SUBJECT CODE : **310251**

As per Revised Syllabus of

SAVITRIBAI PHULE PUNE UNIVERSITY

Choice Based Credit System (CBCS)

T.E. (Computer) Semester - VI

DATA SCIENCE AND BIG DATA ANALYTICS

Iresh A. Dhotre

M.E. (Information Technology)
Ex-Faculty, Sinhgad College of Engineering,
Pune.

Dr. Kalpana V. Metre

(Ph.D.),
Associate Professor,
MET's Institute of Engineering,
Nashik.

TECHNICAL PUBLICATIONS
SINCE 1993
An Up-Thrust for Knowledge

Data Science and Big Data Analytics

Subject Code : 310251

T.E. (Computer Engineering) Semester - VI

Published by :

TECHNICAL PUBLICATIONS
SINCE 1993
An Up-Thrust for Knowledge

Amit Residency, Office No.1, 412, Shaniwar Peth,
Pune - 411030, M.S. INDIA, Ph.: +91-020-24495496/97
Email : sales@technicalpublications.org Website : www.technicalpublications.org

Printer :
Yogiraj Printers & Binders
Sr.No. 10/1A,
Ghule Industrial Estate, Nanded Village Road,
Tal. - Haveli, Dist. - Pune - 411041.

ISBN 978-93-91567-92-7

9 789391 567927

SPPU 19

PREFACE

The importance of **Data Science and Big Data Analytics** is well known in various engineering fields. Overwhelming response to our books on various subjects inspired us to write this book. The book is structured to cover the key aspects of the subject **Data Science and Big Data Analytics.**

The book uses plain, lucid language to explain fundamentals of this subject. The book provides logical method of explaining various complicated concepts and stepwise methods to explain the important topics. Each chapter is well supported with necessary illustrations, practical examples and solved problems. All the chapters in the book are arranged in a proper sequence that permits each topic to build upon earlier studies. All care has been taken to make students comfortable in understanding the basic concepts of the subject.

Representative questions have been added at the end of each section to help the students in picking important points from that section.

The book not only covers the entire scope of the subject but explains the philosophy of the subject. This makes the understanding of this subject more clear and makes it more interesting. The book will be very useful not only to the students but also to the subject teachers. The students have to omit nothing and possibly have to cover nothing more.

We wish to express our profound thanks to all those who helped in making this book a reality. Much needed moral support and encouragement is provided on numerous occasions by our whole family. We wish to thank the **Publisher** and the entire team of **Technical Publications** who have taken immense pain to get this book in time with quality printing.

Any suggestion for the improvement of the book will be acknowledged and well appreciated.

Authors

Dresh. A. Dhotre
Dr. Kalpana. V. Metre

Dedicated to God

SYLLABUS

Data Science and Big Data Analytics - (310251)

Credit :	Examination Scheme :
03	Mid - Sem (TH) : 30 Marks
	End - Sem (TH) : 70 Marks

Unit I Introduction to Data Science and Big Data

Basics and need of Data Science and Big Data, Applications of Data Science, Data explosion, 5 V's of Big Data, Relationship between Data Science and Information Science, Business intelligence versus Data Science, Data Science Life Cycle, Data : Data Types, Data Collection. Need of Data wrangling, Methods : Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization. **(Chapter - 1)**

Unit II Statistical Inference

Need of statistics in Data Science and Big Data Analytics, **Measures of Central Tendency :** Mean, Median, Mode, Mid-range. **Measures of Dispersion :** Range, Variance, Mean Deviation, Standard Deviation. Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t-test.
(Chapter - 2)

Unit III Big Data Analytics Life Cycle

Introduction to Big Data, sources of Big Data, **Data Analytic Lifecycle :** Introduction, Phase 1 : Discovery, Phase 2 : Data Preparation, Phase 3 : Model Planning, Phase 4 : Model Building, Phase 5 : Communication results, Phase 6 : Operationalize. **(Chapter - 3)**

Unit IV Predictive Big Data Analytics with Python

Introduction, Essential Python Libraries, Basic examples. **Data Preprocessing :** Removing Duplicates, Transformation of Data using function or mapping, replacing values, Handling Missing Data. Analytics Types : Predictive, Descriptive and Prescriptive. **Association Rules** : Apriori Algorithm, FP growth. **Regression :** Linear Regression, Logistic Regression. **Classification :** Naïve Bayes, Decision Trees. **Introduction to Scikit-learn,** Installations, Dataset, mat plotlib, filling missing values, Regression and Classification using Scikit-learn.
(Chapter - 4)

Unit V Big Data Analytics and Model Evaluation

Clustering Algorithms : K-Means, Hierarchical Clustering, Time-series analysis. **Introduction to Text Analysis :** Text-preprocessing, Bag of words, TF-IDF and topics. Need and Introduction to social network analysis, Introduction to business analysis. **Model Evaluation and Selection :** Metrics for Evaluating Classifier Performance, Holdout Method and Random Sub sampling, Parameter Tuning and Optimization, Result Interpretation, Clustering and Time-series analysis using Scikit-learn, sklearn. metrics, Confusion matrix, AUC-ROC Curves, Elbow plot. **(Chapter - 5)**

Unit VI Data Visualization and Hadoop

Introduction to Data Visualization, Challenges to Big data visualization, Types of data visualization, Data Visualization Techniques, Visualizing Big Data, Tools used in Data Visualization, Hadoop ecosystem, Map Reduce, Pig, Hive, Analytical techniques used in Big data visualization. **Data Visualization using Python :** Line plot, Scatter plot, Histogram, Density plot, Box- plot. **(Chapter - 6)**

Table of Contents

Unit - I

Unit - II

Unit - III

Unit - IV

Unit - V

Unit - VI

UNIT I

1 Introduction to Data Science and Big Data

Syllabus

Basics and need of Data Science and Big Data, Applications of Data Science, Data explosion, 5 V's of Big Data, Relationship between Data Science and Information Science, Business intelligence versus Data Science, Data Science Life Cycle, Data : Data Types, Data Collection. Need of Data wrangling, Methods : Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization.

Contents

1.1 Basics and Need of Data Science and Big Data

- Data is collection of facts and figures which relay something specific, but which are not organized in any way. It can be numbers, words, measurements, observations or even just descriptions of things. We can say, data is raw material in the production of information.
- Types of data are record data, data matrix, document data, transaction data, graph data and ordered data.

Data science :

- Data science is an interdisciplinary field that seeks to extract knowledge or insights from various forms of data. At its core, data science aims to discover and extract actionable knowledge from data that can be used to make sound business decisions and predictions.
- Data science uses advanced analytical theory and various methods such as time series analysis for predicting future. From historical data, instead of knowing how many products sold in previous quarter, data science helps in forecasting future product sales and revenue more accurately.
- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information and make business decisions. Data science uses complex machine learning algorithms to build predictive models.
- Data science enables businesses to process huge amounts of structured and unstructured big data to detect patterns.

Big data :

- Big data can be defined as very large volumes of data available at various sources, in varying degrees of complexity, generated at different speed i.e., velocities and varying degrees of ambiguity, which cannot be processed using traditional technologies, processing methods, algorithms or any commercial off-the-shelf solutions.
- 'Big data' is a term used to describe collection of data that is huge in size and yet growing exponentially with time. In short, such a data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.
- The processing of big data begins with the raw data that isn't aggregated or organized and is most often impossible to store in the memory of a single computer.

1.1.1 Difference between Data Science and Big Data

Sr. no.	Data science	Big data
1.	It is a field of scientific analysis of data in order to solve analytically complex problems and the significant and necessary activity of cleansing, preparing of data.	Big data is storing and processing large volume of structured and unstructured data that can not be possible with traditional applications.
2.	It is used in Biotech, energy, gaming and insurance.	Used in retail, education, healthcare and social media.
3.	Goals : Data classification, anomaly detection, prediction, scoring and ranking.	Goals : To provide better customer service, identifying new revenue opportunities, effective marketing etc.

1.1.2 Applications of Data Science

- Asking a personal assistant like Alexa or Siri for a recommendation demands data science. So does operating a self - driving car, using a search engine that provides useful results or talking to a chatbot for customer service. These are all real - life applications for data science.
- Following are some main reasons for using data science technology :
  - With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
  - Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, which handle the huge amount of data are using data science algorithms for better customer experience.
  - Data science is working for automating transportation such as creating a self - driving car, which is the future of transportation.
- Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

1. **Healthcare :** Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.
2. **Gaming :** Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.
3. **Image recognition :** Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.
4. **Logistics :** Data science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

5. **Predict future market trends :** Collecting and analysing data on a larger scale can enable to identify emerging trends in market. Tracking purchase data, celebrities and influencers and search engine queries can reveal what products people are interested in.
6. **Recommendation systems :** Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase or browse on their platforms.
7. **Streamline manufacturing :** Another way we can use data science in business is to identify inefficiencies in manufacturing processes. Manufacturing machines gather data from production processes at high volumes. In cases where the volume of data collected is too high for a human to manually analyse it, an algorithm can be written to clean, sort and interpret it quickly and accurately to gather insights.

1.2 Data Explosion

- The essence of computer applications is to store things in the real world into computer systems in the form of data, i.e., it is a process of producing data. Some data are the records related to culture and society and others are the descriptions of phenomena of universe and life. The large scale of data is rapidly generated and stored in computer systems, which is called data explosion.
- Data is generated automatically by mobile devices and computers, think Facebook, search queries, directions and GPS locations and image capture.
- Sensors also generate volumes of data, including medical data and commerce location - based sensors. Experts expect 55 billion IP - enabled sensors by 2021. Even storage of all this data is expensive. Analysis gets more important and more expensive every year.
- Fig. 1.2.1 shows the big data explosion by the current data boom and how critical it is for us to be able to extract meaning from all of this data.

Fig. 1.2.1 Data explosion

- The phenomena of exponential multiplication of data that gets stored is termed as "Data Explosion". Continuous inflow of real - time data from various processes, machinery and manual inputs keeps flooding the storage servers every second.

- Sending emails, making phone calls, collecting information for campaigns; each day we create a massive amount of data just by going about our normal business and this data explosion does not seem to be slowing down. In fact, 90 % of the data that currently exists was created in just the last two years.
- Reason for this data explosion is **Innovation**.

1. **Business model transformation :** Innovation changed the way in which we do business, provide services. The data world is governed by three fundamental trends are business model transformation, globalization and personalization of services.
   - Organizations have traditionally treated data as a legal or compliance requirement, supporting limited management reporting requirements. Consequently, organizations have treated data as a cost to be minimized.
   - The businesses are required to produce more data related to product and provide services to cater each sector and channel of customer.
2. **Globalization :** Globalization is an emerging trend in business where organizations start operating on international scale. From manufacturing to customer service, globalization have changed the commerce of the world. Variety and different formats of data is generated due to globalization.
3. **Personalization of services :** To enhance customer service, the form of one - to - one marketing in the form of personalization of service is opted by customer. Customer expects communication through various channels increases the speed of data generation.
4. **New sources of data :** The shift to online advertising supported by the likes of Google, Yahoo, and others is a key driver in the data boom. Social media, mobile devices, sensor networks and new media are on the fingertips of customer or user. The data generated through this is used by corporations for decision support systems like Business Intelligence and analytics. The growth of technology helped to emerge new business models over the last decade or more. Integration of all the data across the enterprise is used to create business decision support platform.

1.3 5 V's of Big Data

- We differentiate big data characteristics from traditional data by one or more of the five V's : Volume, velocity, variety, veracity and value.

1. **Volume :** Volumes of data are larger than conventional relational database infrastructure can cope with. It consists of terabytes or petabytes of data.

- Fig. 1.3.1 shows big data volume.

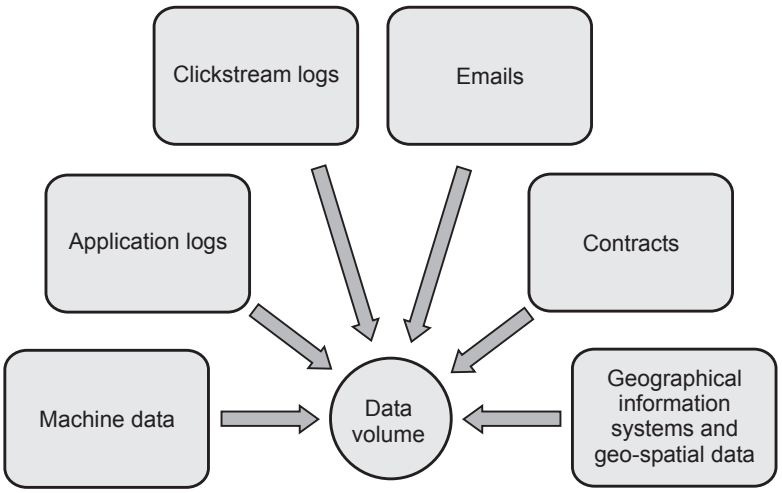


Fig. 1.3.1 Big data volume

2) **Velocity :** The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. It is being created in or near real - time.

3) **Variety :** It refers to heterogeneous sources and the nature of data, both structured and unstructured.

- Fig. 1.3.2 shows big data velocity and data variety.

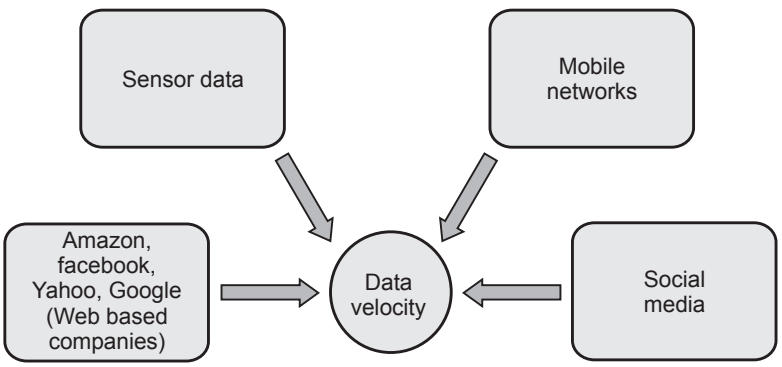


Fig. 1.3.2 (a)

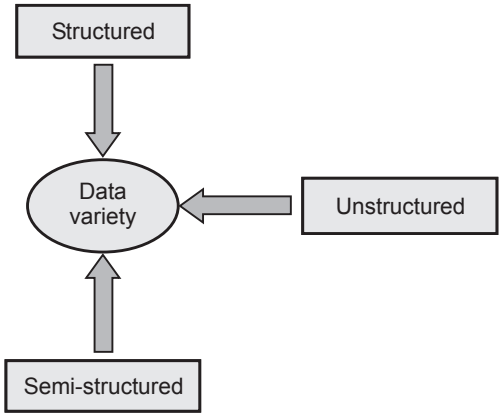


Fig. 1.3.2 (b)

4. **Value :** It represents the business value to be derived from big data.
   - The ultimate objective of any big data project should be to generate some sort of value for the company doing all the analysis.
   - For real - time spatial big data, decisions can be enhance through visualization of dynamic change in such spatial phenomena as climate, traffic, social - media - based attitudes and massive inventory locations.
   - Exploration of data trends can include spatial proximities and relationships. Once spatial big data are structured, formal spatial analytics can be applied, such as spatial autocorrelation, overlays, buffering, spatial cluster techniques and location quotients.
5. **Veracity :** Big data must be fed with relevant and true data. We will not be able to perform useful analytics if many of the incoming data comes from false sources or has errors. Veracity refers to the level of trustiness or messiness of data and if higher the trustiness of the data, then lower the messiness and vice versa. It relates to the assurance of the data's quality, integrity, credibility and accuracy. We must evaluate the data for accuracy before using it for business insights because it is obtained from multiple sources.

1.4 Relationship between Data Science and Information Science

SPPU : Dec.-18

- Data science, as the interdisciplinary field, employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science and computer science. Data science and information science are twin disciplines by nature. The mission, task and nature of data science are consistent with those of information science.

- Data science is heavy on computer science and mathematics. Information science is used in areas such as knowledge management, data management and interaction design.
- Information science is the science and practice dealing with the effective collection, storage, retrieval and use of information. It is concerned with recordable information and knowledge and the technologies and related services that facilitate their management and use.

1.4.1 Business Intelligence versus Data Science

Business Intelligent (BI)	Data Science
Bl tends to provide reports, dashboards, and queries on business questions for the current period or in the past.	Data science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analyzing the present and enabling informed decisions about the future.
Bl systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior quarter or year.	Data science tends to be more exploratory in nature and may use scenario optimization to deal with more open - ended questions.
BI helps monitor the current state of business data to understand the historical performance of a business.	Data science, as used in business, is basically data-driven, where many interdisciplinary sciences are applied together to extract meaning.
BI is designed to handle static and highly structured data.	Data science can handle high-speed, high-volume and complex, multi-structured data from a wide variety of data sources.

1.4.2 Compare Cloud Computing and Big Data

Sr. No.	Cloud computing	Big data
1.	It provides resources on demand.	It provides a way to handle huge volumes of data and generate insights.
2.	It refers to internet services from SaaS, PaaS to IaaS.	It refers to data, which can be structured, semi-structured or unstructured.
3.	Cloud is used to store data and information on remote servers.	It is used to describe huge volume of data and information.
4.	Cloud computing is economical as it has low maintenance costs centralized platform no upfront cost and disaster safe implementation.	Big data is highly scalable, robust ecosystem and cost - effective.

5.	Vendors and solution providers of cloud computing are Google, Amazon Web Service, Dell, Microsoft, Apple and IBM.	Vendors and solution providers of big data are Cloudera, Hortonworks, Apache and MapR.
6.	The main focus of cloud computing is to provide computer resources and services with the help of network connection.	Main focus of big data is about solving problems when a huge amount of data generating and processing.

Review Question

1. Compare BI Vs. data science. **SPPU : Dec.-18 (End Sem), Marks 6**

1.5 Data Science Life Cycle

- A data science life cycle is an iterative set of data science steps you take to deliver a project or analysis. Fig 1.5.1 shows data science life cycle.

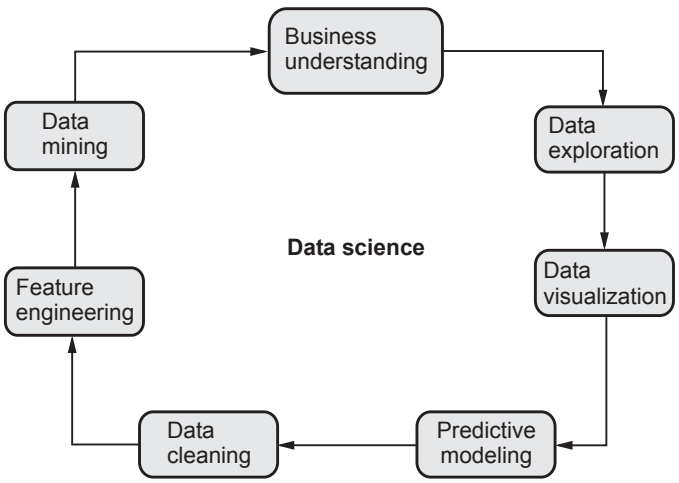


Fig. 1.5.1 Data science life cycle

a) **Business understanding :** Understand the basic problem you are trying to solve.

b) **Data exploration :** Understand the pattern and bias in your data.

c) **Data visualization :** Create and study of the visual representation of data.

d) **Predictive modeling :** It is the stage where the machine learning finally comes into data.

e) **Data cleaning :** Detecting and correcting corrupt or inaccurate records.

f) **Feature engineering :** It is the process of cutting down the features.

g) **Data mining :** Gathering your data from different source.

1.6 Data

- Data is collection of facts and figures which relay something specific, but which are not organized in any way. It can be numbers, words, measurements, observations or even just descriptions of things. We can say, data is raw material in the production of information.
- **Examples of data :** Printed paper, bank account passbook, student attendance, salary sheet etc.
- Data creation is limited because of lack of new technology. After start of using computer, data can be converted into more convenient forms. It starts of using email, e-book, digital audio and video, images etc.
- Data is raw. It simply exists and has no significance beyond its existence. It can exist in any form, usable or not. It does not have meaning of itself. In computer parlance, a spreadsheet generally starts out by holding data.
- Facts and figures which relay something specific, but which are not organized in any way and which provide no further information regarding patterns, context, etc.
- Data represents a fact or statement of event without relation to other things. Example : It is raining. Computer stores data in the form of 0 and 1. This is called digital data. These data are stored in the computer in the form of 0 and 1. Data in this form is called digital data. It is shown in Fig. 1.6.1.

