Data Warehousing & Data Mining

Unit I

B.Tech (CSE) V Semester
2023-27 Batch

Vision and Mission of the Department

Vision

To evolve as a centre of **academic and research excellence** in the area of Computer Science and Engineering.

Mission

The Computer Science & Engineering Department is committed to

1. Utilize **innovative learning methods** for academic improvement.

2. Encourage **higher studies and research** to meet the futuristic requirements of Computer Science and Engineering.

3. Inculcate **ethics and human values** for developing students with good character.

Syllabus

- **Course Outcomes: After Successful completion of the Course, the student will be able to:**

CO1: Discuss the basic concepts and techniques of data warehousing and data mining. **(K2)**

CO2: Demonstrate the types of the data to be mined and apply pre-processing methods on raw data. **(K3)**

CO3: Illustrate various Classification Techniques. **(K3)**

CO4: Discover interesting patterns, analyze supervised and unsupervised models and estimate the accuracy of the algorithms. **(K3)**

CO5: Use different clustering techniques to cluster data. **(K3)**

Syllabus

- **UNIT I: Data Warehousing and Online Analytical Processing:** Basic concepts, Data Warehouse Modeling: Data Cube and OLAP, Data Warehouse Design and Usage, Data Warehouse Implementation, Cloud Data Warehouse, Data Mining and Patten Mining, Technologies, Applications, Major issues, Data Objects & Attribute Types, Basic Statistical Descriptions of Data, Data Visualization, Measuring Data Similarity and Dissimilarity.

- **UNIT II: Data Preprocessing**: An Overview, Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization.

- **UNIT III: Classification:** Basic Concepts, General Approach to solving a classification problem, Decision Tree Induction: Attribute Selection Measures, Tree Pruning, Scalability and Decision Tree Induction, Visual Mining for Decision Tree Induction, Bayesian Classification Methods: Bayes Theorem, Naïve Bayes Classification, Rule-Based Classification, Model Evaluation and Selection.

Syllabus

- **UNIT IV: Association Analysis:** Problem Definition, Frequent Item set Generation, Rule Generation: Confident Based Pruning, Rule Generation in Apriori Algorithm, **Improving the efficiency of Apriori**, Compact Representation of frequent item sets, FP-Growth Algorithm.
-
- **UNIT V: Cluster Analysis:** Overview, Basics and Importance of Cluster Analysis, Clustering techniques, Different Types of Clusters; K-means: The Basic K-means Algorithm, K-means Additional Issues, Bi-secting K Means, Agglomerative Hierarchical Clustering: Basic Agglomerative Hierarchical Clustering Algorithm DBSCAN: Traditional Density Center-Based Approach, DBSCAN Algorithm, Strengths and Weaknesses.

Text Books:

1. Data Mining concepts and Techniques, 3rd edition, Jiawei Han, Michel Kamber, Elsevier, 2011

Third Edition
DATA MINING
Concepts and Techniques
Jiawei Han | Micheline Kamber | Jian Pei

Course Outcomes

After Successful completion of the Course, the student will be able to:

CO1: Discuss the basic concepts and techniques of data warehousing and data mining. (K2)

CO2: Demonstrate the types of the data to be mined and apply pre-processing methods on raw data. **(K3)**

CO3: Illustrate various Classification Techniques. **(K3)**

CO4: Discover interesting patterns, analyze supervised and unsupervised models and estimate the accuracy of the algorithms. **(K3)**

CO5: Use different clustering techniques to cluster data. **(K3)**

CO1: Topics to be covered: Data Warehousing and Online Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- Cloud Data Warehouse
- Data Mining and Patten Mining
- Technologies used
- Kinds of Applications targeted
- Major issues in data mining
- Data Objects
- Attribute types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity

Knowledge
Personalized information
Information
Processed data
Data
Raw numbers and facts

Data, Information, and Knowledge

- **<u>Data</u>:** Unorganized and unprocessed facts; static; a set of discrete facts about events
- **<u>Information</u>:** Aggregation of data that makes decision making easier
- **<u>Knowledge</u>** is derived from information in the same way information is derived from data; it is a person's range of information

Introduction: Need for data mining

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, web
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: bioinformatics, scientific simulation, medical research …
    - Society and everyone: news, digital cameras, …
- Data rich but information poor!
  - What does those data mean?
  - How to analyze data?
- Data mining — Automated analysis of massive data sets

Knowledge Discovery from Data: Data Mining

- **Data mining (knowledge discovery from data)**
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Knowledge Discovery from Data: Data Mining[contd.]

