REVA UNIVERSITY
Bengaluru, India

School of Computing and Information Technology

Course Delivery

B.Tech – 6th Semester CSE
(BTCS15F6410)
Data Mining Techniques

By
Prof Shruthi G
Asst. Professor
shruthig@reva.edu.in

Contents

REVA UNIVERSITY
Bengaluru, India

- Course Objectives
- Course Outcomes
- Text Books/Reference Books
- Syllabus Content Unit-Wise

Course Objectives

- Introduce the basics of data mining, data types, similarity and dissimilarity measures
- Demonstrate association rules and algorithms
- Describe supervised learning algorithms for data categorization
- Illustrate unsupervised learning algorithms for grouping data sets
- Demonstrate the appropriate data mining techniques for decision making

Course Outcomes

REVA UNIVERSITY
Bengaluru, India

- Explain the basics of data mining techniques, data types, identify the similarity and dissimilarity between the data sets.
- Analyze the data sets using the association rules and algorithms
- Characterize and discriminate data sets with classification methods
- Employ the clustering methods in real life problems
- Apply the knowledge for data mining applications

Text books/ Reference Books

REVA UNIVERSITY
Bengaluru, India

Text books:

1. A Pang-Ning Tan, Michael Steinbach and Vipin Kumar, **"Introduction to Data Mining",** Pearson Education, 2007.
2. Jiawei Han and Micheline Kamber, **"Data Mining Concepts and Techniques"** Second Edition, Elsevier, Reprinted 2008.

Reference Books:

1. K.P. Soman, Shyam Diwakar and V. Ajay, **"Insight into Data mining Theory and Practice"**, Easter Economy Edition, Prentice Hall of India, 2006.
2. G. K. Gupta, **"Introduction to Data Mining with Case Studies"**, Easter Economy Edition, Prentice Hall of India, 2006.
3. Data Mining and Knowledge Science – Springer.
4. Inderscience, the International Journal of Datamining, Modelling and Management
5. IEEE, IEEE Transactions on Knowledge and Data Engineering.

Syllabus-Content

REVA UNIVERSITY
Bengaluru, India

Unit – I: Data Mining

Unit – II: Association Analysis

Unit – III: Classification

Unit – IV: Clustering Analysis

REVA UNIVERSITY
Bengaluru, India

Syllabus

of

Unit-1: Data Mining

Syllabus of Unit-1

REVA UNIVERSITY
Bengaluru, India

1. What is data mining ?
2. Motivating Challenges in data mining
3. The Origins of Data Mining
4. Data Mining Tasks
5. Types of Data
6. Data Quality
7. Data Preprocessing
8. Measures of Similarity and Dissimilarity
9. Data Mining Applications
10. Visualization

1.1 What is Data Mining?

What is Data Mining?

- Data Mining is the process of automatically discovering useful information in large data repositories.
- Data mining refers to extracting or mining knowledge from large amounts of data.
- Knowledge mining from data or Knowledge mining or Knowledge extraction.
- Data Mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data.
- It has also opened up exciting opportunities for exploring and analysing new types of data and for analysing old types of data in new ways.
- Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown.
- They also provide capabilities to predict the outcome of a future observation.

❑ Information Retrieval:

- It is the activity of obtaining the information resources relevant to an information need from a collection of information resources.
- Information Retrieval - the ability to query a computer system to return relevant results. The most widely used example is the google web search engine.

❑ Data Mining and Knowledge Discovery:

- Data mining is an integral part of Knowledge discovery in database (KDD) which is the overall process of converting raw data into useful information.
- Data mining is a synonym for another popularly used term, Knowledge discovery from Data or KDD.

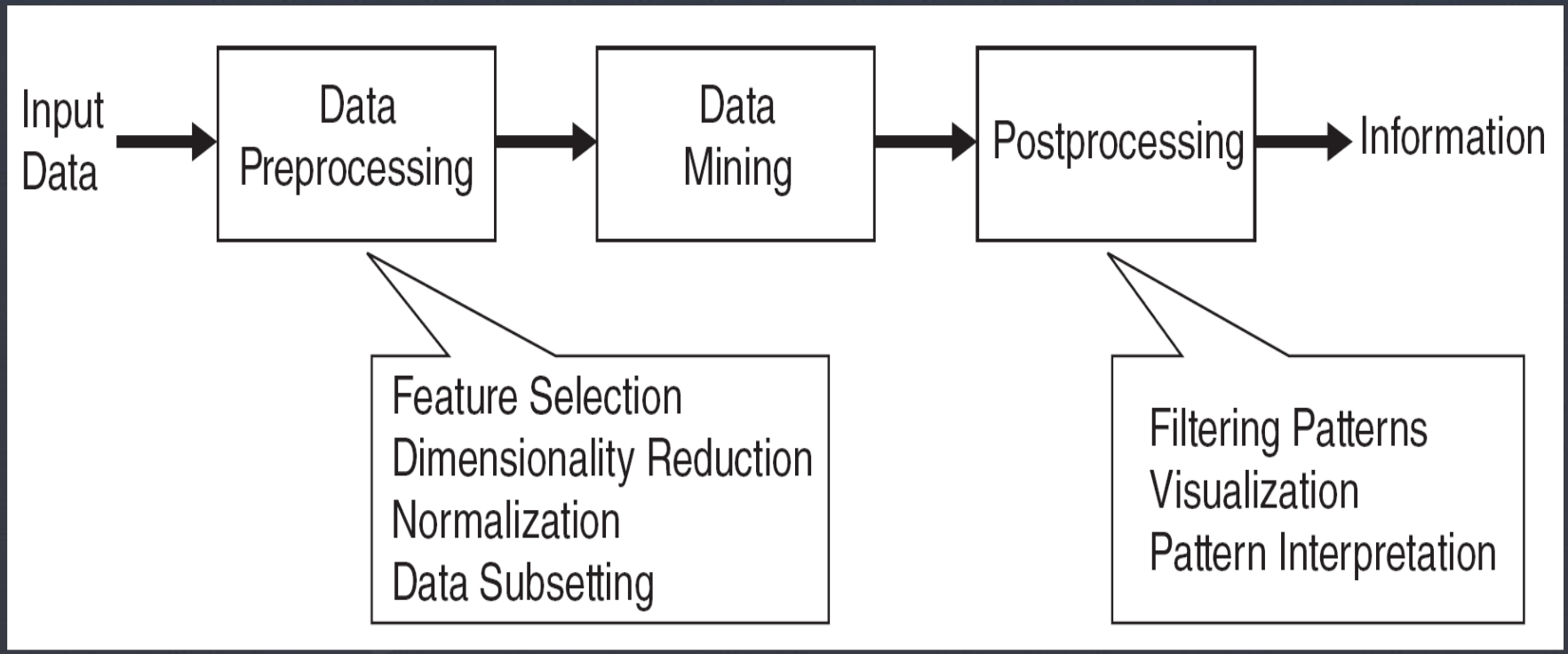


Figure :The process of KDD

❑ Input Data(raw data):

- The input data can be stored in a variety of formats(flat files, spreadsheets or relational tables) and may resides in a centralized data repository or to be distributed across multiple sites.

❑ Data Pre-processing:

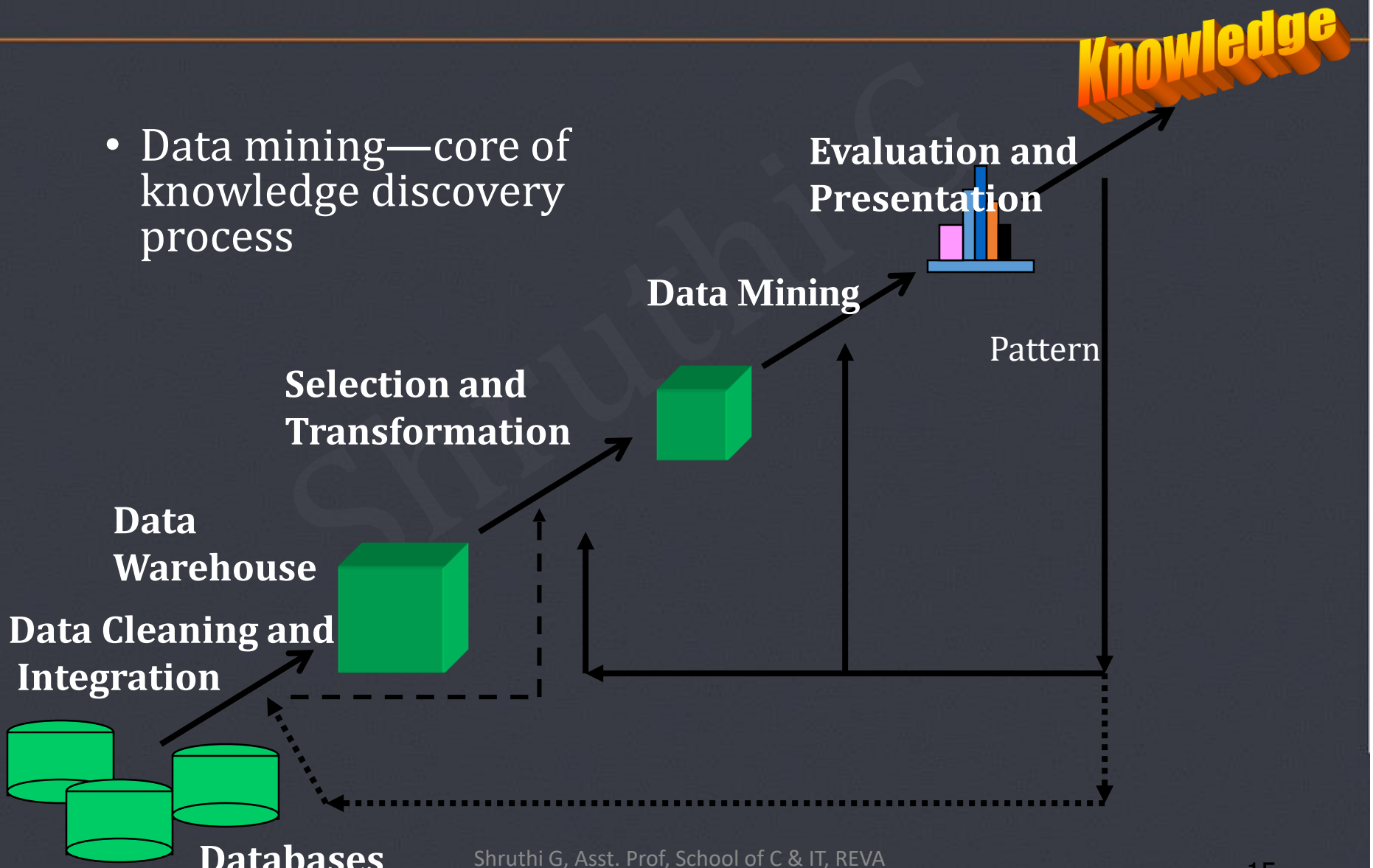
- The purpose of pre-processing is to transform the raw input data into an appropriate format for subsequent analysis.
- The steps involved in data pre-processing include:
  - ✓ Fusing data from multiple sources
  - ✓ Cleaning data to remove noise and duplicate observations
  - ✓ Selecting records and features that are relevant to the data mining task
- The many ways data can be collected and stored , data pre-processing is the most laborious and time consuming step in the overall Knowledge discovery process.

Post Processing:

- It ensures that only valid and useful results are incorporated into the decision support system.
- An example of post processing is visualization, which allows analysts to explore the data and data mining results from a variety of viewpoints.
- Statistical measures or hypothesis testing methods can also be applied during post processing to eliminate spurious data mining results.

Data Mining: A KDD Process

- Data mining—core of knowledge discovery process



- KDD is an iterative process
- Preprocessing of databases consists of Data cleaning and Data Integration.

1 Data Cleaning: To remove noise, inconsistent data and irrelevant data from collection.

2. Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).

3. Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the database.

4. Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

5. Data Mining: Data mining is an essential process where intelligent methods are applied in order to extract data patterns.

6. Pattern Evaluation: To identify the truly interesting patterns representing knowledge based on some interestingness measures.

7. Knowledge presentation: Knowledge presentation is defined as technique which utilizes visualization tools to represent data mining results.

1.2 Motivating Challenges in Data Mining

- The traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by new data sets.
- Following are some of the specific challenges that motivated the development of data mining:

1) Scalability
2) High Dimensionality
3) Heterogeneous and Complex Data
4) Data Ownership and Distribution
5) Non-traditional Analysis

Scalability:

- Because of advances in data generation and collection data sets with sizes of gigabytes, terabytes or even petabytes are becoming common.
- If data mining algorithms are to handle these massive data sets, then they must be scalable.
- Many data mining algorithms employ special search strategies to handle exponential search problems.
- Scalability may require the implementation of novel data structures to access individual records in an efficient manner.
- Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

❑ High Dimensionality(features):

- It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago.
- Data sets with temporal or spatial components also tend to have high dimensionality.
- For Example: consider a data set that contain measurements of temperature at various locations. If the temperature measurements are taken repeatedly for an extended period, the number of dimensions(features)increases in proportion to the number of measurements taken.
- Traditional data analysis techniques that were developed for low dimensional data often do not work well for such high dimensional data.
- For some data analysis algorithms, the computational complexity increases rapidly as the dimensionality increases.

❑ Heterogeneous and Complex Data:

- Traditional data analysis methods often deal with data sets containing attributes of the same type either continuous or categorical.
- As the role of data mining in business, science, medicine and other fields has grown ,so has the need for techniques that can handle heterogeneous attributes.
- Example of such non-traditional types of data include :
  - ✓ Collections of Web pages containing semi structured text and hyperlinks
  - ✓ DNA data with sequential and three-dimensional structure.
  - ✓ Climate data that contains of time series measurements(temperature, pressure etc.) at various locations on he Earth's surface.

Data Ownership and Distribution:

- Sometimes data needed for an analysis is not stored in one location or owned by one organization.
- Instead the data is geographically distributed among resources belonging to multiple entities.
- This requires the development of distributed data mining techniques.
- Challenges faced by distributed data mining algorithms include:
    1) How to reduce the amount of communication needed to perform the distributed computation
    2) How to effectively consolidate the data mining results obtained from multiple sources
    3) How to address data security issues.

❑ Non-traditional Analysis:

- The traditional statistical approach is based on hypothesize and test paradigm.
- A hypothesis is proposed, an experiment is designed to gather the data and then the data is analysed with respect to the hypothesis.
- This process is extremely labor-intensive
- Current data analysis tasks often require the generation and evaluation of thousands of hypotheses and consequently the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.

1.3 The Origins of Data Mining

- Traditional Techniques may be unsuitable due to:
  - ✓ Enormity of data
  - ✓ High dimensionality of data
  - ✓ Heterogeneous nature of data
- Data mining draws upon ideas ,such as:
- Sampling, estimation and hypothesis testing from statistics.
- Search algorithms, modelling techniques and learning theories from artificial intelligence, pattern recognition and machine learning
- Data mining has also been quick to adopt ideas from other areas including optimization, evolutionary computing, information theory, signal processing, visualization and information retrieval.

- Other areas play key supporting roles:
- Database systems are needed to provide support for efficient storage, indexing and query processing.
- Techniques from high performance (parallel) computing are often important in addressing the massive size of some data sets.
- Distributed techniques can also help address the issue of size and are essential when the data cannot be gathered in one location.
- The parallel computing and distributed technology are the two major data addressing issues in data mining to increase the performance

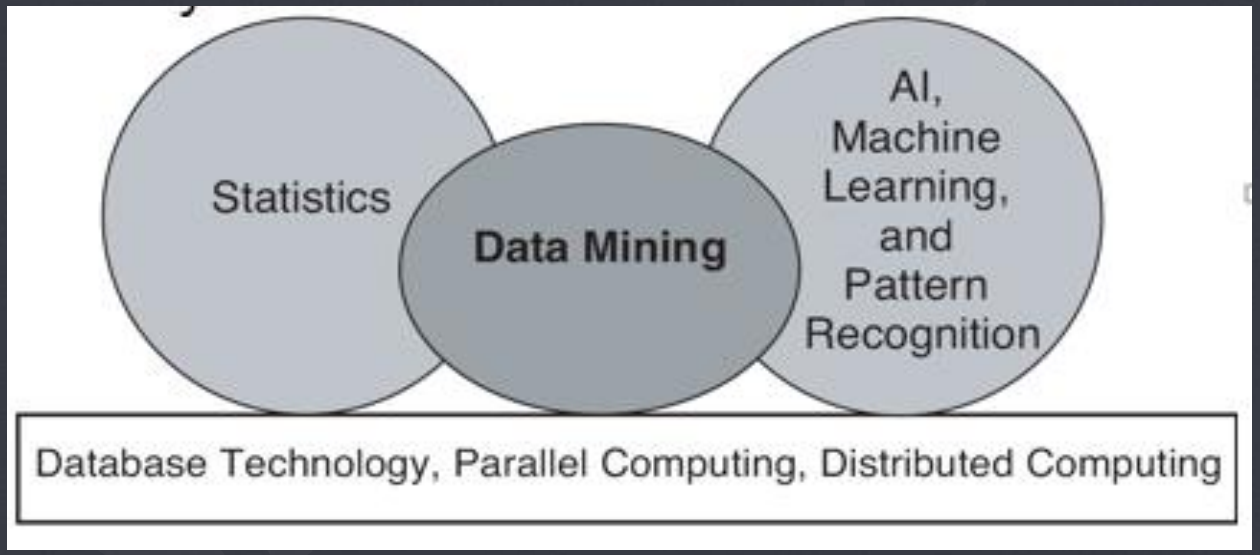


Figure :Data Mining as a confluence of many disciplines

1.4 Data Mining Tasks

- Data Mining tasks are generally divided into 2 major categories:
    1) Predictive tasks
    2) Descriptive tasks

- Predictive tasks: The objective of these tasks is to predict the values of a particular attribute based on the values of other attributes. Use some variables to predict unknown or future values of other variables.

