UNIT - 1 INTRODUCTION

What is Data Mining?

Data Mining is the process of automatically discovering useful information in large repositories.

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 analyzing new types of data and for analyzing old types of data in new ways.

Information Retrieval

It is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Data Mining and Knowledge Discovery

Data Mining is an integral part of Knowledge Discovery in Databases (KDD), which is the overall process of converting raw data into useful information.

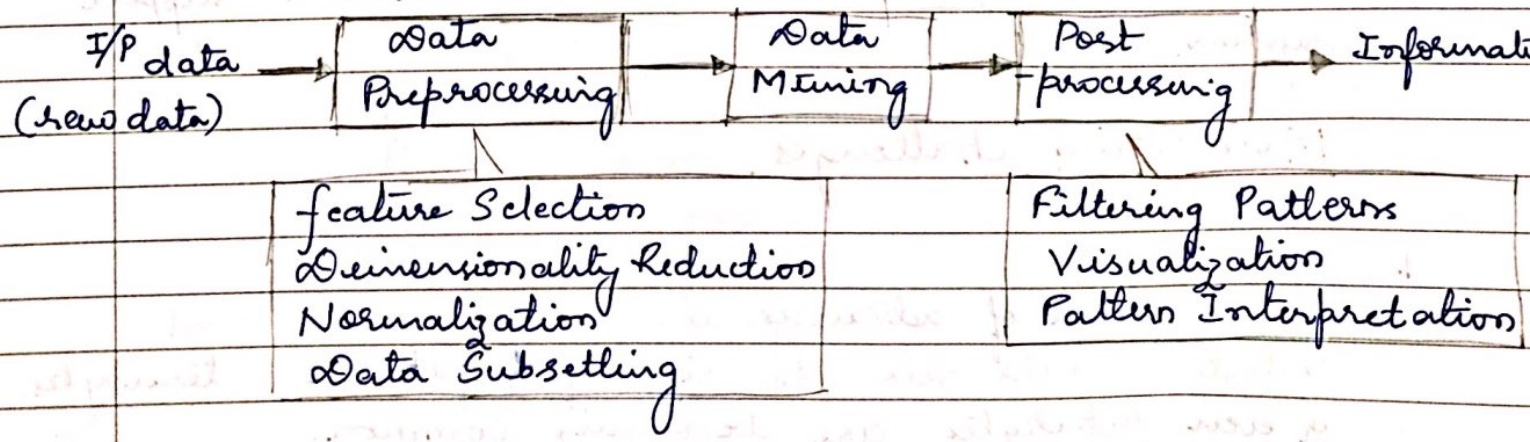


fig: The process of KDD

Input data:

It can be stored in a variety of formats
- flat files
- spread sheets
- relational tables

and may reside in a centralized data repository or be distributed across multiple sites.

Data Preprocessing

The purpose of preprocessing is to transform the raw input data into an appropriate format for subsequent analysis.

The steps involved in data preprocessing includes
- fusing data from multiple sources
- cleaning data to remove noise and duplicate observations
- selecting records and features that are relevant to the datamining task.

Data preprocessing is the most time-consuming step in the overall knowledge discovery process

Postprocessing

This step ensures that only valid & useful results are incorporated into the decision support system.

Motivating challenges

1. **Scalability:**

Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes or even petabytes are becoming common.

If the datamining algorithms are to handle these massive data sets, then they must be scalable.

- Scalability may require special search strategies to handle exponential search problems
- Scalability also require the implementation of new data structure to access individual records in an efficient manner
- Scalability can be improved by using sampling or developing parallel and distributed algorithms

2. High Dimensionality:

Dimensionality (features).

Datasets with temporal and spatial components tend to have high dimentionality.

High dimensionality / high dimensional data would simply mean high number of features or independent variables.

3. Heterogeneous and Complex Data

Traditional data analysis methods often deals with datasets 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.

Eg: of non-traditional types of data

- collection of web pages containing semi-structured text and hyperlinks
- DNA data with sequential and three-dimensional structure
- climate data that consists of time series measurement (temperature, pressure, etc.) at various locations on the Earth's surface.

4. Data ownership and Distribution

Sometimes, the 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 techniques are

1) How to reduce the amount of communication needed to perform the distributed computation.
2) How to effectively consolidate the data mining results obtaining from multiple sources.
3) How to address data security issues.

5. Non-traditional analysis

The traditional statistical approach is based on a hypothesize-and-test paradigm.

In other words, a hypothesis is proposed, an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis.

The current data analysis tasks often require the generation and evaluation of thousands of hypothes and consequently, the development of some data mini techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.

The Origins of Data Mining

Data Mining draws upon ideas from

- Sampling, estimation and hypothesis test from statistics
- Search algorithms, modeling techniques, machine learning, learning theories from AI pattern recognition, statistics database systems.

UNIT - 1 INTRODUCTION

What is Data Mining?

Data Mining is the process of automatically discovering useful information in large repositories.

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 analyzing new types of data and for analyzing old types of data in new ways.

Information Retrieval

It is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Data Mining and Knowledge Discovery

Data Mining is an integral part of Knowledge Discovery in Databases (KDD), which is the overall process of converting raw data into useful information.