1. **Data cleaning** - remove noise and inconsistent data
2. **Data integration** - multiple data sources may be combined
3. **Data selection** - data relevant to the analysis task are retrieved from database
4. **Data transformation** - data transformed or consolidated into forms appropriate for mining
5. **Data mining** - an essential process where intelligent methods are applied to extract data patterns
6. **Pattern evaluation** -identify the truly interesting patterns
7. **Knowledge presentation** - mined knowledge is presented to the user with visualization or representation techniques

Knowledge Discovery from Data: Data Mining [contd.]

Kinds of Data Mined

Data Mining Should be applicable to any kind of information repository. This includes:

1. Database Data
2. Data Warehouses
3. Transactional Databases
4. Others kinds of Data
   - Time-related (temporal) or sequence data
   - Data Streams
   - Spatial Data
   - Engineering design data
   - Hypertext and multimedia data
   - Graph and networked data
   - The World-Wide Web

Kinds of Data Mined

1. Database Data

- **DBMS** – database management system, contains a collection of interrelated databases
  e.g. Faculty database, student database, publications database

- Each database contains a collection of tables and functions to manage and access the data.
  e.g. student_bio, student_graduation, student_parking

- Each table contains columns and rows, with columns as attributes of data and rows as records.

- Tables can be used to represent the relationships between or among multiple tables.

1. Database Data : *A Relational Database for AllElectronics store*

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
. . .	. . .	. . .	. . .	. . .	. . .

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
. . .	. . .	. . .

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

works_at

empl_ID	*branch_ID*
E55	B1
. . .	. . .

1. Database Data [contd.]

- With a relational query language, e.g. SQL, we will be able to find answers to questions such as:
  - How many items were sold last year?
  - Who has earned commissions higher than 10%?
  - What is the total sales of last month for Dell laptops?
- When data mining is applied to relational databases, we can search for trends or data patterns.
- Relational databases are one of the most commonly available and rich information repositories, and thus are a major data form in our study.

2. Data Warehouses

- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.

2. Data Warehouses [contd.]

- Data are organized around major subjects, e.g. customer, item, supplier and activity.
- Provide information from a historical perspective (e.g. from the past 5 – 10 years)
- Typically summarized to a higher level (e.g. a summary of the transactions per item type for each store)
- User can perform drill-down or roll-up operation to view the data at different degrees of summarization

2. Data Warehouses [contd.]

3. Transactional Databases

- Consists of a file where each record represents a transaction
- A transaction typically includes a unique transaction ID and a list of the items making up the transaction.

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
. . .	. . .

- Either stored in a flat file or unfolded into relational tables
- Easy to identify items that are frequently sold together

Others kinds of Data

- **Time-related (temporal) or sequence data :** *eg. Historical records, stock exchange data, time-series and biological sequence data*
- **Data Streams :** *eg. video surveillance and sensor data*
- **Spatial Data :** *eg. maps*
- **Engineering design data :** *eg. the design of buildings, system components, or integrated circuits*
- **Hypertext and multimedia data :** *eg. including text, image, video, and audio data*
- **Graph and networked data :** *eg. social and information networks*
- **The World-Wide Web :** *eg. characterize and classify web pages and uncover web dynamics*

Kinds of Patterns Mined: Data mining functionalities

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:

1. Descriptive
2. Predictive

1. **Descriptive**: Characterize the general properties of the data in the Database.
2. **Predictive:** Perform inference on the current data in order to make predictions.