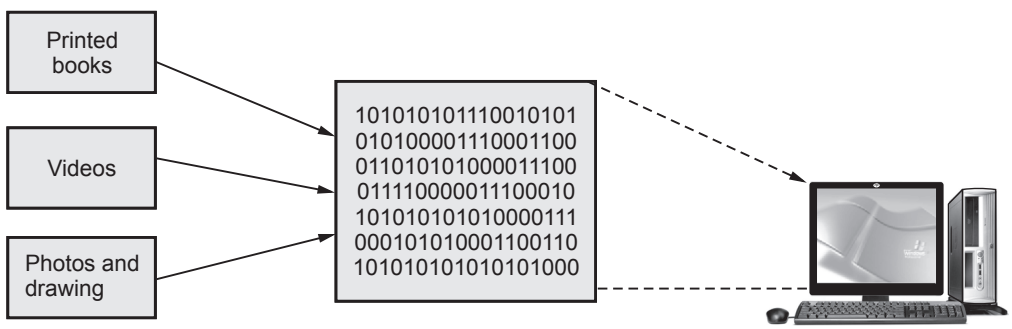


Fig. 1.6.1 Digital data

- Data is raw. It has not been shaped, processed or interpreted. Once data has been processed and turned into information. Computers need data. Humans need information. Data is a building block. Information gives meaning and context.
- Information systems are combinations of hardware, software and telecommunications networks that people build and use to collect, create and distribute useful data, typically in organizational settings.

- Company offers their products to customers for making money. These products can be goods or services. Large number of different types of data is accumulated in the organization. It contains data of products, customer data, employee's data, data of the delivery of products and data of other sources.
- These data must be stored, managed and processed. For performing these operations on data, information system can play an important role. Because this data is raw data. There is no unique format.
- Because of new technology, generation of new data and sharing of data has increased exponentially. The factors that affects the growth of digital data :
  1. Cost of digital storage device is decreases.
  2. Data processing capabilities is also increases.
  3. Faster access of data

1.6.1 Data Types

- Data are of two types : Structured and unstructured.

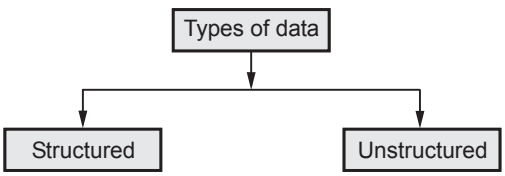


Fig. 1.6.2

Structured data :

- Structured data is arranged in rows and column format. It helps for application to retrieve and process data easily. Database management system is used for storing structured data.
- The term structured data refers to data that is identifiable because it is organized in a structure. The most common form of structured data or records is a database where specific information is stored based on a methodology of columns and rows.
- Structured data is also searchable by data type within content. Structured data is understood by computers and is also efficiently organized for human readers.

Unstructured data

- Unstructured data is data that does not follow a specified format. Row and columns are not used for unstructured data. Therefore it is difficult to retrieve required information. Unstructured data has no identifiable structure.

- The unstructured data can be in the form of text : (Documents, email messages, customer feedbacks), audio, video, images. Email is an example of unstructured data.
- Even today in most of the organizations more than 80 % of the data are in unstructured form. This carries lots of information. But extracting information from these various sources is a very big challenge.
- Characteristics of unstructured data :
  1. There is no a structural restriction or binding for the data.
  2. Data can be of any type.
  3. Unstructured data does not follow any structural rules.
  4. There are no predefined formats, restriction or sequence for unstructured data.
  5. Since there is no structural binding for unstructured data it is unpredictable in nature.

Examples of machine generated unstructured data :

1. **Satellite images :** This includes weather data or the data that the government captures in its satellite surveillance imagery.
2. **Scientific data :** This includes atmospheric data and high energy physics.
3. **Photographs and video :** This include security, surveillance and traffic video.

1.6.2 Difference between Structured and Unstructured Data

Parameters	Structured data	Unstructured data
Representation	It is in discrete form. i.e. stored in row and column format.	Unstructured data is data that does not follow a specified format.
Metadata	Syntax	Semantics
Storage	Database management system	Unmanaged file structure
Standard	SQL, ADO.net, ODBC	Open XML, SMTO, SMS
Tools for integration	ETL	Batch processing or manual data entry.
characteristics	With a structured document, certain information always appears in the same location on the page.	In unstructured document information can appear in unexpected places on the document.
Used by organizations	Low volume operations	High volume operations

1.6.3 Difference between Information and Data

Sr. No.	Information	Data
1.	It is processed data.	It is raw data.
2.	Information is specific.	Data is not specific.
3.	Information depends on data.	Data does not depend on information.
4.	Information is output of computer.	Data is input to the computer.
5.	Information simply refers to the knowledge of value obtained through the collection, interpretation and/or analysis of data.	Data refer to facts, measurements, characteristics, or traits of an object of interest.

1.6.4 Qualitative and Quantitative Data

- Data can broadly be divided into following two types :
  1. Qualitative data
  2. Quantitative data

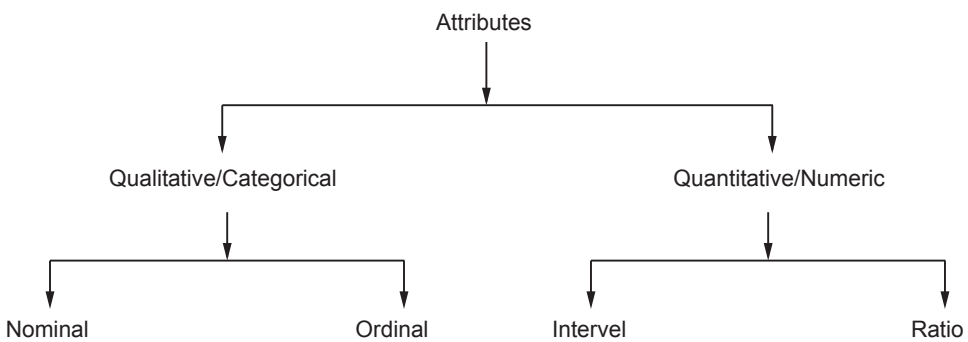


Fig. 1.6.3

Qualitative data :

- **Qualitative data** provides information about the quality of an object or information which cannot be measured. Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, economic status, religious preference are usually considered to be qualitative data.
- **Qualitative data** is data concerned with descriptions, which can be observed but cannot be computed. Qualitative data is also called categorical data. Qualitative data can be further subdivided into two types as follows :
  1. Nominal data
  2. Ordinal data

Nominal data

- A nominal data is the 1^{st} level of measurement scale in which the numbers serve as "tags" or "labels" to classify or identify the objects.
- A nominal data usually deals with the non-numeric variables or the numbers that do not have any value. While developing statistical models, nominal data are usually transformed before building the model.
- It is also known as categorical variables.

Characteristics of nominal data :

1. A nominal data variable is classified into two or more categories. In this measurement mechanism, the answer should fall into either of the classes.
2. It is qualitative. The numbers are used here to identify the objects.
3. The numbers don't define the object characteristics. The only permissible aspect of numbers in the nominal scale is "counting."

- Example :
  1. Gender : Male, Female, Other.
  2. Hair color : Brown, Black, Blonde, Red, Other.

Ordinal data

- Ordinal data is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.
- Ordinal represents the "order." Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.
- Characteristics of the ordinal data :
  a) The ordinal data shows the relative ranking of the variables.
  b) It identifies and describes the magnitude of a variable.
  c) Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables.
  d) The interval properties are not known.
  e) The surveyors can quickly analyze the degree of agreement concerning the identified order of variables.
- Examples :
  a) University ranking : 1^{st}, 9^{th}, 87^{th}…
  b) Socioeconomic status : Poor, middle class, rich.
  c) Level of agreement : Yes, maybe, no.
  d) Time of day : Dawn, morning, noon, afternoon, evening, night.

Quantitative data

- **Quantitative data** is the one that focuses on numbers and mathematical calculations and can be calculated and computed.
- **Quantitative data** are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject. These data may be represented by ordinal, interval or ratio scales and lend themselves to most statistical manipulation.
- There are two types of quantitative data : Interval data and Ratio data

Interval data :

- Interval data corresponds to a variable in which the value is chosen from an interval set.
- It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.
- Characteristics of interval data :
  a) The interval data is quantitative as it can quantify the difference between the values.
  b) It allows calculating the mean and median of the variables.
  c) To understand the difference between the variables, you can subtract the values between the variables.
  d) The interval scale is the preferred scale in statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.
- Examples :
  1. Celsius temperature.
  2. Fahrenheit temperature.
  3. Time on a clock with hands.

Ratio data :

- Any variable for which the ratios can be computed and are meaningful is called ratio data.
- It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.
- Characteristics of ratio data :
  a) Ratio scale has a feature of absolute zero.
  b) It doesn't have negative numbers, because of its zero - point feature.

c) It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, divided. Mean, median, and mode can be calculated using the ratio scale.

d) Ratio data has unique and useful properties. One such feature is that it allows unit conversions like kilogram - calories, gram - calories, etc.

- Examples : Age, Weight, Height, Ruler measurements, Number of children

1.6.5 Difference between Qualitative and Quantitative Data

Qualitative data	Quantitative data
Qualitative data provides information about the quality of an object or information which cannot be measured	Quantitative data relates to information about the quantity of an object; hence it can be measured
Types : Nominal data and Ordinal data	Types : Interval data and Ratio data
Narratives often make use of adjectives and other descriptive words to refer to data on appearance, color, texture, and other qualities	Measure's quantities such as length, size, amount, price, and even duration.
They are descriptive rather than numerical in nature	Expressed in numerical form.
For example : • The team is well prepared. • The leaf feels waxy. • The river is peaceful.	For example : • The team has 7 players. • The leaf weighs 2 ounces. • The river is 25 miles long.

1.6.6 Data Collection

- Data collection is the systematic approach to gathering and measuring information from a variety of sources to get a complete and accurate picture of an area of interest. The big data includes information produced by humans and devices.
- The big data collection is focused on the following types of data :

a) Network data : This type of data is gathered on all kinds of networks, including social media, information and technological networks, the Internet and mobile networks, etc.

b) Real - time data : They are produced on online streaming media, such as YouTube, Twitch, Skype or Netflix.

c) Transactional data : They are gathered when a user makes an online purchase (information on the product, time of purchase, payment methods, etc.).

d) Geographic data : Location data of everything, humans, vehicles, building, natural reserves and other objects are continuously supplied with satellites.

e) Natural language data : These data are gathered mostly from voice searches that can be made on different devices accessing the Internet.

f) Time series data : This type of data is related to the observation of trends and phenomena taking place at this very moment and over a period of time, for instance, global temperatures, mortality rates, pollution levels, etc.

1.7 Data Wrangling

- Data wrangling is the process of getting the data from its raw format into something suitable for more conventional analytics. Data wrangling can be defined as the process of cleaning, organizing and transforming raw data into the desired format for analysts to use for prompt decision - making.
- The data is often stored in a special purpose database that requires specialized tools to access. Fig 1.7.1 shows data wrangling process.

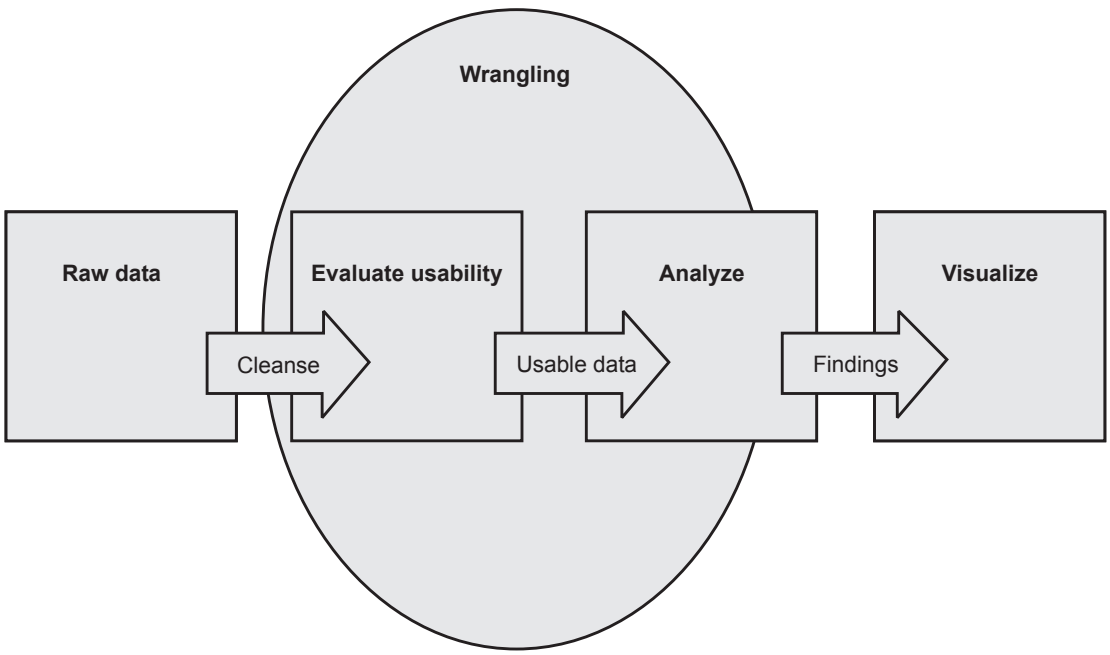


Fig. 1.7.1 Data wrangling process

- Data wrangling is the process of cleaning, structuring and enriching raw data into a desired format for better decision making in less time.
- The **necessity for data wrangling** is often a by - product of poorly collected or presented data. Data that is entered manually by humans is typically fraught with errors; data collected from websites is often optimized to be displayed on websites, not to be sorted and aggregated.

- Tasks in data wrangling include :
  a) Merging multiple datasets into one large dataset for analysis.
  b) Examining missing/gaps in data.
  c) Removing outliers or anomalies in datasets.
  d) Standardizing inputs.
- **Goals of data wrangling :**
  a) Creating valid and novel data out of messy data to drive decision - making in businesses.
  b) Standardizing raw data into formats that big data systems can ingest.
  c) Reducing the time spent by data analysts when creating data models by presenting orderly data.
  d) Creating consistency, completeness, usability and security for any dataset consumed or stored in a data warehouse.
- There are typically six iterative steps that make up the data wrangling process :
  1. **Discovering :** Before you can dive deeply, you must better understand what is in your data, which will inform how you want to analyze it. How you wrangle customer data, for example, may be informed by where they are located, what they bought or what promotions they received.
  2. **Structuring :** This means organizing the data, which is necessary because raw data comes in many different shapes and sizes. A single column may turn into several rows for easier analysis. One column may become two. Movement of data is made for easier computation and analysis.
  3. **Cleaning :** What happens when errors and outliers skew user data ? User clean the data. What happens when state data is entered as AP or Andhra Pradesh or Arunachal Pradesh ? You clean the data. Null values are changed and standard formatting implemented, ultimately increasing data quality.
  4. **Enriching :** Here you take stock in data and strategize about how other additional data might augment it. Questions asked during this data wrangling step might be : What new types of data can derive from what already have or what other information would better inform my decision making about this current data ?
  5. **Validating :** Validation rules are repetitive programming sequences that verify data consistency, quality and security. Examples of validation include ensuring uniform distribution of attributes that should be distributed normally (e.g., birth dates) or confirming accuracy of fields through a check across data.

6. **Publishing :** Analysts prepare the wrangled data for use downstream - Whether by a particular user or software and document any particular steps taken or logic used to wrangle said data. Data wrangling gurus understand that implementation of insights relies upon the ease with which it can be accessed and utilized by others.

1.7.1 Benefits of Data Wrangling

a) Data wrangling helps to improve data usability as it converts data into a compatible format for the end system.

b) It helps to quickly build data flows.

c) Integrates various types of information and their sources.

d) Help users to process very large volumes of data easily and easily share data - flow techniques.

e) Improved efficiency when it comes to data - driven decision making.

1.8 Data Cleaning

- Sometimes real - world data is incomplete, noisy, and inconsistent. Data cleaning methods are used for making useable data.
- Data cleaning tasks are as follows :
  1. Data acquisition and metadata
  2. Fill in missing values
  3. Unified date format
  4. Converting nominal to numeric
  5. Identify outliers and smooth out noisy data
  6. Correct inconsistent data
- Data cleaning is a first step in data pre-processing techniques which is used to find the missing value, smooth noise data, recognize outliers and correct inconsistent.

1.8.1 Missing Value

- These dirty data will affects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines.

How to handle noisy data in data mining ?

- Following methods are used for handling noisy data :
  1. **Ignore the tuple :** Usually done when the class label is missing. This method is not good unless the tuple contains several attributes with missing values.
  2. **Fill in the missing value manually :** It is time - consuming and not suitable for a large data set with many missing values.

3. **Use a global constant to fill in the missing value :** Replace all missing attribute values by the same constant.
4. **Use the attribute mean to fill in the missing value :** For example, suppose that the average salary of staff is ₹ 65000/-. Use this value to replace the missing value for salary.
5. Use the attribute mean for all samples belonging to the same class as the given tuple.
6. Use the most probable value to fill in the missing value.

1.8.2 Noisy Data

- **Noise :** Random error or variance in a measured variable.
- For numeric values, box plots and scatter plots can be used to identify outliers. To deal with these anomalous values, data smoothing techniques are applied, which are described below.

1. **Binning :** Using binning methods smooths sorted value by using the values around it. The sorted values are then divided into 'bins'. There are various approaches to binning. Two of them are smoothing by bin means where each bin is replaced by the mean of bin's values, and smoothing by bin medians where each bin is replaced by the median of bin's values.

Binning methods for data smoothing :

a) **In smoothing by bin means :** Each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 5, 9 and 13 in
Bin is 9. Therefore, each original value in this bin is replaced by the value 9.

b) **Smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median.

c) **Smoothing by bin boundaries :** The minimum and maximum bin values are stored at the boundary while intermediate bin values are replaced by the boundary value to which it is more closer.

2. **Regression :** Linear regression and multiple linear regression can be used to smooth the data, where the values are conformed to a function.
3. **Outlier analysis :** Approaches such as clustering can be used to detect outliers and deal with them.

1.9 Data Integration and Transformation

1.9.1 Data Integration

- Data integration combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute toward smooth data integration.
- With the increasing volume of data collected through a variety of sources and at a much faster velocity every day, it is very much clear that Data is and has been the most valuable possession.
- Data integration is important as it provides a unified view of the scattered data not only this it also maintains the accuracy of data.
- Issues in data integration : While integrating the data we have to deal with several issues.

1. Entity identification problem

- As we know the data is unified from the heterogeneous sources then how can we 'match the real-world entities from the data'. For example, we have customer data from two different data source.
- An entity from one data source has **customer_id** and the entity from the other data source has **customer_number**. Now how does the data analyst or the system would understand that these two entities refer to the same attribute ? The schema integration can be achieved using metadata of each attribute.

2. Redundancy

- Redundancy is one of the big issues during data integration. Redundant data is an unimportant data or the data that is no longer needed. It can also arise due to attributes that could be derived using another attribute in the data set.
- For example, one data set has the customer age and other data set has the customers date of birth then age would be a redundant attribute as it could be derived using the date of birth.
- The redundancy can be discovered using correlation analysis. The attributes are analyzed to detect their interdependency on each other thereby detecting the correlation between them.

1.9.2 Data Transformation

- In data transformation, the data are transformed or consolidated into forms appropriate for mining.

- Data transformation can involve the following :
  1. **Smoothing :** It removes noise from the data. Such techniques include binning, regression and clustering.
  2. **Aggregation :** An aggregation or summary operation is applied to the data.
  3. **Generalization** of the data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
  4. **Normalization :** The attribute data are scaled so as to fall within a small specified range.
  5. **Attribute construction :** New attributes are constructed and added from the given set of attributes to help the mining process.
- An attribute is normalized by scaling its values so that they fall within a small specified range. There are many methods for data normalization. They are min-max normalization, z-score normalization, and normalization by decimal scaling.
  a) Min-max normalization performs a linear transformation on the original data. It will scale the data between the 0 and 1.

Example :

Marks
8
10
15
20

Min : Minimum value of the given attribute. Here min is 8.

Max : Maxing value of the given attribute. Here max is 20.

V : V is the respective value of attribute.

For example :

V_1 : 8, V_2 = 10, V_3 = 15 and V_4 = 20.