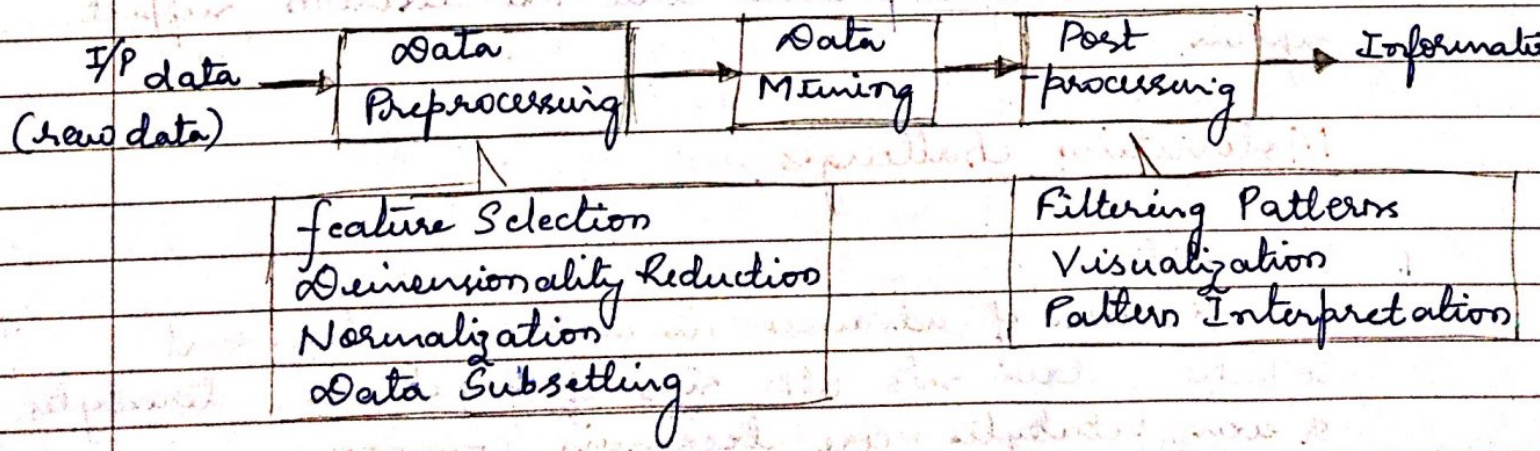


fig: The process of KDD

Traditional techniques may be unsuitable due to
- Enormity of data
- High dimensionality of data
- Heterogeneous nature of data

Data Mining also had been quickly to adopt ideas from other areas including
- optimization
- Evolutionary Computing
- Signal processing
- Information Theory

Database systems are needed to provide support for efficient storage, indexing, query processing

The parallel computing and distributed technology are two major data addressing issues in data mining to increase the performance

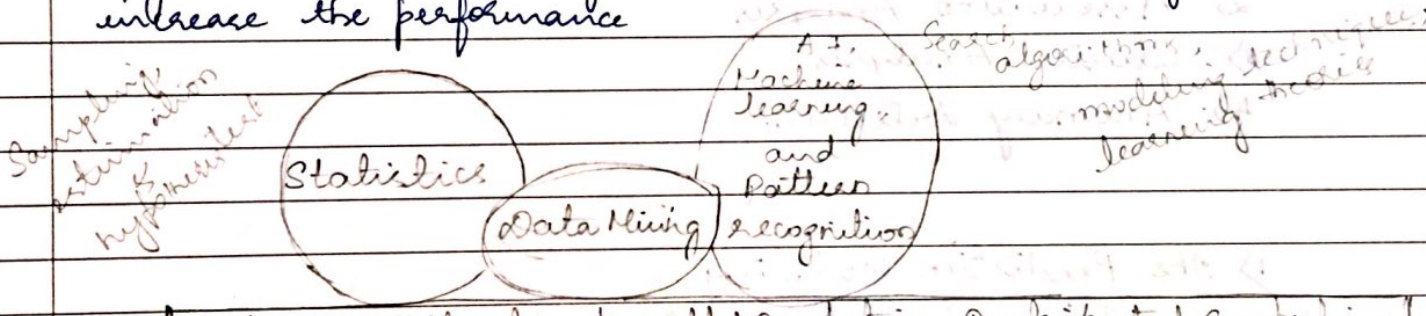


fig: Datamining as a confluence of many disciplines

Data Mining Tasks

Data Mining tasks are generally divided in two major categories

1) Predictive Tasks:

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attributes to be predicted is commonly

known as the target or dependent variable, while the attributes used for making the prediction are known as the exploratory or independent variables

2) Descriptive tasks

The objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the relationships in data.

Descriptive data mining tasks are often exploratory in nature and frequently require post processing techniques to validate and explain the results.

Core Data Mining Tasks

Four of the core data mining tasks are

1) Predictive modeling
2) Association Analysis
3) cluster Analysis
4) Anomaly Detection.

1) The Predictive Modeling

This refers to the task of building a model for the target variable as a function of the explanatory variable

The goal is to learn a model that minimizes the error between the predicted and true values of the target variable.

There are two types of predictive modeling tasks

i) classification: used for discrete target variables

Eg:- web user will make purchase at an online bookstore is a classification task, because the target variable is a binary valued.

ii) Regression: used for continuous target variables
Eg:- forecasting the future price of a stock is regression task because price is a continuous value attribute.

2) Association Analysis

This is used to discover patterns that describe strongly associated features in the data.

The goal is to extract the most interesting patterns in an efficient manner.

Useful applications include,
- finding groups of genes that have related functionality or
- identifying web pages that are accessed together

3) Cluster Analysis

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 belong to other clusters.

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.

4) Anomaly Detection / outlier detection

This 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 is to
- discover the real anomalies and
- avoid falsely labeling normal objects as anomalous

Applications include the detection of fraud, network intrusions, and unusual patterns of diseases.

Data

The Types of Data

* A data-set refers to a collection of data-objects and their attributes.
* Other names for a data-object are
  - record, transactions, vector, event, entity, sample or observation
* Data objects are described by a number of attributes such as
  - mass of a physical object or
  - time at which an event occured
* other names for an attribute are
  - dimension, variable; field, feature or characteristic

What is an Attribute?

An attribute is a property or characteristics of an object that may vary; either from one object to another or from one time to another.

1. Eg:- Eye colour varies from person to person

Eye colour is a symbolic attribute with a small number of possible values {brown, black, blue, green}.

2. Eg:- Student information

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

A dataset is a file, in which the objects are records (or rows) in the file and each field (or column) corresponds to an attribute.

In the above example, each row corresponds to a student and each column is an attribute that describes some aspect of a student, such as GPA, or ID.

Properties of Attribute values

* The type of an attribute depends on which of the following properties it possesses.
  1. Distinctness = and ≠
  2. Order <, ≤, >, and ≥
  3. Addition + and −
  4. Multiplication * and /

* Nominal attributes: Uses only distinctness
  Eg: ID numbers, eye color, pincodes

* Ordinal attributes: Uses distinct and order
  Eg: Grades in {SC, FC, FCD}
  Shirt size in {S, M, L, XL}

* Interval attribute: Uses distinctness, order & addition
  Eg: Calendar dates
  Temperature in Celsius or Fahrenheit

* Ratio attribute: Uses all four properties
  Eg: Temperature in Kelvin, length, time, counts

Table: Different attribute types

Attribute Type		Description	Examples	Operations
categorical (qualitative)	Nominal	The values of a nominal attribute are just different names i.e., nominal values provide only enough information to distinguish one object from another ($=, \neq$)	Zip codes, Employee ID, eye color, gender	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, percentile, rank correlation, run tests, sign tests
Numeric (quantitative)	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

Date No.