Data Mining functionalities, and the kinds patterns they can discover are described below:

1. Concept / Class Description: Characterization and Discrimination
2. Mining Frequent Patterns, Associations and Correlations
3. Classification and Regression for Predictive Analysis
4. Cluster Analysis
5. Outlier Analysis

Kinds of Patterns Mined: Data mining functionalities[contd.]

1. Concept / Class Description: Characterization and Discrimination

- Data can be associated with classes or concepts.
  - E.g. classes of items – computers, printers, …

    concepts of customers – bigSpenders, budgetSpenders, …
  - How to describe these items or concepts?
- Descriptions can be derived via
  - **Data characterization** – summarizing the general characteristics of a target class of data.
    - E.g. summarizing the characteristics of customers who spend more than $1,000 a year at *AllElectronics*. Result can be a general profile of the customers, such as 40 – 50 years old, employed, have excellent credit ratings.

Kinds of Patterns Mined: Data mining functionalities[contd.]

- **Data discrimination** – comparing the target class with one or a set of comparative classes
  - E.g. Compare the general features of software products whole sales increase by 10% in the last year with those whose sales decrease by 30% during the same period
- **Or both of the above**

2. Mining Frequent Patterns, Associations and Correlations

- **Frequent Patterns**: (a set of items, subsequences, substructures, etc.) that occurs frequently in a transactional data set (e.g. milk and bread)

Kinds of Patterns Mined: Data mining functionalities[contd.]

- **Association Analysis**: Searches for interesting relationships among items in a given dataset.
  - E.g. a sample analysis result – an association rule:

    buys(X, “computer”) => buys(X, “software”) [support = 1%, confidence = 50%]

    (if a customer buys a computer, there is a 50% chance that she will buy software. 1% of all of the transactions under analysis showed that computer and software are purchased together.)
  - Associations rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.

- **Correlation Analysis**: additional analysis to find statistical correlations between associated pairs

Kinds of Patterns Mined: Data mining functionalities[contd.]

3. Classification and Prediction

- **Classification** (Supervised Learning)
  - The process of finding a model that describes and distinguishes **the 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 may be represented in various forms, such as,
    - Classification (IF-THEN) rules
    - Decision Tree
    - neural networks, etc.

Kinds of Patterns Mined: Data mining functionalities[contd.]

3. Classification and Regression

(a)

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")

(b)

(c)

A classification model can be represented in various forms:
(a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Kinds of Patterns Mined: Data mining functionalities[contd.]

3. Classification and Prediction

- **Prediction**
  - Used to estimate the future data based on past and current data.
  - The future values may be predicted by using **regression (or) Time-Series Analysis.**

Kinds of Patterns Mined: Data mining **functionalities**[contd.]

4. Cluster Analysis (Unsupervised Learning)

- Class label is unknown: group data to form new classes
- It is a collection of data objects that are similar to one another with in the same cluster and dissimilar objects are in another cluster.
- The objects are clusterd or grouped based on the principle of
  - *maximizing intra-class similarity &*
  - *minimizing interclass similarity*

Kinds of Patterns Mined: Data mining **functionalities**[contd.]

4. Cluster Analysis (Unsupervised Learning)

A 2-D Plot of customer data w.r.to customer locations in a city showing 3 data clusters. Each cluster “center” is marked with a “+”.

Kinds of Patterns Mined: Data mining functionalities[contd.]

5. Outlier Analysis

- Data that do no comply with the general behavior or model.
- Outliers are usually discarded as noise or exceptions.
- Useful for fraud detection.
  - E.g. Detect purchases of extremely large amounts

Technologies used

Data mining adopts techniques from many domains

Technologies used [contd.]

1. Statistics

- **Statistics** studies the collection, analysis, interpretation or explanation, and presentation of data.
- Data mining has an inherent connection with 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.

Technologies used [contd.]