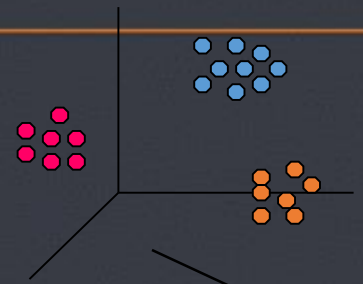
- The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as explanatory or independent variables.

- Descriptive tasks: The objective is to derive patterns(correlations, trends, clusters, trajectories and anomalies) that summarize the underlying relationships in data.
- This tasks are often exploratory in nature and frequently require post processing techniques to validate and explain the results.
- Find human-interpretable patterns that describe the data.

Core Data Mining Tasks:

There are four core data mining tasks:

1) Predictive Modeling
2) Association Analysis
3) Cluster Analysis
4) Anomaly Detection

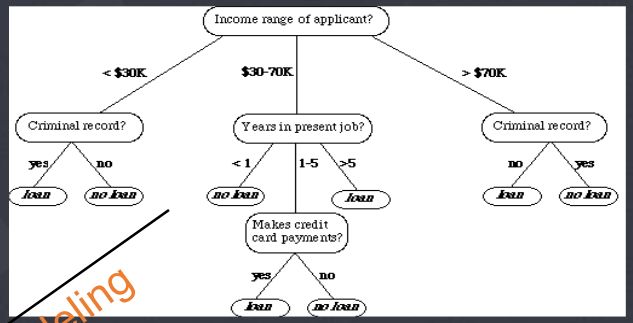


Clustering

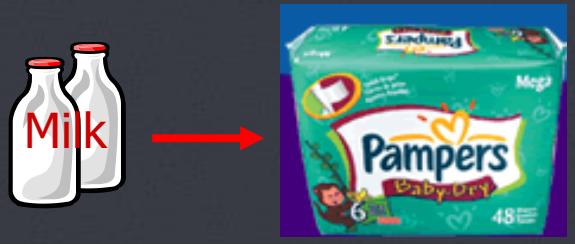
Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

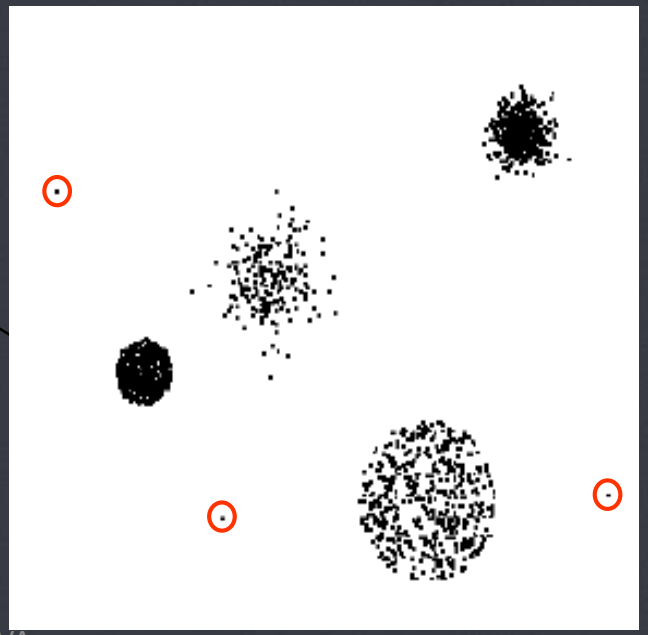
Predictive Modeling



Association Rules



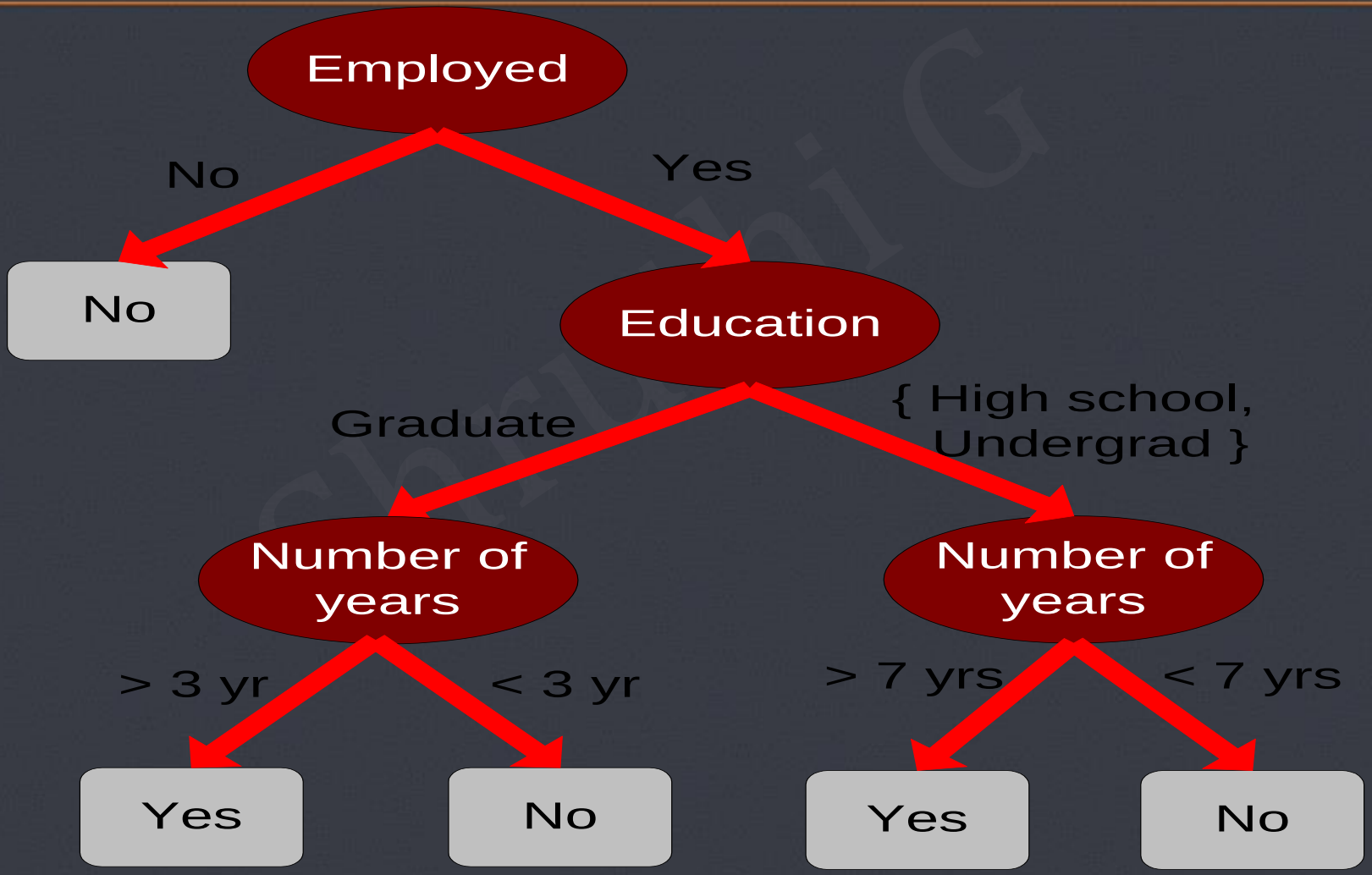
Anomaly Detection



Predictive Modeling

REVA UNIVERSITY
Bengaluru, India

- It refers to the task of building a model for the target variable as a function of the explanatory variables.
- There are two types of predictive modelling tasks:
  1) Classification
  2) Regression
- Classification: is used for discrete target variables. Example: Predicting whether a Web user will make a purchase at an online bookstore is a classification task because the target variable is binary-valued.


Employed
No
Yes
No
Education
Graduate
{ High school, Undergrad }
Number of years
Number of years
> 3 yr
< 3 yr
> 7 yrs
< 7 yrs
Yes
No
Yes
No

- Regression: is used for continuous target variables. Example: Forecasting the future price of a stock is a regression task because price is a continuous-valued attribute.
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- The goal of both the tasks is to learn a model that minimizes the error between the predicted and true values of the target variable.
- Predictive modelling can be used to identify customers that will respond to a marketing campaign, or judge whether a patient has a particular disease based on the results of medical tests.

Association Analysis

REVA UNIVERSITY
Bengaluru, India

- It is used to discover patterns that describe strongly associated features in the data.
- The discovered patterns are typically represented in the form of implication rule or feature subsets.
- The goal of association analysis is to extract the most interesting pattern in an efficient manner.
- Useful applications includes:
  - ✓Finding groups of genes that have related functionality
  - ✓Identifying web pages that are accessed together
- Example: Market Basket Analysis: association analysis can be applied to find items that are frequently bought together by customers. Rule {Diapers}->{Milk}

Cluster Analysis

REVA UNIVERSITY
Bengaluru, India

- This seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belongs to other cluster.
- Clustering has been used to:
  - ✓Group sets of related customers.
  - ✓Find areas of the ocean that have a significant impact on the Earth's climate .
- Example: Document clustering : The collection of news articles can be grouped based on their respective topics. Each article is represented as a set of word-frequency pairs(w, c),where w is a word and c is the number of times the word appears in the article. There are 2 natural clusters in the data set. The first cluster consists of news about the economy while the second cluster contains news about health care. A good clustering algorithm should be able to identify these 2 clusters based on the similarity between words that appear in the articles.

- Cluster 1:
- Dollar:1, industry:4, country:2, loan:3, deal:2, government:2,labor:3,market:4,country:1,job:5,
- jobless:4

- Cluster 2:
- Patient:4,symptom:2,drug:3,health:2,clinic:2,doctor:2,flu:3,cancer:9,health:3,medical:2

Anomaly Detection

REVA UNIVERSITY
Bengaluru, India

- It is the task of identifying observations whose characteristics are significantly different from the rest of the data. Such observations are known as anomalies or outliers.
- The goal of an anomaly detection algorithm is to discover the real anomalies and avoid falsely labelling normal objects as anomalous.
- Applications includes:
  - ✓Detection of fraud
  - ✓Network Intrusions
  - ✓Unusual patterns of disease
  - ✓Ecosystem disturbances.

1.5 Types of Data

What is Data set?

REVA UNIVERSITY
Bengaluru, India

- A data set is collection of data objects
  - Object is also known as record, point, case, vector, pattern, sample, entity, tuples, observation or instance
- Data objects are described by a number of attributes such as mass of a physical object or the time at which an event occurred.
- An attribute is a property or characteristic of an object or features of a data object.
- An attribute is a data field.
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable (statistician), field, characteristic, dimension(data warehouse) or feature(ML)
- A collection of attributes describe an object
- Attribute vector is a set of attributes used to describe a given object.

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Example: Student Information

REVA UNIVERSITY
Bengaluru, India

- A data set is a file, in which the objects are records(or rows) in the file and each field(or column) corresponds to an attribute. Below table shows a data set that consists of student information.
- Each row corresponds to a student and each column is an attribute that describes some aspects of a student ,such as grade point average(GPA) or ID.
- These record based data sets are common, either in flat files or relational database systems there are other important types of data sets and systems for storing data.

Student ID	Year	Grade point Average (GPA)
	:	
	.	
1034262	Senior	3.24
1052663	Second year	3.51
1082246	Freshman	3.62

- What is an attribute?
- An attribute is a property or characteristic of an object that may vary either from object to another or from one time to another.

Ex: 1. eye colour varies from person to person.

2. Temperature of an object varies over time.

- Note that eye colour is a symbolic attribute with a small number of possible values {brown, black,blue,green,hazel etc.}
- Temperature is a numerical attribute with a unlimited number of values.
- Measurement scale is a rule(function) that associates a numerical or symbolic value with an attribute of an object

Different types of attributes(properties of attribute values)

- The type of an attribute depends on which of the following properties(operations) it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /
- Using these operations (properties) we can define 4 types of attributes.

Types of Attributes

- Nominal Attribute (qualitative)
- Binary Attribute (Boolean attribute) (qualitative)(
  - ✓Symmetric binary (both outcome are equally important ex:gender)
  - ✓Asymmetric binary(outcomes are not equally important ex:medical test(positive vs negative))
- Ordinal Attribute (qualitative)(values have a meaningful order or ranking but magnitude b/w successive values is not known ex size={small, medium, large}
- Numeric Attribute(quantitative)
  - ✓ Interval-scaled: measured on a scale of equal size units.
  - ✓Ratio -scaled

Types of Attributes

REVA UNIVERSITY
Bengaluru, India

- There are different types of attributes
  - Nominal: Related to name, name of things, values, represent category
    - Examples: ID numbers, eye color, zip codes, hair color, Marital status, Occupation
  - Ordinal: Represent a meaningful order or ranking
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval : measured on equal sized unit, linear scaled
    - Examples: calendar dates, GRE score, temperatures in Celsius or Fahrenheit.
  - Ratio : non linear scale
    - Examples: temperature in Kelvin, length, time, counts

- Nominal, Binary, Ordinal
- Qualitative attributes: describes a feature of an object without giving an. actual size or quantity
- Quantitative is a measurable quantity, represented in integer or real values.
- Interval-scaled: measured on a scale of equal size units, values have ordered it can be +ve 0,-ve, no exact zero point. Ex: calendar 2002 and 2010 are 8 yrs apart. Practically its not possible to give temp and calendar as zero. It does not have true zero point this is the drawback.
- Ratio scaled attribute :It has inherent zero point. Ex :In kelvin ex 10Kelvin is twice as high as 5Kelvin ie 0-5 and 0-5

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation


18 19 20 21 22 26 23 24 25 27 28 29 30 cm
35 6 37 8
RATIO SCALE
VS
INTERVAL SCALE

Interval scale and ratio scale are the two variable measurement scales where they define the attributes of the variables quantitatively. The difference between interval and ratio scales is that, while interval scales are void of absolute or true zero for example temperature can be below 0 degree Celsius (-10 or -20), ratio scales have a true zero value, for example, height or weight it will always be measured between 0 to maximum but never below 0.

In an interval scale, all the quantitative attributes can be measured. Any measurement belonging to this category of interval scale can be ranked, counted, subtracted, added but by no means it will give any sense of ratio between the two measurements.

A ratio scale is a measurement scale which has more or less all the properties of an interval scale. Ratio data on this scale has measurable intervals. Where the ratio scale differs is, it has a zero point or character of origin.

Properties of Attribute Values

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties

- Nominal and ordinal attributes are collectively referred to as categorical or qualitative attributes. Qualitative attributes such as employee ID lack most of the properties of numbers. Even if they represented by numbers i.e. integers they should be treated more like symbols.
- Interval and ratio are collectively referred to as quantitative or numeric attributes. Quantitative attributes are represented by numbers and have most of the properties of numbers. The quantitative attributes can be integer-valued or continuous.

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countable infinite set of values. Need not to be an integer.
  - These attributes can be categorical.
  - Examples: zip codes or ID numbers ,or numeric such as counts, profession, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes and assume only 2 values ex true/false, yes/no, male/female,0/1.
  - Binary attribute is a special case of discrete attribute
  - Binary attribute are often represented as Boolean variables or as integer variables that only take the values 0 or 1.
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.(precision)
  - Continuous attributes are typically represented as floating-point variables.

- Nominal and ordinal attributes are binary or discrete while interval and ratio attribute are continuous.
- Asymmetric attributes
- Only presence of a non zero attribute value is regarded as important.
- Eg: consider soft cores or electives. If a student takes a particular course then mark 1 corresponding to that attribute else 0 and only 1 becomes important.
- Binary attribute where only non zero values are important are called asymmetric binary attribute.