New max : 1

Now min : 0

$$V' = \frac{V - Min_A}{Max_A - Min_A}(New - Max_A - New - Min_A) + New - Min_A$$

For mark 8 :

$$\text{minmax} = \frac{\text{V} - \text{min marks}}{\text{Max marks} - \text{Min marks}}(\text{New marks} - \text{New min}) + \text{New min}$$

$$\text{minmax} = \frac{8-8}{20-8} \times (1-0) + 0$$

$$\text{minmax} = \frac{(0)}{12} \times 1$$

For mark 10 :

$$\text{minmax} = \frac{(10-8)}{20-8} \times (1-0) + 0$$

$$\text{minmax} = \frac{2}{12} \times 1$$

$$\text{minmax} = 0.25$$

For mark 15 :

$$\text{minmax} = \frac{(15-8)}{20-8} \times (1-0) + 0$$

$$\text{minmax} = \frac{(3)}{12} \times 1$$

$$\text{minmax} = 0.25$$

For mark 20 :

$$\text{minmax} = \frac{(20-8)}{20-8} \times (1-0) + 0$$

$$\text{minmax} = \frac{12}{12} \times 1$$

$$\text{mimmax} = 1$$

Marks	Marks after min-max normalization
8	0
10	0.16
15	0.25
20	1

b) Decimal scaling : Decimal scaling is a data normalization technique. In this technique we move the decimal point of values of the attribute. His movement of decimal points totally depends on the maximum value among all values in the attribute.

Formula : A value V if attribute A can be obtained by normalization by the following formula.

Normalized value of attribute : = $(V^i/10^j)$

Example :

CGPA	Formula	CGPA normalized after decimal scaling
2	2/10	0.2
3	3/10	0.3

We will check maximum value among our attribute CGPA. Here maximum value is 3 so we can convert it into decimal by dividing with 10.

Example 1.9.1 *1) Minimum salary is ₹ 20,000 and maximum salary is ₹ 1,70,000. Map the salary ₹ 1,00,000 in new range of ₹ (60,000, 2,60,000) using min-max normalization method.*

2) If mean salary is ₹ 54,000 and standard deviation is ₹ 16,000 then find z score value of ₹ 73,600 salary.

Solution :

Solution 1 :

Old range = (20000, 1,70,000)

max = 1,70,000

min = 20000

New range = (60000, 260000)

new_max = 260000

new_min = 60000

V_i = 100000

$$V_i' = [\{(V_i - \min)(\max - \min)\} \times \{new\_max - new\_min\}] + new\_min$$

$$= [\{(80000/150000) \times 200000\}] + 60000$$

$$= [106666] + 60000$$

$$= 166666$$

Salary ₹ 100000 in old range is equal to salary ₹ 166666 in the new range.

Solution 2 :

$$\text{mean} = ₹\ 54{,}000$$

$$\text{Standard deviation} = ₹\ 16{,}000$$

$$\text{Z-score value of } 76{,}300 = \frac{(76{,}300 - \text{Mean})}{\text{Standard deviation}} = \frac{(54000)}{16000}$$

$$= \frac{54000}{16000} = 3.375$$

Z-score value of ₹ 73,600 salary is 3.375.

Example 1.9.2 *Use min-max normalization method to normalize the following group of data by setting min = 0 and max = 1, 200, 300, 400, 600, 1000.*

Solution :

i) Min-max normalization by setting min = 0 and max = 1.

Original data	200	300	400	600	1000
0, 1 normalized	0	0.125	0.25	0.5	1

ii) Z-score normalization

Original data	200	300	400	600	1000
0, 1 normalized	– 1.06	– 0.7	– 0.35	0.35	1.78

Example 1.9.3 *Suppose that the minimum and maximum values for the attribute income are $ 73,600 and $ 98,000, respectively. Normalize income value $ 73,600 to the range [0:0 ; 1 : 0] using min-max normalization method.*

Solution : The min-max normalization to transform value 73,600 onto the range [0.0, 1.0].

Given data : $\min_A = 12000$, $\max_A = 98000$, $\text{new\_min}_A = 0.0$,

$\text{new\_max}_A = 1.0$, $v = 73600$, $v' = ?$

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

$$v' = \frac{73600 - 12000}{98000 - 12000} \times (1.0 - 0.0) + 0.0$$

$$v' = 0.716$$

Example 1.9.4 *Consider the following group of data 200, 300, 400, 600, 1000.*

i) Use the min-max normalization to transform value 600 onto the range [0.0,1.0]

ii) Use the decimal scaling to transform value 600.

Solution : i) The min-max normalization to transform value 600 onto the range [0.0, 1.0].

Given data : $\min_A = 200$, $\max_A = 1000$, $\text{new\_min}_A = 0.0$,

$\text{new\_max}_A = 1.0$, $v = 600$, $v' = ?$

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

$$v' = \frac{600 - 200}{1000 - 200} \times (1.0 - 0.0) + 0.0$$

$$v' = 0.5$$

ii) Decimal scaling to transform value 600

$$v = 600, j = 3$$

$$v = \frac{v}{10^j} = \frac{600}{10^3} = 0.6$$

1.10 Data Reduction

- Data reduction is nothing but obtaining a reduced representation of the data set that is much smaller in volume but yet produces the same analytical results.
- Data reduction is a process that reduced the volume of original data and represents it in a much smaller volume. Data reduction techniques ensure the integrity of data while reducing the data.
- Data reduction does not affect the result obtained from data mining that means the result obtained from data mining before data reduction and after data reduction is the same.
- Data reduction strategies are as follows :
  1. **Data cube aggregation :** Aggregation operations are applied to the data in the construction of a data cube.
  2. **Dimensionality reduction :** In dimensionality reduction redundant attributes are detected and removed which reduce the data set size.
  3. **Data compression :** Encoding mechanisms are used to reduce the data set size.
  4. **Numerosity reduction :** In numerosity reduction where the data are replaced or estimated by alternative.

5. **Discretisation and concept hierarchy generation :** Where raw data values for attributes are replaced by ranges or higher conceptual levels.

- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Mining on the reduced data set should be more efficient yet produce the same analytical results.
- Strategies for data reduction include the following :
  1. **Data cube aggregation :** It is lowest level of a data cube. Summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to computer monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
  2. **Attribute subset selection :** Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes. The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.
- **Irrelevant attributes :** It contains no information that is useful for the data mining task at hand. For example; Student's roll number is often irrelevant to the task of predicting student marks or CGPA.
- **Redundant attributes :** Duplicate much or all of the information contained in one or more other attributes. For example : purchase price of a product and the amount of GST paid.
  3. **Dimensionality reduction :** Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.
- If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called **lossless**. If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called **lossy**.
- Lossy dimensionality reduction methods are Principal Components Analysis (PCA) and wavelet transforms.
- Principal Component Analysis (PCA) is to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retains most of the sample's information and useful for the compression and classification of data.
- In PCA, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that feature carries.

- Hence, PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.
- A Discrete Wavelet Transform (DWT) is a transform that decomposes a given signal into a number of sets, where each set is a time series of coefficients describing the time evolution of the signal in the corresponding frequency band.

4. **Numerosity reduction :** The numerosity reduction reduces the volume of the original data and represents it in a much smaller form. This technique includes two types parametric and non-parametric numerosity reduction.

- Log - linear models is an example of parametric method and nonparametric methods are histograms, clustering and sampling.
- Regression and log-linear linear regression models a relationship between the two attributes by modelling a linear equation to the data set.
- Log - linear regression analysis involves using a dependent variable measured by frequency counts with categorical or continuous independent predictor variables.

Clustering :

- Cluster analysis is a popular data discretization method. A clustering algorithm can be applied to discretize a numerical attribute (A) by partitioning the values of A into clusters or groups.
- Clustering takes the distribution of A into consideration, as well as the closeness of data points, and therefore is able to produce high-quality discretization results.
- Clustering can be used to generate a concept hierarchy for A by following either a top down splitting strategy or a bottom-up merging strategy, where each cluster forms a node of the concept hierarchy.

Sampling

- The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent item sets in S instead of D.
- Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random sample of the data. Different methods of sampling are Simple Random Sampling (SRS), stratified sampling, cluster sampling, systematic sampling and multistage sampling.

1.11 Data Discretization

Data Discretization :

- Data discretization means diving the range of continuous attribute into intervals. Actual data values are replaced by interval labels.

- It reduces the number of values for a given continuous attribute. Some classification algorithms only accept categorical attributes. It helps to a concise, easy-to-use, knowledge-level representation of mining results.
- Data discretization techniques can be categorized based on class information and which direction it proceeds. Class information is divided into two types : Supervised and unsupervised discretization. Categorized based on which direction it proceeds are of two types : Top - down and bottom - up.
- Discretization techniques can be classified as supervised and unsupervised discretization. Supervised discretization uses class information and unsupervised discretization does not use class information.
- **Top - down :** If the process starts by first finding one or a few points to split the entire attribute range, and then repeats this recursively on the resulting intervals. It is also called splitting.
- **Bottom - up :** It starts by considering all of the continuous values as potential split- points, removes some by merging neighbourhood values to form intervals, and then recursively applies this process to the resulting intervals. It is also called merging.
- Discretization can be performed recursively on an attribute to provide a hierarchical or multiresolution partitioning of the attribute values, known as a concept hierarchy. Concept hierarchies are useful for mining at multiple levels of abstraction.
- Discretization and concept hierarchy generation for numerical data uses following methods :
  a) Binning
  b) Histogram analysis
  c) Clustering analysis
  d) Entropy-based discretization
  e) Segmentation by natural partitioning

1.11.1 Concept Hierarchy Generation for Categorical Data

- Categorical data are discrete data. Categorical attributes have a finite number of distinct values, with no ordering among the values.
- Example : geographic location, job category and item type.
- Various methods are used for the generation of concept hierarchies for categorical data :
  a) Specification of a partial ordering of attributes explicitly at the schema level by users or experts
  - Example : A relational database or a dimension location of a data warehouse may contain the following group of attributes : Street, city, province or state and country.

- A user or expert can easily define a concept hierarchy by specifying ordering of the attributes at the schema level.
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level, such as : *street < city < province or state < country.*

b) Specification of a portion of a hierarchy by explicit data grouping

- We can easily specify explicit groupings for a small portion of intermediate - level data.
- For example, after specifying that area and country form a hierarchy at the schema level, a user could define some intermediate levels manually, such as : {India, Maharashtra, Pune} < SPPU.

c) Specification of a set of attributes, but not of their partial ordering

- A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.
- The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept.
- Example : Suppose a user selects a set of location-oriented attributes, street, country, state and city, from the database, but does not specify the hierarchical ordering among the attributes.
- Fig. 1.11.1 shows automatic generation of a schema concept hierarchy based on the number of distinct attribute values.

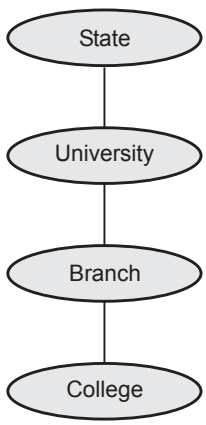


Fig. 1.11.1 Automatic generation of a schema concept hierarchy

1.12 Multiple Choice Questions

Q.1 Which of the following are NOT data reduction techniques ?

a Data cube aggregation
b Numerosity reduction
c Attribute subset selection
d Decimal scaling

Q.2 _____ normalization performs a linear transformation on the original data.

a z-score
b Min-max
c Smoothing
d Aggregation

Q.3 Which of the following methods to perform dimension reduction ?

a Missing values
b Decision tree
c Random forest
d All of these

Q.4 _____ is usually performed as an iterative two-step process consisting of discrepancy detection and data transformation.

a Data reduction
b Data integration
c Data cleaning
d Data transformation

Q.5 A _____ measure is a measure that must be computed on the entire data set as a whole.

a distributive
b holistic
c unimodal
d bimodal

Q.6 KDD stands for _____.

a Knowledge Database in Discovery
b Known Discovery in Databases
c Known Distributed Databases
d Knowledge Discovery in Databases

Q.7 Which of the following is NOT data preprocessing techniques ?

a Data cleaning
b Data integration
c Data handling
d Data transformation

Q.8 **Concept hierarchies are a form of data _____ that can also be used for data smoothing.**

a binaryzation b discretization

c missing d all of these

Q.9 **Strategies for data reduction include _____.**

a attribute subset selection b numerosity reduction

c data cube aggregation d all of these

Answer Keys for Multiple Choice Questions :

Q.1	d	**Q.2**	b	**Q.3**	d
Q.4	c	**Q.5**	b	**Q.6**	d
Q.7	c	**Q.8**	b	**Q.9**	d

❑❑❑

UNIT II

2 Statistical Inference

Syllabus

Need of statistics in Data Science and Big Data Analytics, ***Measures of Central Tendency :*** *Mean, Median, Mode, Mid - range.* ***Measures of Dispersion :*** *Range, Variance, Mean Deviation, Standard Deviation. Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi - Square Tests, t - test.*

Contents

2.1 Need of Statistics in Data Science and Big Data Analytics

- Statistics is the science of collecting, analyzing and understanding data and accounting for the relevant uncertainties. As such, it permeates the physical, natural and social sciences, public health, medicine, business and policy.
- As such, statistics is a fundamental tool of data scientists, who are expected to gather and analyze large amounts of structured and unstructured data and report on their findings.
- Statistics is a way to get information from data. Statistics is a tool for creating an understanding from a set of numbers. Statistics is the science of collecting, organizing, summarizing, analyzing, and interpreting information.
- Descriptive statistics involves computing values which summarize a set of data. This typically includes statistics like the mean, standard deviation, median, min, max, etc. which are called summary statistics.
- Fig. 2.1.1 shows descriptive statistics

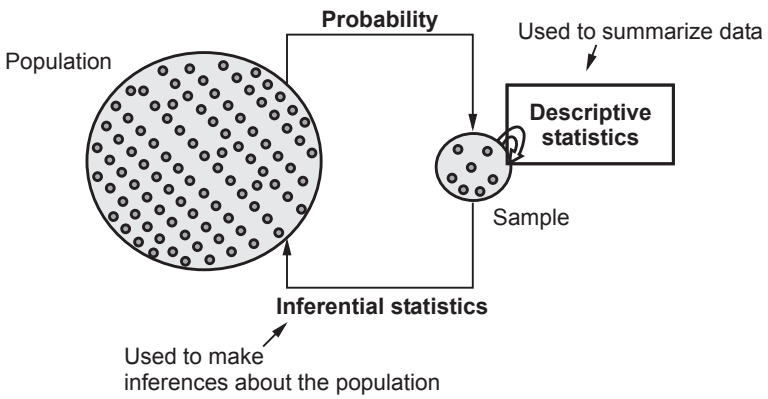


Fig. 2.1.1 Descriptive statistics

- Descriptive statistics is key because it allows us to present large amounts of raw data in a meaningful way. This enables a better interpretation of data.
- Descriptive statistics are also used to represent data graphically. Histograms, pie charts, bar graphs, and so on are nothing but a visual way of representing the data. This helps put it in the form that allows for analysis and interpretation.
- Big data is the collection and analysis of data sets that are complex in terms of the volume and variety, and in some cases the velocity at which they are collected. Big data are especially challenging because some of them were not collected to address a specific scientific question.

- Data scientists use a combination of statistical formulas and computer algorithms to notice patterns and trends within data. Then, they use their knowledge of social sciences and a particular industry or sector to interpret the meaning of those patterns and how they apply to real-world situations. The purpose is to generate value for a business or organization.
- While big data focuses on providing tools and techniques for managing and processing large and diverse quantities of data, it is not as focused on interpreting the data processing results to support decision making. This is where the area of data science continues and focuses on using advanced statistical techniques to analyze big data and interpret the results in a domain - specific context.
- Not everything is a table and not everything has a predefined schema. The massive data generation originating from social networks, mobile devices, sensors, and other data sources raised challenges that motivated the creation of novel tools and techniques.
- Initially, big data was characterized by 3 Vs : volume, variety, and velocity. Thus, enormous volumes of different and fast-growing data challenged the RDBMSs that do not scale easily due to ACID properties and fixed schema requirements.
- Many new Vs have emerged since then, such as data variability, veracity, value, etc. This caused the creation of new tools and frameworks that have the purpose of addressing one or more of the novel challenges.

2.2 Measures of Central Tendency

We look at various ways to measure the central tendency of data, include : Mean, Weighted mean, Trimmed mean, Median, Mode and Midrange.

1. Mean :

- The mean of a data set is the average of all the data values. The sample mean $\bar{x}$ is the point estimator of the population mean μ.

$$\text{Sample mean } \bar{x} = \frac{\text{Sum of the values of the n observations}}{\text{Number of observations in the sample}} = \frac{\sum x_i}{n}$$

$$\text{Population mean } \mu = \frac{\text{Sum of the values of the N observations}}{\text{Number of observations in the population}} = \frac{\sum x_i}{n}$$

2. Median :

- The median of a data set is the value in the middle when the data items are arranged in ascending order. Whenever a data set has extreme values, the median is the preferred measure of central location.

- The median is the measure of location most often reported for annual income and property value data. A few extremely large incomes of property values can inflate the mean.
- For an off number of observations :

$$\text{7 observations} = 26, 18, 27, 12, 14, 29, 19$$

$$\text{Numbers in ascending order} = 12, 14, 18, 19, 26, 27, 29$$

- The median is the middle value.

$$\text{Median} = 19$$

- For an even number of observations :

$$\text{8 observations} = 26, 18, 29, 12, 14, 27, 30, 19$$

$$\text{Numbers in ascending order} = 12, 14, 18, 19, 26, 27, 29, 30$$

The median is the average of the middle two values.

$$\text{Median} = \frac{(19+26)}{2} = 22.5$$

3. Mode :

- The mode of a data set is the value that occurs with greatest frequency. The greatest frequency can occur at two or more different values. If the data have exactly two modes, the data have exactly two modes, the data are bimodal. If the data have more than two modes, the data are multimodal.
- **Weighted mean :** Sometimes, each value in a set may be associated with a weight, the weights reflect the significance, importance, or occurrence frequency attached to their respective values.
- **Trimmed mean :** A major problem with the mean is its sensitivity to extreme (e.g., outlier) values. Even a small number of extreme values can corrupt the mean. The trimmed mean is the mean obtained after cutting off values at the high and low extremes.
- For example, we can sort the values and remove the top and bottom 2 % before computing the mean. We should avoid trimming too large a portion (such as 20 %) at both ends as this can result in the loss of valuable information.
- **Holistic measure** is a measure that must be computed on the entire data set as a whole. It cannot be computed by partitioning the given data into subsets and merging the values obtained for the measure in each subset.

2.3 Measures of Dispersion

- An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population.
- **First quartile (Q_1) :** The first quartile is the value, where 25 % of the values are smaller than Q_1 and 75 % are larger.
- **Third quartile (Q_3) :** The third quartile is the value, where 75 % of the values are smaller than Q_3 and 25 % are larger.
- The **box plot** is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles. If the lower quartile is Q_1 and the upper quartile is Q_3, then the difference (Q_3 - Q_1) is called the interquartile range or IQ.
- **Range :** Difference between highest and lowest observed values.

Variance :

- The variance is a measure of variability that utilizes all the data. It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$) for a sample, μ for a population.
- The variance is the average of the squared between each data value and the mean.

$$\text{Sample variance : } S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{Population variance : } \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Standard deviation :

- The standard deviation of a data set is the positive square root of the variance. It is measured in the same in the same units as the data, making it more easily interpreted than the variance.
- The standard deviation is computed as follows :

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

$$\text{Sample standard deviation} = S = \sqrt{S^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Difference between standard deviation and variance

Sr. No.	Standard deviation	Variance
1.	Standard deviation is a measure of dispersion of the values of a data set from their mean.	It is the statistical measure of how far the numbers are spread in a data set from their average.
2.	It is a common term in statistical theory to calculate central tendency.	Variance is primarily used for statistical probability distribution to measure volatility from the mean.
3.	It measures the absolute variability of the dispersion.	It helps determine the size of the data spread.
4.	It is calculated by taking the square root of the variance.	It is calculated by taking the average of the squared deviation of each value in the data set from the mean.
5.	The standard deviation is symbolized by the Greek letter sigma "σ" as in lower case sigma.	The notation for the variance of a variable is "σ^2" sigma squared.
6.	$\sigma = \sqrt{\frac{\sum (x - M)^2}{n}}$ where M = mean, x = a values in a data set and n = number of values	$\sigma^2 = \frac{\sum (x - M)^2}{n}$ where M = mean, x = each value in the data set, n = number of values in the data set
7.	Used in finance sector as a measure of market and security volatility.	Used in asset allocation.