2. Machine Learning

- Machine learning investigates how computers can learn (or improve their performance) based on data
- **For example,** a typical machine learning problem is to program a computer so that it can automatically recognize handwritten postal codes on mail after learning from a set of examples
- Classic problems in machine learning that are highly related to data mining:
  - **Supervised learning** is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set

Technologies used [contd.]

2. Machine Learning

- **Unsupervised learning** is essentially a synonym for clustering
- **Semi-supervised learning** is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model
- **Active learning** is a machine learning approach that lets users play an active role in the learning process

Technologies used [contd.]

3. Database Systems and Data Warehouses

- Database systems research focuses on the creation, maintenance, and use of databases
- A data warehouse integrates data originating from multiple sources and various timeframes. It consolidates data in multidimensional space to form partially materialized data cubes. The data cube model not only facilitates Online Analytical Processing (OLAP) in multidimensional databases but also promotes multidimensional data mining

Technologies used [contd.]

4. Information Retrieval

- Information retrieval (IR) is the science of searching for documents or information in documents.
- Documents can be text or multimedia, and may reside on the Web.
- Information retrieval assumes that:

  (1) the data under search are unstructured; and

  (2) the queries are formed mainly by keywords, which do not have complex structures

Kinds of Applications targeted

1. **Business Intelligence(BI)**

- 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 BI. Online analytical processing tools in business intelligence rely on data warehousing and multidimensional data mining.

- Classification and prediction techniques are the core of predictive analytics in BI, for which there are many applications in analyzing markets, supplies, and sales.

Kinds of Applications targeted [contd.]

2. Web Search Engines

- 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.
- Some search engines also search and return data available in public databases or open directories.
- Web search engines are essentially very large data mining applications. Search engines pose grand challenges to data mining.

1) They have to handle a huge and ever-growing amount of data.

2) Web search engines often have to deal with online data.

3) Web search engines often have to deal with queries that are asked only a very small number of times.

Major Issues in Data Mining

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Ad hoc data mining and data mining query languages
  - Presentation and visualization of data mining results

Major Issues in Data Mining [contd.]

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods
- Diversity of data 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

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.

Examples:

  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attribute Types

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- **Types:**
  - Nominal
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

1. **Nominal Attributes :** The values of a Nominal attribute are names of things, some kind of symbols. Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as **categorical attributes** and there is no order (rank, position) among values of the nominal attribute
   - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
   - marital status, occupation, ID numbers, zip codes

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

2. **Binary Attributes:** Binary data has only 2 values/states. **For Example** yes or no, affected or unaffected, true or false.
   - •**Symmetric:** Both values are equally important (Gender).
   - •**Asymmetric:** Both values are not equally important (Result).

Attribute	Values
Gender	Male , Female

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

Attribute Types [contd.]

3. Ordinal Attributes

- Values have a meaningful order (ranking) but magnitude between successive values is not known.
- *Size* = {*small, medium, large*}, grades, army rankings

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Attribute Types [contd.]

Quantitative Attributes:

1. **Numeric:** A **numeric attribute** is q*uantitative*; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be *interval-scaled* or *ratio-scaled*.

- **interval-scaled**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio -scaled**
  - is a numeric attribute with a fix **zero-poin**t .
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Attribute Types [contd.]

2. Discrete:

- Has only a finite or countably infinite set of values.
  - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- **Note:** Binary attributes are a special case of discrete attributes

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

3. Continuous : Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 .....etc

Basic Statistical Descriptions of Data

- It can be used to identify the properties of the data and highlight which data values should be treated as noise or outliers.
- The main purpose of this is to analyze/interpret /process large amount of data

  Ex: Population of India.
- **Three areas of Basic Statistical Descriptions are:**
  1. **Measuring the central Tendency:** Mean, Median and Mode
  2. **Measuring the Dispersion of Data:**
     1. Range
     2. Quartile
     3. Variance
     4. Standard Deviation
     5. Interquartile Range
  3. **Graphic Displays of Basic Statistical Descriptions of Data:**
     1. Quantile plot
     2. Quantile – Quantile Plot
     3. Histogram
     4. Scatter Plots and Data Correlation

Basic Statistical Descriptions of Data

Measuring the Central Tendency: Mean, Median and Mode

- Central Tendency refers to the “Middle” or “Central position” Value.
- **Mean, Median and Mode** are all valid measures of Central Tendency.
- **Mean:** It can be used with discrete and continuous data.