An independent way of distinguishing between attributes is by the number of values they can take.

1. Discrete:

* A discrete attribute has a finite or countably infinite set of values.

Eg: ZIP codes, ID numbers etc.

They are often represented as integer variables

* Binary attributes are special case of discrete attributes and assume only two values.

Eg: true/false, yes/no...

2. Continuous:

A continuous attribute is one whose values are real numbers

Eg: temperature, height or weight

They are typically represented as floating point variab

General characteristics of Data Sets

1. Dimensionality

* Dimensionality of a dataset is the number of attributes that the objects in the dataset possess
* Difficulties associated with analysing high-dimensio data are referred to as the curse of dimensionality

2. Sparsity:

For some datasets, most attributes of an object hav values of 0: fewer than 1% of the entries are non-zero. This is called sparsity.

3. Resolution:

It is frequently possible to obtain data of differen levels of resolutions, and often the properties of the

data are different at different resolutions (Eg:- Surface of earth).

Types of Data Sets

1. Record - Data
   - Transaction Data
   - Data Matrix
   - Sparse Data Matrix

2. Graph - Based Data
   - Data with relationships among objects
   - Data with objects that are graphs

3. Ordered Data
   - Sequential Data
   - Time - Series Data
   - Spatial Data
   - (genetic) Sequence Data

Record Data

* It is a data that consists of a collection of records, each of which consists of a fixed set of attributes
* Different types of record data are

i) Transaction / Market - Basket Data
- Here each record involves a set of items
- Each record is called a transaction
- Eg:- Consider a grocery store, the set of products purchased by a customer during one shopping trip constitute a tradction, while the individual products that were purchased are the items.

ii) Data Matrix
- If data objects have the same fixed set of numeric attributes, then the data objects can be

thought of as points in a multidimensional space, where each dimension represents a distinct attribute.

- Such datasets can be represented by an m x n matrix where,
  m → rows for each object
  n → columns for each attribute

- These matrices are called Data Matrix / Pattern matrix.

iii) Space Data Matrix

- It is a special type of data matrix in which the attributes are of the same type and are assymetric i.e., only non-zero values are important
  Eg: ① is transaction data that has only 0-1 entries ② Document Data

Each document becomes a 'term' vector
- Each term is a component (attribute) of the vector.
- Value of each component is the no. of terms the corresponding terms occurs in the document

Tid	Refund	Marital Status	Taxable income	Defaulted Borrower
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

(a) Record Data

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

(b) Transaction Data

projection of x load	projection of y load	Distance	load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data Matrix

	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

(d) Documentation - term Matrix

fig: Different variations of record data

Graph - Based Data

* Data is represented as Graph
There are two specific cases:
<1> The graph captures relationships among data objects
<2> The data objects themselves are represented as graphs

(i) Data with relationships among objects

The data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links between objects

Eg: web pages on WWW

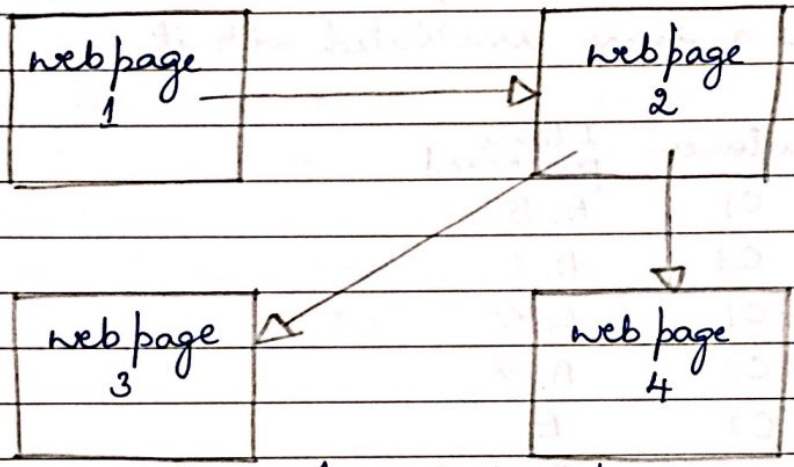
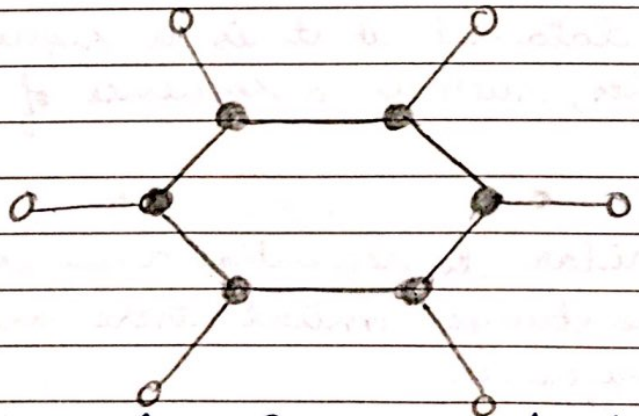


fig: Linked web pages

(ii) Data with objects that are graphs

- If objects have structure, i.e., the objects contain subobjects that have relationships, then such objects are frequently represented as graphs

Eg: structure of chemical compounds can be represented by graphs, where the nodes are atoms, and the links between nodes are chemical bonds



C_6H_6
Benzene

fig: Benzene Molecule

Ordered Data

* Here, the attributes have relationships that involve order in time & space

Different types of ordered data are,

(i) **Sequential Data / Temporal Data**

It is an extension of record data, where each record has a time associated with it.

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

(oo)

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)

(ii) **Sequence Data**

* It consists of a data set that is a sequence of individual entities, such as a 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.

Eg:- Genetic information of plants / animals can be represented in the form of sequences of nucleotides called genes.