Name	DM	USP	IA	xyz	blah
Anand	1	0	0	0	0
Anitha	0	1	0	0	0
Araav	0	0	1	0	0
Aravind	1	0	0	0	0

- Symmetric attributes
  - Here all the attribute values need to be considered

Name	Male	Female
Anand	Y	N
Anitha	N	Y
Araav	Y	N

  - Binary Attributes (have only 2 values) 0/1, yes/no etc

Important Characteristics of Structured Data Sets

REVA UNIVERSITY
Bengaluru, India

- **Dimensionality:**

- The dimensionality of a data sets is the number of attributes that the objects in the data set possess.
- The difficulties are associated with analyzing high dimensional data are referred as Curse of Dimensionality. Increases computation time.
- Because of this important motivation in preprocessing the data is dimensionality reduction.

- **Sparsity:**

- For some data sets(asymmetric features), most attributes of an object have values of 0; fewer than 1% of the entries are non zero.
- Directly proportional. Sparsity is an advantage Coz only non zero values need to be stored and manipulated.
- Saves computation time and storage.
- Some Data mining algorithms work well only for sparse data.

- **Resolution:**

- It is frequently possible to obtain data at different levels of resolution and often the properties of the data are different at different resolution. Ex: surface of the earth
- Patterns in the data also depend on the level of resolution. Sometime high resolution and sometimes low resolution of data is needed

Types of data sets

REVA UNIVERSITY
Bengaluru, India

- **Record Data**
  - **Transaction Data**
  - **Data Matrix**
  - **Sparse Data Matrix**
- **Graph**
  - **World Wide Web**
  - **Molecular Structures**
- **Ordered**
  - **Sequential data/Temporal data**
  - **Sequence data (Genetic Sequence Data/genomic sequence data)**
  - **Time series data**
  - **Spatial data**

Record Data

- Data that consists of a collection of records(data objects), each of which consists of a fixed set of attributes (fields).
- Record data is stored either in flat files or in relational databases.
- Different types of record data are:
  - ✓ Transaction or Market Basket Data
  - ✓ Data Matrix
  - ✓ Sparse Data Matrix or document term matrix

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Transaction Data

REVA UNIVERSITY
Bengaluru, India

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Data Matrix

REVA UNIVERSITY
Bengaluru, India

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points(vectors) in a multi-dimensional space, where each dimension represents a distinct attribute describing the object.
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute(A representation that has data objects as columns and attributes as rows is also fine).

- A data matrix is a variation of record data, but because it consists of numeric attributes, standard matrix operation can be applied to transform and manipulate the data. The data matrix is the standard data format for most statistical data.
- This matrix is called a data matrix or a pattern matrix.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Sparse Data Matrix

REVA UNIVERSITY
Bengaluru, India

- A sparse data matrix is a special case of a data matrix in which the attributes are of the same type and are asymmetric i.e. only non-zero values are important.
- Ex: 1. Transaction data that has only 0-1 entries
- 2. Document data
- Each document becomes a `term' vector
- If the order of the terms(words) in a document is ignored, then a document can be represented as a term vector, where each term is a component(attribute) of the vector and the value of each component is the number of times the corresponding term occurs in the document.
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

- This representation of a collection of document is often called a document-term matrix.
- Documents are the rows.
- Terms are the columns.
- Only non zero entries of sparse data matrices are stored.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Graph-Based Data

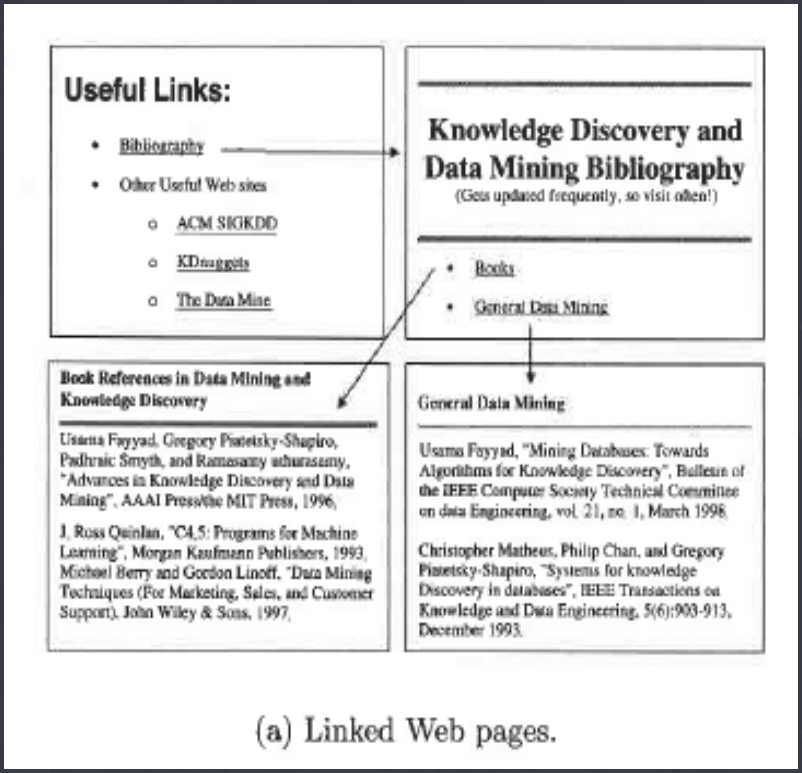
REVA UNIVERSITY
Bengaluru, India

- A graph can sometimes be a convenient and powerful representation for data.
- There are 2 specific cases:

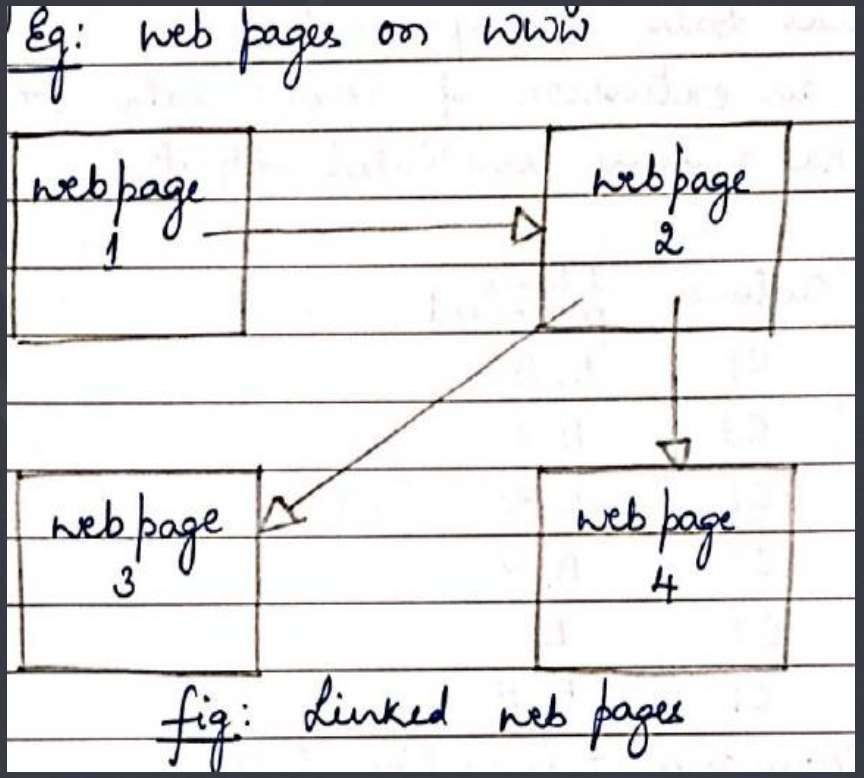
1. The graph captures relationships among data objects
2. The data objects themselves are represented as graphs

1. Data with relationships among objects:

- The relationship among objects frequently convey important information.
- Data is often represented as a graph. The data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links between objects and link properties such as direction and weight.
- Consider web pages on WWW, which contain both text and links to other pages. In order to process search queries, web search engines collect and process web pages to extract their contents.
- Examples: Generic graph and HTML Links

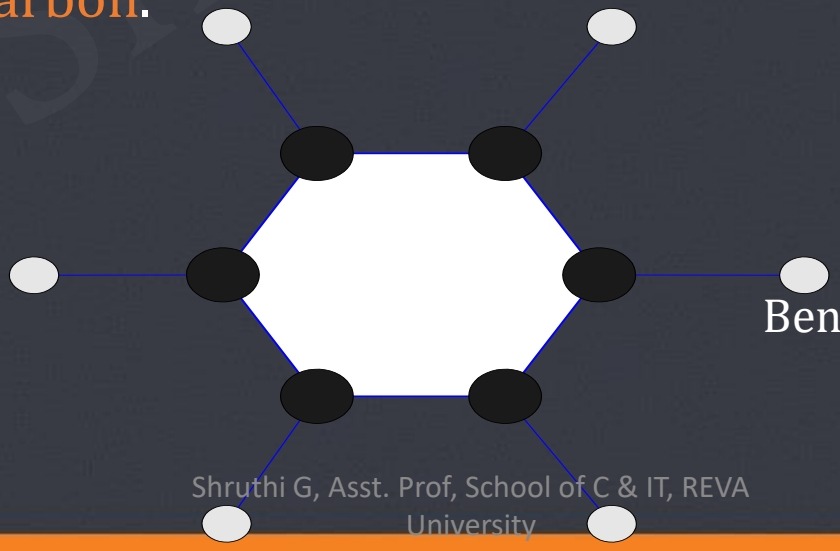


(a) Linked Web pages.



2. Data with objects that are Graphs

- If objects have structure ,that is the objects contain sub objects that have relationships, then such objects are frequently represented as graphs.
- Example: the structure of chemical compounds can be represented as graphs, where the nodes are atoms and the links between nodes are chemical bonds.
- The benzene molecule is composed of six carbon atoms joined in a ring with one hydrogen atom attached to each. As it contains only carbon and hydrogen atoms, benzene is classed as a hydrocarbon.



Benzene Molecule: C_6H_6

Ordered Data

REVA UNIVERSITY
Bengaluru, India

- Here the attributes have relationships that involve order in time or space.
- Different types of ordered data are:
  - ✓Sequential data/Temporal data
  - ✓Sequence data
  - ✓Time series data
  - ✓Spatial data

- Sequential transaction data
  - Also called temporal data, can be thought of an extension of record data, where each record has time associated with each object.
  - Consider the retail transaction data set that also stores the time at which the transaction took place.
  - During winter sales of sweaters/ shawls increases

Time	Customer	Items purchased
T1	C1	A,B
T2	C2	A,D
T3	C3	A,E
T4	C4	B,D

Time	Customer	Items purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

(a) Sequential transaction data

(OR)

Customer	Time and Items Purchased
C1	(t1: A, B) (t2: C, D) (t5: A, E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Sequence Data

- It consists of a data set that is sequence of individual entities, such as sequence of words or letters.
- It is quite similar to sequential data, except that there are no time stamps; instead there are positions in an ordered sequence.
- Ex: Genetic information of plants and animals can be represented in the form of sequences of nucleotides that are known as genes.
- Genomic sequence data.
- Human genetic code

Expressed using the four

Nucleotides from which all

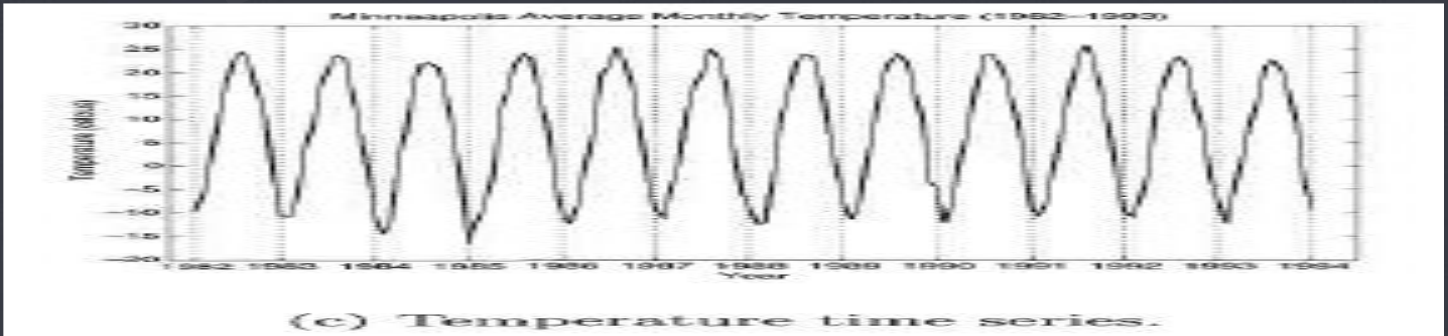
DNA is constructed: A,T,G and C

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Time series data

REVA UNIVERSITY
Bengaluru, India

- It is a special type of sequential data in which each record is a time series. Here the measurements are taken with respect to time.
- When working with temporal data ,it is important to consider temporal autocorrelation(serial correlation)(successive values of the same variable) i.e. if 2 measurements are close in time, then the values of those measurements are often very similar
  - Eg: stock market(financial data set might contain objects that are time series of the daily prices of various stocks), temperature etc
  - Average monthly temperature for Minneapolis during the year 1982 to 1994.

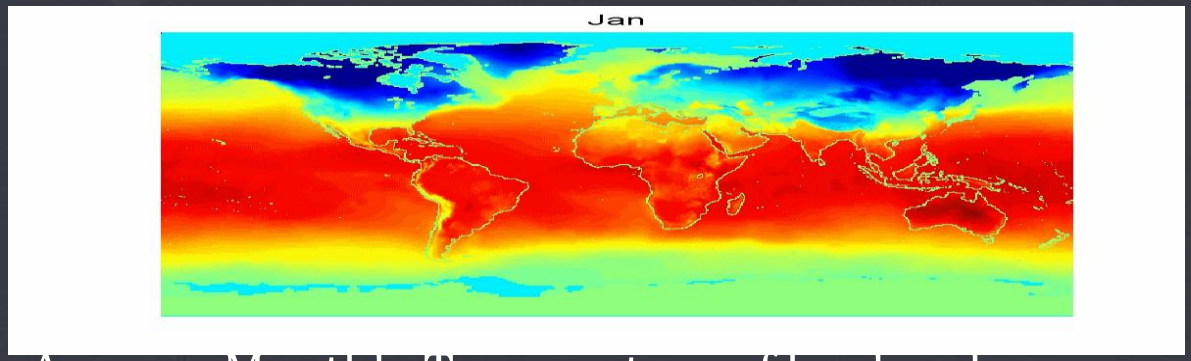


(c) Temperature time series.

Spatial Data

- Some objects have spatial attributes such as positions or areas .
- Ex: weather data(precipitation, temperature, pressure) that is collected for a variety of geographical locations.
- An important aspect of spatial data is spatial autocorrelation i.e. objects that are physically close tend to be similar in other ways as well.
- Thus, 2 points on the earth that are close to each other usually have similar values for temperature and rainfall.
- Spatial autocorrelation measures the correlation of a variable with itself through space. Spatial autocorrelation can be positive or negative. Positive spatial autocorrelation occurs when similar values occur near one another. Negative spatial autocorrelation occurs when dissimilar values occur near one another.

- Important examples of spatial data are the science and engineering data sets that are the result of measurements or model output taken at regularly or irregularly distributed points on a two or three dimensional grid or mesh.
- Ex: Earth Science data sets record the temperature or pressure measured at points(grid cells) on latitude-longitude spherical grids of various resolutions



Average Monthly Temperature of land and ocean

1.6 Data Quality

Data Quality

REVA UNIVERSITY
Bengaluru, India

- Data mining applications are often applied to data that was collected for another purpose, or for future, but unspecified applications.
- So data mining cannot usually take advantage of the significant benefits of "addressing quality issues at the source"
- Preventing data quality problems is not possible. Hence data mining focuses on:
  - ✓Detection and correction of data quality problems. This step is called data cleaning.
  - ✓The use of algorithms that can tolerate poor data quality.

Measurement and data collection issues:

REVA UNIVERSITY
Bengaluru, India

- It is unrealistic to expect that data will be perfect.
- There may be problems due to human error, limitations of measuring devices, or flaws in the data collection process.
- Values or even entire data objects may be missing.
- There may be spurious or duplicate objects i.e. multiple data objects that all corresponds to a single “real” object.
- Ex: there might be 2 different records for a person who has recently lived at 2 different addresses.
- Even if all the data is present and “looks fine”, there may be inconsistencies –a person has a height of 2 meters, but weighs only 2 kilogram

1) Measurement and data collection errors
2) Noise and Artifacts
3) Precision, Bias and Accuracy
4) Outliers
5) Missing Values
6) Inconsistent Values
7) Duplicate data

- A variety of problems involve measurement error: noise, artifacts, bias, precision and accuracy.
- Data quality issues that may involve both measurement and data collection problems: outliers, missing and inconsistent values and duplicate data.