Example 2.3.1 *Find sample mean and sample standard deviation for the following data set : 5, 10, 15, 20.*

Solution : Sample mean ($\bar{x}$) :

$$\bar{x} = \frac{\sum x}{n} = \frac{5 + 10 + 15 + 20}{4} = \frac{50}{4} = 12.5$$

Sample standard deviation :

Data	$x - \bar{x}$	$(x - \bar{x})^2$
5	$5 - (12.5) = -7.5$	$(-7.5)^2 = 56.25$
10	$10 - (12.5) = -2.5$	$(-2.5)^2 = 6.25$
15	$15 - (12.5) = 2.5$	$(2.5)^2 = 6.25$
20	$20 - (12.5) = 7.5$	$(7.5)^2 = 56.25$
		$\sum (x - \bar{x})^2 = 125.01$

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{125.01}{4 - 1}} \approx 6.455$$

Example 2.3.2 *The heights at the shoulders are : 600 mm, 470 mm, 170 mm, 430 mm and 300 mm. Find out the mean and the variance.*

Solution :

Your first step is to find the Mean : $\frac{600 + 470 + 170 + 430 + 300}{5} = 394$

Variance :

$600 - 394 = 206$

$470 - 394 = 76$

$170 - 394 = -224$

$430 - 394 = 36$

$300 - 394 = -94$

$$\text{Variance} = \frac{(206)^2 + (76)^2 + (-224)^2 + (36)^2 + (-94)^2}{5}$$

$$\text{Variance} = \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} = 21704$$

2.4 Bayes Theorem

- Bayes' theorem is a method to revise the probability of an event given additional information. Bayes's theorem calculates a conditional probability called a posterior or revised probability.
- Bayes' theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, P(A | B) denotes the conditional probability of A occurring, given that B occurs. The two conditional probabilities P(A | B) and P(B | A) are in general different.
- Bayes' theorem gives a relation between P(A | B) and P(B | A). An important application of Bayes' theorem is that it gives a rule how to update or revise the strengths of evidence-based beliefs in light of new evidence a posteriori.
- A **prior probability** is an initial probability value originally obtained before any additional information is obtained.
- A **posterior probability** is a probability value that has been revised by using additional information that is later obtained.

- Suppose that $B_1, B_2, B_3, \ldots, B_n$ partition the outcomes of an experiment and that A is another event. For any number, k, with $1 \leq k \leq n$, we have the formula :

$$P(B_k/A) = \frac{P(A/B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A/B_i) \cdot P(B_i)}$$

Example 2.4.1 *A mechanical factory production line is manufacturing bolts using three machines, A, B and C. The total output, machine A is responsible for 25 %, machine B for 35 % and machine C for the rest. The machines that 5 % of the output from machine A is defective, 4 % from machine B and 2 % from machine C. A bolt is chosen at random from the production line and found to be defective. What is the probability that it came from : i. machine A ii. machine B iii. machine C ?*

Solution : Let,

D = {bolt is defective},

A = {bolt is from machine A},

B = {bolt is from machine B},

C = {bolt is from machine C}.

Given data : P(A) = 0.25, P(B) = 0.35, P(C) = 0.4.

$P(D|A) = 0.05$, $P(D|B) = 0.04$, $P(D|C) = 0.02$.

From the Bayes' Theorem :

$$P(A/D) = \frac{P(D/A) \times P(A)}{P(D/A) \times P(A) + P(D/B) \times P(B) + P(D/C) \times P(C)}$$

$$= \frac{0.05 \times 0.25}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= \frac{0.0125}{0.0125 + 0.014 + 0.008}$$

$$\mathbf{P(A/D) = 0.3621}$$

Similarly :

$$P(B/D) = \frac{P(D/B) \times P(B)}{P(D/A) \times P(A) + P(D/B) \times P(B) + P(D/C) \times P(C)}$$

$$= \frac{0.04 \times 0.35}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= \frac{0.014}{0.0125 + 0.014 + 0.008} = \frac{0.014}{0.0345}$$

$$\mathbf{P(B/D) = 0.4057}$$

$$P(C/D) = \frac{P(D/C) \times P(C)}{P(D/A) \times P(A) + P(D/B) \times P(B) + P(D/C) \times P(C)}$$

$$= \frac{0.02 \times 0.4}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= \frac{0.008}{0.0125 + 0.014 + 0.008} = \frac{0.008}{0.0345}$$

$$\mathbf{P(C/D) = 0.2318}$$

Example 2.4.2 *At a certain university, 4 % of men are over 6 feet tall and 1 % of women are over 6 feet tall. The total student population is divided in the ratio 3 : 2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman ?*

Solution : Let us assume following :

M = {Student is Male},

F = {Student is Female},

T = {Student is over 6 feet tall}.

Given data : $P(M) = 2/5$,

$P(F) = 3/5$,

$P(T|M) = 4/100$

$P(T|F) = 1/100$.

We require to find $P(F|T)$?

Using Bayes' Theorem we have :

$$P(F/T) = \frac{P(T/F)\,P(F)}{P(T/F)\,P(F) + P(T/M)\,P(M)}$$

$$= \frac{\frac{1}{100} \times \frac{3}{5}}{\frac{1}{100} \times \frac{3}{5} + \frac{4}{100} \times \frac{2}{5}} = \frac{\frac{3}{500}}{\frac{3}{500} + \frac{8}{500}}$$

$$P(F/T) = \frac{3}{11}$$

Example 2.4.3 *A pair of dice is rolled. If the sum of 9 has appeared, find the proabability that one of the dice shows 3.*

Solution : Let A = The event that the sum is 9

B = The event the one of dice shows 3.

Exhaustive cases = $6^2 = 36$.

Favorable cases of the event A = (3, 6), (6, 3), (4, 5), (5, 4).

So $P(A) = 4.36$

$$P(A) = \frac{1}{9}$$

Favorable case for the event $A \cap B$ = (3, 6), (6, 3)

Hence $$P(A \cap B) = \frac{2}{36} = \frac{1}{18}$$

But $$P(A \cap B) = P(A) \times P(B/A)$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

$$P(B/A) = \frac{1/18}{1/9} = \frac{1}{18} \times \frac{9}{1}$$

$$P(B/A) = 1/2$$

Example 2.4.4 *At a certain university, 4 % of men are over 6 feet tall and 1 % of women are over 6 feet tall. The total student population is divided in the ratio 3 : 2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman ?*

Solution :

Let M = {Student is Male},

F = {Student is Female},

(note that M and F partition the sample space of students),

T = {Student is over 6 feet tall}.

We know that

$P(M) = 2/5$,

$P(F) = 3/5$,

$$P(T|M) = 4/100$$

$$P(T|F) = 1/100.$$

We require $P(F|T)$.

Using Bayes' Theorem we have :

$$P(F|T) = \frac{P(T/F)P(F)}{P(T/F)\,P(F) + P(T/M)\,P(M)} = \frac{\frac{1}{100} \times \frac{3}{5}}{\frac{1}{100} \times \frac{3}{5} + \frac{4}{100} \times \frac{2}{5}} = \frac{3}{11}$$

2.5 Hypothesis

SPPU : Aug.-18, Oct.-19

- General definition of a hypothesis : **"A hypothesis is a statement of a relationship between two or more variables."** A statistical hypothesis is simply a particular kind of hypothesis.
- A hypothesis is a statement or claim regarding a characteristic of one or more populations. Hypothesis testing is a procedure, based on sample evidence and probability, used to test claims regarding a characteristic of one or more populations.
- A statistical hypothesis is either
  1. A statement about the value of a population parameter (e.g., mean, median, mode, variance, standard deviation, proportion, total), or
  2. A statement about the kind of probability distribution that a certain variable obeys.
- The null hypothesis, denoted H_0 (read "H-naught"), is a statement to be tested. The null hypothesis is assumed true until evidence indicates otherwise. The alternative hypothesis, denoted, H_1 (read "H-one"), is a claim to be tested. We are trying to find evidence for the alternative hypothesis.
- Examples of statistical hypotheses :
  a. The mean age of all college students is 20.4 years. **(simple hypothesis)**
  b. The proportion of college students who are men is 60 %. **(simple hypothesis)**
  c. The proportion of books in the college library whose heights exceed 30 cm is less than or equal to 0.13. **(composite hypothesis)**
- A statistical hypothesis that specifies a single value for a population parameter is called a simple hypothesis; every statistical hypothesis that is not simple is called composite.

2.5.1 Hypothesis Testing

- A statistical hypothesis test is a procedure for deciding between two possible statements about a population. The phrase significance test means the same thing as the phrase "hypothesis test."
- A hypothesis test is a statistical method that uses sample data to evaluate a hypothesis about a population. The general goal of a hypothesis test is to rule out chance as a plausible explanation for the results from a research study.
- The goal in hypothesis testing is to analyze a sample in an attempt to distinguish between population characteristics that are likely to occur and population characteristics that are **unlikely** to occur.

Basic assumption of hypothesis testing

- If the treatment has any effect, it is simply to add or subtract a constant amount to each individual's score.
- Remember that adding or subtracting constant changes the mean, but not the shape of the distribution for the population and/or the standard deviation.
- The population after treatment has the same shape and standard deviation as the population prior to treatment.
- If the individuals in the sample are noticeably different from the individuals in the original population, we have evidence that the treatment has an effect.

The purpose of the hypothesis test is to decide between two explanations :

1. The difference between the sample and the population can be explained by sampling error.
2. The difference between the sample and the population is too large to be explained by sampling error.

Steps in hypothesis testing

1. Specify the null hypothesis.
2. Specify the alternative hypothesis
3. Set the significance level (?)
4. Calculate the test statistic and corresponding P-value.
5. Display the conclusion.

Step 1 : Formulate the hypothesis

- A null hypothesis is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.
- An alternative hypothesis is one in which some difference or effect is expected.

- The null hypothesis refers to a specified value of the population parameter, not a sample statistic.

Step 2 : Select an appropriate test

- The **test statistic** measures how close the sample has come to the null hypothesis.
- The test statistic often follows a well-known distribution (e.g., normal, t, or chi-square).
- Calculate Z statistic.

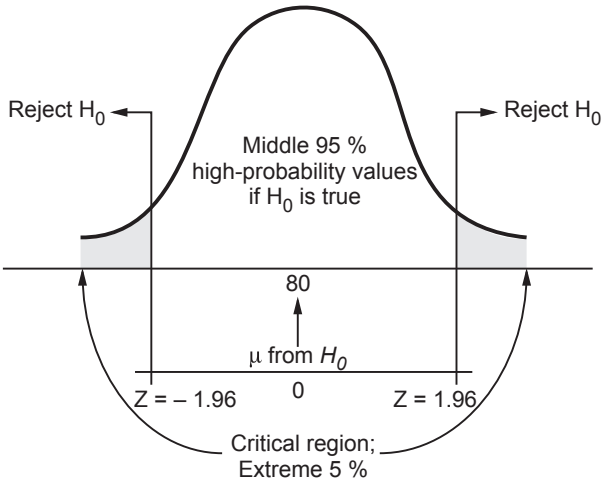


Fig. 2.5.1

Step 3 : Choose level of significance

Type I Error

- Occurs if the null hypothesis is rejected when it is in fact true.
- The probability of type I error (α) is also called the **level of significance.**

Type II Error

- Occurs if the null hypothesis is not rejected when it is in fact false.
- The probability of type II error is denoted by β.
- Unlike α, which is specified by the researcher, the magnitude of β depends on the actual value of the population parameter (proportion).
- **It is necessary to balance the two types of errors.**
- The power of a test is the probability $(1 - \beta)$ of rejecting the null hypothesis when it is false and should be rejected. Although β is unknown, it is related to α.

Step 4 : Collect data and calculate test statistic

- The required data are collected and the value of the test statistic computed. The test statistic z can be calculated as follows :

$$Z_{cal} = \frac{\hat{P} - \pi}{\sigma_P}$$

Step 5 : Determine probability value/critical value

- Using standard normal tables.
- Note, in determining the critical value of the test statistic, the area to the right of the critical value is either α or $\alpha/2$. It is α for a one-tail test and $\alpha/2$ for a two-tail test.

- If the prob associated with the calculated value of the test statistic (Z_{cal}) is less than the level of significance (α), the null hypothesis is rejected.
- Alternatively, if the calculated value of the test statistic is greater than the critical value of the test statistic (z_α), the null hypothesis is rejected.

1. **Two-tailed alternative :** If the alternative states that a population parameter is different from a specific value. The corresponding test is called a two-tailed test.
2. **Right-tailed alternative :** If the alternative states that a population parameter is greater than a specific value. The corresponding test is called a right-tailed test.
3. **Left-tailed alternative :** If the alternative states that a population parameter is less than a specific value. The corresponding test is called a left-tailed test.

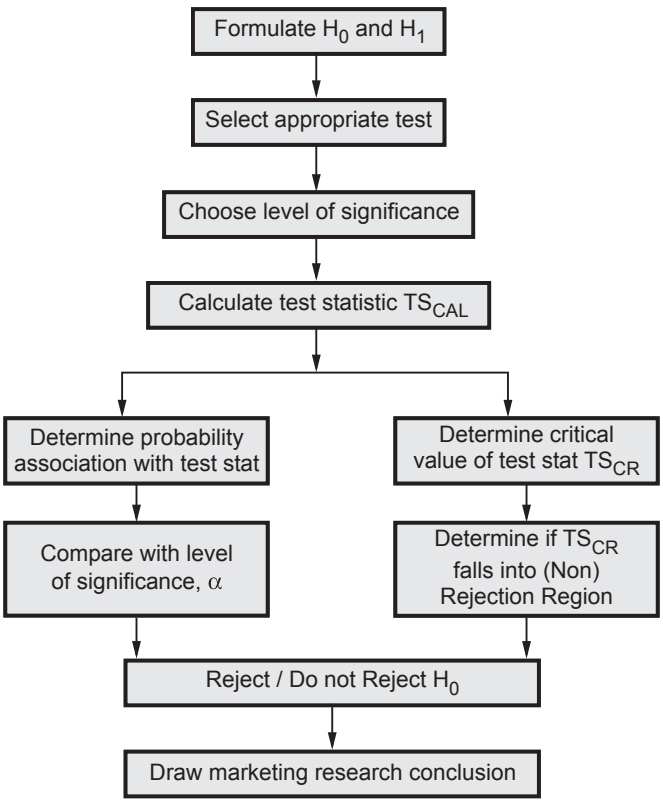


Fig. 2.5.2

Decide the rejection region of the test

- Based on the test statistic and a given confidence level, we can determine the rejection region, the acceptance region, and the critical value of the test.
- Rejection region is the region in which we can reject the null-hypothesis when the test statistics falls in this region. Acceptance region is simply the complement of the rejection region.

- Critical value is the value on the boundary of the rejection region and acceptance region.

1) For arbitrary population, acceptance and rejection regions are shown in Fig. 2.5.3.

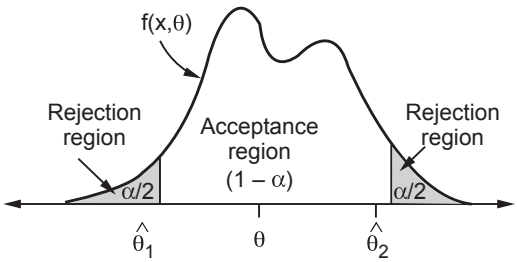


Fig. 2.5.3 Arbitrary population

2) For normal population, acceptance and rejection regions are shown in Fig. 2.5.4.

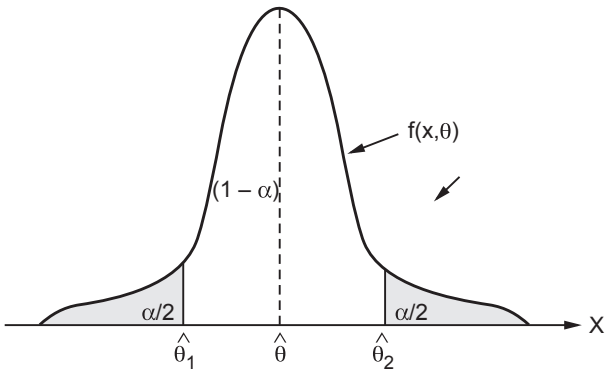


Fig. 2.5.4 Normal population

P-value and hypotheses testing

- As an alternative approach to the rejection/acceptance-region approach, we can calculate a probability related to the test statistic, called P-value, and base our decision of rejection/acceptance on the magnitude of the P-value.
- P-value is the probability to observe a value of the test statistic as extreme as the one observed, if the null hypothesis is true. So a small P-value indicates that the null hypothesis is not true and hence should be rejected.

In a hypothesis testing problem :

a) The null hypothesis will not be rejected unless the data are not unusual (given that the hypothesis is true).

b) The null hypothesis will not be rejected the P-value indicates the data are very unusual (given that the hypothesis is true).

c) The null hypothesis will not be rejected only if the probability of observing the data provides convincing vidence that it is true.

d) The null hypothesis is also called the research hypothesis; the alternative hypothesis often represents the status quo.

e) The null hypothesis is the hypothesis that we would like to prove; the alternative hypothesis is also called the research hypothesis.

2.5.2 Null and Alternative Hypothesis

1. Null hypothesis (H_0)

- The null hypothesis states that there is no change in the general population before and after an intervention. In the context of an experiment, H_0 predicts that the independent variable had no effect on the dependent variable.
- The null hypothesis is the stated or assumed value of a population parameter. When trying to identify the population parameter needed for your solution, look for the following phrases :
  i. "It is known that…"
  ii. "Previous research shows…"
  iii. "The company claims that…"
  iv. "A survey showed that…"
- When writing the null hypothesis, make sure it includes an "=" symbol.

Null hypothesis : Population characteristic = Hypothesized value

2. Alternate hypothesis (H_1)

- The alternative hypothesis states that there is a change in the general population following an intervention. In the context of an experiment, predicts that the independent variable did have an effect on the dependent variable.
- The alternative hypothesis must be true when the null hypothesis is false.
- The alternate hypothesis is the stated or assumed value of a population parameter if the null hypothesis is rejected. When trying to identify the information needed for alternate hypothesis statement, look for the following phrases :
  i. "Is it reasonable to conclude…"
  ii. "Is there enough evidence to substantiate…"
  iii "Does the evidence suggest…"
  iv. "Has there been a significant…"
- When writing the alternate hypothesis, make sure it never includes an "=" symbol.
- **Alternative hypothesis :** Three possibilities
  a. Upper tailed test : H_1 : Population characteristic > Hypothesised value
  b. Lower tailed test : H_1 : Population characteristic < Hypothesised value
  c. Two tailed test : H_1 : Population characteristic ≠ Hypothesised value

3. Multiple hypothesis testing :

- The multiple hypothesis testing problem is the situation when we wish to consider many hypotheses simultaneously. For example, suppose we have n genes and data about expression levels for each gene among healthy individuals and those with prostate cancer.
- However, consider a case where you have 20 hypotheses to test and a significance level of 0.05. What's the probability of observing at least one significant result just due to chance ?

$$\begin{aligned} P(\text{at least one significant result}) &= 1 - P(\text{no. significant results}) \\ &= 1 - (1 - 0.05)^{20} \\ &= 0.64 \end{aligned}$$

- So, with 20 tests being considered, we have a 64 % chance of observing at least one significant result, even if all of the tests are actually not significant.
- In genomics and other biology-related fields, it's not unusual for the number of simultaneous tests to be quite a bit larger than 20 and the probability of getting a significant result simply due to chance keeps going up.

2.5.3 Difference between Null Hypothesis and Alternative Hypothesis

Sr. No.	Null hypothesis	Alternative hypothesis
1.	Represented by H_0.	Represented by H_1.
2.	Statement about the value of a population parameter.	Statement about the value of a population parameter that must be true if the null hypothesis is false.
3.	Always stated as an equality.	Stated in on of three forms : $>$, $<$, $\neq$
4.	This is the hypothesis or claim that is initially assumed to be true.	This is the hypothesis or claim which we initially assume to be false but which we may decide to accept if there is sufficient evidence.
5.	Independent variable had **no effect** on the dependent variable.	Independent variable **did have an effect** on the dependent variable.

Review Questions

1. *Explain hypothesis testing with example.* **SPPU : Aug.-18 (In Sem), Marks 4**
2. *Explain Type 1 and Type 2 errors.* **SPPU : Aug.-18 (In Sem), Marks 2**
3. *Compare Type - I and Type - II errors.* **SPPU : Oct.-19 (In Sem), Marks 5**
4. *Explain hypothetical testing in detail with example.* **SPPU : Oct.-19 (In Sem), Marks 5**

2.6 Pearson Correlation

- The Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value r = 1 means a perfect positive correlation and the value r = – 1 means a perfect negative correlation.
- Correlation means to find out the association between the two variables and Correlation coefficients are used to find out how strong the is relationship between the two variables. The most popular correlation coefficient is Pearson's correlation coefficient. It is very commonly used in linear regression.
- Pearson's Correlation coefficient is represented as 'r', it measures how strong is the linear association between two continuous variables using the formula :

$$r = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where, r = Pearson correlation coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ = mean of values in y variable

- Fig. 2.6.1 shows various correlation with scatter diagram.