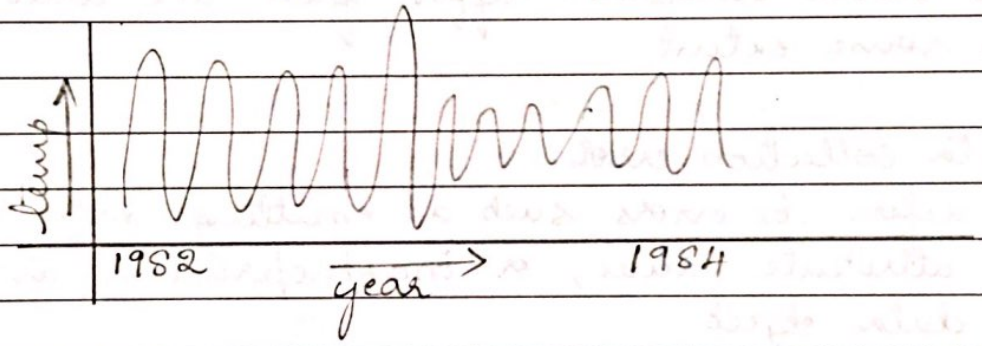
GGTTCCG CCTTCAG CCCCG CGCC
CGCAGGG GCCCG CCCCG CGCCGTC
GAGAAGGGGCCCG CCTGGCGGGCG
⋮

(b) Genomic sequence data

(iii) Time-Series Data

It is a special type of sequential data in which each record is a time series i.e., a series of measurements taken over time

Eg:- Average monthly temperature for Minneapolis during the years 1982 to 1994.



(iv) Spatial Data

Some objects have spatial attributes such as positions or areas

Eg:- Weather data (Precipitation, temperature, pressure) that is collected for a variety of geographical locations.

Data Quality

Preventing data quality problems is not possible. Hence Data Mining focuses on

1. Selection and correction of data quality problems This step is called Data Cleaning
2. Use of algorithms that can tolerate poor data quality.

Measurement and Data Collection Issues

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

1. Measurement and Data Collection Errors

* **Measurement error**

refers to any problem resulting from the measurement process. A common problem is that the value recorded differs from the true value to some extent.

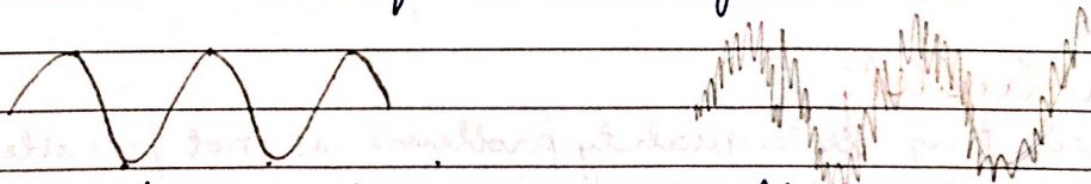
* **Data collection error**

refers to errors such as omitting data objects or attribute values, or inappropriately including a data object.

2. Noise and Artifacts

* **Noise**

It is the random component of a measurement error. It may involve the distortion of a value or the addition of spurious objects.



(a) time series (b) time series with noise

* Elimination of noise is difficult. Therefore DM focuses on devising robust algorithm 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 data are often referred to as artifacts

3. Precision, Bias, and Accuracy

Precision:

The closeness of repeated measurements (of the same quantity) to one another

Bias:

A systematic variation of measurements from the quantity being measured

Accuracy:

Closeness of measurements to the true value of the quantity being measured.

4. Outliers

Outliers are either

1. Data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set (or)
2. Values of an attribute that are unusual w.r.t the typical values for that attribute

5. Missing Values

* Reasons for missing values
  - Information was not collected.
    Eg:- some people decline to give their age or weight
  - Some attributes are not applicable for all objects
    Eg:- annual income is not applicable to children

* 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 data are often referred to as artifacts

3. Precision, Bias, and Accuracy

Precision:

The closeness of repeated measurements (of the same quantity) to one another

Bias:

A systematic variation of measurements from the quantity being measured

Accuracy:

Closeness of measurements to the true value of the quantity being measured.

4. Outliers

Outliers are either

1. Data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set (or)
2. Values of an attribute that are unusual w.r.t the typical values for that attribute

5. Missing Values

* Reasons for missing values
  - Information was not collected.
    Eg:- Some people decline to give their age or weight
  - Some attributes are not applicable for all objects
    Eg:- 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
   - A related strategy is to eliminate attributes that have missing values

2. Estimate missing values
* Sometimes missing data can be reliably estimated.
  Eg: In case of time series, wave 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 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

6. Inconsistent Values
* Data can contain inconsistent values
  Eg:- in address field, specified zip code area is not contained in that city

* Regardless of cause of inconsistent values, it is important to detect and correct (if possible) such problems

* Some types of inconsistencies are easy to detect
  Eg:- Person's height should not be negative
  In other cases, its necessary to consult an external source of information.

7. **Duplicate Data**

* A dataset may include data objects that are duplicates or almost duplicates of one another.

* To detect and eliminate such duplicates, two main issues must be addressed.
  1. If there are two 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 used to refer to the process of dealing with these issues.

Issues related to Applications

1. **Timeliness:**
   Some data starts to age as soon as it has been collected
   Eg:- Snapshot of some ongoing process represents reality for only a limited time.

2. **Relevance:**
   The available data must contain the information necessary for the application

3. **Knowledge about the data:**
* Ideally, the data sets are accompanied by documentation that describes different aspects of data, the quality of this documentation can either aid or hinder the subsequent analysis.
* If the documentation is poor, then our analyses of the data may be faulty.

MATRIKAS

Data Preprocessing

The purpose of preprocessing:

To transform the raw input data into an appropriate format for subsequent analysis in the process of Knowledge Discovery in Databases (KDD)

Data preprocessing topics are
Aggregation
Sampling
Dimensionality reduction
Feature subset selection
Feature creation
Discretization and binarization
Variable (or attribute) transformation.

1. Aggregation

Definition:

Combining two or more objects into a single object

Example:

Consider dataset consisting of transactions (data objects) recording the daily sales of products in various store locations for different days over the course of the year

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

The key principle for effective sampling is
- Using a sample will work almost as well as using the entire data sets, if the sample is representative
- In turn, 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)

1. Simple Random sampling
   Here, there is an equal probability of selecting any particular item.

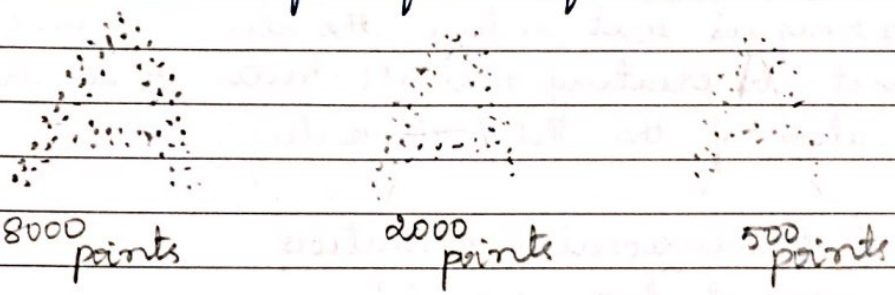
2. Sampling without replacement
   As each item is selected, it is removed from the set of all objects that together constitutes the population