1. Measurement and data collection errors

REVA UNIVERSITY
Bengaluru, India

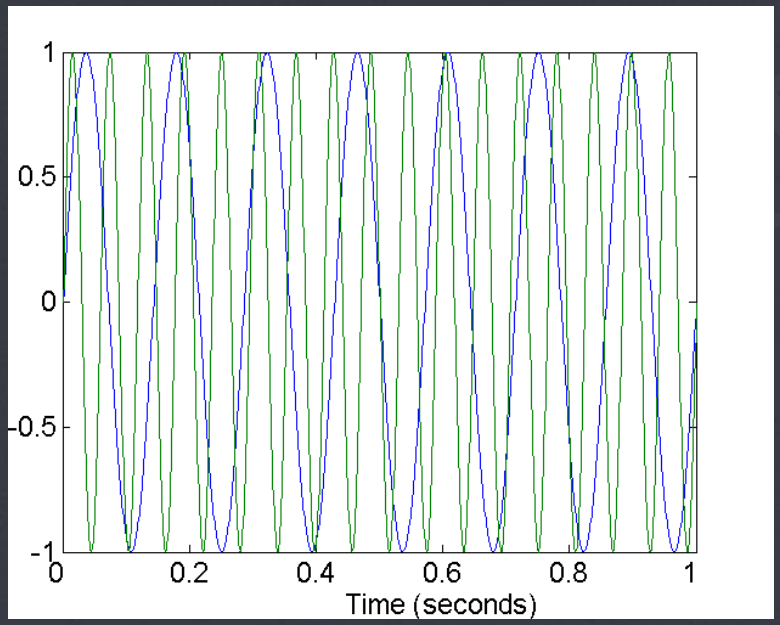
- Measurement Error:
- It refers any problem resulting from the measurement process. A common problem is that the value recorded differs from the true value to some extent.
- For continuous attributes, the numerical difference of the measured and true value is called the error.
- Data Collection error:
- It refers to errors such as omitting data objects or attribute values or inappropriately including a data object.
- Ex: a study of animals of a certain species might include animals of a related species that are similar in appearance to the species of interest.
- Both the measurement errors and data collection errors can be either systematic or random.

2. Noise and Artifacts

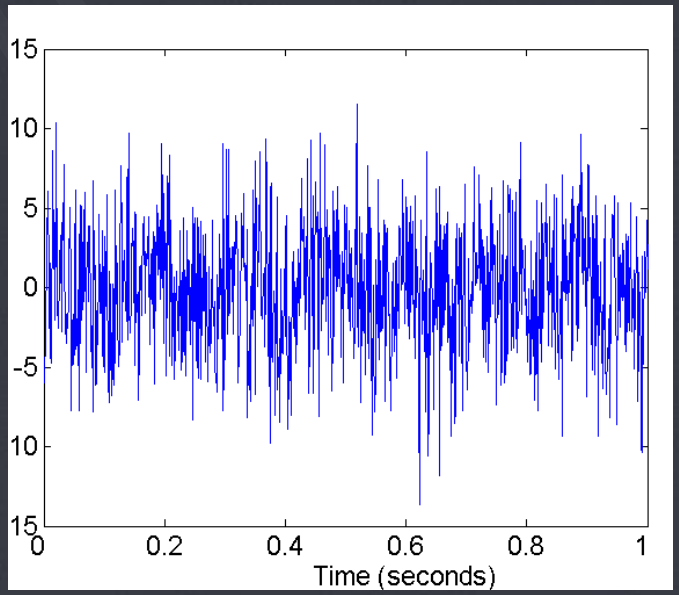
REVA UNIVERSITY
Bengaluru, India

- It is the random component of a measurement error.
- It may involve the distortion of a value or the addition of spurious objects.
- Elimination of noise is difficult. Therefore DM focuses on devising robust algorithms that produce acceptable results even when noise is present.
- Data errors may be the result of a more deterministic phenomenon such as a streak in the same place on a set of photographs. Such deterministic distortions of the data are often referred to as artifacts.
- Example: a small crack in the lens of a camera produces images with distortion or a mark at that place.

- Noise refers to modification of original values
  - Example: distortion of a person's voice when talking on a poor phone



Two Sine Waves



Two Sine Waves + Noise

3. Precision, Bias and Accuracy

REVA UNIVERSITY
Bengaluru, India

- Precision
  - The closeness of repeated measurements (of the same quantity) to one another.
  - Precision is often measured by the standard deviations of a set of values.
- Bias
  - A systematic variation of measurements from the quantity being measured.
  - Bias is measured by taking the difference between the mean of the set of values and known value of the quantity being measured.
- Accuracy
  - The closeness of measurements to the true value of the quantity being measured.
  - Accuracy depends on precision and bias but since it is a general concept there is no specific formula for accuracy in terms of these 2 quantities.
  - Accuracy can be increased by making use of significant digits.

4. Outliers

- Outliers are either:
- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
- Or values of an attribute that are unusual with respect to typical values for that attribute

5. Missing Values

REVA UNIVERSITY
Bengaluru, India

- Reasons for missing values:
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Several strategies to Handle missing values are:

1.Eliminate Data Objects or attributes:

- Eliminate objects with missing values.
- Even a partially specified data object contain some information and if many objects have missing values, then a reliable analysis can be difficult or impossible
- If a data set has only a few objects that have missing values, then it may be expedient to omit them.
- A related strategy is to eliminate attributes that have missing values.

2. Estimate Missing Values:

- Sometimes missing data can be reliably estimated.
- Ex: consider a time series having smooth changes, we can estimate its value at a specific time by observing previous values.
- If the attribute is continuous, then the average attribute value of the nearest neighbors is used.
- If the attribute is categorical, then the most commonly occurring attribute value can be taken.

3. Ignore the Missing Value During Analysis:

- Many data mining approaches can be modified to ignore missing values.
- Ex: suppose that objects are being clustered and the similarity between pairs of data objects needs to be calculated. If one or both objects of a pair have missing values for some attributes then the similarity can be calculated by using only the attributes that do not have missing values.

6. Inconsistent data

REVA UNIVERSITY
Bengaluru, India

- Data can contain inconsistent values.
- Ex: In an address field, where both a zip code and city are listed, but the specified zip code area is not contained in that city.
- It is important to detect and if possible to correct inconsistent data
  - Example: Age or height of a person in negative numbers
  - Some inconsistent data are easy to detect and correct but some need external source. Ex : when an insurance company processes claims for reimbursement, it checks the names and addresses on the reimbursement forms against a database of its customers.

7. Duplicate Data

REVA UNIVERSITY
Bengaluru, India

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogonous sources
- Examples:
  - Same person with multiple email addresses
- To detect and eliminate such duplicates, 2 main issues must be addressed:

1. If there are 2 objects that actually represent a single object, then the values of corresponding attributes may differ and these inconsistent values must be resolved.
2. Care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates ,such as two distinct people with identical names.

- The term deduplication is often used to refer to the process of dealing with these issues.
- In some cases, two or more objects are identical with respect to the attributes measured by the database, but still represent different objects. Here, the duplicates are legitimate, but may still cause problems for some algorithms if the possibility of identical objects is not specifically accounted for in their design.
- Data cleaning
  - Process of dealing with duplicate data issues

Issues related to applications

1) **Timeliness:** some data starts to age as soon as it has been collected. Ex: Data provides a snapshot of some ongoing phenomenon or process, such as the purchasing behavior of customers or web browsing patterns, then this snapshot represents reality for only a limited time.
2) **Relevance:** The available data must contain the information necessary for the application. Ex: consider the task of building a model that predicts the accident rate for drivers. If information about the age and gender of the driver is omitted, then it is likely that the model will have limited accuracy unless this information is indirectly available through other attributes.
3) **Knowledge about the data:** Data sets are accompanied by documentation that describes different aspects of the data: the quality of this documentation can either aid or hinder the subsequent analysis. Ex: If the documentation identifies several attributes as being strongly related these attributes are likely to provide highly redundant information and we may decide to keep just one. If the documentation is poor, then our analysis of the data may be faulty.

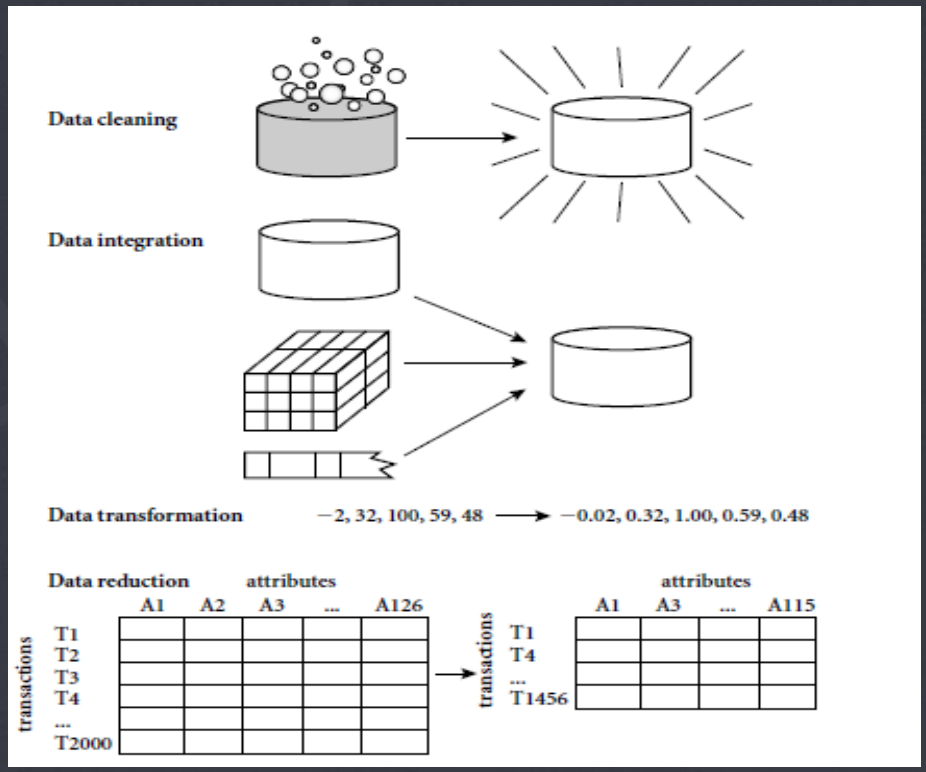
1.7 Data Preprocessing

Data preprocessing

REVA UNIVERSITY
Bengaluru, India

The purpose of pre-processing:

To transfer the raw input data into an appropriate format for subsequent analysis in the process of knowledge discovery in databases (KDD). Data pre processing is a broad area and consists of a number of different strategies and techniques.



Data Preprocessing

REVA UNIVERSITY
Bengaluru, India

- It address the issue of which preprocessing steps should be applied to make the data more suitable for data mining.

1. Aggregation
2. Sampling
3. Dimensionality Reduction
4. Feature subset selection
5. Feature creation
6. Discretization and Binarization
7. Attribute (or variable) Transformation

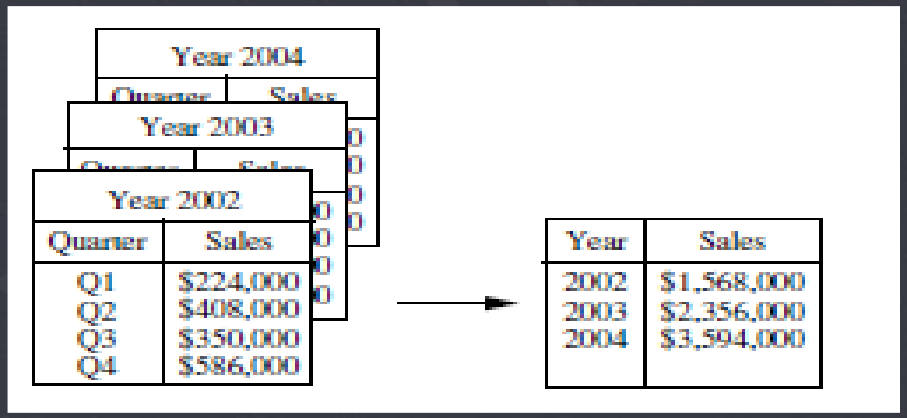
These items fall into 2 categories:

1. Selecting data objects and attributes for the analysis
2. Creating/changing the attributes

In both the cases the goal is to improve the data mining analysis wrt time, cost and quality.

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)



- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
  - More "stable" data
    - Aggregated data tends to have less variability

- Ex: consider dataset consisting of transactions(data objects) recording the daily sales of products in various store locations for different days over the course of a year.

Transaction ID	Item	Store location	Date	Price
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
101	watch	chicago	9/6/04	$29.99
102	shoes	ottowo	9/6/04	$31.44
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮

- One way to aggregate transactions for this data set is to replace all the transactions of a single store with a single storewide transaction. This reduces the hundreds or thousands of transactions that occur daily at a specific store to a single daily transaction and the number of data objects is reduced to the number of stores.

- Quantitative attributes, such as price, are aggregated by taking a sum or an average.
- Qualitative attribute such as item can either be omitted or summarized as the set of all the items that were sold at that location.
- Motivations for aggregation:

1) The smaller data sets resulting from data reduction require less memory and processing time.
2) It act as a change of scope or scale by providing a high level view of the data instead of a low-level view.

- A disadvantage of aggregation is the potential loss of interesting details. Ex : aggregating over months loses information about which day of the week has the highest sales.

Sampling

REVA UNIVERSITY
Bengaluru, India

- **It is used for selecting a subset of the data objects to be analyzed.**
- **Sampling is the main technique employed for data selection.**
  - **In statistics, It is often used for both the preliminary investigation of the data and the final data analysis in Statistics.**
- **Motivations for sampling in statistics and data mining are often different.**
- **Statisticians use sampling because obtaining entire set of data of interest is too expensive or time consuming.**
- **Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.**
- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data
  - For instance, if the mean (average) of the data objects is the property of interest, then a sample is representative if it has a mean that is close to that of the original data.

Types of Sampling or Sampling Approaches

REVA UNIVERSITY
Bengaluru, India

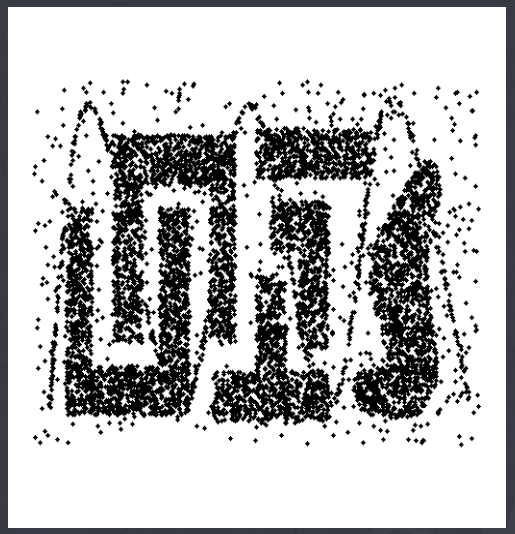
- Simple Random Sampling
  - There is an equal probability of selecting any particular item. There are 2 variations on random sampling:

1) Sampling without replacement
   - As each item is selected, it is removed from the set of all objects that together constitute the population

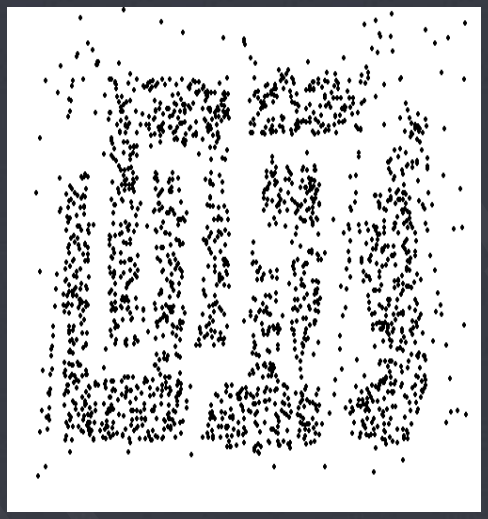
2) Sampling with replacement
   - Objects are not removed from the population as they are selected for the sample.
     - In sampling with replacement, the same object can be picked up more than once

- Stratified sampling
  - Split the data into several partitions(i.e groups); then draw random samples from each partition

Sample Size



8000 points



2000 Points

500 Points

- Once the sampling technique has been selected, it is necessary to choose the sample size.
- Larger sample sizes increase the probability
- If the sample size is smaller, patterns may be missed or erroneous pattern can be detected.
- Most of the structure of this data set is present in the sample of 2000 points, much of the structure is missing in the sample of 500 points.