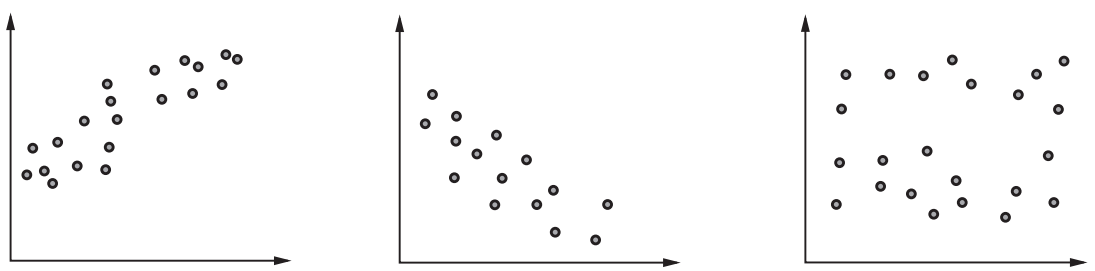


Fig. 2.6.1 Various correlation with scatter diagram

- Value of 'r' ranges from '– 1' to '+ 1'. Value '0' specifies that there is no relation between the two variables. A value greater than '0' indicates a positive relationship between two variables where an increase in the value of one variable increases the value of another variable. Value less than '0' indicates a negative relationship between two variables where an increase in the value of one decreases the value of another variable.

- Requirements for Pearson's correlation coefficient
  a) Scale of measurement should be interval or ratio.
  b) Variables should be approximately normally distributed.
  c) The association should be linear.
  d) There should be no outliers in the data.
- When an investigator has collected two series of observations and wishes to see whether there is a relationship between them, he or she should first construct a scatter diagram. The vertical scale represents one set of measurements and the horizontal scale the other.
- If one set of observations consists of experimental results and the other consists of a time scale or observed classification of some kind, it is usual to put the experimental results on the vertical axis. These represent what is called the "dependent variable". The "independent variable", such as time or height or some other observed classification, is measured along the horizontal axis, or baseline.

Example 2.6.1 *Given below are the monthly income and their net savings of a sample of 10 supervisory staff belonging to a firm. Calculate the correlation coefficient.*

Employee No.	1	2	3	4	5	6	7	8	9	10
Monrhly income (₹)	780	360	980	250	750	820	900	620	650	390
Net savings	84	51	91	60	68	62	86	58	53	47

Solution : Following table list out some calculations.

Monthly income X	Net savings, y	X^2	Y^2	XY
780	84	608400	7056	65520
360	51	129600	2601	18360
980	91	960400	8281	89180
250	60	62500	3600	15000
750	68	562500	4624	51000
820	62	672400	3844	50840
900	86	810000	7396	77400
620	58	384400	3364	35960
650	53	422500	2809	34450

390	47	152100	2209	18330
$\sum X = 6500$	$\sum Y = 660$	$\sum X^2 = 4764800$	$\sum Y^2 = 45784$	$\sum XY = 456040$

Here, n = 10,

$$\overline{X} = \frac{\sum X}{n} = \frac{6500}{10} = 650$$

and

$$\overline{Y} = \frac{\sum Y}{n} = \frac{660}{10} = 66$$

Correlation coefficient is given as,

$$r(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum XY - \overline{X}\,\overline{Y}}{\sqrt{\left(\frac{1}{n}\sum X^2 - \overline{X}^2\right)\left(\frac{1}{n}\sum Y^2 - \overline{Y}^2\right)}}$$

Putting values in above equation

$$r(X, Y) = \frac{\frac{456040}{10} - 650 \times 66}{\sqrt{\left(\frac{4764800}{10} - 650^2\right)\left(\frac{45784}{10} - 66^2\right)}} = 0.7804$$

Example 2.6.2 *Following is the score of seven students in management accounting (X) and business statistics (Y). Calculate correlation between the score in two subjects.*

Student no.	1	2	3	4	5	6	7
Score X	40	70	84	74	26	78	48
Score Y	64	74	100	60	50	48	80

Solution : Following table lists out some calculations.

X	Y	X^2	Y^2	XY
40	64	1600	4096	2560
70	74	4900	5476	5180
84	100	7056	10000	8400
74	60	5476	3600	4440

26	50	676	2500	1300
78	48	6084	2304	3744
48	80	2304	6400	3840
$\sum X = 420$	$\sum Y = 476$	$\sum X^2 = 28096$	$\sum Y^2 = 34376$	$\sum XY = 29464$

Here, $$\overline{X} = \frac{\sum X}{n} = \frac{420}{7} = 60$$

and $$\overline{Y} = \frac{\sum Y}{n} = \frac{476}{7} = 68$$

Correlation coefficient is given as,

$$r(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum XY - \overline{X}\,\overline{Y}}{\sqrt{\left(\frac{1}{n}\sum X^2 - \overline{X}^2\right)\left(\frac{1}{n}\sum Y^2 - \overline{Y}^2\right)}}$$

Putting values in above equation

$$r(X, Y) = \frac{\frac{29464}{7} - 60 \times 68}{\sqrt{\left(\frac{28096}{7} - 60^2\right)\left(\frac{34376}{7} - 68^2\right)}} = 0.3749$$

Example 2.6.3 *The competitors in a beauty contest are ranked by three judges in the following order. Use rank correlation coefficient to discuss which pair of judges has nearest approach to beauty.*

1st Judge	1	5	4	8	9	6	10	7	3	2
2nd Judge	4	8	7	6	5	9	10	3	2	1
3rd Judge	6	7	8	1	5	10	9	2	3	4

Solution : Here n = 10.

Let us form the following table of calculations.

1st Judge x	2nd Judge y	3rd Judge z	d_1 x − y	d_2 y − z	d_3 z − x	d_1^2	d_2^2	d_3^2
1	4	6	− 3	− 2	5	9	4	25
5	8	7	− 3	1	2	9	1	4
4	7	8	− 3	− 1	4	9	1	16
8	6	1	2	5	− 7	4	25	49
9	5	5	4	0	− 4	16	0	16
6	9	10	− 3	− 1	4	9	1	16
10	10	9	0	1	− 1	0	1	1
7	3	2	4	1	− 5	16	1	25
3	2	3	1	− 1	0	1	1	0
2	1	4	1	− 3	2	1	9	4
			$\sum d_1 = 0$	$\sum d_2 = 0$	$\sum d_3 = 0$	$\sum d_1^2 = 74$	$\sum d_2^2 = 44$	$\sum d_3^2 = 156$

From above table, correlations are calculated as follows :

$$\rho(x, y) = 1 - \frac{6\sum d_1^2}{n(n^2 - 1)} = 1 - \frac{6 \times 74}{10\,(100 - 1)} = 0.5515$$

$$\rho(y, z) = 1 - \frac{6\sum d_2^2}{n(n^2 - 1)} = 1 - \frac{6 \times 44}{10\,(100 - 1)} = 0.7333$$

$$\rho(z, x) = 1 - \frac{6\sum d_3^2}{n(n^2 - 1)} = 1 - \frac{6 \times 156}{10\,(100 - 1)} = 0.0545$$

Since $\rho(y, z)$ is maximum 2nd and 3rd judge have nearest approach to beauty.

2.7 Chi - square Tests

- There are many theoretical distributions, both continuous and discrete. We use 4 of these a lot : z (unit normal), t, chi-square and F.
- Z and t are closely related to the sampling distribution of means; chi-square and F are closely related to the sampling distribution of variances.

The definition of χ^2

- Scientists will often use the Chi-square (χ^2) test to determine the goodness of fit between theoretical and experimental data. In this test, we compare **observed values** with **theoretical or expected values.**
- Observed values are those that the researcher obtains empirically through direct observation; theoretical or expected values are developed on the basis of some hypothesis.
- The test statistic for comparing observed and expected frequencies is χ^2 defined as follows :

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

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

- If V independent variables x_i are each normally distributed with mean μ_i and variance σ_i^2 then the quantity known as chi-square^2 is defined by

$$\chi^2 = \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} + \ldots + \frac{(x_v-\mu_v)^2}{\sigma_v^2}$$

$$= \sum_{i=1}^{v} \frac{(x_i-\mu_i)^2}{\sigma_i^2}$$

Where, O = Observed value

E = Expected value

k = Number of categories, groupings, or possible outcomes

- Chi-square is the distribution of a sum of squares. Each squared deviation is taken from the unit normal : N(0,1). The shape of the chi-square distribution depends on the number of squared deviates that are added together.

The critical values for the χ^2 distribution

- The use of the χ^2 distribution in hypothesis testing is analogous to the use of the t and F distributions. A null hypothesis is stated, a test statistic is computed, the observed value of the test statistic is compared to the critical value, and a decision is made whether or not to reject the null hypothesis.

2.7.1 Characteristics

1. It is not symmetric.
2. The shape of the chi-square distribution depends upon the degrees of freedom, just like Student's t-distribution.
3. As the number of degrees of freedom increases, the chi-square distribution becomes more symmetric as is illustrated in Fig. 2.7.1.
4. The values are non-negative. That is, the values of are greater than or equal to 0.

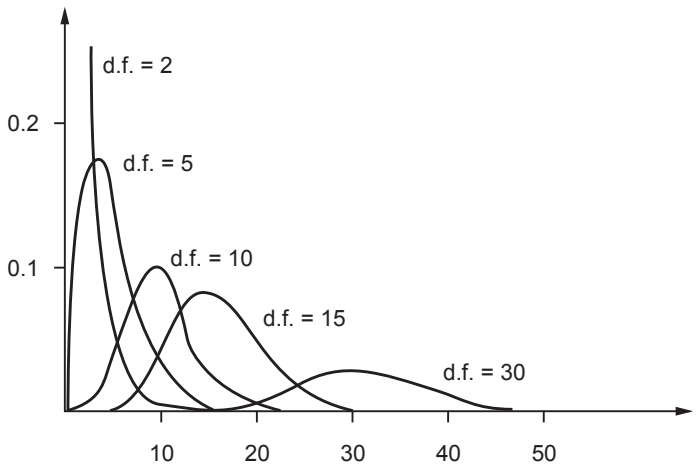


Fig. 2.7.1

- A **goodness-of-fit test** is an inferential procedure used to determine whether a frequency distribution follows a claimed distribution.
- The level of significance is used to determine the critical value. All chi-square goodness of fit tests are right tailed tests, so the critical value is with k–1 degree of freedom. In the Fig. 2.7.2, shaded region represents the critical region.
- The exact P-value for a non-directional test is the sum of probabilities for the table having a test statistic greater than or equal to the value of the observed test statistic.

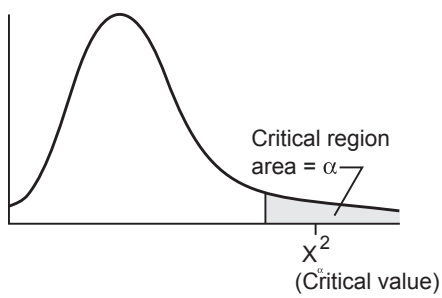


Fig. 2.7.2

1. High P-value : High probability that test statistic > Observed test statistic. **Do not reject** null hypothesis.
2. Low P-value : Low probability that test statistic > Observed test statistic. **Reject** null hypothesis.

Example 2.7.1 *A firm manufacturing rivets wants to limit variations in their length as much as possible. The length (in cms) of 10 rivets manufactured by a new process are*

2.15	1.99	2.05	2.12	2.17
2.01	1.98	2.03	2.25	1.93

Examine whether the new process can be considered superior to the old if the old population has standard deviation 0.145 cm ?

Solution : Given data : $n = 10, \sigma_0 = 0.145$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2.15+1.99+2.05+2.12+2.17+2.01+1.98+2.03+2.25+1.93}{10}$$

$$\bar{x} = \frac{20.68}{10} = 2.068$$

$$S^2 = \frac{\sum (x-\bar{x})^2}{n}$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
2.15	0.082	0.006724
1.99	– 0.078	0.006084
2.05	– 0.018	0.000324
2.12	0.052	0.002704
2.17	0.102	0.010404
2.01	– 0.058	0.003364
1.98	– 0.088	0.007744
2.03	– 0.038	0.001444
2.25	0.182	0.033124
1.93	– 0.138	0.019044
		0.09096

$$S^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{0.09096}{10} = 0.009096$$

Steps :

1. Null hypothesis $(H_0) : \sigma^2 = \sigma_0^2$
2. Alternative hypothesis $(H_1) : \sigma^2 > \sigma_0^2$
3. Level of significance $\alpha = 0.05$
4. Test statistic

$$X^2 = \frac{n s^2}{\sigma_0^2} = \frac{10 \times 0.009096}{(0.145)^2} = 4.326$$

Degree of freedom = n – 1 = 10 – 1 = 9

Tabulated X^2 for 9 d.f. at 5 % level of significance is 16.919.

Calculated (4.326) < Tabulated (16.919)

Null hypothesis (H_0) is accepted.

New process cannot be considered superior to the old process.

2.7.2 Chi - square Test for Goodness of Fit

- The chi-square test is used to test if a sample of data came from a population with a specific distribution.
- An attractive feature of the chi-square goodness-of-fit test is that it can be applied to any uni-variate distribution for which you can calculate the cumulative distribution function.
- The chi-square goodness-of-fit test is applied to binned data.
- The chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson.
- The chi square goodness of fit test begins by hypothesizing that the distribution of a variable behaves in a particular manner. For example, in order to determine daily staffing needs of a retail store, the manager may wish to know whether there are an equal number of customers each day of the week.
- To begin, an hypothesis of equal numbers of customers on each day could be assumed, and this would be the null hypothesis.
- Suppose that a variable has a frequency distribution with k categories into which the data has been grouped. The frequencies of occurrence of the variable, for each category of the variable, are called the observed values.
- The manner in which the chi square goodness of fit test works is to determine how many cases there would be in each category if the sample data were distributed exactly according to the claim.

- These are termed the expected number of cases for each category. The total of the expected number of cases is always made equal to the total of the observed number of cases.
- The null hypothesis is that the observed number of cases in each category is exactly equal to the expected number of cases in each category.
- The alternative hypothesis is that the observed and expected number of cases differ sufficiently to reject the null hypothesis.
- Chi-square goodness of fit test is a non-parametric test that is used to find out how the observed value of a given phenomena is significantly different from the expected value.
- In chi-square goodness of fit test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution.
- Chi-square goodness of fit test determines how well theoretical distribution (such as normal, binomial, or Poisson) fits the empirical distribution.
- In chi-square goodness of fit test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval.
- Let X be a random variable with unknown probability distribution function assuming the values x_1, x_2, …..
- Suppose that the values, say x_1, x_2, ……., x_i, ……., x_n of size n, have occupied with frequencies $(Of)_1$, $(Of)_2$, $(Of)_3$, ……, $(Of)_i$, ……, $(Of)_n$ respectively, where (Of) stands for observed frequency and $\sum n\ i = 1\ (Of)i = N$.

1. State null hypothesized proportions for each category (p_i). Alternative is that at least one of the proportions is different than specified in the null.
2. Calculate the expected counts for each cell as npi .
3. Calculate the X^2 statistic :

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

4. Compute the p-value as the proportion above the X^2 statistic for either a randomization distribution or a X^2 distribution with df = (number of categories - 1) if expected counts all > 5
5. Interpret the p-value in context.

2.7.3 Chi - square Test for Independence of Attributes

- The chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables.

- The frequency of each category for one nominal variable is compared across the categories of the second nominal variable.
- The data can be displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable.
- For example, say a researcher wants to examine the relationship between gender (male vs. female) and empathy (high vs. low). The chi-square test of independence can be used to examine this relationship.
- The null hypothesis for this test is that there is no relationship between gender and empathy.
- The alternative hypothesis is that there is a relationship between gender and empathy (e.g. there are more high-empathy females than high-empathy males).
- This test is also known as chi-square Test of Association. This test utilizes a contingency table to analyze the data.
- A contingency table is an arrangement in which data is classified according to two categorical variables. The categories for one variable appear in the rows, and the categories for the other variable appear in columns.
- Each variable must have two or more categories. Each cell reflects the total count of cases for a specific pair of categories.
- From the sample of N observations,
  1. $(Of)_{11}$ be the observed frequency of children of low income level going to government schools.
  2. $(Of)_{12}$ be the observed frequency of children of low income level going to private schools.
  3. $(Of)_{21}$ be the observed frequency of children of high income level going to government schools.
  4. $(Of)_{22}$ be the observed frequency of children of high income level going to private schools.
- This can be represented by 2×2 contingency table :

Income Level	Government	Private	Total
Low	$(Of)_{11}$	$(Of)_{12}$	R_1
High	$(Of)_{21}$	$(Of)_{22}$	R_2
Total	C_1	C_2	N

- Here R_1 and R_2 is row totals and C_1 and C_2 are column totals such that $C_1 + C_2 = R_1 + R_2 = N$, the total frequency.

	B_1	B_2	Row Total
A_1	a	b	a + b r_1
A_2	$a\|c$	$b\|d$	c + d r_2
Column Total	c_1	c_2	$a\|b\|c\|d = n$

- In case of 2×2 contigency table X^2 can be directly found using the short cur formula.

$$X^2 = \frac{n(ad\ bc)^2}{c_1 \cdot c_2 \cdot r_1 \cdot r_2}$$

2.7.4 Strength and Limitation of Chi - square Test

Strength :

1. It is easier to compute than some statistics.
2. Chi-square makes no assumptions about the distribution of the population.

Limitations :

1. The chi-square test does not give us much information about the strength of the relationship or its substantive significance in the population.
2. The chi-square test is sensitive to sample size. The size of the calculated chi-square is directly proportional to the size of the sample, independent of the strength of the relationship between the variables.
3. The chi-square test is also sensitive to small expected frequencies in one or more of the cells in the table.

Example 2.7.2 *If a sample of 900 units has mean of width of 3.5 cm and standard deviation of 2.61 cm then test whether this sample has come from a large population of mean width 3.25 cm and standard deviation 2.61 cm.*

Solution : Given data :

Sample size (n) = 900 (large sample)

Sample mean ($\bar{x}$) = 3.50

Sample standard deviation (S) = 2.61

Population mean $\mu = \mu_0$ = 3.25 cm

Population standard deviation (σ) = 2.61 cm

Step 1 : Null hypothesis $(H_0) : \mu = \mu_{0 = 3.25}$

Step 2 : Alternative hypothesis $(H_1) : \mu \neq \mu_{0 = 3.25}$

Step 3 : Level of significance : No specific level of significance (α) is proposed. It can be assumed as either 5 % or 1 %.

When α = 5 % or α = 0.05, we have for the two sided test

$-Z_{\alpha/2} = -Z_{0.025} = -1.96$ and $+Z_{\alpha/2} = +Z_{0.025} = +1.96$

When α = 1% or α = 0.01, we have for the two sided test

$-Z_{\alpha/2} = -Z_{0.025} = -2.58$ and $+Z_{\alpha/2} = +Z_{0.025} = +2.58$

Step 4 : Test statistics is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{3.50 - 3.25}{2.61 - \sqrt{900}} = 2.87 = Z_2$$

Step 5 : Calculated Z value is greater than the tabulated Z values. So hypothesis is rejected at 5 % and 1 % level of significance.

Example 2.7.3 *Four methods are under development for making discs of a superconducting material. 50 discs are made by each method, and they are checked for superconductivity when cooled with liquid .*

	Method				
	M1	***M2***	***M3***	***M4***	***Total***
Superconducters	*31*	*42*	*22*	*25*	*120*
Failures	*19*	*8*	*28*	*25*	*80*
Total	*50*	*50*	*50*	*50*	*200*

Test the significance difference between the proportions of superconductors under different methods at 0.05 level using the chi-square test.

Solution :

Step 1 : Null hypothesis (H_0) : the proportions of semiconductors are equal.

Step 2 : Alternative hypothesis (H_1) : the proportions of semiconductors are not equal.