3. Sampling with replacement
   Objects are not removed from the population as they are selected for the sample. Hence, same object can be picked more than once

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

NOTE 1: 2 and 3 are variations of simple random sampling itself

NOTE 2: Example of loss of structure with sampling



NOTE 3: Sample size is necessary to get atleast one object from each of ten groups of points fig(a)

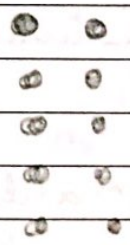
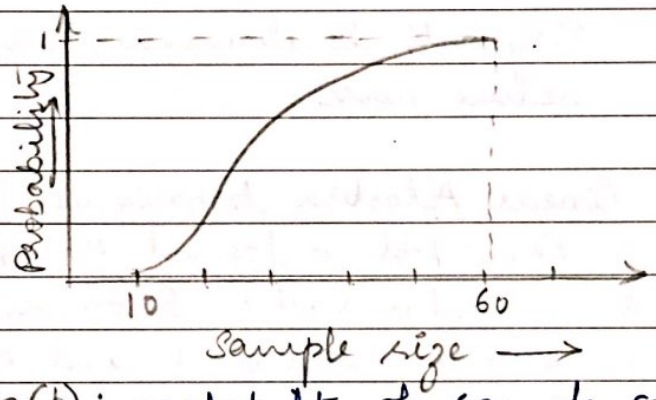


fig (a)



fig(b): probability of sample contains points from each of 10 groups

5. Progressive / Adaptive Sampling

It is used because proper sample size can be difficult to determine

These approaches start with a small sample, and then increase in the sample size until a sample of sufficient size has been obtained.

3. Dimensionality Reduction

Curse of Dimensionality

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

MATRIKAS

* Dimensionality Reduction is often reffred to those techniques that reduce the dimensionality of a dataset by creating new attributes that are a combination of the old attributes.

* Purpose of Dimensionality Reduction
  - Avoid curse of dimensionality
  - Reduce amount of memory and time 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
  1. Principal Component Analysis (PCA)
  2. Singular Value Decomposition
  3. Others: Supervised and Non-linear techniques

4. Feature Subset Selection

* Another way to reduce the dimensionality 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.
    Eg:- Purchase price of product and amount of sales tax paid.
  - Irrelevant features: contain almost no useful information for the data mining task at hand.
    Eg:- Students ID are irrelevant to the task of predicting students grade point average.

* 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.

Disadvantages: No. of subsets involving n attributes is 2^n

The three standard approach of feature selection

1. Embedded Approaches

Feature selection occurs normally as part of the DM algorithm

2. Filter approaches

Features are selected before the DM algorithm is run. using some approach that is independent of the datamining task.

3. Wrapper Approach

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

NOTE: An architecture for feature selection

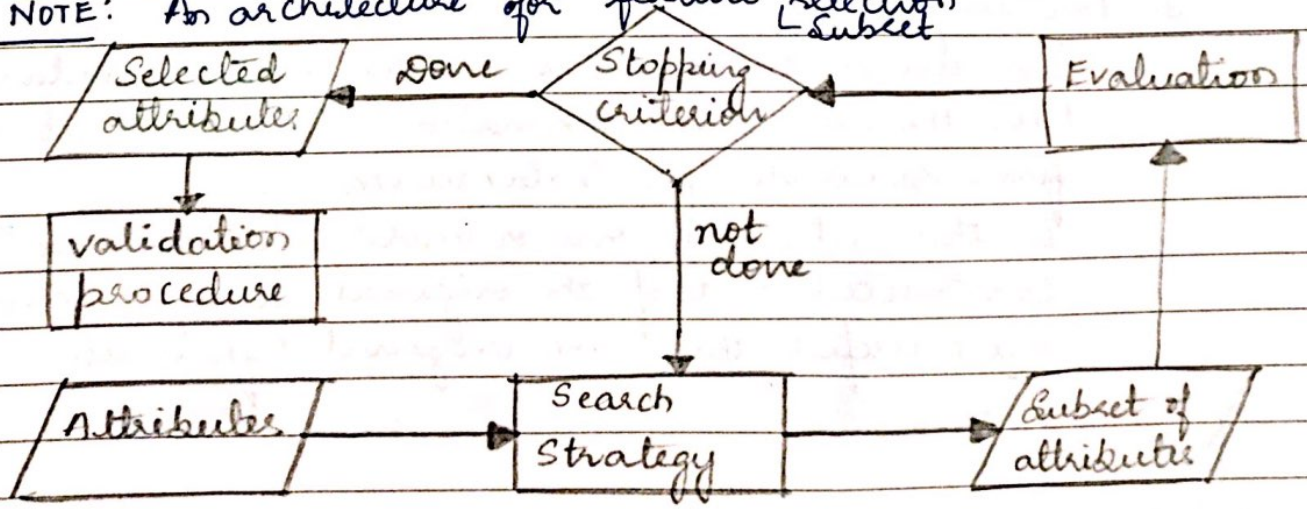


fig: flowchart of a feature seubset selected process

MATRIKAS

NOTE 2: Feature weighting 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.

5. Feature Creation

Definition: (Meaning)

Create a new set of attributes (from the original attributes) that captures the important information in a dataset much more effectively. Furthermore, the no. of new attributes can be smaller than the no. of original attributes (dimensionality reduction).

Three related methodologies for creating new attributes

1. **Feature Extraction**
   - Creation of new set of features from the original raw data is known as feature extraction
   - It is highly domain specific

2. **Mapping the Data to a New Space**
   - Fourier transform
   - Wavelet transform

   Fourier transformation produces a new object whose attributes are related to frequency

3. **Feature Construction**
   - Sometimes the features in the original datasets have the necessary information, but it is not in a form suitable for datamining algorithm
   - In this situation, one or more new features constructed out of the original features can be more useful than the original features.

6. Discretization and Binarization

Discretization:

A process of transforming a continuous attribute into a categorical attribute.

Binarization:

A process of transforming both continuous and discrete attributes into one or more binary attributes

NOTE: 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 $n = \log_2 m$ binary digits are required.

Eg:-

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

fig:- Conversion of a categorical attribute to three binary attribute

Such a transformation can cause two complications

1. Creates unintended relationships among the transformed attributes.
   Eg:- attributes x_2 and x_3 are correlated because information about the "good" value is encoded using both attributes.