- Progressive sampling /Adaptive:
- It is used because proper sample size can be difficult to determine.
- Start with small samples and then increase the number of sample size until a sample of sufficient size has been obtained or there is no change or increase in accuracy.
- Other name is adaptive sampling

Dimensionality Reduction

REVA UNIVERSITY
Bengaluru, India

- Dimensionality Reduction is often referred to those techniques that reduce the dimensionality of a dataset by creating new attributes that are a combination of the old attributes.
- A key benefit of DR is that many data mining algorithms work better if the dimensionality-the number of attributes in the data is lower. This is because it can eliminate irrelevant features and reduce noise.
- Another benefit is that it can lead to a more understandable model because the model may involve fewer attributes and also allow the data to be more easily visualized.
- The amount of time and memory required by the data mining algorithm is reduced with the reduction in dimensionality.
- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

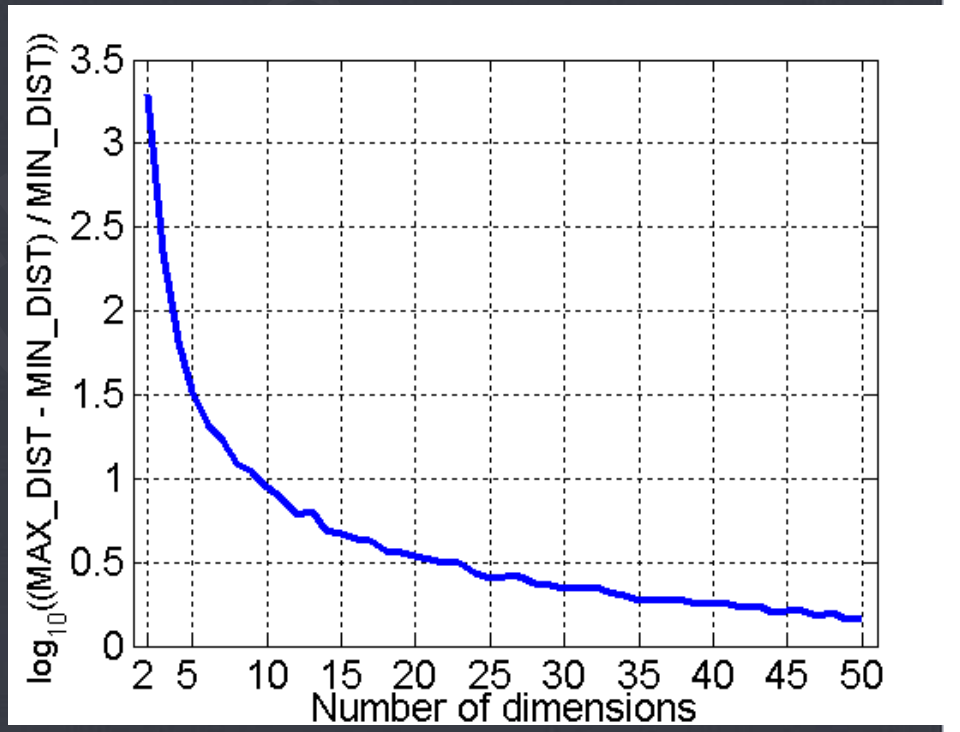
- Linear Algebra Techniques for Dimensionality Reduction:
  - Principle Component Analysis(PCA)
  - Singular Value Decomposition(SVD)
  - Others: supervised and non-linear techniques
- Curse of Dimensionality: It refers to the phenomenon that many types of data analysis becomes significantly harder as the dimensionality of the data increases. Specifically, as dimensionality increases ,the data becomes increasingly sparse in the space that it occupies.

Curse of Dimensionality

REVA UNIVERSITY
Bengaluru, India

- When dimensionality increases, data becomes increasingly sparse(thinly dispersed or scattered) in the space that it occupies

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

- PCA is a linear algebra technique for continuous attributes that finds new attributes(principal component) that:

1) Are linear combinations of the original attributes
2) Are orthogonal (perpendicular) to each other
3) Capture the maximum amount of variation in the data.

Feature Subset Selection

REVA UNIVERSITY
Bengaluru, India

- Another way to reduce dimensionality of data is to use only a subset of the features. This approach is useful when redundant and irrelevant features are present in the dataset.
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA(grade point averages)
- Redundant and irrelevant features can reduce classification accuracy and quality.

Feature Subset Selection

REVA UNIVERSITY
Bengaluru, India

- Techniques /Approaches for feature subset selection:
- Ideal approach: is to try all possible subsets of features as input to DM algorithm of interest, and then take the subset that produces the best result that is referred to as Brute-force approach.
- Disadvantage: No of subsets involving n attributes in 2n
- The 3 standard approach of feature selection:
  1) Embedded approaches:
     - Feature selection occurs naturally as part of the data mining algorithm. During the operation of the DM algorithm, the algorithm itself decides which attributes to use and which to ignore.

2) Filter approaches:

- Features are selected before data mining algorithm is run using some approach that is independent of data mining task.

3) Wrapper approaches:

- Use the target data mining algorithm as a black box to find best subset of attributes, in a way similar to that of ideal algorithm (brute force), but typically without enumerating all possible subsets.

- Feature Weighting: it is an alternative to keeping or eliminating features. More important features are assigned a higher weight, while less important features are given a lower weight.

Architecture for feature subset selection

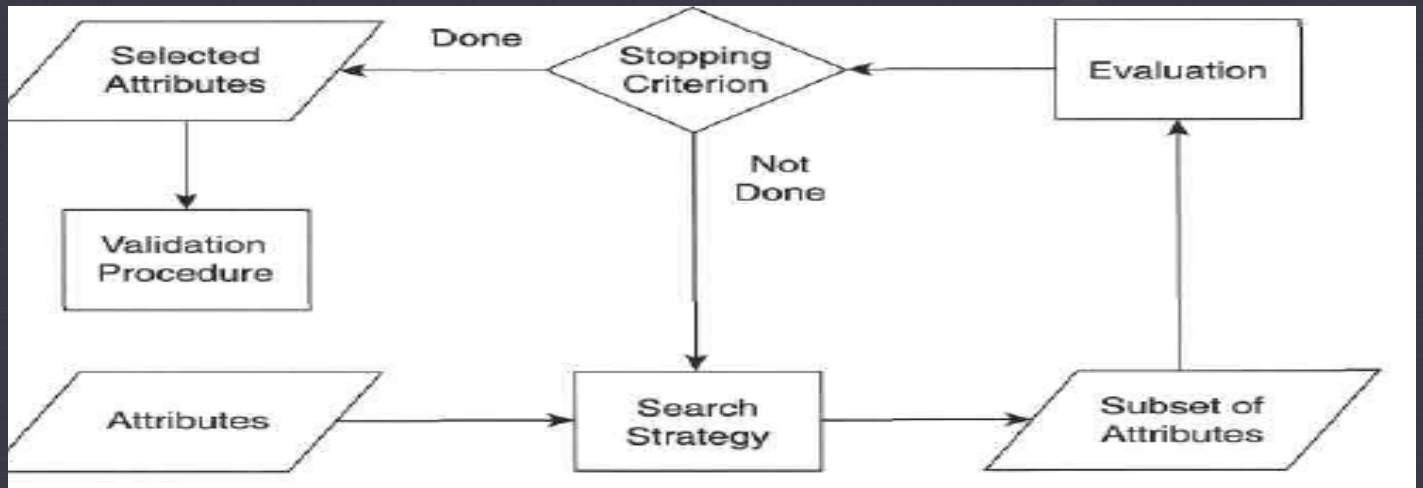


Figure 2.11. Flowchart of a feature subset selection process.

It consists of 4 parts: a measure for evaluating a subset, a search strategy that controls the generation of a new subset of features, a stopping criterion and a validation procedure.

Filter and wrapper methods differ only in the way in which they evaluate a subset of features.

Feature Creation

REVA UNIVERSITY
Bengaluru, India

- Create new set of attributes (from the original attributes)that can capture the important information in a data set much more efficiently than the original attributes. Furthermore, the no. of new attributes can be smaller than the no. of original attributes(dimensionality reduction).

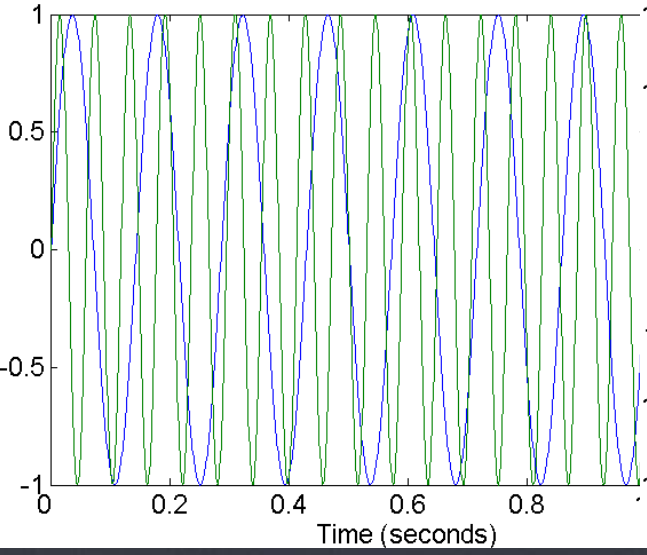
- Three general methodologies for creating new attributes:
  - Feature Extraction
    - Creation of new set of features from the original raw data is known as feature extraction. It is domain-specific ex: extracting edges from pixels of an image.
    - For a particular field, such as image processing, various features and the techniques to extract them have been developed over a period of time, and often these techniques have limited applicability to other fields.

- Mapping Data to New Space:
  - A totally different view of the data can reveal important and interesting features.
  - Ex: time series data, which often contains periodic patterns. If there is only a single periodic pattern and not much noise, then the pattern is easily detected. If there are a number of periodic patterns and amount of noise is present then these patterns are hard to detect. Such patterns often be detected by applying a Fourier Transform.
  - Wavelet transform has also proven very useful for time series and other types of data.
- Feature Construction:
- Sometimes the features in the original data sets have the necessary information, but it is not in a form suitable for data mining algorithm.
- In this situation, one or more new features constructed out of the original features can be more useful than the original features.
  - combining features (example density = mass/volume)

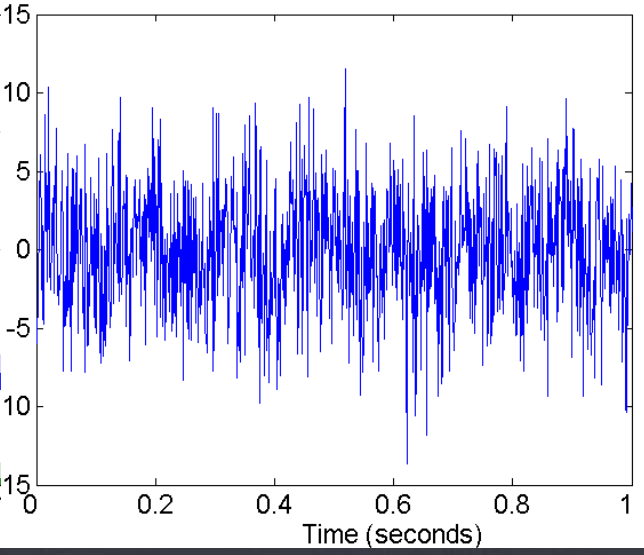
Mapping Data to a New Space

REVA UNIVERSITY
Bengaluru, India

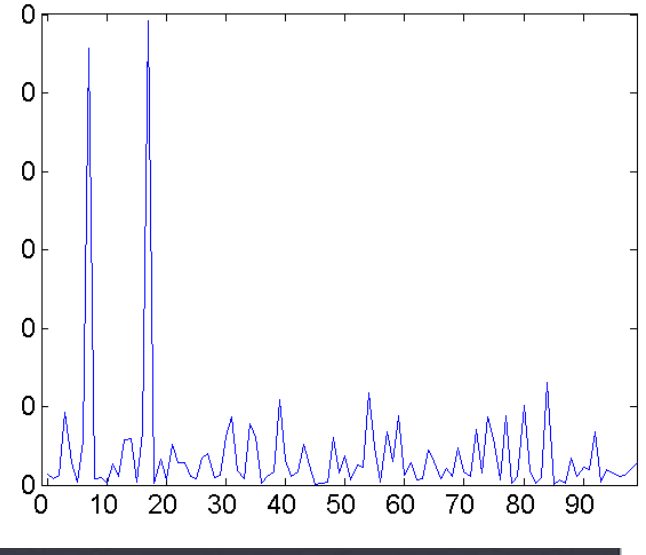
- Fourier transform
- Wavelet transform



Two Sine Waves



Two Sine Waves + Noise



Frequency

Discretization and Binarization

- Some data mining algorithm i.e. classification algorithms require that the data be in the form of categorical attributes. Algorithms that find association patterns require that the data be in the form of binary attributes.
- Discretization is a process of transforming a continuous attribute into a categorical attribute.
  - Procedure involves 2 steps
  - 1. Decide the number of categories
  - 2. Determining how to map values of continuous attributes to these categories.
  - Sort the attributes and then divide them into the categories
  - Assign a discrete value to all the attributes in one category
- Binarization is a process of transforming both continuous and discrete attributes into one or more binary attributes.
  - If ordinal, maintain order.

As with feature selection, the best discretization and binarization approach is the one that "produces the best result for the DM algorithm that will be used to analyze the data".

- A simple technique to binarize a categorical attribute is as following:

✓ If there are m categorical attributes(values),then uniquely assign each original value to an integer in the interval {0,m-1},if the attribute is ordinal, then order must be maintained by the assignment.

✓ Convert each of these m integers to a binary number n= log 2m binary digits are required.

$$n = \log_2 m$$

Binarization

Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Such a transformation can cause 2 complications:

1. Creates unintended relationships among the transformed attributes. Ex: attributes x2 and x3 are correlated because information about the "good" value is encoded using both attributes.

2.Leads to symmetric binary attributes, but association analysis require asymmetric binary attributes, where only presence of the attribute (value=1) is important.

- Solution: Introduce one binary attribute for each categorical value.

Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Discretization of Continuous attributes

REVA UNIVERSITY
Bengaluru, India

- This process has two subtasks:

1. Deciding how many categories to have: After the values of continuous attributes are sorted, they are then divided into 'n' intervals by specifying n-1 split points.
2. Determining how to map the values of the continuous attributes to these categories: All the values of one interval are mapped to the same categorical value.

Conclusion: Problem of discretization is on deciding how many split points to choose and where to place them. The result can be represented either as a set of intervals {(x0,x1),(x1,x2)……(xn-1,xn)},where, x0 and xn may be +∞ or -∞ respectively, or equivalently as a series of inequalities.

X0<x<=x1……xn-1 < x < xn.

- Discretization may be either:

1) Unsupervised (here, class information is not used)
2) Supervised (here , class information is used)

- Unsupervised discretization:
- Here class information is not used.
- Relatively simple approach.
- Ex: Equal width approach divides the range of the attribute into a user-specified number of intervals each having the same width.
- Disadvantage: Badly affected by outliers.
- Equal depth (equal frequency) approach tries to put same number of objects into each interval.
- Clustering method such as K-means.

- Supervised Discretization: Here the class information is used while constructing an interval.
- Conceptually, simple approach is to place the splits in a way that maximizes the purity of the intervals.
- Entropy based approaches:
- One of the most promising approaches to discretization and a simple approach is entropy.

- Definition of Entropy:

The entropy of the i^{th} interval

$$e_i = \sum_{i=1}^{k} P_{ij} \log_2 P_{ij}$$

where,

k – no. of different class labels

$P_{ij} = \frac{m_{ij}}{m_i}$ – is the probability (fraction of values) of class j in the i^{th} interval

m_i – no. of values in i^{th} interval of a partition

m_{ij} – no. of values of class j in interval i.

Total Entropy (e)

The total entropy of the partition is the weighed average of the individual interval entropies.

$$e = \sum_{i=1}^{n} w_i e_i$$

where,

$w_i = \frac{m_i}{m}$

m - no. of values

n - no. of intervals

- Entropy of an interval:
- It is a measure of the purity of an interval. If an interval contains only values of one class (is perfectly pure), then the entropy is 0 and it contributes nothing to the overall entropy if the classes of values in an interval occur equally often (the interval is an impure as possible), then the entropy is maximum.
- A simple approach for partitioning a continuous attribute starts by bisecting the initial values so that the resulting two intervals gives minimum entropy.
- The splitting process is then repeated with another interval, typically choosing the interval with the worst(highest) entropy, until a user specified no. of intervals is reached, or a stopping criterion is satisfied.