Step 3 : Computations of expected frequencies (E_f) is

Observed frequency (Of)	Expected frequencies (Ef)
(Of)11 = 31	(Ef)11 = (50 X 120) / 200 = 30
(Of)12 = 42	(Ef)12 = (50 X 120) / 200 = 30
(Of)13 = 22	(Ef)13 = (50 X 120) / 200 = 30
(Of)14 = 25	(Ef)14 = (50 X 120) / 200 = 30
(Of)21 = 19	(Ef)21 = (50 X 80) / 200 = 20
(Of)22 = 8	(Ef)22 = (50 X 80) / 200 = 20
(Of)23 = 28	(Ef)23 = (50 X 80) / 200 = 20
(Of)24 = 25	(Ef)24 = (50 X 80) / 200 = 20

Step 4 : Determination of degree of freedom = (m – 1) (n – 1) = (2 – 1) (4 – 1) = 3

Step 5 : Chi-square statistic

$$\chi^2 = \frac{\left[(Of)_{ij} - (Ef)_{ij}\right]^2}{(Ef)_{ij}}$$

$$= \frac{(31-30)^2}{30} + \frac{(42-30)^2}{30} + \frac{(22-30)^2}{30} + \frac{(25-30)^2}{30}$$

$$+ \frac{(19-20)^2}{20} + \frac{(8-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(25-20)^2}{20} = 19.50$$

Step 6 : Calculated chi-square test is greater than the tabulated value (7.815). We reject null hypothesis. Therefore, the proportions of superconductors are not equal.

2.8 t - test

SPPU : Aug.-18, May-19, Oct.-19, Dec.-18, 19

- Student's t-test assumes that distributions of the two populations have equal but unknown variances. A t-test is used as a hypothesis testing tool, which allows testing an assumption applicable to a population.
- The t-test is a parametric test of difference, meaning that it makes the same assumptions about your data as other parametric tests. The t-test assumes your data :
  a) are independent
  b) are (approximately) normally distributed.
  c) have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)

- If your data do not fit these assumptions, you can try a nonparametric alternative to the t-test, such as the Wilcoxon Signed-Rank test for data with unequal variances.
- Suppose n_1 and n_2 samples are randomly and independently selected from two populations, pop1 and pop2, respectively.
- If each population is normally distributed with the same mean and with the same variance, then T (the t-statistic), given as

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where, $$S_P^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}$$

Properties of Student's t-Distribution :

1. The t-distribution is different for different degrees of freedom.
2. The t-distribution is centered at 0 and symmetric about 0.
3. The total area under the curve is 1. The area to the left of 0 is 1/2 and the area to the right of 0 is 1/2.
4. As the magnitude of t increases the graph approaches but never equals 0.
5. The area in the tails of the t-distribution is larger than the area in the tails of the normal distribution.
6. The shape of the t-distribution is dependent on the sample size n.
7. As sample size n increases, the distribution becomes approximately normal.
8. The standard deviation is greater than 1.
9. The mean, median and mode of the t-distribution are equal to zero.

2.8.1 Wilcoxon Rank - sum Test

- The Wilcoxon rank-sum test is a nonparametric alternative to the two sample t-test which is based solely on the order in which the observations from the two samples fall.
- It is used to test the null hypothesis that the median of a distribution is equal to some value.

- It can be used -
  a) In place of a one-sample t-test
  b) In place of a paired t-test
  c) For ordered categorial data where a numerical scale is inappropriate but where it is possible to rank the observations.
- The logic behind the Wilcoxon test is quite simple. The data are ranked to produce two rank totals, one for each condition.
- If there is a systematic difference between the two conditions, then most of the high ranks will belong to one condition and most of the low ranks will belong to the other one.
- As a result, the rank totals will be quite different and one of the rank totals will be quite small.
- On the other hand, if the two conditions are similar, then high and low ranks will be distributed fairly evenly between the two conditions and the rank totals will be fairly similar and quite large.
- The Wilcoxon test statistic "W" is simply the smaller of the rank totals. The SMALLER it is taking into account how many participants you have then the less likely it is to have occurred by chance.
- A table of critical values of W shows you how likely it is to obtain your particular value of W purely by chance.
- There are N observations in all, where $N = n_1 + n_2$. Rank all N observations. The sum W of the ranks for the first sample is Wilcoxon rank sum statistic.
- If the two populations have the same continuous distribution, then W has mean,

$$\mu_w = \frac{n_1(N+1)}{2}$$

and standard deviation

$$\sigma_w = \sqrt{\frac{n_1 n_2 (N+1)}{2}}$$

- The **Wilcoxon rank sum test** rejects the hypothesis that the two populations have identical distributions when the rank sum W is far from its mean.

Example 2.8.1 *Consider following data for wilcoxon rank-sum test :*

Participant	Left ear	Right ear
1	25	32
2	29	30
3	10	7
4	31	36
5	27	20
6	24	32
7	27	26
8	29	33
9	30	32
10	32	32
11	20	30
12	5	32
median :	**24.08**	**32.00**

a) Find the difference between each pair of scores.

b) Rank these differences, ignoring any "0" differences and ignoring the sign of the difference (i.e. whether it is a positive or negative difference).

Solution :

Participant	Left ear	Right ear	Difference (d)
1	25	32	– 7
2	29	30	– 1
3	10	7	3
4	31	36	– 5
5	27	20	7
6	24	32	– 8
7	27	26	1
8	29	33	– 4

9	30	32	– 2
10	32	32	0
11	20	30	– 10
12	5	32	– 27

To rank the differences :

- Give the lowest rank to the smallest difference-score, ignoring whether it's a positive or negative difference.
- If two or more difference-scores are the same, this is a "tie": tied scores get the average of the ranks that those scores would have obtained, had they been different from each other.
- Here, ignoring the sign of the difference, the lowest difference is – 1.
- However there are two instances of this score (one positive and one negative).
- Therefore we add up the ranks that these scores would have had, if they had been different from each other (the ranks of 1 and 2), and then divide the sum of these (1 + 2 = 3) by the number of ranks involved (2).
- This gives us an "average" rank, 1.5, that we allocate to both of these two scores.
- The next lowest difference-score is – 2. We have now used up the ranks of 1 and 2, so this difference-score gets the ranks of 3.
- After that, ranking is straightforward until we get to the two difference scores of – 7 and 7. These would have got the ranks of 7 and 8, but instead get the average rank of 7.5 (7 + 8 = 15; 15/2 = 7.5).
- This "uses up" the ranks of 7 and 8, so the next highest difference-score (– 8) gets the rank of 9.

Participant	Left ear	Right ear	Difference (d)	Ranked difference
1	25	32	– 7	7.5
2	29	30	– 1	1.5
3	10	7	3	4
4	31	36	– 5	6
5	27	20	7	7.5
6	24	32	– 8	9
7	27	26	1	1.5

8	29	33	– 4	5
9	30	32	– 2	3
10	32	32	0	ignore
11	20	30	– 10	10
12	0	32	– 27	11

- Add together the ranks belonging to scores with a positive sign (shaded in the table above) :

$$4 + 7.5 + 1.5 = 13$$

- Add together the ranks belonging to scores with a negative sign (unshaded in the table above) :

$$7.5 + 1.5 + 6 + 9 + 5 + 3 + 10 + 11 = 53$$

- Whichever of these sums is the smaller, is our value of W. So, W = 13.

Review Questions

1. *Explain Wilcoxon rank sum test.*
   SPPU : Aug.-18 (In Sem), Marks 2, Oct.-19 (In Sem), Dec.-19 (End Sem), Marks 5
2. *How Wilcoxon rank - sum test works ?* **SPPU : May-19 (End Sem), Marks 5**
3. *When do we use Wilcoxon rank-sum test ? Write steps in the test.*
   SPPU : Dec.-18 (End Sem), Marks 6

2.9 Multiple Choice Questions

Q.1 **The _____ of a data set is the average of all the data values.**

a mean b median

c mode d all of these

Q.2 **An _____ is an observation that lies an abnormal distance from other values in a random sample from a population.**

a range b quartiles

c outlier d none

Q.3 **The standard deviation of a data set is the positive square root of the _____.**

a sample b variance

c population d outlier

Q.4 **Pearson's correlation coefficient is represented by _____.**

a p b q

c r d σ

Q.5 **The Wilcoxon Rank Sum Test is often described as the _____ version of the two-sample t-test.**

a parametric b non-parametric

c regular d none

Q.6 **Chi-square test is a _____ method.**

a decision tree b clustering

c hypothesis testing d correlation testing

Q.7 **The mean, median and mode of the t-distribution are equal to _____.**

a zero b one

c null d minus one

Q.8 **In hypothesis testing, the hypothesis which is tentatively assumed to be true is called the _____.**

a correct hypothesis b null hypothesis

c alternative hypothesis d level of significance

Answer Keys for Multiple Choice Questions :

Q.1	a	**Q.2**	c	**Q.3**	b
Q.4	c	**Q.5**	b	**Q.6**	c
Q.7	a	**Q.8**	b		

❑❑❑

Notes

UNIT III

3 Big Data Analytics Life Cycle

Syllabus

Introduction to Big Data, sources of Big Data, ***Data Analytic Lifecycle :*** *Introduction, Phase 1 : Discovery, Phase 2 : Data Preparation, Phase 3 : Model Planning, Phase 4 : Model Building, Phase 5 : Communication results, Phase 6 : Operation alize.*

Contents

3.1 Introduction to Big Data

SPPU : Aug.-18, May-19, Oct.-19, Dec.-19

- Big data can be defined as very volumes of data available at various sources, in varying degrees of complexity, generated at different speed i.e. velocities and varying degrees of ambiguity, which cannot be processed using traditional technologies, processing methods, algorithms or any commerical off-the-shelf solutions.
- 'Big data' is a term used to describe collection of data that is huge in size and yet growing exponentially with time. In short, such a data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.
- The processing of big data begins with the raw data that isn't aggregated or organized and is most often impossible to store in the memory of a single computer.
- Big data processing is a set of techniques or programming models to access large-scale data to extract useful information for supporting and providing decisions. Hadoop is the open-source implementation of MapReduce and is widely used for big data processing.

3.1.1 Big Data Requirement

- We can classify **Big Data requirements** based on its five main characteristics :

1. Volume :
   - Size of data to be processed is large-it needs to be broken into manageable chunks.
   - Data needs to be processed in parallel across multiple systems.
   - Data needs to be processed across several program modules simultaneously.
   - Data needs to be processed once and processed to completion due to volumes.
   - Data needs to be processed from any point of failure, since it is extremely large to restart the process from the beginning.
2. Velocity :
   - Data needs to be processed at streaming speeds during data collection.
   - Data needs to be processed for multiple acquisition points.
3. Variety :
   - Data of different formats needs to be processed.
   - Data of different types needs to be processed.
   - Data of different structures needs to be processed.
   - Data from different regions needs to be processed.

4. Ambiguity :
   - Big data is ambiguous by nature due to the lack of relevant metadata and context in many cases. An example is the use of M and F in a sentence, it can mean, respectively, monday and friday, male and female or mother and father.
   - Big data that is within the corporation also exhibits this ambiguity to a lesser degree. For example, employment agreements have standard and custom sections and the latter is ambiguous without the right context.
5. Complexity :
   - Big data complexity needs to use many algorithms to process data quickly and efficiently.
   - Several types of data need multi-pass processing and scalability is extremely important.

3.1.2 Benefits of Big Data Processing

Benefits of big data processing

1. Improved customer service.
2. Business can utilize outside intelligence while taking decision.
3. Reducing maintenance costs.
4. Re-develop your products : Big data can also help you understand how others perceive your products so that you can adapt them or your marketing, if need be.
5. Early identification of risk to the product/services, if any.
6. Better operational efficiency.

3.1.3 Big Data Challenges

- Collecting, storing and processing big data comes with its own set of challenges :
  1. Big data is growing exponentially and existing data management solutions have to be constantly updated to cope with the three Vs.
  2. Organizations do not have enough skilled data professionals who can understand and work with big data and big data tools.

3.1.4 Data Analytical Architecture

- Analytics architecture refers to the systems, protocols and technology used to collect, store and analyze data. Analytics architecture also focuses on multiple layers, starting with data warehouse architecture, which defines how users in an organization can access and interact with data.
- Fig. 3.1.1 shows data analytical architecture.

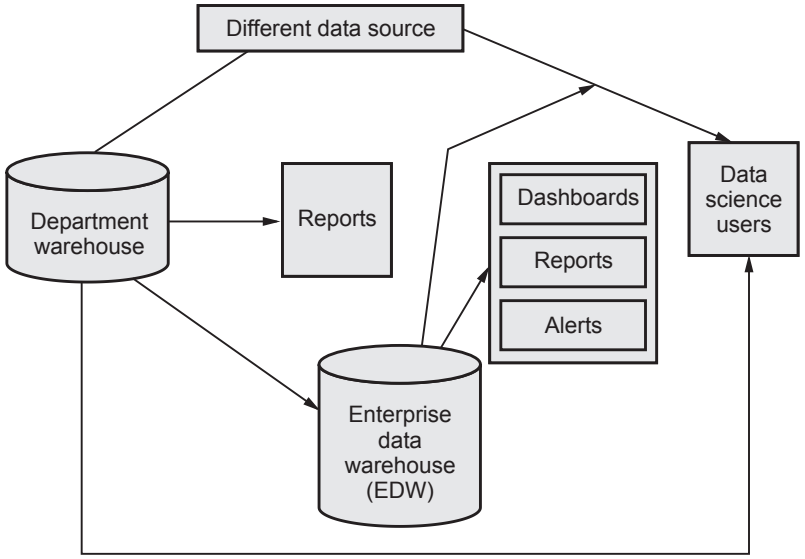


Fig. 3.1.1 Data analytical architecture

- Data to be loaded into the data warehouse. It must be well understood structured and normalized with the appropriate data type. Centralization provides security, backup facility. Also provides significant pre-processing and checkpoints facility before storing data.
- Required level of control on the EDM with additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. Sometimes local data marts allow users to do some level of more in-depth analysis.
- Once in the data warehouse, data is read by additional applications across the enterprise for Bl and reporting purposes. These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.
- At last, analysts get data provisioned for their downstream analytics. Many times, these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset.

3.1.5 Big Data Ecosystem

- Big data ecosystem is the comprehension of massive functional components with various enabling tools. Capabilities of the big data ecosystem are not only about computing and storing big data, but also the advantages of its systematic platform and potentials of big data analytics.
- Organizations and data collectors that can gather data from individuals start realizing that a new economy is emerging as a data business. Depending on the evolution of this economy, a new ecosystem is rising.

- Many organizations and data collectors are depend on the data they can gather from individuals which contains great value and, as a result, a new economy is emerging.
- This new digital economy continues to evolve; the market sees the introduction of data vendors and data cleaners that use crowdsourcing to test the outcomes of machine learning techniques.
- As the ecosystem has been growing, groups of interest have been formed. Currently there are four main groups which are associated with each other : Data devices, data collectors, data aggregators and data users and buyers.
- Fig. 3.1.2 shows emerging big data ecosystem.

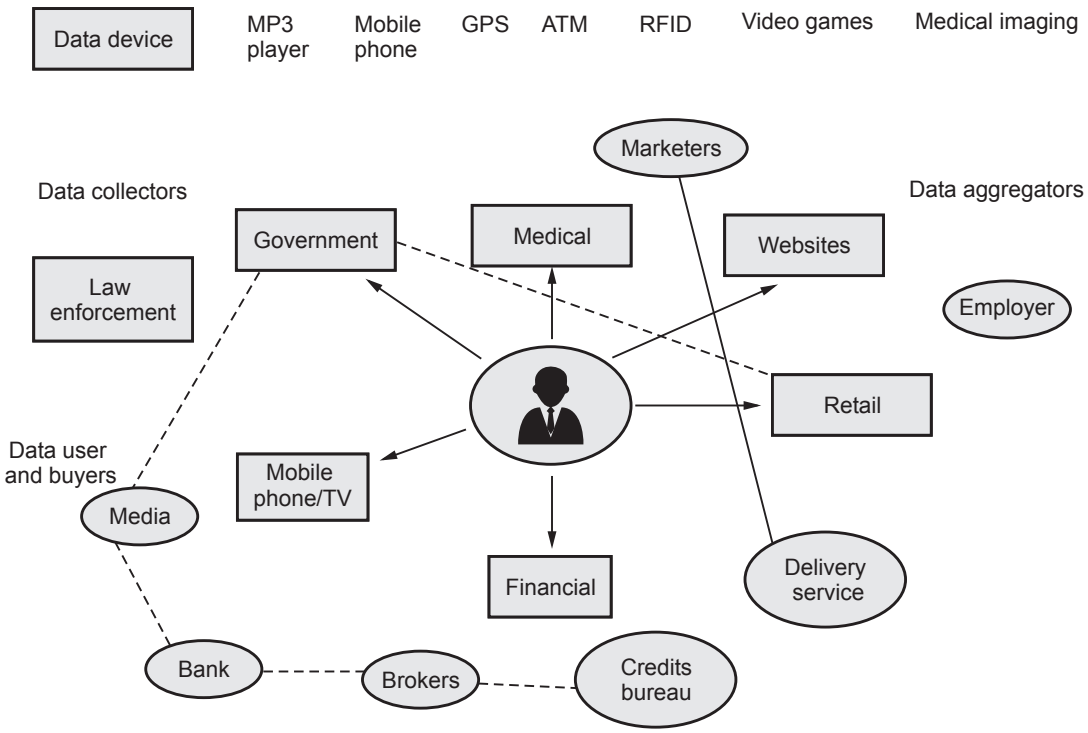


Fig. 3.1.2 Emerging big data ecosystem

1. Data device :

- Data device and sensor network gather data from various locations and continuously generate new data.
- **Example of data devices :** Playing games, smart phone and retail **shopping**.
- **Sensor data :** Growing network of sensor devices generate data based on montoring environmental coditions, such as temperature, sound, pressure, power water levels etc.

- This data can have a wide range of practical application if collected, aggregated, analyzed and acted upon. Examples include, water lavel monitoring and smart home monitoring.
- Mobile networks : Mobile network generates large number of data to share picture, video, audio file and text. These data is process at every mobile tower with associated demographics, location latencies etc. Sometime mobile network may crash for large number of data movement takes place.
- Retail shopping loyalty cards records not just the amount an individual spends, but the location of stores that person visits, the kinds of product purchased, the store where goods are purchased most often and combinations of product purchased together.

2. Data collectors :

- Data collectors : It includes samples entities that collect data from the device and user.
- Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foot traffic using geospatial data collected from the RFID chips.
- Data collectors example : Government, retail stores, cable TV provider.
- Cable TV provider which tracks the shows that a person watches.

3. Data aggregators :

- The entities which process collected data from the first layer make them understandable. They give them additional value to prepare them for the handing over process. Now the data is ready to be offered on the market.
- Typically, one of these data aggregators can transform and package the data as products to sell to list brokers that might want to generate marketing lists of people who may be good targets for specific ad compaigns.

4. Data users and buyers :

- The enetities represent a group of the final layer from the big data ecosystem. This group has the final benefits from the collected and aggregated data offered by the data aggregators.
- Data users may wanto to track or prepare for natural disaster by identifying which areas a hurricane will affect first hand. It can be observed by tracking tweets it or discussing it in social media.

Review Questions

1. *What is Big data ? Explain characteristics of Big Data.* **SPPU : Aug.-18 (In Sem), Marks 4**
2. *Explain current analytics architecture with suitable diagram.* **SPPU : Aug.-18 (In Sem), Marks 6**
3. *Explain bigdata ecosystem.* **SPPU : Aug.-18 (In Sem), Marks 4**
4. *What is big data ? Explain 3V's of big data.* **SPPU : May-19 (End Sem), Marks 5**
5. *Draw and explain current analytical architecture.* **SPPU : Oct.-19 (In Sem), Marks 5**
6. *Enlist and explain various users involved to make successful analytical project.* **SPPU Oct.-19 (In Sem), Marks 5**
7. *Discuss with example data devices and data collectors of emerging big data ecosystem.* **SPPU : Oct.-19 (In Sem), Marks 5**
8. *Draw and Explain big data ecosystem.* **SPPU : Dec.-19 (End Sem), Marks 5**
9. *Explain current analytical architecture with diagram.* **SPPU : Dec.-19 (End Sem), Marks 5**

3.2 Sources of Big Data

- Machine data consists of information generated from industrial equipment, real-time data from sensors that track parts and monitor machinery and even web logs that track user behavior online.
- At Arcplan client CERN, the largest particle physics research center in the world, the Large Hadron Collider (LHC) generates 40 terabytes of data every second during experiments.
- Regarding transactional data, large retailers and even B2B companies can generate multitudes of data on a regular basis considering that their transactions consist of one or many items, product IDs, prices, payment information, manufacturer and distributor data and much more.
- Some of the examples of big data are :
  1. **Social media :** Social media is one of the biggest contributors to the flood of data we have today. Facebook generates around 500+ terabytes of data everyday in the form of content generated by the users like status messages, photos and video uploads, messages, comments etc.
  2. **Stock exchange :** Data generated by stock exchanges is also in terabytes per day. Most of this data is the trade data of users and companies.
  3. **Aviation industry :** A single jet engine can generate around 10 terabytes of data during a 30 minute flight.
  4. **Survey data :** Online or offline surveys conducted on various topics which typically have hundreds and thousands of responses and need to be processed

for analysis and visualization by creating a cluster of population and their associated responses.