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

Solution

Introduce one binary attribute for each categorical value

Eg:

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

fig: Conversion of a categorical attribute into five asymmetric binary attribute

Discretization of Continuous attributes

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 value 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 $\{(x_0, x_1), (x_1, x_2) \cdots (x_{n-1}, x_n)\}$, where, x_0 & x_n may be $+\infty$ or $-\infty$ respectively, or equivalently as a series of inequalities

$$x_0 < x \leq x_1 \cdots\cdots x_{n-1} < x < x_n$$

Discretization may be either
- Unsupervised (Here, class information is not used)
- Supervised (Here, class information is used)

Unsupervised Discretization

* Here, class information is not used
* Relatively simple approach

Eg:
- 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

There are one of the most promising approaches to discretization and it is explained below

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 as 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

It refers to the transformation that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.

Two important types of variable transformation

1. **Simple functions**

A simple mathematical function is applied to each value individually.

Eg: x^k, $\log(x)$, e^x, $|x|$, $\sqrt{x}$, $\sin x$

2. **Normalization or Standardization**

The goal of normalization or standardization 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

MATRIKAS

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

Measures of Similarity and Dissimilarity

Definitions

Similarity:

Similarity between two objects is a numerical measure of the degree to which the two objects are alike

* It is higher when objects are more alike
* often falls in the range [0, 1]
  - 0 → no similarity
  - 1 → Complete similarity

Dissimilarity:

Numerical measure of the degree to which the two objects are different

* Lower when objects are more alike
* Minimum dissimilarity is often 0
* Upper limit varies i.e., $[0, \infty]$

Proximity:

refers to either similarity or Dissimilarity.

NOTE: Transformations are often applied to convert a similarity to a dissimilarity, or vice versa, or a proximity measure to fall within a particular range such as [0, 1]

Similarity and Dissimilarity between simple attributes

Attribute type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x == y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x == y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x-y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the no. of values)	$s = 1-d$
Interval or ratio	$d = \lvert x-y \rvert$	$s = -d$, $s = 1/(1+d)$, $s = e^{-d}$; $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarities between Data objects

Various kinds of dissimilarities

Distances

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

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

n - no. of dimensions

x_K, y_K - K^{th} attributes of x & y

Minkowski Distance

It is a generalization of Euclidean distance. and it is given by,

Given by,

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

$r \rightarrow$ parameter

The following are the three most common examples of Minkowski distance.

→ $r=1$ City block (Manhattan, taxicab, L1 norm) distance

A common example of this is the Hamming distance, which is just the no. of bits that are different between two binary vectors

→ $r=2$ Euclidean distance (L_2 norm)

→ $r=\infty$ "Supremum" (L_{max} norm, L_∞ norm) distance

This is the maximum difference between any attribute of the object.

More formally L_∞ distance is defined by

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

Distances such as 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) $d(x,x) \geq 0$ 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 $x, y, \& z$

Measures that satisfy all three properties are known as metrics

Similarities between Data objects

* for similarities, the triangle inequality typically does not hold, but symmetry and positivity typically do

* if $s(x,y)$ is the similarity between points x and y then the typical properties of similarities are
  1. $s(x,y) = 1$ only if $x = y$ $(0 \leq s \leq 1)$
  2. $s(x,y) = s(y,x)$ for all $x \& y$ (symmetry)

Examples of proximity Measures

This section provides specific examples of some similarity and dissimilarity (proximity).

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

NOTE:

Similarity measures between data objects that contain only binary attributes are called Similarity Coefficients

Let $x \& y$ be two objects that consists of n binary attributes. The comparisions of two such objects i.e., two binary vectors leads to the following four attributes (frequencies)

f_{00} = no. of attributes where $x=0, y=0$

f_{01} = no. of attributes where $x=0, y=1$

f_{10} = no. of attributes where, $x=1, y=0$
f_{11} = no. of attributes where, $x=1, y=1$

1. Simple Matching Coefficient (SMC)

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

$$SMC = \frac{\text{no. of matching attribute values}}{\text{no. of attributes}}$$

$$= \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

This measure counts both presence & absences equally.

2. Jaccard Coefficient (J)

Jaccard coefficient is frequently used to handle objects consisting of asymmetric binary attributes.

$$J = \frac{\text{no. of matching presences}}{\text{no. of attributes not involved in 00 matches}}$$

$$= \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

3. Cosine Similarity

With transaction data, similarity should not depend on the no. 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, & also must be able to handle non-binary vectors.

The Cosine similarity is defined as one of the most common measure of document similarity. If x and y are two document vectors, then

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

where, $\cdot$ indicates the vector dot product

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

and $\|x\|$ is the length of the vector x,

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

4. **Extended Jaccard Coefficient (Tanimoto Coefficient) - (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.

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

5. **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:

$$Corr(x, y) = \frac{covariance(x, y)}{standard\_dev(x) * standard\_dev(y)} = \frac{s_{xy}}{s_x s_y}$$

MATRIKAS

where we are using the following standard statistical notation and definitions:

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

$$\text{standard\_deviation}(x) \rightarrow s_x$$

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

$$\text{standard\_deviation}(y) \rightarrow s_y$$

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

Example Problems

① The SMC and Jaccard Similarity Coefficients

Calculate SMC and J for the following binary vectors.

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

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

Soln:

$f_{01} = 2$

$f_{10} = 1$

$f_{00} = 7$

$f_{11} = 0$

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

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

(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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(b) $x = (1, 1, 1, 1)$

$y = (2, 2, 2, 2)$

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

Soln:

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

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

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

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

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

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

~~$S_{xy} = 2.66$~~ $\qquad S_{xy} = 0$ ✓

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

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

$$= \sqrt{\frac{1}{3}(4)} = \sqrt{\frac{4}{3}} \qquad S_x = 0$$

~~$S_x = 0.66$~~

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

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

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

$$S_y = 0$$

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

Excercise Problems

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: $Cos(x,y) = 0$

5. Find $Corr(x,y)$ for the following
   $x = (1,1,1,1)$ $y = (2,2,2,2)$
   Ans: $Corr(x,y) = \frac{0}{0}$
   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: $Corr(x,y) = 0$
   $Cos(x,y) = 0$

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

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

$$= \sqrt{4}$$

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

Visualization

In this section, 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)

discrete	Continuous
- only take certain values (like whole numbers)	- take any value (within a range)
- This is counted	- This is measured