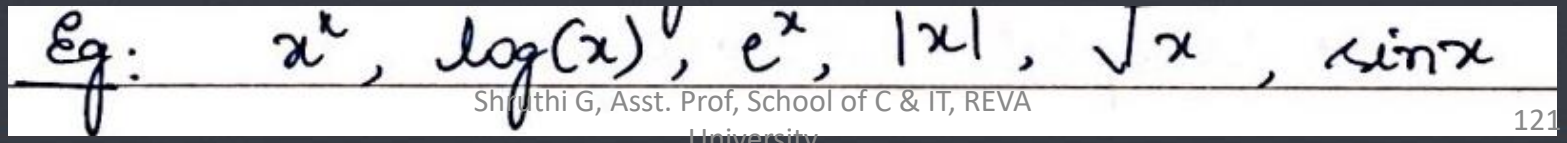
Variable Transformation

REVA UNIVERSITY
Bengaluru, India

- Refers to a transformation that is applied to all the values of a variable.
- Main methods are standardization and normalization
- One method is subtract the mean from each attribute and divide by standard deviation
  - (X - X_mean)/std_deviation
  - Standardization or normalization makes the entire set of values have a particular property.
- Two important types of variable transformation:

1. Simple functions:

- A simple mathematical function is applied to each value individually. If x is a variable, then ex of such transformation include:



2. Normalization or standardization: The goal is to make an entire set of values have a particular property.

Eg: "Standardizing a variable" in statistics if $\bar{x}$ is the mean (average) of the attribute values and S_x is their standard deviation, then the transformation $x' = \frac{x - \bar{x}}{S_x}$ creates a new variable that has mean of 0 and a standard deviation of 1.

Advantages: avoids a variable having large values dominate the results of calculation.

Disadvantage: Strongly affected by outliers

Solution: Mean is replaced by median(middle value)

Standard deviation is replace by absolute standard deviation.

Absolute std. dev. of x

$$\sigma_A = \sum_{i=1}^{m} |x_i - \mu|$$

x_i – i^{th} value of the variable

m – no. of objects

μ – mean or median

1.8 Measures of Similarity and dissimilarity

Similarity and Dissimilarity

- Similarity and dissimilarity are important because they are used by a number of data mining techniques, such as clustering, nearest neighbor classification and anomaly detection.
- In many cases, the initial data set is not needed once these similarities or dissimilarities have been computed.
- Such approaches can be viewed as transforming the data to a similarity(dissimilarity) space and then performing the analysis.
- Proximity is used to refer to either similarity or dissimilarity.
- Proximity between two objects is a function of the proximity between the corresponding attributes of the two objects.
- This includes measures such as correlation and Euclidean distance, which are useful for dense data such as time series or 2 dimensional points, as well as the Jaccard and cosine similarity measures, which are useful for sparse data like documents.

Similarity and Dissimilarity

REVA UNIVERSITY
Bengaluru, India

- Similarity
  - Similarity between 2 objects is a Numerical measure of the degree to which the 2 objects are alike .
  - It Is higher when objects are more alike.
  - Similarities are usually non-negative.
  - Often falls in the range [0,1],0 is no similarity, 1 is complete similarity.
- Dissimilarity
  - Numerical measure of the degree to which the 2 objects are different.
  - Lower when objects are more alike
  - The term distance is used as synonym for dissimilarity
  - Distance often used to refer to a special class of dissimilarities.
  - Sometimes it fall in the interval [0,1]
  - Minimum dissimilarity is often 0
  - Upper limit varies i.e..[0,∞]

Transformation

REVA UNIVERSITY
Bengaluru, India

- Transformations are often applied to convert a similarity to a dissimilarity or vice versa or a proximity measure to fall with in a particular range such as [0,1].
- For instance, we may have similarities that range from 1 to 10, but the particular algorithm or software package that we want to use may be designed to only work with dissimilarities , or it may only work with similarities in the interval [0,1].
- Frequently, proximity measures , especially similarities are defined or transformed to have values in the interval [0,1].
- Informally the motivation for this is to use a scale in which a proximity value indicates the fraction of similarity (or dissimilarity) between 2 objects. Such a transformation is often relatively straightforward.

- Ex: if the similarities between objects range from 1(not at all similar) to 10 (completely similar), we can make them fall within the range [0,1] by using transformation s'=(s-1)/9 where s and s' are the original and new similarity values.
- In general case the transformation of similarities to the interval [0,1] is given by the expression :

s'=(s-min_s)/(max_s-min_s) where max_s and min_s are the maximum and minimum similarity values.

- Likewise dissimilarity measures with a finite range can be mapped to the interval [0,1] by using formula:

d'=(d-min_d)/(max_d-min_d).

- There can be various complications in mapping proximity measures to the interval [0,1]:

1. The proximity measure takes values in the interval [0,∞] then non linear transformation is needed and values will not have the same relationship to one another on the new scale.

Ex: consider the transformation d'=d/(1+d) for a dissimilarity measure that ranges from 0 to ∞. The dissimilarities 0, 0.5, 2, 10, 100 and 1000 will be transformed into new dissimilarities 0, 0.33, 0.67, 0.90, 0.99 and 0.999. Larger values on the original dissimilarity scale are compressed into the range of values near 1, but whether or not this is desirable depends on application.

2. Meaning of the proximity measure may be changed.

Ex: correlation is a measure of similarity that takes values in the interval

[-1,1]. Mapping these values to interval[0,1] by taking absolute value loses information about sign, which can be important in some applications.

- Transforming similarities to dissimilarities and vice versa is also relatively straightforward, although we face the issues of preserving meaning and changing a linear scale into a non-linear scale.
- If the similarity(or dissimilarity) falls in the interval [0,1], then the dissimilarity can be defined as d=1-s or (s=1-d).
- Another simple approach is to define similarity as the negative of the dissimilarity (or vice versa).
- Ex: Dissimilarities 0,1,10 and 100 can be transformed into the similarities 0,-1,-10 and -100.

- The similarities resulting from the negation transformation are not restricted to the range [0,1] but if that is desired, then transformations such as s=1/(d+1), s=e to the power of –d or

  s=1-(d-min_d/(max_d-min_d)).

- Ex : The dissimilarities 0,1,10,100 then s=1/d+1 are transformed to 1, 0.5 , 0.09, 0.01.

s=e to the power of –d they become 1.00,0.37,0.00,0.00

s=1-(d-min_d/(max_d-min_d)) they become 1.00,0.99,0.00,0.00.

Similarity/Dissimilarity between Simple Attributes

- The proximity of objects with a number of attributes is defined by combining the proximities of individual attributes.
- Nominal attribute: It convey information about the distinctness of objects. In this case similarity defined as 1 if attribute values match and as 0 otherwise. A dissimilarity would be defined in the opposite way 0 if the attribute values match and 1 if they do not.
- Ordinal attribute: It gives the information about the order should be taken into account. Ex: consider an attribute that measures the quality of a product i.e. a candy bar on the scale{poor, fair, OK, good, wonderful}.If the product P1 is rated wonderful, would be closer to a product P2 which is rated good, than it would be to a product P3, which is rated OK. The values of the ordinal attribute are often mapped to successive integers beginning at 0 or 1 {poor=0,fair=1,OK=2,good=3,wonderful=4}. Then d(P1,P2)=3-2=1 or the dissimilarity to fall between 0 and 1, d(P1,P2)=3-2/4=0.25. A similarity defined as s=1-d.

- The definition of similarity(dissimilarity) for an ordinal attribute is uneasy since this assumes equal intervals and this is not.
- The difference between the values is probably not same. But in practice our options are limited and in the absence of more information., this is the standard approach for defining proximity between ordinal attributes.
- Interval and Ratio attribute: The measure of dissimilarity between two objects is the absolute difference of their values. Ex: compare our current weight and our weight a year ago. The similarity of interval or ratio attributes is typically expressed by transforming a similarity into a dissimilarity. The dissimilarity range from 0 to ∞ rather than 0 to 1.
- Table summarizes the discussion, x and y are two objects that have one attribute of the indicated type. Also d(x,y) and s(x,y) are the dissimilarity and similarity between x and y.

Similarity/Dissimilarity between Simple Attributes

REVA UNIVERSITY
Bengaluru, India

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Dissimilarities between data Objects

REVA UNIVERSITY
Bengaluru, India

- It is various kinds of dissimilarities.

Distances:

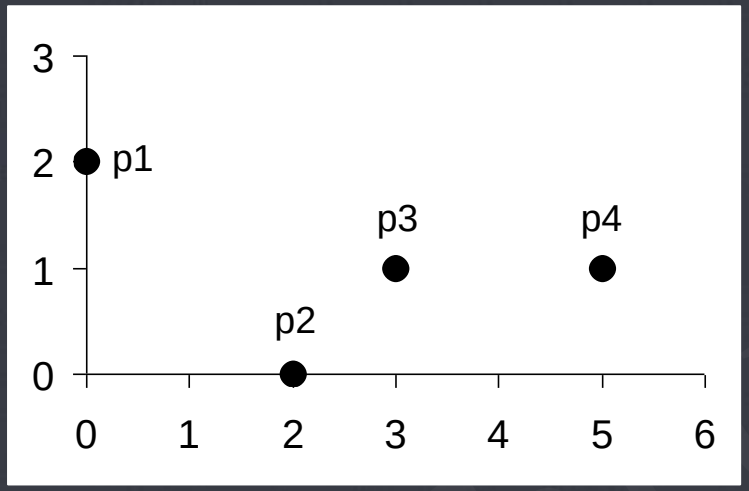
- Distances are dissimilarities with certain properties. The Euclidean Distance d between two points x & y, in 0ne-,two-,three- or higher dimensional space, is given by the following formula.

$$d(x,y) = \sqrt{\sum_{K=1}^{n} (x_k - y_k)^2}$$

Where n is the number of dimensions (attributes) and xk and yk are, respectively, the kth attributes (components) or data objects x and y.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance: Examples

- It is a generalization of Euclidean distance and it is given by:

$$d(x, y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

- Where r is a parameter.
- The following are the three most common examples of Minkowski distance:
- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance (L2 norm)

- r=∞ "Supremum" (L max norm or L∞ norm) distance.

This is the maximum difference between any attribute of the object. More formally L∞ distance is defined by

$$d(x, y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Common Properties of a Distance

REVA UNIVERSITY
Bengaluru, India

- Distances, such as the Euclidean distance, have some well known properties.
- If d(x,y) is the distance between two points x and y then the following properties hold:

1. Positivity:
   a) d(x,x)>=0 for all x and y
   b) d(x,y)=0 only if x=y

2. Symmetry:
   d(x, y) = d(y, x) for all x and y.

3. Triangle Inequality:
   $d(x, z) \leq d(x, y) + d(y, z)$ for all points x, y, and z. where d(x, y) is the distance (dissimilarity) between points (data objects), x and y.

- A distance that satisfies these properties is a metric

Similarities between data objects

REVA UNIVERSITY
Bengaluru, India

- Similarities, also have some well known properties.
- If s(x,y) is the similarity between points x and y then the typical properties of similarities are:

1. *s(x, y) = 1* (or maximum similarity) only if *x = y (0<=s<=1)*.
2. *s(x, y) = s(y, x)* for all *x* and *y*. (Symmetry)

where *s(x, y)* is the similarity between points (data objects), *x* and *y*.

Examples of Proximity Measures

REVA UNIVERSITY
Bengaluru, India

- Similarity measures between objects that contain only binary attributes are called similarity coefficients, and typically have values between 0 and 1.
- A value of 1 indicates that the two objects are completely similar, while a value of 0 indicates that the objects are not at all similar.
- Let x and y be two objects that consists of n binary attributes. The comparison of two objects, i.e., two binary vectors leads to the following four quantities(frequencies).

Similarity Between Binary Vectors

- Common situation is that objects, *x* and *y*, have only binary attributes
- Compute similarities using the following quantities

  f_{01} = the number of attributes where x was 0 and y was 1
  f_{10} = the number of attributes where x was 1 and y was 0
  f_{00} = the number of attributes where x was 0 and y was 0
  f_{11} = the number of attributes where x was 1 and y was 1

- Examples of some similarity and dissimilarity(proximity).

1. Simple Matching Coefficient(SMC)
2. Jaccard coefficient
3. Cosine Similarity
4. Extended Jaccard Coefficient(Tanimato Coefficient)
5. Correlation
6. Bregman Divergence

Simple Matching Coefficients (SMC)

- One commonly used similarity coefficient is the SMC, which is defined as:

SMC = number of matching attribute values / number of attributes

$$SMC = (f11 + f00) / (f01 + f10 + f11 + f00)$$

- This measure counts both presence and absence equally.
- The SMC could be used to find students who had answered questions similarly on a test that consisted only of true/false questions.

Jaccard Coefficients (J)

- Jaccard coefficient is frequently used to handle objects consisting of asymmetric binary attributes.
- J = number of matching presences/ number of attributes not involved in 00 matches

  = (f11) / (f01 + f10 + f11)

SMC versus Jaccard: Example

REVA UNIVERSITY
Bengaluru, India

p = 1 0 0 0 0 0 0 0 0 0 (two binary vectors)
q = 0 0 0 0 0 0 1 0 0 1

f_{01} = 2 (the number of attributes where x was 0 and y was 1)
f_{10} = 1 (the number of attributes where x was 1 and y was 0)
f_{00} = 7 (the number of attributes where x was 0 and y was 0)
f_{11} = 0 (the number of attributes where x was 1 and y was 1)

SMC = $(f_{11} + f_{00})/(f_{01} + f_{10} + f_{11} + f_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (f_{11}) / $(f_{01} + f_{10} + f_{11})$ = 0 / (2 + 1 + 0) = 0

Cosine Similarity

REVA UNIVERSITY
Bengaluru, India

- Documents are often represented as vectors, where each attribute represents the frequency with which a particular term(word) occurs in the document.
- Documents have thousands or tens of thousands of attributes(terms) each document is sparse since it has few non zero attributes.
- With transaction data, similarity should not depend on the number of shared 0 values since any two documents are likely to "not contain" many of these words, and therefore if 0-0 matches are counted, most documents will be highly similar to most other documents.
- Therefore a similarity measure for documents needs to ignore 0-0 matches and also must be able to handle non-binary vectors.
- The cosine similarity is defined as one of the most common measure of document similarity.

Cosine Similarity

REVA UNIVERSITY
Bengaluru, India

- If x and y are two document vectors ,then

$\cos(x, y) = (x \bullet y) / ||x|| \, ||y||$,

where • indicates vector dot product and $||x||$ is the length of vector x.

$$x \cdot y = \sum_{k=1}^{n} x_k y_k$$

$$||x|| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x \cdot x}$$

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$x \bullet y$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

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

$||y||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

$\cos(x,y)$ = 0.3150

Find $\cos(x, y)$ for the following problems

1. $x = (0, 1, 0, 1)$ $y = (1, 0, 1, 0)$

   Ans: $\cos(x, y) = 0$

2. $x = (0, -1, 0, 1)$ $y = (1, 0, -1, 0)$

   Ans: $\cos(x, y) = 0$

3. $x = (1, 1, 0, 1, 0, 1)$ $y = (1, 1, 1, 0, 0, 1)$

   Ans: $\cos(x, y) = \frac{3}{4}$

4. $x = (2, -1, 0, 2, 0, -3)$ $y = (-1, 0, -1, 0, 0, -1)$

   Ans: $\cos(x, y) = 0$

Extended Jaccard Coefficient (Tanimoto)(EJ)

- The extended Jaccard coefficient can be used for document data and that reduces to the Jaccard coefficient in the case of binary attributes.
- The extended Jaccard coefficient is also known as the Tanimoto coefficient.

$$EJ(x, y) = \frac{x \cdot y}{\|x\|^2 + \|y\|^2 - x \cdot y}$$

Correlation

- The correlation between two data objects that have binary or continuous variables is a measure of the linear relationship between the attributes of the objects.
- Pearson's correlation coefficient between two data objects, x and y is defined by the following equation:

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}$$

- Correlation measures the linear relationship between objects

We are using the following standard statistical notation and definitions:

$$\text{covariance}(\mathbf{x},\mathbf{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$$