5. **Compliance data :** Many organizations like healthcare, hospitals, life sciences, finance etc, has to file compliance reports.

3.2.1 Data Repository

- Data repository is also known as a data library or data archive. This is a general term to refer to a data set isolated to be mined for data reporting and analysis. The data repository is a large database infrastructure, several databases that collect, manage and store data sets for data analysis, sharing and reporting.
  1. Spreadsheets :
     - Spreadsheets enabled business user to crate simple logic on data structured in rows and columns and create their own analyses of business problems.
     - Database administrator training is not required to create spreadsheets : They can be set up to do many things quickly and independently of Information Technology (IT) groups.
  - Spreadsheets are easy to share and en users have control over the logic involved.
  2. Enterprise Data Warehouse (EDWs) are critical for reporting and 81 tasks and solve many of the problems that profilerating spreadsheets introduce, such as which of the multiple versions of a spreadsheet is correct.
     - From an analyst perpective, EDW and 81 solve problems related to data accuracy and availability.
     - But EDW and BI introduce new problems related to flexibility and agility, which were less pronounced when dealing with spreadsheets.
  3. Analytic sandbox : Analytic sandbox is solution to this problem. It attempts to resolve the conflict for analysts and data scientists with EDW and more formally managed corporate data.
  4. In this model, the IT group may strill manage the analytic sandboxes, but they will be purposefully designed to enable robust analytics, while centrally managed and secured.

3.2.2 Example of Data Repository

1. Data warehouse is a large repository that aggregates data usually from multiple sources or segments of a business, without the data being necessarily related.
2. Data lake is a large repository that stores unstructed data that is classified and tagged with metadata.

3. Data marts are subsets of the data repository. These data marts are more targeted to what the data user needs and easier to use.
4. Metadata repositories store data about data and databases. The metadata explains where the data source, how it was captured and what it represents.
5. Data cubes are lists of data with three or more dimensions stored as a table.

3.2.3 Advantages and Disadvantages of Data Repository

Advantages :

1. Data is preserved and archived.
2. Data isolation allows for easier and faster data reporting.
3. Data adminitratrors have easier time tracking problems.
4. There is value to storing and analyzing data.

Disadvantages :

1. Growing data sets could slow down systems.
2. A system crash could affect all the data.
3. Unauthorized users can access all sensitive data more easily than if it was distributed across several locations.

3.2.4 Analytic Sandbox

- An analytical sandbox is used as a tool for data storage manipulation and exploration separated from a production environment due to getting flexibility and freelance that is caused by this separation.
- The policy of accessing a production environment is tightly controlled due to its importance and need for financial reporting.
- An analytical sandbox sometimes mentioned as a workspace.
- An analytcial sandbox is primarily assigned for data analysts who require access to al data throughout the organization.
- Analytic sandbox are designed to enable teams to explore many datasets in a controlled fashion and are not typically used for enterprise level finanical reporting ans sales dashboards.
- Many times, analytics sandboxes enable high-performance computing using in-database processing, the analytics occur within the database itself.

- Analytical sandboxes attempt to provide a less restricted environment to enable robust analytics being centrally managed and secured.
- Rather than the typical structured data in the EDW, analystic sandboxes can house a greater variety of data, such as raw data, textual data and other kinds of unstructured data, without interfering with critical production databases.

3.2.5 Factor Responsible for Data Volume in Big Data

1. **Machine data :** Machine data contains a definitive record of all activity and behavior of your customers, users, transitions, applications, servers, networks, factory machinery and so on. It's configuration data, data from APIs and message queues, change events, the output of diagnostic commands and call detail records, sensor data from remote equipment and more.
2. **Application log :** Most homegrown and packaged applications write local logfiles, logging services built into application severs like Web Logic, Web Sphere and JBoss. These files are critical for day-to-day debugging of production applications by developers and application support. When developers put timing information into their log events, they can also be used to monitor and report on application performance.
3. **Business process logs :** Complex events processing and business process management system logs are treasure troves of business and IT relevant data. These logs are will generally include definitive records of customer activity across multiple channels such as the web, IVR/contact center or retail.
4. **Clickstream data :** User activity on the Internet is captured in clickstrem data. This provides insight into a user's website and web page activity. The informations is valueble for usability analysis, marketing and general research.
5. **Thrid party data :** The sensititve data that's not in databases is on file systems. In some industries such as healthcare, the biggest data leakage risk is consumer records on shared file systems. Different OS, third-party tools and storage technologies provide different options for auditing read access to sensitive data at the file systems. Different options for auditing read access to sensitive data at the file system level. This audit data is a vital data source for monitoring and investigating access to sensitive data.
6. **Electronic mails :** Every company have large collection of emails generated by customers, employees and executives on daily basis. These email communications are an important asset to an organization, which are audited case-by case basis and entire life cycle management of emails is done.

3.3 Data Analytic Lifecycle

SPPU : Aug.-18, Dec.-18, May-19, Oct.-19

- The data analytic lifecycle is designed for big data problems and data science projects. With six phases the project work can occur in several phases simultaneously. The cycle is iterative to portray a real project. Work can return to earlier phases as new information is uncovered.
- According to Dietrich (2013), it is a cyclical life cycle that has iterative parts in each of its six steps :
  1) Discovery
  2) Pre-processing data
  3) Model planning
  4) Model building
  5) Communicate results
  6) Operationalize
- Fig. 3.3.1 shows data analytic lifecycle.

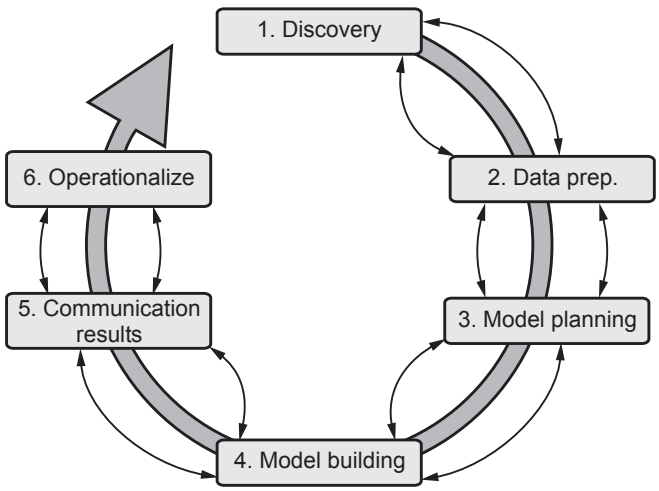


Fig. 3.3.1 Data analytic life cycle

3.3.1 Phase 1 : Discovery

- This phase is all about defining the data's purpose and how to achieve it by the end of the data analytics lifecycle. The stage consists of identifying critical objectives a business is trying to discover by mapping out the data.
- During this process, the team learns about the business domain and checks whether the business unit or organization has worked on similar projects to refer to any learnings.
- In this phase, the team also evaluates technology, people, data and time. For example, while dealing with a small dataset, the team can use excel.

- Phase 1 contains following process :
  a) Learning the business domain
  b) Resources
  c) Framing the problem
  d) Identifying key stakeholders
  e) Interviewing the analytics sponsor
  f) Developing initial hypotheses
  g) Identifying potential data sources

3.3.2 Phase 2 : Data Preparation

- This stage involves collecting, processing and cleaning data. Here the focus shifts from business requirements to data requirements. In this early phase, data is collected but not analyzed.
- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often.

a) Preparing the analytic sandbox :
- Create the analytic sandbox. It also called a workspace. It allows the team to explore data without interfering with live production data.
- Sandbox collects all kinds of data. The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics.

b) Performing ETLT (Extract, Transform, Load, Transform) :
- The team needs to execute Extract, Load and Transform (ELT) to get data into the sandbox.
- Extract, Transform, Load (ETL) : It transforms the data based on a set of business rules before loading it into the sandbox.
- Extract, Load, Transform (ELT) : It loads the data into the sandbox and then transforms it based on a set of business rules.
- Extract, Transform, Load, Transform (ETLT) : It's the combination of ETL and ELT and has two transformation levels.

c) Learning about the data :
- Data is captured through three main ways :
  i. Data acquisition : Obtaining existing data from outside sources.
  ii. Data entry : Creating new data values from data inputted within the organization.

iii. Signal reception : Capturing data created by devices.

d) Data Conditioning :

- Data conditioning includes cleaning data, normalizing datasets and performing transformations. It is often viewed as a preprocessing step prior to data analysis; it might be performed by data owner, IT department, DBA, etc.
- Best to have data scientists involved and data science teams prefer more data than too little.

e) Common tools for data preparation :

- Hadoop can perform parallel ingest and analysis.
- Alpine miner provides a graphical user interface for creating analytic workflows.
- OpenRefine is a free, open source tool for working with messy data.
- Similar to OpenRefine, data wrangler is an interactive tool for data cleansing and transformation.

3.3.3 Phase 3 : Model Planning

- The team determines the methods, techniques and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.
- Activities to consider :

  a) Assess the structure of the data - this dictates the tools and analytic techniques for the next phase.

  b) Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses.

  c) Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow.

  d) Research and understand how other analysts have approached this kind or similar kind of problem.

a) Data exploration and variable selection

- Explore the data to understand the relationships among the variables to inform selection of the variables and methods. A common way to do this is to use data visualization tools.

- Often, stakeholders and subject matter experts may have ideas. For example, some hypothesis that led to the project.
- Aim for capturing the most essential predictors and variables. This often requires iterations and testing to identify key variables.
- If the team plans to run regression analysis, identify the candidate predictors and outcome variables of the model.

b) Model selection

- The main goal is to choose an analytical technique or several candidates, based on the end goal of the project. We observe events in the real world and attempt to construct models that emulate this behavior with a set of rules and conditions.
- A model is simply an abstraction from reality. Determine whether to use techniques best suited for structured data, unstructured data or a hybrid approach.
- Teams often create initial models using statistical software packages such as R, SAS or Matlab. Which may have limitations when applied to very large datasets.
- The team moves to the model building phase once it has a good idea about the type of model to try.

c) Common tools for the model planning phase

- R programming language has a complete set of modeling capabilities. It contains about 5000 packages for data analysis and graphical presentation.
- SQL analysis services can perform in-database analytics of common data mining functions, involved aggregations and basic predictive models.
- SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connections.

3.3.4 Phase 4 : Model Building

- The team develops datasets for testing, training and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
- Building a model involves two phases :
  - **a) Design the model :** Identify a suitable model. This step can involve a number of different modeling techniques to identify a suitable model. These may include decision trees, regression techniques and neural networks.
  - **b) Execute the model :** The model is run against the data to ensure that the model fits the data.
- Common commercial tools for the model building phase :
  a. SAS enterprise miner used for building enterprise-level computing and analytics.

b. SPSS modeler (IBM) provides enterprise-level computing and analytics.

c. Matlab is a high-level language for data analytics, algorithms, data exploration.

d. Alpine miner provides GUI frontend for backend analytics tools.

e. STATISTICA and MATHEMATICA is popular data mining and analytics tools.

3.3.5 Phase 5 : Communicate Results

- This phase aims to determine whether the project results are a success or failure and start collaborating with significant stakeholders.
- The team identifies the vital findings of their analysis, measures the associated business value and creates a summarized narrative to convey the stakeholders' results.
- Communicate and document the key findings and major insights derived from the analysis. This is the most visible portion of the process to the outside stakeholders and sponsors.

3.3.6 Phase 6 : Operationalize

- This final phase moves data from the sandbox into a live environment. Data is monitored and analyzed to see if the generated model is creating the expected results. If the results aren't as expected, you can return to any of the preceding phases to tweak the data.
- The team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way. Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout.
- During the pilot project, the team may need to execute the algorithm more efficiently in the database rather than with in-memory tools like R, especially with larger datasets.
- To test the model in a live setting, consider running the model in a production environment for a discrete set of products or a single line of business. Monitor model accuracy and retrain the model if necessary.
- Key outputs from successful analytics project
  a) Business user tries to determine business benefits and implications.
  b) Project sponsor wants business impact, risks, ROI.
  c) Project manager needs to determine if project completed on time, within budget, goals met.
  d) Business intelligence analyst needs to know if reports and dashboards will be impacted and need to change.

e) Data engineer and DBA must share code and document.

f) Data scientist must share code and explain model to peers, managers, stakeholders.

Review Questions

1. *Explain different phases of data analytics life cycle.* **SPPU : Aug.-18 (In Sem), Marks 6**
2. *Explain data analytic life cycle.* **SPPU : Dec.-18 (End Sem), Marks 8**
3. *Draw data analytics lifecycle and give brief description about all phases.* **SPPU : May-19 (End Sem), Marks 5**
4. *Why communication is important in data analytics lifecycle projects ?* **SPPU ; May-19, (End Sem), Marks 8**
5. *Demonstrate the overview of data analytics life cycle.* **SPPU : Oct.-19 (In Sem), Marks 5**

3.4 Multiple Choice Questions

Q.1 What are the different features of Big data analytics ?

a Open-source b Scalability

c Data recovery d All of above

Q.2 _____ data has internal structure but is not structured via pre-defined data models or schema.

a Structured b Semi-structured

c Unstructured d All of these

Q.3 In big data, _____ refer to heterogeneous sources and the nature of data, both structured and unstructured.

a volume b variety

c velocity d all of these

Q.4 Type of data analytics are _____.

a descriptive model b predictive model

c prescriptive model d all of these

Q.5 Data is collection of data objects and their _____.

a information b attributes

c characteristics d none

Q.6 **Data frame is used for storing data _____.**

a value b numbers

c tables d all of these

Q.7 **Data frames is a collection of vectors that all have the _____ length.**

a same b variable

c short d all of these

Q.8 **Following which method is NOT used for handling missing values.**

a Eliminate data objects

b Estimate missing values

c Ignore the missing value during analysis

d Replace with all error values

Q.9 **Data _____ means removing the inconsistent data or noise and collecting necessary information of a collection of interrelated data.**

a preprocessing b cleaning

c transforming d none

Q.10 **The process of converting the integrating data into correct format is called _____.**

a data cleaning b data preprocessing

c data transformation d data handling

Answer Keys for Multiple Choice Questions :

Q.1	d	**Q.2**	c
Q.3	b	**Q.4**	d
Q.5	b	**Q.6**	c
Q.7	a	**Q.8**	d
Q.9	b	**Q.10**	c

❑❑❑

Notes

UNIT IV

4 Predictive Big Data Analytics with Python

Syllabus

Introduction, *Essential Python Libraries, Basic examples.* ***Data Preprocessing :*** *Removing Duplicates, Transformation of Data using function or mapping, replacing values, Handling Missing Data. Analytics Types : Predictive, Descriptive and Prescriptive.* ***Association Rules :*** *Apriori Algorithm, FP growth.* ***Regression :*** *Linear Regression, Logistic Regression.* ***Classification :*** *Naïve Bayes, Decision Trees.* ***Introduction to Scikit-learn,*** *Installations, Dataset, mat plotlib, filling missing values, Regression and Classification using Scikit-learn.*

Contents

4.1 Introduction of Python

- Python is a high-level scripting language which can be used for a wide variety of text processing, system administration and internet-related tasks.
- Python is a true object-oriented language, and is available on a wide variety of platforms.
- Python was developed in the early 1990's by Guido van Rossum, then at CWI in Amsterdam, and currently at CNRI in Virginia.
- Python 3.0 was released in Year 2008.
- Python statements do not need to end with a special character.
- Python relies on modules, that is, self-contained programs which define a variety of functions and data types.
- A module is a file containing Python definitions and statements. The file name is the module name with the suffix .py appended.
- Within a module, the module's name (as a string) is available as the value of the global variable __name__.
- If a module is executed directly however, the value of the global variable __name__ will be "__main__".
- Modules can contain executable statements aside from definitions. These are executed only the first time the module name is encountered in an import statement as well as if the file is executed as a script.
- Integrated Development Environment (IDE) is the basic interpreter and editor environment that you can use along with Python. This typically includes an editor for creating and modifying programs, a translator for executing programs, and a program debugger. A debugger provides a means of taking control of the execution of a program to aid in finding program errors.
- Python is most commonly translated by use of an interpreter. It provides the very useful ability to execute in interactive mode. The window that provides this interaction is referred to as the Python shell.
- Python support two basic modes : *Normal mode and interactive mode.*
- Normal mode : The normal mode is the mode where the scripted and finished .py files are run in the Python interpreter. This mode is also called as script mode.
- Interactive mode is a command line shell which gives immediate feedback for each statement, while running previously fed statements in active memory.
- Start the Python interactive interpreter by typing python with no arguments at the command line.

- To access the Python shell, open the terminal of your operating system and then type "python". Press the enter key and the Python shell will appear.

```
C:\Windows\system32>python
Python 3.5.0 (v3.5.0:374f501f4567, Sep 13 2015, 02:27:37) [MSC v.1900 64 bit (AMD64)] on
win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

- The >>> indicates that the Python shell is ready to execute and send your commands to the Python interpreter. The result is immediately displayed on the Python shell as soon as the Python interpreter interprets the command.
- For example, to print the text "Hello World", we can type the following :