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

Where *N* is number of observations

Mean. 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.

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

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is $58,000. ■

Basic Statistical Descriptions of Data

Measuring the Central Tendency: Mean, Median and Mode

- **Weighted arithmetic mean:**

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

Example: Mean for Discrete Series

Marks	No. of Students
10	5
20	10
30	20
40	30
50	40

=50+200+600+1200+2000/105
= 38.57

- **Trimmed mean: chopping extreme values**

Basic Statistical Descriptions of Data

Measuring the Central Tendency

- **Median**:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$$

  - where L_1 is lower boundary of median interval, N is the number of values in the entire data set, $(\sum freq)_l$ is the sum of frequencies of all the intervals that are lower than the median interval, $freq_{median}$ is the frequency of median interval and width is the width of median interval

Basic Statistical Descriptions of Data [contd.]

Measuring the Central Tendency

- **Mode:**
  - Value that occurs most frequently in the data.
  - Data sets with one, two, or three modes are respectively called **Unimodal, bimodal, trimodal** .
  - Data set with two or more modes is **multimodal .**
  - For unimodal numeric data that are moderately skewed (asymmetrical), we have the following **empirical relation**:

$$mean - mode = 3 \times (mean - median)$$

- **Midrange:** used to assess the central tendency of a numeric data set.

$$Midrange = \frac{\min + \max}{2}$$

Basic Statistical Descriptions of Data [contd.]

Measuring the Central Tendency

- **Symmetric vs. Skewed Data:**
  - Data in most real applications are not symmetric.
  - Median, mean and mode of symmetric, positively and negatively skewed data

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data: Range, Quartiles and Interquartile Range

- How the data scattered / spread, this is not clear with central tendency.
- Measures to asses the dispersion (or) spread of numeric data. The Measures include range, quantiles, quartiles, percentiles.
- The five – number summary, which can be displayed as a boxplot, is useful in identifying outliers.
- Variance and Standard Deviation indicate the spread of a data distribution.

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data: Range, Quartiles and Interquartile Range

- Let $x_1, x_2, : : : , x_N$ be a set of observations for some numeric attribute, X.

- **Range**: The range of the set is the difference between the largest (max()) and smallest (min()) values.
  - **Range(R)** = H – L
  - **Coefficient of R** = H – L / H + L

Example:

x= 3 4 8 7 6 9 4

R= 9 – 3 = 6

Coe of R =9 - 3/ 9 + 3 = 6/12=0.5

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.
- **Quartiles**: The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles.

  For ex. Q_1 (25th percentile), Q_2 (50th percentile) Q_3 (75th percentile)
- The 100-quantiles are more commonly referred to as **percentiles**
- **Inter-quartile range**: IQR = $Q_3 - Q_1$

A plot of the data distribution for some attribute X.

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Formula For Quartile:**

Let's say that we have a data set with N data points:

X – {X1, X2, X3…….. XN}

The formula for quartiles is given by:

Lower Quartile (Q1) = (N+1) * 1 / 4

Middle Quartile (Q2) = (N+1) * 2 / 4

Upper Quartile (Q3)= (N+1) * 3 / 4

Interquartile Range = Q3 – Q1

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Example 1:** 2,3,1,4,6,8,9,10,12,3,4

Step 1: Put the numbers in order: 1, 2, 3, 3, 4, 4, 6, 8, 9, 10, 12.