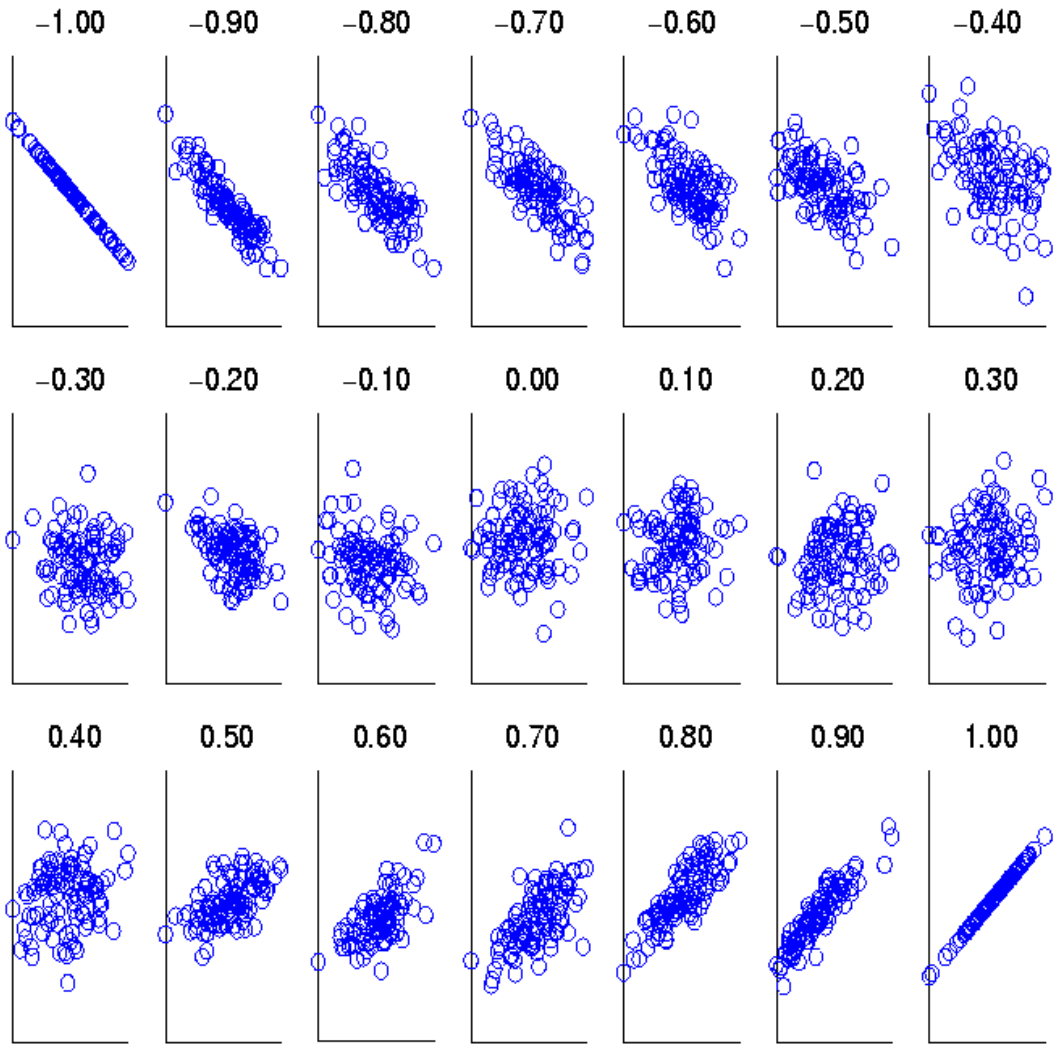
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

5. Find Corr(x, y) for the following
   x = (1, 1, 1, 1) y = (2, 2, 2, 2)
   Ans: Corr(x, y) = 0/0
   Euclidean (x, y) = 2

6. x = (0, 1, 0, 1) y = (1, 0, 1, 0)
   Ans: Corr(x, y) = -1
   Euclidean(x, y) = 2

7. x = (0, -1, 0, 1) y = (1, 0, -1, 0)
   Ans: Corr(x, y) = 0
   Euclidean(x, y) = 2

8. x = (1, 1, 0, 1, 0, 1) y = (1, 1, 1, 0, 0, 1)
   Ans: Corr(x, y) = 0.25
   Euclidean(x, y) = 0.75

9. x = (2, -1, 0, 2, 0, -3) y = (-1, 1, -1, 0, 0, -1)
   Ans: Corr(x, y) = 0
   Cos(x, y) = 0

Bregman Divergence

REVA UNIVERSITY
Bengaluru, India

- It is possible to construct general data mining algorithms, such as clustering algorithms, that work with any Bregman divergence. Ex : K-means clustering algorithm.
- Bregman Divergence are loss or distortion functions.
- $D(x,y)=\phi(x)-\phi(y)-(\nabla\phi(y),(x-y))$
- Where $\nabla\phi(y)$is the gradient of ϕ evaluated at y
- x-y is the vector difference between x and y
- $(\nabla\phi(y),(x-y))$ is the inner product between $\nabla\phi(x)$ and (x-y).
- For points in Euclidean space, the inner product is just the dot product.
- D(x,y) can be written as $D(x,y)=\phi(x)-L(x)$, where $L(x)=\phi(y)+(\nabla\phi(x),(x-y))$.

Example Problems

① The SMC and Jaccard Similarity Coefficients

Calculate SMC and J for the following binary vectors.

$x = (1,0,0,0,0,0,0,0,0,0)$

$y = (0,0,0,0,0,0,1,0,0,1)$

ls:

$f_{01} = 2$

$f_{10} = 1$

$f_{00} = 7$

$f_{11} = 0$

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = \frac{0+7}{2+1+0+7} = 0.7$$

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2+1+0} = 0.$$

(2) **Cosine Similarity**

Calculate the cosine similarity for the following two data objects, which represent document vectors.

$x = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$

$y = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$

$x \cdot y = 3*1 + 2*0 + 0 + 0 + 0 + 0 + 0 + 2*1 + 0 + 0$

$= 3 + 2$

$x \cdot y = 5$

$\|x\| = \sqrt{3*3 + 2*2 + 0 + 5*5 + 0 + 0 + 0 + 2*2 + 0 + 0}$

$= \sqrt{9 + 4 + 25 + 4}$

$= \sqrt{42}$

$= 6.48$

$\|y\| = \sqrt{1*1 + 0 + 0 + 0 + 0 + 0 + 0 + 1*1 + 0 + 2*2}$

$= \sqrt{1 + 1 + 4}$

$= \sqrt{6}$

$\|y\| = 2.24$

$\cos(x, y) = \frac{5}{6.48 * 2.24} = \frac{5}{14.51}$

$\boxed{\cos(x, y) = 0.31}$

(3) (a) Given $x = (-3, 6, 0, 3, -6)$

$y = (1, -2, 0, -1, 2)$

find $\text{Cov}(x, y)$

Soln:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = 0 \quad , \quad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = 0$$

$$\bar{x} = \frac{(-3+6+0+3+(-6))}{5} \quad , \quad \bar{y} = \frac{1+(-2)+0+(-1)+2}{5}$$

$$\bar{x} = 0 \qquad \bar{y} = 0$$

$$S_{xy} = \frac{1}{n-1}\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$= \frac{1}{5-1}\left[(-3)(1) + (6)(-2) + 0 + 3(-1) + (-6)(2)\right]$$

$$= \frac{1}{4}\left[-3 - 12 - 3 - 12\right]$$

$$S_{xy} = \frac{-15}{2}$$

$$S_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$= \sqrt{\frac{1}{4}(9+36+0+9+36)}$$

$$S_x = \sqrt{\frac{45}{2}}$$

$$S_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$= \sqrt{\frac{1}{4}(1+4+0+1+4)}$$

$$S_y = \sqrt{\frac{5}{2}}$$

$$Corr(x,y) = \frac{S_{xy}}{S_x \cdot S_y} = \frac{-15}{2} \Big/ \sqrt{\frac{45}{2}} \cdot \sqrt{\frac{5}{2}}$$

$$= \frac{-15 \cdot \sqrt{2} \cdot \sqrt{2}}{2 \cdot \sqrt{45} \cdot \sqrt{5}}$$

$$\boxed{Corr(x,y) = -1}$$

(6) $x = (1,1,1,1)$
$y = (2,2,2,2)$
find corr(x,y)

$\bar{x} = \frac{(1+1+1+1)}{4} = \frac{4}{4} = 1$

$\bar{y} = \frac{(2+2+2+2)}{4} = \frac{8}{4} = 2$

$S_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$

$= \frac{1}{4-1} [(1)(2) + (1)(2) + (1)(2) + (1)(2)]$

$= \left[\frac{1}{3}[0+0+0+0]\right] \quad \frac{1}{3}(0)$

$= \left[\frac{1}{3}[0] = \frac{0}{3}\right] \quad \frac{0}{3}$

$S_{xy} = 0$ ✓

$S_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$

$= \sqrt{\frac{1}{3}(1+1+1+1)}$

$= \sqrt{\frac{1}{3}(4)} = \sqrt{\frac{4}{3}}$

$S_x = 0.66$

$S_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$

$= \sqrt{\frac{1}{3}(0)}$

$S_x = 0$

$S_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2}$

$= \sqrt{\frac{1}{3}[(4-4) + (4-4) + (4-4) + (4-4)]}$

$= \frac{1}{3}(0)$

$S_y = 0$

$S_{xy} = \frac{2.66}{0.66 * 0} = \frac{0}{0} = 0$

Problems on Euclidean distance

Find Euclidean(x,y) for the following values

① $x = (1,1,1,1)$ $y = (2,2,2,2)$

$$d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2 + (x_4-y_4)^2}$$

$$= \sqrt{(1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2}$$

$$= \sqrt{4}$$

$$d(x,y) = 2$$

② $x = (0,1,0,1)$ $y = (1,0,1,0)$

$$d(x,y) = \sqrt{(-1)^2 + (1)^2 + (-1)^2 + (1)^2}$$

$$= \sqrt{4}$$

$$d(x,y) = 2$$

1.9 Data Mining Applications

- Data mining is widely used in diverse areas.
- Data Mining Applications:
- Here is the list of areas where data mining is widely used:

1) Financial Data Analysis
2) Retail Industry
3) Telecommunication Industry
4) Biological Data Analysis
5) Other Scientific Applications
6) Intrusion Detection

Financial Data Analysis:

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows –

1) Design and construction of data warehouses for multidimensional data analysis and data mining.
2) Loan payment prediction and customer credit policy analysis.
3) Classification and clustering of customers for targeted marketing.
4) Detection of money laundering and other financial crimes.

Retail Industry:

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry –

1) Design and Construction of data warehouses based on the benefits of data mining.
2) Multidimensional analysis of sales, customers, products, time and region.
3) Analysis of effectiveness of sales campaigns.
4) Customer Retention.
5) Product recommendation and cross-referencing of items.

Telecommunication Industry:

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services –

1) Multidimensional Analysis of Telecommunication data.
2) Fraudulent pattern analysis.
3) Identification of unusual patterns.
4) Multidimensional association and sequential patterns analysis.
5) Mobile Telecommunication services.
6) Use of visualization tools in telecommunication data analysis.

Biological Data Analysis:

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis –

1) Semantic integration of heterogeneous, distributed genomic and proteomic databases.
2) Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
3) Discovery of structural patterns and analysis of genetic networks and protein pathways.
4) Association and path analysis.
5) Visualization tools in genetic data analysis.

Other Scientific Applications

Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications –

1) Data Warehouses and data preprocessing.
2) Graph-based mining.
3) Visualization and domain specific knowledge.

Intrusion Detection

- Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection –

1) Development of data mining algorithm for intrusion detection.
2) Association and correlation analysis, aggregation to help select and build discriminating attributes.
3) Analysis of Stream data.
4) Distributed data mining.
5) Visualization and query tools.

1.10 Visualization

- Visualization:
- We will learn different data Visualization techniques.
- Data: Data is a collection of facts such as numbers, words, measurements, observations or even just description of things.
- Data can be qualitative and quantitative.
- Qualitative data is descriptive information(it describes something)
- Quantitative data is numerical information (numbers).

1. Discrete: only take certain values(like whole numbers),this is counted

2. Continuous: take only value(within a range), this is measured.

Definition of Visualization:

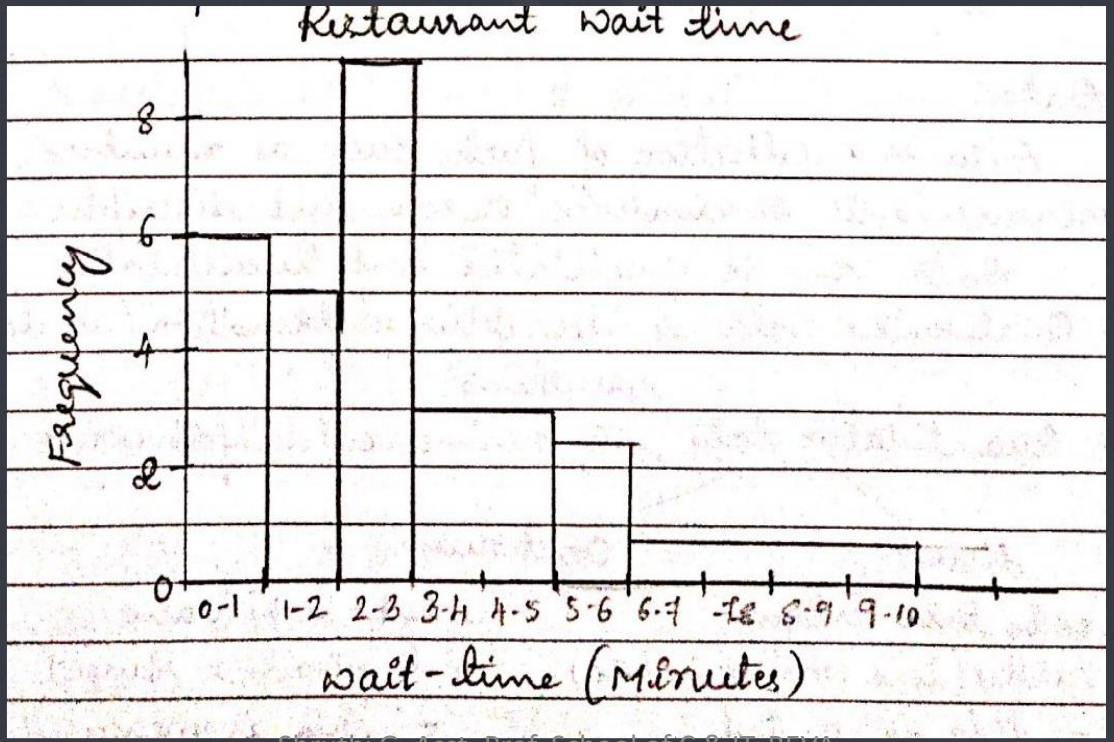
- Visualization is the conversion of data into a visual or tabular format, so that the characteristics of the data and the relationships among data items or attributes can be analysed.
- Visualization is most powerful and appealing techniques for data exploration.

Visualization Techniques:

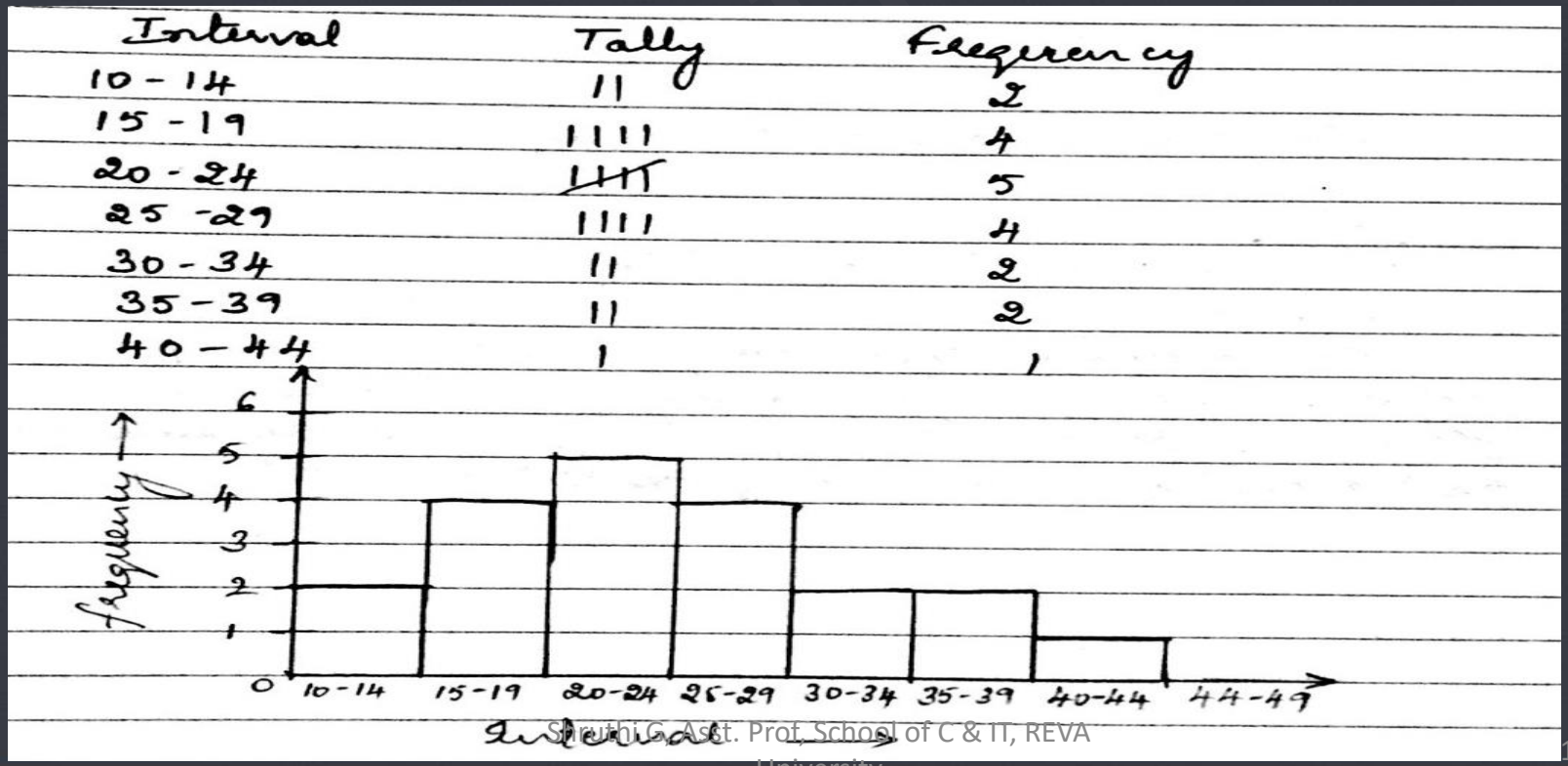
1. Histograms
2. Two-dimension Histograms
3. Box plots
4. Scatter plots
5. Contour plots
6. Matrix Plots(for higher-dimensional data)

Histograms

- A histogram is bar graph that represents a frequency distribution. The width represents the interval and the height represents the corresponding frequency. There are no spaces between the bars.



