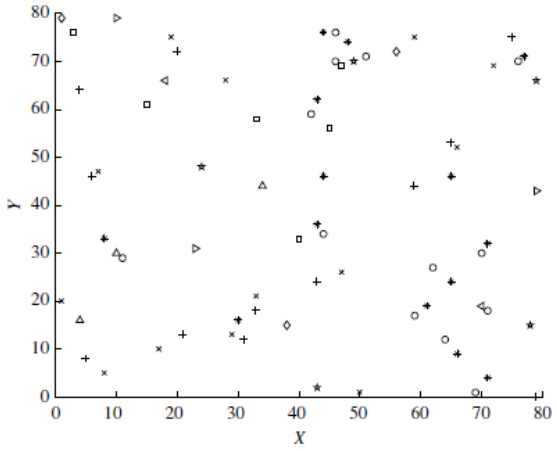


The ***circle segment technique*** uses windows in the shape of segments of a circle, as illustrated in Figure. This technique can ease the comparison of dimensions because the dimension windows are located side by side and form a circle.



2. Geometric Projection Visualization Techniques: Geometric projection techniques help users find interesting projections of multidimensional data sets. The central challenge the geometric projection techniques try to address is how to visualize a high-dimensional space on a 2-D display.

Visualization of a 2-D data set using a scatter plot.



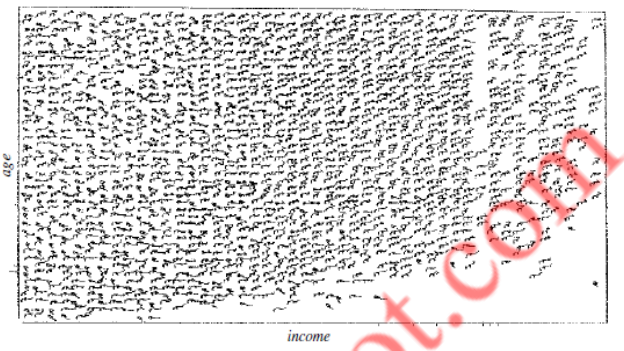
A **scatter plot** displays 2-D data points using Cartesian coordinates. A third dimension can be added using different colors or shapes to represent different data points. Figure 2.13 shows an example, where *X* and *Y* are two spatial attributes and the third dimension is represented by different shapes. Through this visualization, we can see that points of types "+" and "_" tend to be colorcated.

2. Icon-Based Visualization Techniques: It uses small icons to represent multidimensional data values. Two popular icon-based techniques: *Chernoff faces* and *stick figures.*

Chernoff faces were introduced in 1973 by statistician Herman Chernoff. They display multidimensional data of up to 18 variables (or dimensions) as a cartoon human face (shown in figure). Chernoff faces help reveal (make known) trends in the data. Components of the face, such as the *eyes*, *ears*, *mouth*, and *nose*, represent values of the dimensions by their shape, size, placement,

and orientation. For example, dimensions can be mapped to the following facial characteristics: *eye size, eye spacing, nose length, nose width, mouth curvature, mouth width, mouth openness, pupil size, eyebrow slant, eye eccentricity,* and *head eccentricity.*

The **stick figure** visualization technique maps multidimensional data to five-piece stick figures, where each figure has four limbs and a body. Two dimensions are mapped to the display (*x* and *y*) axes and the remaining dimensions are mapped to the angle and/or length of the limbs. **Figure shows** census data, where *age* and *income* are mapped to the display axes, and the remaining dimensions (*gender*, *education*, and so on) are mapped to stick figures.



3. Hierarchical Visualization Techniques: Visualization of the data using a hierarchical partitioning into subspaces.

- Methods
  - **Dimensional Stacking**
  - **Worlds-within-Worlds**
  - **Tree-Map**
  - **Cone Trees**
  - **InfoCube**

1. **Dimensional Stacking:** Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other.
2. **"Worlds-within-Worlds,"** also known as *n*-Vision, is a representative hierarchical visualization method.
3. **Tree-maps** display hierarchical data as a set of nested rectangles.
4. **Cone Trees:** *3D cone tree* visualization technique works well for up to a thousand nodes or so First build a *2D circle tree* that arranges its nodes in concentric circles centered on the root node.
5. **InfoCube:** A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes. The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on.

1.11. Measuring Data Similarity and Dissimilarity: A **cluster** is a collection of data objects such that the objects within a cluster are *similar* to one another and *dissimilar* to the objects in other clusters. The term *proximity* is used to refer to either similarity or dissimilarity.

Data Matrix: In the central tendency, dispersion and variance & Standard deviation, the data is handled through a single attribute (i.e. one dimensional array). In *Data Matrix,* attribute handle multiple data (i.e. multi dimensional array).

For example: Suppose that we have *n* objects (e.g., persons, items, or courses) described by *p* attributes (also called *measurements* or *features*, such as age, height, weight, or gender).

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}.$$

The objects are $x1 = .x_{11}, x_{12}, \ldots, x_{1p})$, $x2 = .x_{21}, x_{22}, \ldots, x_{2p})$ and so on, where xij is the value for object xi of the jth attribute.

- The Data Matrix (or *object-by-attribute structure*) structure stores the n data objects in the form of a relational table, or n-by-p matrix (n objects _p attributes):

Dissimilarity matrix: (or *object-by-object structure*): This structure stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table:

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix}$$

where $d.(i, j)$ is the measured **dissimilarity** or "difference" between objects i and j. In general, $d.(i, j)$ is a non-negative number that is close to 0 when objects i and j are highly similar or "near" each other, and becomes larger the more they differ. Note that $d(.i, i) = 0$; that is, the difference between an object and itself is 0.

Measures of similarity can often be expressed as a function of measures of dissimilarity. For example, for nominal data, $sim(.i, j) = 1 - d(.i, j)$, where $sim(.i, j)$ is the similarity between objects i and j.

"How is dissimilarity computed between objects described by nominal attributes?"

: The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i, j) = \frac{p - m}{p},$$

where m is the number of *matches* (i.e., the number of attributes for which i and j are in the same state), and p is the total number of attributes describing the objects. Weights can be assigned to increase the effect of m or to assign greater weight to the matches in attributes having a larger number of states.

Dissimilarity between nominal attributes. Suppose that we have the sample data of **Table,** except that only the *object-identifier* and the attribute *test-1* are available, where *test-1* is nominal. (We will use *test-2* and *test-3* in later examples.) Let's compute the dissimilarity matrix (Equation), that is, Since here we have one nominal attribute, *test-1*, we set $p = 1$ in Eq. So that $d(i, j)$ evaluates to 0 if objects i and j match, and 1 if the objects differ. Thus, we get

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

From this, we see that all objects are dissimilar except objects 1 and 4 (i.e., $d(.4,1) = 0$).

Dissimilarity between binary attributes: Suppose that a patient record table (Table 2.4) contains the attributes *name, gender, fever, cough, test-1, test-2, test-3,* and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary.

Relational Table Where Patients Are Described by Binary Attributes

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

For asymmetric attribute values, let the values *Y* (*yes*) and *P* (*positive*) be set to 1, and the value *N* (*no* or *negative*) be set to 0. Suppose that the distance between objects is computed based

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67,$$

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33,$$

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75.$$

only on the asymmetric attributes. According to Eq. (2.14), the distance between each pair of the three patients—Jack, Mary, and Jim—is