Step 2:

Q1= (n+1)*1/4 = (11+1)*0.25 = 12/4=3=@3rd Position

Q2= (n+1)*1/2 = (11+1)*0.5 = 6=@6th Position

Q3= (n+1)*3/4 = (11+1)*0.75 = 3/4 *12=9=@9th Position

IQR=Q3-Q1=9-3 = 6

Example 2: 1 2 3 3 4 5 6 6 7 8 8 9

Q1= (n+1)*1/4 = (12+1)*0.25 = 13/4=3.25= average of 3+3/2=3

Q2= (n+1)*1/2 = (12+1)*0.5 = 13/2=6.5 = 5+6 /2 =5.5

Q3= (n+1)*3/4 = (12+1)*0.75 = 3/4 *13=9.75 = 15/2 = 7.5

IQR=Q3-Q1=7.5-3=4.5

NOTE: IQR gives the Central Half of the Data

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data: Five number summary, Boxplot and Outliers

Five number summary: It Contains

1. Minimum Value,
2. Q1 (First Quartile)
3. Q2 (Median)
4. Q3 (Third Quartile)
5. Maximum Value

These can be displayed graphically as boxplot (also called a box and whisker plot)

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data: Five number summary, Boxplot and Outliers

- **Boxplot**:
- It gives the Visual representation of the five-number summary.
- A boxplot incorporates the five-number summary as follows:
  - Data is represented with a box
  - The median is marked by a line within the box.
  - The horizontal line that extends out from the box is called **whiskers** and the actual box itself is called **Interquartile Range**.
  - Two lines (Called whiskers) outside the box extended to the smallest (Minimum) and largest (maximum) observation.

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data: Five number summary, Boxplot and Outliers

- **Outlier**: points beyond a specified outlier threshold, plotted individually usually, a value higher/lower than 1.5 x IQR

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Boxplot example**

Boxplot for the unit price data for items sold at four branches of AllElectronics during a given time period

Basic Statistical Descriptions of Data [contd.]

Measuring the Dispersion of Data

- **Variance and standard deviation :** Variance and standard deviation 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, x1,x2,...,xN , for a numeric attribute X is

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

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

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

Quantile plot: It is a simple and effective way to have a first look at a univariate data distribution.

- It displays all of the data for the given attribute
- It plots quantile information
- Each observation x_i is paired with a percentage f_i , indicating that approximately f_i *100 % of data are below the value x_i

A Quantile plot for the unit price data

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one 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.

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

- **Histograms (or frequency histograms):** Graph display of tabulated frequencies, shown as bars.
- "**Histos"** means pole(or) mast and **"gram"** means chart, so a histogram is a chart of poles.
- It summarizes the distribution of a given attribute
- The range of values for *X (numeric)* is partitioned into disjoint consecutive subranges.
- The subranges, referred to as *buckets* or *bins*, are disjoint subsets of the data distribution for *X*. The range of a bucket is known as the **width**.

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

- **Scatter plots :** This is used to show a relationship 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 .

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

Data Correlation : Describe the type of relationship between the two sets of data.

1. Positive Correlation: If the pattern of plotted points 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.

2. Negative Correlation : 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 .

3. No Correlation : No relation between the two attributes (that is X and Y).

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

Data Correlation

Positively correlated data

Negative correlated data

Basic Statistical Descriptions of Data [contd.]

Graphic Displays of Basic Statistical Descriptions

Data Correlation

Uncorrelated Data

Measuring Similarity and Dissimilarity

- Most important - used in Clustering, some Classification, anomaly detection.
- **Similarity**
  - Numerical measure of how alike two data objects are
  - The higher the similarity value, the greater the similarity between objects.
  - Often falls in the range [0,1] *(**Returns:**0 if they are unalike, 1 , if they are alike)*
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different are two data objects
  - The higher the dissimilarity value, the greater the dissimilarity between objects.
  - Minimum dissimilarity is often 0 *(**Returns**:0 if they are alike, 1 , if they are unlike)*
- **Proximity** means measuring the similarity or dissimilarity between the attributes.

Measuring Similarity and Dissimilarity [contd.]