- Construct a histogram for the following details of average gas mileage of twenty cars.
- Average Gas Mileage(in miles/gallon)
- 24,17,14,22,25,26,38,42,24,12,28,19,32,21,35,28,31, 21,18,19



Two-dimension Histograms

REVA UNIVERSITY
Bengaluru, India

- Each attribute is divided into intervals and the two sets of intervals define two-dimensional rectangles of values.
- Two-dimensional histograms can be used to discover interesting facts about how the values of two attributes co-occur, they are visually more complicated.

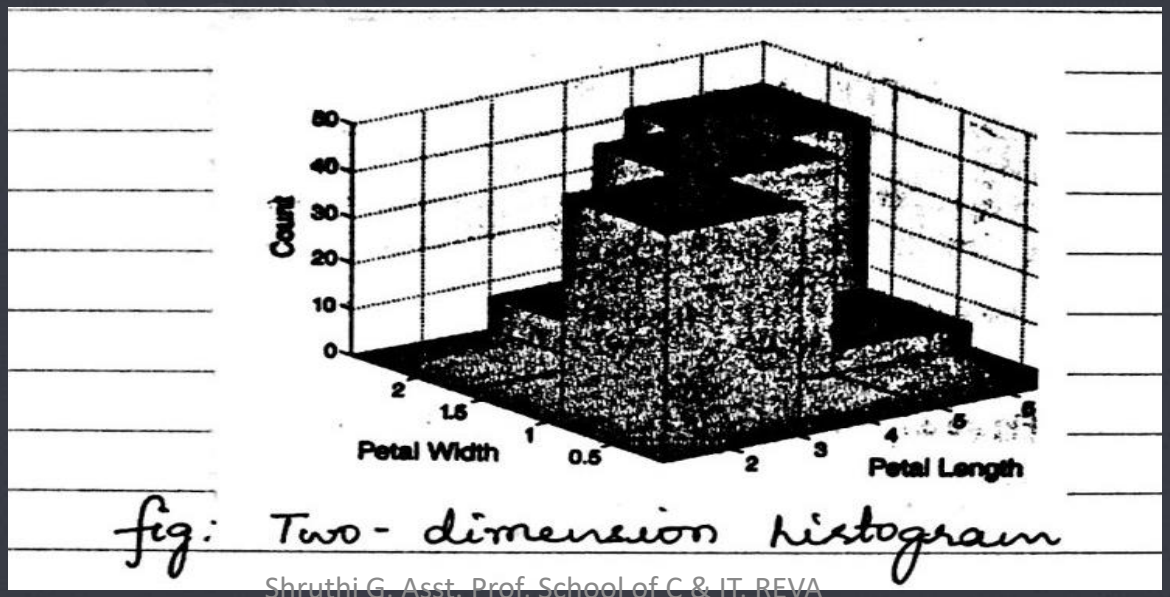


fig: Two-dimension histogram

Box plots

- There are another method for showing the distribution of the values of a single numerical attribute.
- The following figure shows a labelled box plot for sepal length. The lower ends of the box indicate the 25th and 75th percentiles respectively, while the line inside the box indicates the values of the 50th percentile.

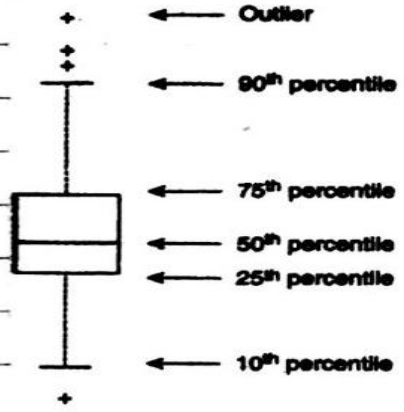


Figure 3.10. Description of box plot for sepal length.

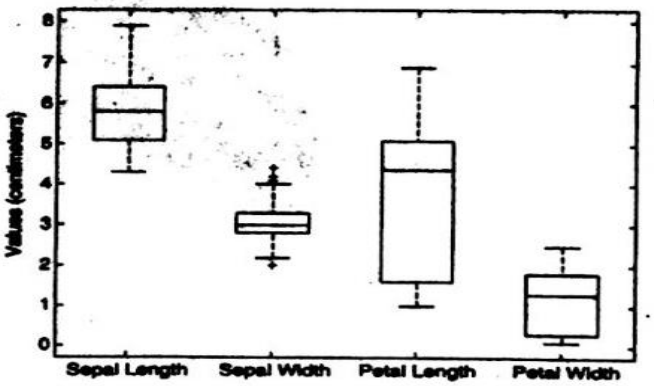
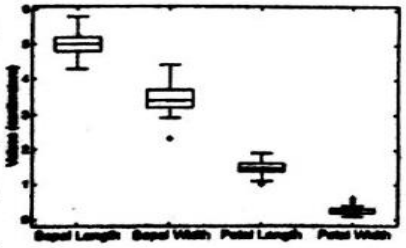
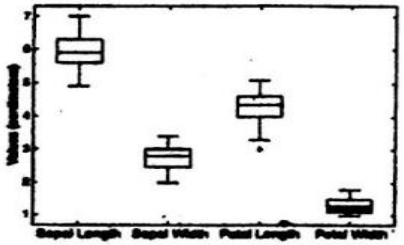


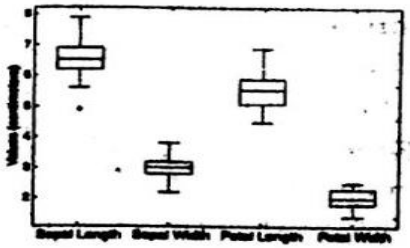
Figure 3.11. Box plot for iris attributes.



(a) Setosa.



(b) Versicolour.



(c) Virginica.

The top and bottom lines of the tails indicate the 90th and 10th percentiles respectively. Outliers are shown by '+' marks.

Pie Chart

- A pie chart is similar to a histogram, but is typically used with categorical attributes that have relatively small number of values.

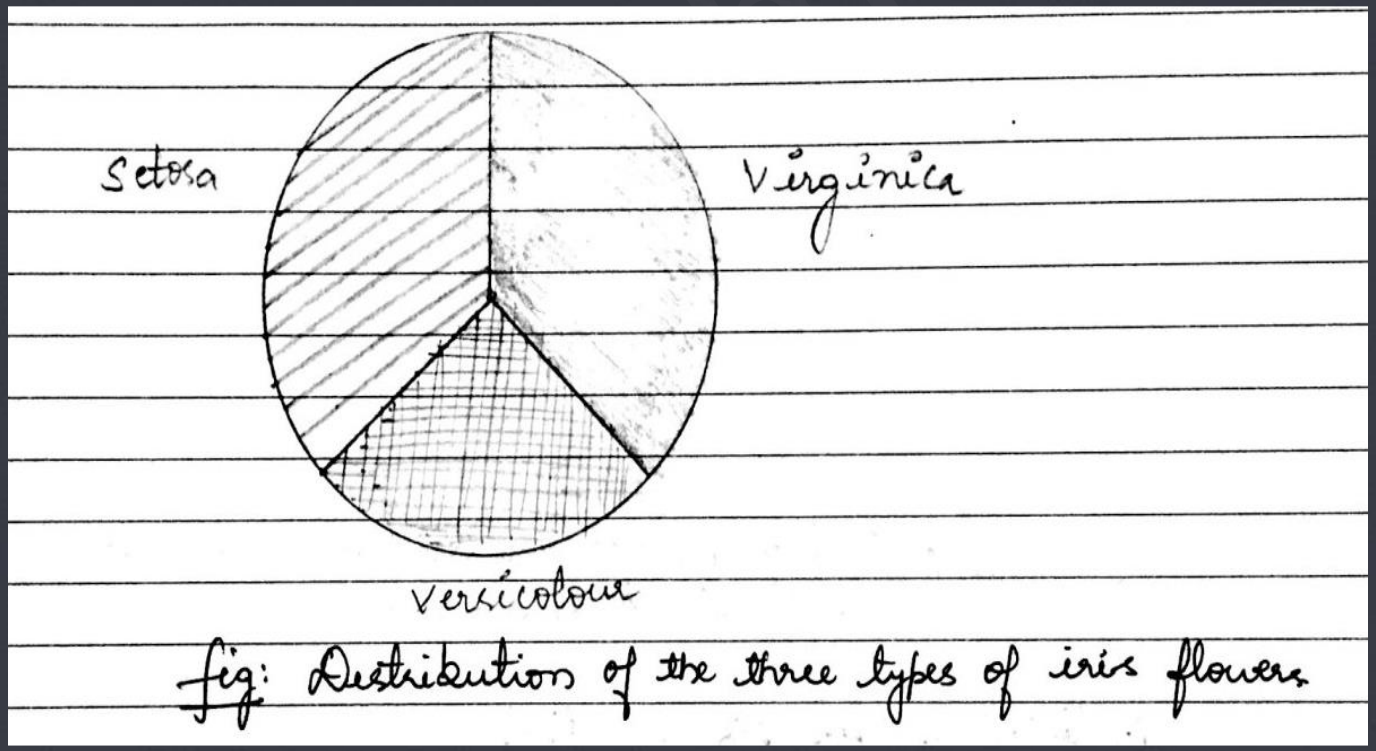
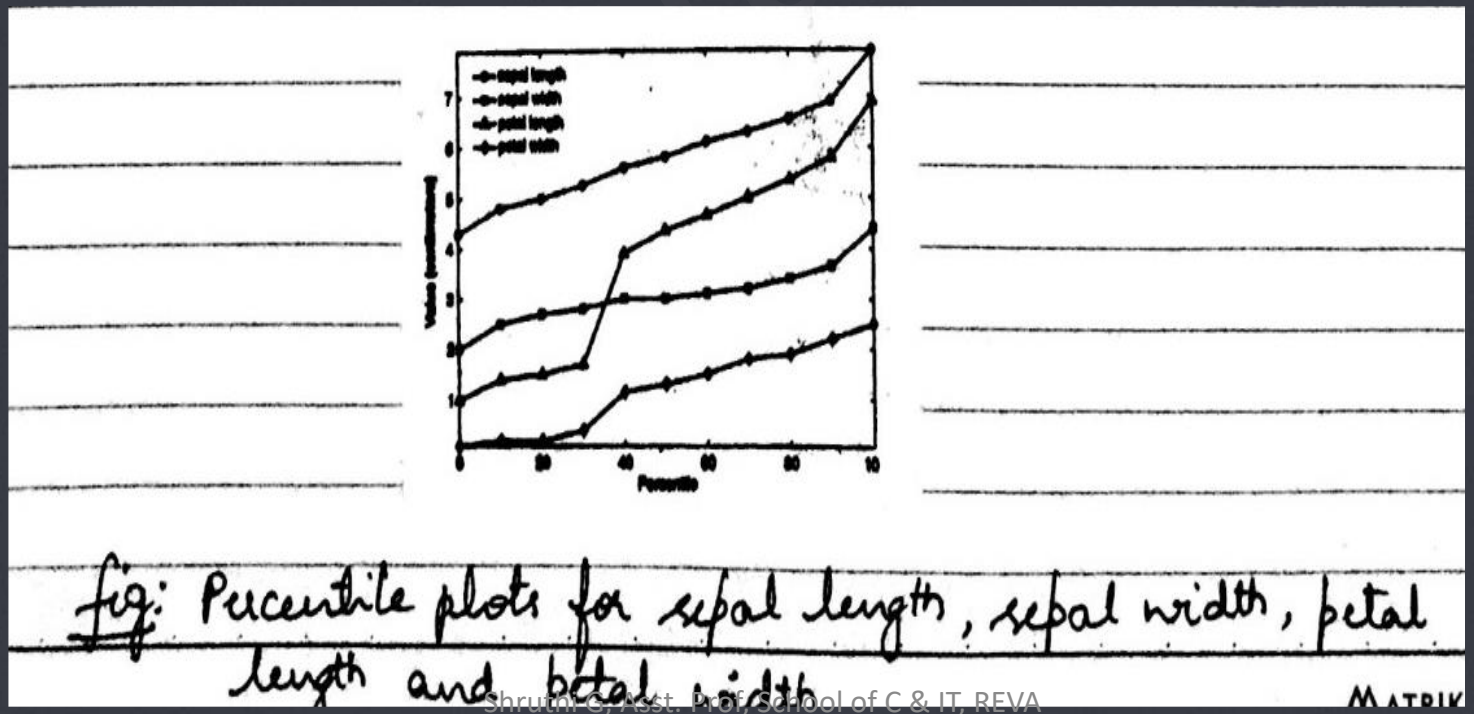


fig: Destribution of the three types of iris flowers.

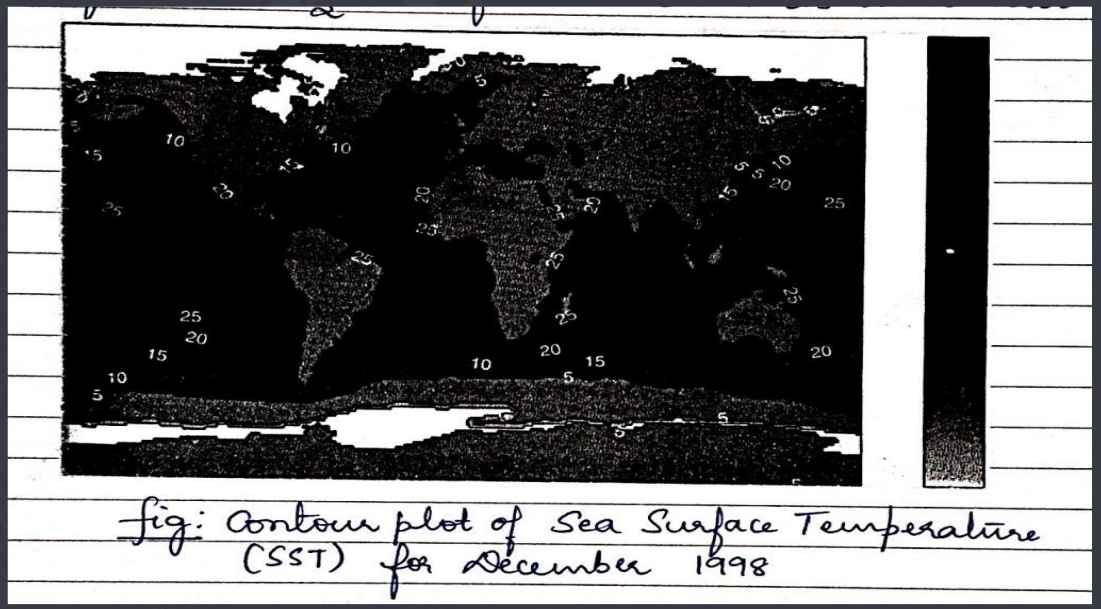
Scatter plots

- Scatter plots are used to illustrate linear correlation. Each data object is plotted as a point in the plane using the values of the two attributes as x and y coordinates. It is assumed that the attributes are either integer or real-valued.



Contour plots

- For some three-dimensional data, two attributes specify a position in a plane, while the third has a continuous value, such as temperature or elevation. A useful visualization for such data is a contour plot.



Motivation for visualization

- People can quickly absorb large amounts of visual information and find patterns in it.
- Helps in deciding which attributes contain useful information.

Iris Dataset

- Iris data set is available from the University of California at Irvine(UCI) Machine Learning Repository at http://www.ics.uci.edu/~mlearn.
- It consists of information on 150 Iris flowers, 50 each from one of three Iris species:
- 1.Setosa
- 2. Versicolor
- 3.Virginica
- In addition to the species of a flower, this data set contain 5 other attributes: sepal width in centimeters, sepal length in centimeters, petal length in centimeters ,petal width in centimeters and class(Setosa, Versicolor, Virginica).

Thank You