MATRIKAS

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 analyzed

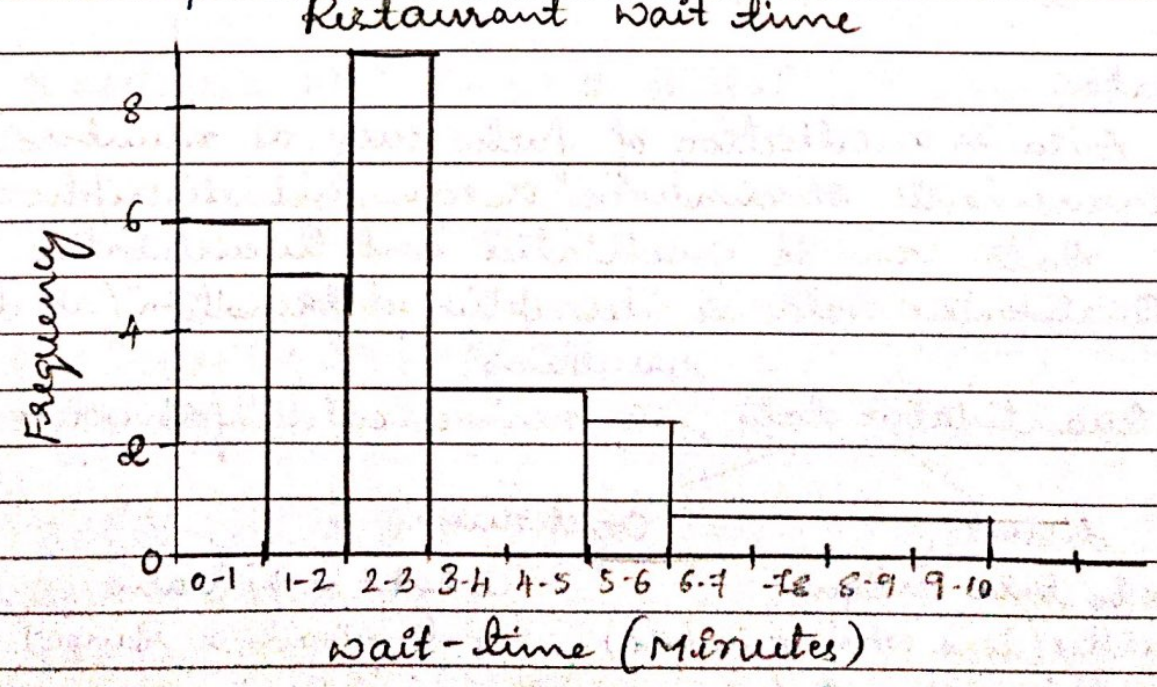
Visualization is most powerful and appealing techniques for data exploration.

Visualization Techniques

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

1. Histograms

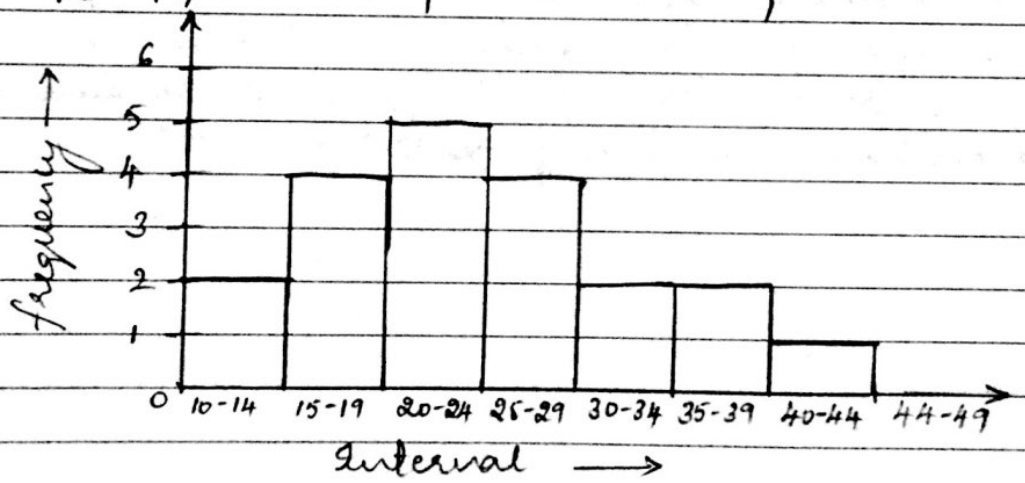
A histogram is bar graph that represents a frequency distribution. The width represents the interval and the height represents the corresponding frequency. There are no spaces between the bars.



Create (Construct) a histogram for the following details of average gas mileage of twenty cars.

Average Gas Milege: (in miles/gallon)
24, 17, 14, 22, 25, 26, 38, 42, 24, 12, 28, 19, 32, 21, 35, 28, 31, 21, 18, 19

Interval	Tally	Frequency
10-14	II	2
15-19	IIII	4
20-24	~~IIII~~	5
25-29	IIII	4
30-34	II	2
35-39	II	2
40-44	I	1



2. Two-Dimension Histogram

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

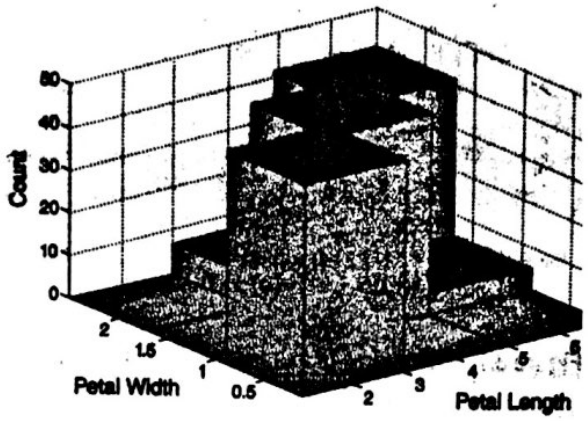


fig: Two-dimension histogram

3. Box Plots

There are another method for showing the distributions of the values of a single numerical attribute.

The following figure shows a labeled 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.

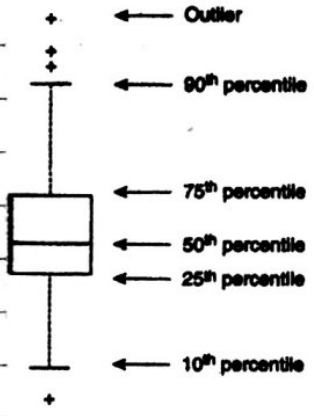


Figure 3.10. Description of box plot for sepal length.