Data matrix Vs. Dissimilarity matrix

•Two data structures that are commonly used in the above types of applications: the ***data matrix*** (used to store the data objects) and the ***dissimilarity matrix*** (used to store dissimilarity values for pairs of objects).

•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).

•The objects are $x1 = (x11,x12,...,x1p)$,
$x2 = (x21,x22,...,x2p)$, and so on,

where xij is the value for object xi of the jth attribute

Data matrix Vs. Dissimilarity matrix

- **Data matrix (*or* *object-by-attribute structure*):** This structure stores the n data objects in the form of a relational table, or *n-by-p matrix* (n objects ×p attributes).
  - Each row corresponds to an object. As part of our notation, we may use f to index through the p attributes.

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

Measuring Similarity and Dissimilarity [contd.]

Data matrix Vs. Dissimilarity matrix

- 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:

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

where d(i, j) is the measured dissimilarity or “difference” between objects i and j.

Note: d(i, i) = 0; that is, the difference between an object and itself is 0.

Measuring Similarity and Dissimilarity [contd.]

- **Measure of similarity** can be expressed as a function of measures of dissimilarity

  For example, for nominal data, ***sim(i, j)= 1-d(i, j)***

- A data matrix is made up of two entities or “things,” namely rows (for objects) and columns (for attributes). Therefore, the data matrix is often called a **two-mode** matrix .

- The dissimilarity matrix contains one kind of entity (dissimilarities) and so is called a **one-mode** matrix.

- Many clustering and nearest-neighbor algorithms operate on a dissimilarity matrix. Data in the form of a data matrix can be transformed into a dissimilarity matrix before applying such algorithms.

Measuring Similarity and Dissimilarity [contd.]

Proximity Measure for Nominal Attributes

- A **nominal attribute** can take on two or more states. ***For example***, *map_color* is a nominal attribute that may have, say, five states: *red, yellow, green, pink*,and *blue .*
- Let the number of states of a nominal attribute be *M*. The states can be denoted by letters, symbols, or a set of integers, such as 1, 2, … , *M*

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

where,

- ***m* is the number of *matches*** (i.e., the number of attributes for which i and j are in the same state)
- ***p is the total number of attributes*** describing the objects

Measuring Similarity and Dissimilarity [contd.]

Proximity Measure for Nominal Attributes

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

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

- Here p=1, d(i, j) evaluates to 0 if objects i and j match, and 1 if the objects differ :

Now the dissimilarity matrix is:

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

Measuring Similarity and Dissimilarity [contd.]

Proximity Measure for Nominal Attributes

Similarly,

$$sim(i,j) = 1 - d(i,j)$$
$$= 1 - m / p$$

$$sim(i, j) = \begin{bmatrix} 0 & & & \\ 1-1 & 0 & & \\ 1-1 & 1-1 & 0 & \\ 1-0 & 1-1 & 1-1 & 0 \end{bmatrix}$$

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

Measuring Similarity and Dissimilarity [contd.]

Proximity Measures for Binary Attributes

- ***Binary Attribute has only one of two states***: 0 and 1, where 0 means that the attribute is absent, and 1 means that it is present .

- Dissimilarity and Similarity measures for objects described by either ***symmetric*** or ***asymmetric*** binary attributes.

- Given the attribute smoker describing a patient, for instance, 1 indicates that the patient smokes, while 0 indicates that the patient does not.

Measuring Similarity and Dissimilarity [contd.]

Proximity Measures for Binary Attributes

- Let i, j are two objects and ‘p’ total number of attributes :

$$p = q + r + s + t$$

where,

q – 1 for both objects
r – 1 for i, 0 for j
s - 0 for i, 1 for j
t – 0 for both i and j

Contingency Table for Binary Attributes

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Measuring Similarity and Dissimilarity [contd.]