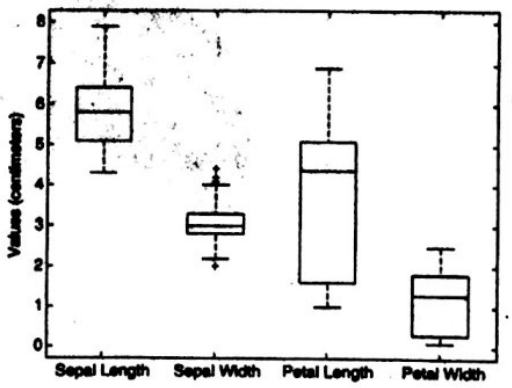
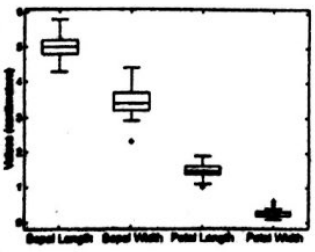
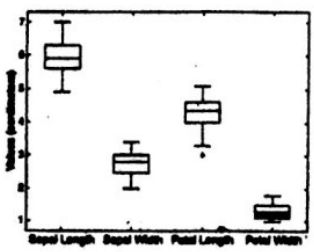


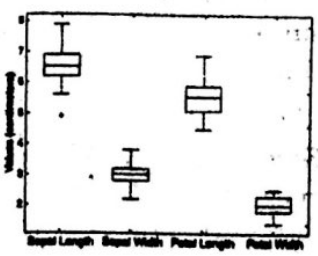
Figure 3.11. Box plot for iris attributes.



(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.

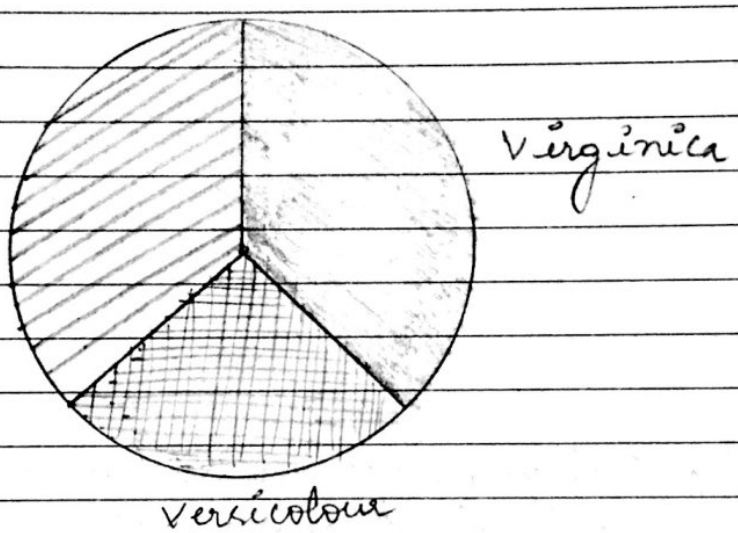


fig: Distribution of the three types of iris flowers.

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.

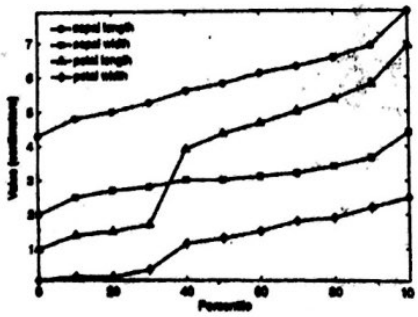


fig: Percentile plots for sepal length, sepal width, petal length and petal width

MATRIKAS

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.

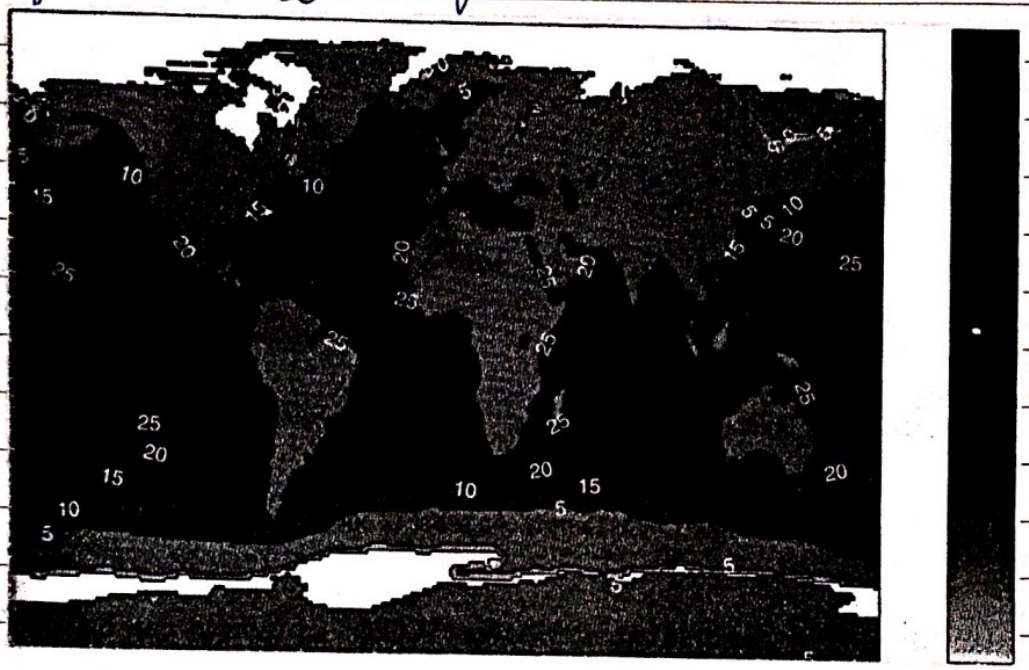


fig: Contour plot of Sea Surface Temperature (SST) for December 1998

Motivation for Visualization

1. People can quickly absorb large amounts of visual information and find Patterns in it

2. Helps in deciding which attributes contain useful information.

The Iris Data Set

Iris Data set is available from the University of California at Irvine (UCI) Machine Learning Repository.

It consists of information on 150 Iris flowers, 50 each from one of three Iris species:

- Setosa
- Versicolour
- Virginica

Each flower is characterized by five attributes:

1. Sepal length in centimeters
2. Sepal width in centimeters
3. Petal length in centimeters
4. Petal width in centimeters
5. class (Setosa, Versicolour, Virginica)