Proximity Measures for Binary Attributes

- **Dissimilarity for symmetric binary variables:**

$$d(i, j) = \frac{r + s}{q + r + s + t}.$$

- **Dissimilarity for asymmetric binary variables:**

$$d(i, j) = \frac{r + s}{q + r + s}.$$

- **Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):**

$$sim(i, j) = \frac{q}{q + r + s} = 1 - d(i, j).$$

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity between Binary Variables

- **Example :** 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

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

  - Gender is a symmetric attribute
  - The remaining attributes are asymmetric binary
  - Let the values Y and P be 1, and the value N be 0

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

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

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

Of the three patients jack and Mary are the most likely to have a similar disease.

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity of Numeric Data : Minkowski Distance

- The most popular distance measure is **Euclidean distance**

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+\ldots+|x_{i_p}-x_{j_p}|^2)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ be two objects described by p numeric attributes

- Another well-known measure is the **Manhattan (or city block) distance**

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+\ldots+|x_{i_p}-x_{j_p}|$$

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity of Numeric Data : Minkowski Distance

- Both the Euclidean and the Manhattan distance satisfy the following mathematical properties:
- **Non-negativity**: $d(i, j) \geq 0$ if $i \neq j$, Distance is a non-negative numberd
- **Identity of indiscernibles:** $d(i, i) = 0$, distance of an object to itself is 0
- **Symmetry**: $d(i, j) = d(j, i)$, Distance is a symmetric function
- **Triangle Inequality**: $d(i, j) \leq d(i, k) + d(k, j)$, Going directly from object i to object j in space is no more than making a detour over any other object k.
- A measure that satisfies these conditions is known as a metric

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity of Numeric Data : Minkowski Distance

- **Minkowski distance** is a generalization of the Euclidean and Manhattan distances

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects,

h is a real number such that $h \geq 1$ (the distance so defined is also called L-h norm)

Measuring Similarity and Dissimilarity [contd.]

Special Cases of Minkowski Distance

- h = 1: Manhattan (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

- h = 2: (L_2 norm) Euclidean distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

Measuring Similarity and Dissimilarity [contd.]

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Measuring Similarity and Dissimilarity [contd.]

Proximity Measures for Ordinal Variables

- The values of an ordinal attribute have a meaningful order or ranking about them,yet the magnitude between successive values is unknown
- An example includes the sequence *small*, *medium*, *large* for a *size* attribute
- Let M represent the number of possible states that an ordinal attribute can have. These ordered states define the ranking 1, … , M_f .
- Dissimilarity can be calculated in following steps:

1. The value of f for the ith object is x_{if} , and f has M_f ordered states, representing the ranking 1, …, M_f, Replace each x_{if} by its corresponding rank

$$r_{if} \in \{1,\ldots,M_f\}$$

2. Map the range of each attribute onto [0.0, 1.0] so that each attribute has equal weight, using

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

3. Dissimilarity can then be computed using any of the distance measures

Measuring Similarity and Dissimilarity [contd.]

Dissimilarity for Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing, or (2) $x_{if} = x_{jf} = 0$ and attribute *f* is asymmetric binary;

otherwise, $\delta_{ij}^{(f)} = 1$

- *If f* is numeric, $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$ where h runs over all nonmissing objects for attribute f

- *If f* is nominal or binary:

  $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise

- *If* is ordinal

  Compute ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$ and treat z_{if} as numeric

Measuring Similarity and Dissimilarity [contd.]

Cosine Similarity

- It is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words.
- A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as a keyword) or phrase in the document
- **Cosine measure**: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

where $\bullet$ indicates vector dot product, $||d||$: the length of vector d

Measuring Similarity and Dissimilarity [contd.]

Example: Cosine Similarity

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Ex: Find the **similarity** between documents 1 and 2.

d_1 **=** (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

d_2 **=** (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2$ = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25

$||d_1||$= $(5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5}$ = 6.481

$||d_2||$= $(3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5}$ = 4.12

$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$ = 25/(6.481*4.12) = 0.94

Reference:

- Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei,3rd Edition, Morgan Kaufmann Publishers