```
>>> print("Hello World")
Hello World
>>>
```

- In script mode, a file must be created and saved before executing the code to get results. In interactive mode, the result is returned immediately after pressing the enter key.
- In script mode, you are provided with a direct way of editing your code. This is not possible in interactive mode.

4.1.1 Features of Python Programming

1. Python is a high-level, interpreted, interactive and object-oriented scripting language.
2. It is simple and easy to learn.
3. It is portable.
4. Python is free and open source programming language.
5. Python can perform complex tasks using a few lines of code.
6. Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc
7. It provides a vast range of libraries for the various fields such as machine learning, web developer, and also for the scripting.

4.1.2 Advantages and Disadvantages of Python

Advantages of Python

- Ease of programming
- Minimizes the time to develop and maintain code

- Modular and object-oriented
- Large community of users
- A large standard and user-contributed library

Disadvantages of Python

- Interpreted and therefore slower than compiled languages
- Decentralized with packages

4.2 Essential Python Libraries

- A library is a collection of files (called modules) that contains functions for use by other programs. A Python library is a reusable chunk of code that you may want to include in your programs.
- Many popular Python libraries are NumPy, SciPy, Pandas and SciKit-Learn. Python visualization libraries are matplotlib and Seaborn.

4.2.1 NumPy

- NumPy has risen to become one of the most popular Python science libraries and just secured a round of grant funding.
- NumPy's multidimensional array can perform very large calculations much more easily and efficiently than using the Python standard data types.
- To get started, NumPy has many resources on their website, including documentation and tutorials.
- NumPy (Numerical Python) is a perfect tool for scientific computing and performing basic and advanced array operations.
- The library offers many handy features performing operations on n-arrays and matrices in Python. It helps to process arrays that store values of the same data type and makes performing math operations on arrays easier. In fact, the vectorization of mathematical operations on the NumPy array type increases performance and accelerates the execution time.

4.2.2 Pandas

- Pandas is one of the most popular Python libraries in data science.
- The pandas library provides support for data structures and data analysis tools. The library is optimized to perform data science tasks especially fast and efficiently.

- The basic principle behind pandas is to provide data analysis and modeling support for Python that is similar to other languages, such as R.
- Pandas is best suited for structured, labelled data, in other words, tabular data, that has headings associated with each column of data.
- Pandas has two core data structures used to store data : The Series and the DataFrame.
- The series is a one-dimensional array-like structure designed to hold a single array (or 'column') of data and an associated array of data labels called an index.
- The DataFrame represents tabular data, a bit like a spreadsheet. DataFrames are organised into colums and each column can store a single data-type, such as floating point numbers, strings, boolean values etc. DataFrames can be indexed by either their row or column names.

4.2.3 SciPy

- SciPy contains many different packages and modules to assist in mathematics and scientific computing.
- It's difficult to state a single use case for SciPy considering that it contains so many different useful packages (including Numpy).
- Some of the important packages include :
  1. SciPy library : One of the core packages of the SciPy stack. This includes assistance with scientific computing, including those for numerical integration and optimization.
  2. Matplotlib : A 2D plotting library that can be used in Python scripts, the Python and IPython shell, web application servers, and more.
  3. IPython : An interactive console that runs your code like the Python shell, but gives you even more features, like support for data visualizations.

4.2.4 SciKit-Learn

- Scikit-learn is probably the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- In scikit-learn, an estimator for classification is a Python object that implements the methods fit (X, y) and predict (T).
- An example of an estimator is the class sklearn.svm. SVC, which implements support vector classification. The estimator's constructor takes as arguments the model's parameters.

- Scikit-learn comes loaded with a lot of features. Here are a few of them for understanding :
  1. Supervised learning algorithms : Think of any supervised learning algorithm you might have heard about and there is a very high chance that it is part of scikit-learn.
  2. Cross-validation : There are various methods to check the accuracy of supervised models on unseen data.
  3. Unsupervised learning algorithms : Again, there is a large spread of algorithms in the offering - starting from clustering, factor analysis, principal component analysis to unsupervised neural networks.
  4. Various toy datasets : This came in handy while learning scikit-learn. For example : IRIS dataset, Boston House prices dataset.
  5. Feature extraction : Useful for extracting features from images and text (e.g. Bag of words).

4.3 Data Pre-processing

- Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Aim to reduce the data size, find the relation between data and normalized them.
- Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing.

Why Data Pre-processing ?

- Data which capture from various sources is not pure. It contains some noise. It is called dirty data or incomplete data. In this data, there is lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. For example : occupation= " "
- Noisy data which contains errors or outliers. For example : Salary="-10"
- Inconsistent data which contains discrepancies in codes or names.
  For example : Age="51" Birthday="03/08/1998"
- Incomplete, noisy, and inconsistent data are commonplace properties of large real-world databases and data warehouses. Incomplete data can occur for a number of reasons.

Steps during pre-processing :

1. **Data cleaning :** Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

2. **Data integration :** Data with different representations are put together and conflicts within the data are resolved.
3. **Data transformation :** Data is normalized, aggregated and generalized.
4. **Data reduction :** This step aims to present are duced representation of the data in a datawarehouse.
5. **Data discretization :** Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

4.3.1 Removing Duplicates

- Removing Duplicates in the context of data quality is where an organisation looks to identify and then remove instances where there is more than one record of a single person. With large scales of data, this will often be done using tools that find and merge duplicate records in an existing database and prevent new ones from entering it based on similarities in specific fields.
- Preparing a dataset before designing a machine learning model is an important task for the data scientist. If there are more duplicates then making machine learning model is useless or not so accurate. Therefore, you must know to remove the duplicates from the dataset.
- Data set may include data objects that are duplicates, or almost duplicates of one another. Major issue when merging data from heterogeneous sources. Examples is same person with multiple email addresses. Data cleaning is the process of dealing with duplicate data issues.
- A dataset may include data objects which are duplicates of one another. It may happen when say the same person submits a form more than once. The term deduplication is often used to refer to the process of dealing with duplicates.
- In most cases, the duplicates are removed so as to not give that particular data object an advantage or bias, when running machine learning algorithms.

4.3.2 Handling Missing Data Values

- Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. The various methods for handling the problem of missing values in data tuples are as follows :
  1. **Ignoring the tuple :** This is usually done when the class label is missing. This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

2. **Manually filling in the missing value :** This approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.
3. **Using a global constant to fill in the missing value :** Replace all missing attribute values by the same constant, such as a label like "Unknown," or - _. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common that of "Unknown."
4. Using a measure of central tendency for the attribute, such as the mean (for symmetric numeric data), the median (for asymmetric numeric data), or the mode (for nominal data).
5. Using the attribute mean for numeric values or attribute mode nominal values, for all samples belonging to the same class as the given tuple.

4.3.3 Transformation of Data using Function or Mapping

- Data transformation is the process of converting data from one format or structure into another format or structure. Data transformation is critical to activities such as data integration and data management.
- Common reasons to transform data :
  a) Moving data to a new data store.
  b) Users want to join unstructured data or streaming data with structured data so users can analyze the data together.
  c) Users want to add information to data to enrich it, such as performing lookups, adding geolocation data, or adding timestamps.
  d) Users want to perform aggregations, such as comparing sales data from different regions or totalling sales from different regions.
- There are a few different ways to transform data :
  1. Scripting : Some companies perform data transformation via scripts using SQL or Python to write the code to extract and transform the data.
  2. On-premise ETL tools : ETL (Extract, Transform, Load) tools can take much of the pain out of scripting the transformations by automating the process. These tools are typically hosted on your company's site, and may require extensive expertise and infrastructure costs.
  3. Cloud-based ETL tools : These ETL tools are hosted in the cloud, where you can leverage the expertise and infrastructure of the vendor.

4.4 Analytics Types

- Business Analytics (BA) is the iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies that are committed to making data-driven decisions.
- Business analytics combines the fields of management, business and computer science. The business aspect requires both a high-level understanding of the business as well as the practical limitations that exist. The analytical part requires an understanding of data, statistics and computer science.
- Business analytics is the process of making sense of gathered data, measuring business performance and producing valuable conclusions that can help companies make informed decisions on the future of the business, through the use of various statistical methods and techniques.
- Business analytics utilizes big data, statistical analysis and data visualization to implement organization changes. Predictive analytics is an important aspect of this work as it involves available data to create statistical models.
- These models can be used to predict outcomes and inform decision making. By learning from existing data, business analytics can make concrete recommendations to solve problems and improve businesses.
- Companies use Business Analytics (BA) to make data-driven decisions. The insight gained by BA enables these companies to automate and optimize their business processes. In fact, data-driven companies that utilize business analytics achieve a competitive advantage because they are able to use the insights to :
  1. Conduct data mining.
  2. Complete statistical analysis and quantitative analysis to explain why certain results occur.
  3. Test previous decisions using A/B testing and multivariate testing.
  4. Make use of predictive modeling and predictive analytics to forecast future results.
- Challenges with developing and implementing business analytics are as follows :
  1. Executive ownership - Business analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
  2. IT involvement - Technology infrastructure and tools must be able to handle the data and Business Analytics processes.
  3. Available production data vs. Cleansed modeling data - Watch for technology infrastructure that restrict available data for historical modeling and know the difference between historical data for model development and real-time data in production.

4. Project Management Office (PMO) - The correct project management structure must be in place in order to implement predictive models and adopt an agile approach
5. End user involvement and buy-in - End users should be involved in adopting Business Analytics and have a stake in the predictive model.
6. Change management - Organizations should be prepared for the changes that business analytics bring to current business and technology operations.

- Data-driven decision-making process uses the following steps :
  1. Identify the problem or opportunity for value creation.
  2. Identify primary as well secondary data sources.
  3. Pre-process the data for issues such as missing and incorrect data. Generate derived variables and transform the data if necessary. Prepare the data for analytics model building.
  4. Divide the data sets into subsets training and validation data sets.
  5. Build analytical models and identify the best model(s) using model performance in validation data.
  6. Implement solution / Decision / Develop product.
- Descriptive analytics tells you what happened in the past.
- Predictive analytics tells you what could happen in the future.
- Prescriptive analytics tells you how you should react to possible future outcomes

4.4.1 Predictive

- Predictive analytics helps your organization predict with confidence what will happen next so that you can make smarter decisions and improve business outcomes.
- The purpose of the predictive model is finding the likelihood different samples will perform in a specific way.
- The predictive model typically calculates live transactions multiple times to help evaluate the benefit of a customer transaction.
- Predictive models typically utilize a variety of variable data to make the prediction. The variability of the component data will have a relationship with what it is likely to predict.
- Predictive analytics can be used throughout the organization, from forecasting customer behavior and purchasing patterns to identifying trends in sales activities.
- They also help forecast demand for inputs from the supply chain, operations and inventory.

- Process involved in predictive analytics :
  1. **Project definition :** Identify what shall be the outcome of the project, the deliverables, business objectives and based on that go towards gathering those data sets that are to be used.
  2. **Data collection :** This is more of the big basket where all data from various sources are binned for usage. This gives a picture about the various customer interactions as a single view item.
  3. **Analysis :** Here the data is inspected, cleansed, transformed and modelled to discover if it really provides useful information and arriving at conclusion ultimately.
  4. **Statistics :** This enables to validate if the findings, assumptions and hypothesis are fine to go ahead with and test them using statistical model.
  5. **Modelling :** Through this accurate predictive models about the future can be provided. From the options available the best option could be chosen as the required solution with multi model evaluation.
  6. **Deployment :** Through the predicative model deployment an option is created to deploy the analytics results into everyday effective decision. This way the results, reports and other metrics can be taken based on modelling.
  7. **Monitoring :** Models are monitored to control and check for performance conformance to ensure that the desired results are obtained as expected.

Examples of predictive analytics :

1. **Retail :** Probably the largest sector to use predictive analytics, retail is always looking to improve its sales position and forge better relations with customers. One of the most ubiquitous examples is Amazon's recommendations. When you make a purchase, it puts up a list of other similar items that other buyers purchased.
2. **Weather :** Weather forecasting has improved by leaps and bounds thanks to predictive analytics models. Today's five-day forecast is as accurate as a one-day forecast from the 1980s. Forecasts as long as nine to 10 days are now possible, and more important, 72 - hour predictions of hurricane tracks are more accurate than 24-hour forecasts from 40 years ago.
3. **Social media analysis :** Online social media is a fundamental shift of how information is being produced, particularly as relates to businesses. Tracking user comments on social media outlets enables companies to gain immediate feedback and the chance to respond quickly. Nothing makes a local business jump like a bad review on yelp or makes a merchant respond like a bad review on Amazon. This means collecting and sorting through massive amounts of social media data and creating the right models to extract the useful data.

4. **Health care :** Usage of predictive analytics in the health care domain can aid to determine and prevent cases and risks of those developing certain health related complications like diabetics, asthma and other life threatening ailments. Through the administering of predictive analytics in health care better clinical decisions can be made.
5. **Fraud detection :** Predictive analytics can aid to spot inaccurate credit application, deviant transactions leading to frauds both online and offline, identity thefts and false insurance claims saving financial and insurance institutions of lots of security issues and damages to their operations.

4.4.2 Descriptive

- It is simple method and used in first phase of analytics, involves gathering, organizing tabulating and depicting data then the characteristics of what we are studying.
- The descriptive model shows relationships between the customer and product/service with the acquired data. This model can be used to organize a customer by their personal preferences for example.
- Descriptive statistics are useful to show things like, total stock in inventory, average dollars spent per customer and year over year change in sales.
- Common examples of descriptive analytics are reports that provide historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers.
- While business intelligence tries to make sense of all the data that's collected each and every day by organizations of all types, communicating the data in a way that people can easily grasp often becomes an issue.
- Data visualization evolved because data displayed graphically allows for an easier comprehension of the information, validating the old adage, "a picture is worth a thousand words."
- In business, proper data visualization provides a different approach to show potential connections, relationships, etc. which are not as obvious in data that's non-visual.
- A business intelligence dashboard is an information management tool that is used to track KPIs, metrics and other key data points relevant to a business, department or specific process.
- Through the use of data visualizations, dashboards simplify complex data sets to provide users with at a glance awareness of current performance.
- Dashboards provide sleek, real-time visibility to your team.
- Combining business intelligence data with dashboards gives your team the at-a-glance view of their performance that they need to run smoothly.

- BI dashboards must be designed carefully though. If the data being fed into the visualizations is not reliable, no matter how easy the dashboard itself is to read and analyze, the dashboard will be useless.
- The goal of BI dashboards is to help business individuals make more informed decisions by enabling companies to gather, analyze, build dashboards and create reports on their most important and business-driving data.

4.4.3 Prescriptive

- This model suggests a course of action. Prescriptive analytics assists users in finding the optimal solution to a problem or in making the right choice/decision among several alternatives.
- The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take.
- A prescriptive analysis is typically not just with one individual response but is, in fact, a host of other actions.
- An example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road and crucially, the current traffic constraints.
- Another example might be producing an exam time-table such that no students have clashing schedules.
- Larger companies are successfully using prescriptive analytics to optimize production; scheduling and inventory in the supply chain to make sure that are delivering the right products at the right time and optimizing the customer experience.
- Operations Research (OR) techniques form the core of prescriptive analytics.
- With known parameters, prescriptive analytics not only can anticipate what will happen and when, but it also explains why it will happen. It can automatically improve prediction accuracy and inform the best next step because it can continually take in new data to re-predict and re-prescribe.
- Fig. 4.4.1 shows relation between all analysis. (See Fig. 4.4.1 on next page).
- Organisations can take advantage of the following benefits :
  1. Helps make decisions for the future before decisions have to be made.
  2. Can assist in mitigating risk.
  3. Continuously processes new data to give better options.
  4. Improve operations - optimise planning, reduce inefficiencies, etc.
  5. Optimise production.
  6. Schedule inventory and optimise supply chain.

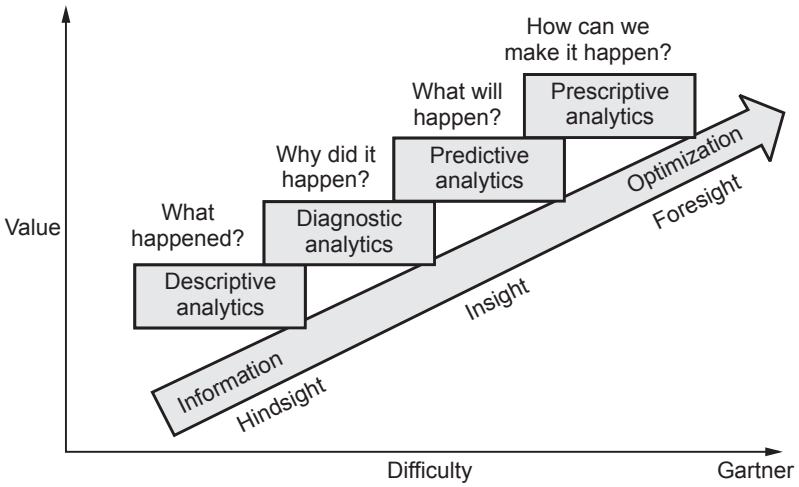


Fig. 4.4.1 Relation between descriptive, predictive and prescriptive analytics

4.4.4 Difference between Descriptive, Predictive and Prescriptive Data Analytics Model

Descriptive model	Predictive model	Prescriptive model
It use data aggregation and data mining to provide insight into the past and asnwer.	Use statistical models and forecasts techniques to understand the future and answer.	Use optimization and simulation algorithms to advice on possible outcomes and answer.
What has happned ?"	What could happened ?	What should we do ?
Descriptive analytics is the analysis of past or historical data to understand trends and evaluate metrics over time.	Predictive analytics predicts future trends.	Prescriptive analytics showcases viable solutions to a problem and the impaact of considering a solution on future trend.
Examples of tools used : Data aggregation and data mining.	Examples of tools used : Machine learning, statistical models and simulation.	Examples of tools used : Optimization and heuristics.
Used when user want to sumarize results for all or part of your business.	Used when user want to make an educated guess at likely results.	Used when user have importance iterdependent, complex or time-sensitive decisitions to make.
Limitation : Snapshot of the past, often with limited ability to help guide decisions.	Limitation : Guess at the future, helps inform low complexity decisions.	Limitation : Most effective where user have some control over what is being modeled.

4.5 Association Rules

SPPU : Aug.-18, May-19, Oct.-19, Dec.-18, 19

- Association analysis is useful for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules or sets of frequent items.

4.5.1 Market Basket Analysis

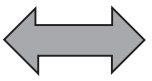
- **Market basket analysis** is an example of frequent itemset mining. The purpose of market basket analysis is to determine what products customers purchase together. It takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market basket") during grocery shopping.
- Market Basket Analysis is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence. A co-occurrence is when two or more things take place together.
- One basket tells you about what one customer purchased at one time.
  Fig. 4.5.1 shows shopping cart.
- Market Basket Analysis creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased.
- The rules are probabilistic in nature or, they are derived from the frequencies of co-occurrence in the observations.
- Frequency is the proportion of baskets that contain the items of interest. The rules can be used in pricing strategies, product placement, and various types of cross-selling strategies.
- Market basket analysis takes data at transaction level, which lists all items bought by a customer in a single purchase.

Fig. 4.5.1 Shopping cart

- The technique determines relationships of what products were purchased with which other product(s). These relationships are then used to build profiles containing If-Then rules of the items purchased.
- The rules could be written as : **If {A} Then {B}**
- The If part of the rule (the {A} above) is known as the antecedent and the THEN part of the rule is known as the consequent (the {B} above).

- The antecedent is the condition and the consequent is the result. The association rule has three measures that express the degree of confidence in the rule, Support, Confidence, and Lift.
- Association rule mining is market basket analysis we have some large number of items like "bread, butter, eggs, cereal", customer their market basket with some subset of the items and we get to know what item people buy together, even if we don't know who they are.
- Market basket analysis is just one form of frequent pattern mining. There are many kind of frequent pattern association rules in frequent mining can be classified in various way based on the following criteria.
  i) Association values handles in the rule.
  ii) Association multidimensional rule.
  iii) Association on the kinds of rules.
  iv) Association on abstraction rules.
  v) Association on complementness of pattern rule.
- Market basket data can be represented by binary format. Fig. 4.5.2 shows binary representation of market basket data.

TID	Items
1	Bread, milk
2	Bread, diaper, beer, eggs
3	Milk, diaper, beer, coke
4	Bread, milk, diaper, beer
5	Bread, milk, diaper, coke



	Beer	Bread	Milk	Diaper	Eggs	Coke
T_1	0	1	1	0	0	0
T_2	1	1	0	1	1	0
T_3	1	0	1	1	0	1
T_4	1	1	1	1	0	0
T_5	0	1	1	1	0	1

Fig. 4.5.2 Binary representation of market basket data

- **Support :** Support is the number of transactions that include items in the {A} and {B} parts of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transaction.

$$\text{Support} = \frac{A + B}{\text{Total}}$$

- **Confidence :** Confidence of the rule is the ratio of the number of transactions that include all items in {B} as well as the number of transactions that include all items in {A} to the number of transactions that include all items in {A}.

$$\text{Confidence} = \frac{A + B}{A}$$

- **Lift or lift ratio :** It is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.

$$\text{Lift Ratio} = \frac{(A + B) \ / \ A}{(B \ / \ \text{Total})} = \frac{\text{Confidence}}{(B \ / \ \text{Total})}$$

- **Leverage :** Leverage measures the difference in the probability of X and Y appearing together compared to statistical independence.

$$\text{Leverage}(X \rightarrow Y) = \text{Support}(X \wedge Y) - \text{Support}(X) * \text{Support}(Y)$$

- Leverage = 0 if X and Y are statistically independent
- Leverage > 0 indicates degree of usefulness of rule
- **Conviction :** The conviction of a rule is defined as :

$$\text{conv}(X \rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \rightarrow Y)}$$

- The conviction of the rule X => Y can be interpreted as the ratio of the expected frequency that **X** occurs without **Y** if **X** and **Y** were independent divided by the observed frequency of incorrect predictions.

4.5.2 Association Rule

- Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.
- Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository.
- An example of an association rule would be "If a customer buys a dozen eggs, he is 80 % likely to also purchase milk."
- Association rule mining can be viewed as a two-step process :
  1. **Find all frequent item sets :** By definition, each of these item sets will occur atleast as frequently as a predetermined minimum support count, min sup.
  2. **Generate strong association rules from the frequent item sets :** By definition, these rules must satisfy minimum support and minimum confidence.
- An association rule is commonly understood to be an expression of the form : $X \Rightarrow Y$ where X and Y are sets of items such that $X \cap Y = \theta$.
- The association rule $X \Rightarrow Y$ means that transactions containing items from set X tend to contain items from set Y.

- Association rules shows attribute value conditions that occur frequently together in a given data set. A typical example of association rule mining is market basket analysis.
- Data is collected using bar-code scanners in supermarkets. Such market basket databases consist of a large number of transaction records.
- Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together.
- They could use this data for adjusting store layouts, for cross-selling, for promotions, for catalog design, and to identify customer segments based on buying patterns.
- Association rules provide information of this type in the form of if-then statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature.
- In addition to the antecedent (if) and the consequent (then), an association rule has two numbers that express the degree of uncertainty about the rule.
- In association analysis, the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common).
- The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule.
- The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent, as well as the antecedent (the support) to the number of transactions that include all items in the antecedent.

4.5.3 Application of Market Basket Analysis

1. **Retail :** In retail, market basket analysis can help determine what items are purchased together, purchased sequentially, and purchased by season. This can assist retailers to determine product placement and promotion optimization. Does it make sense to sell soda and chips or soda and crackers ?
2. **Telecommunications :** Market basket analysis can be used to determine what services are being utilized and what packages customers are purchasing. They can use that knowledge to direct marketing efforts at customers who are more likely to follow the same path. Telecommunications these days is also offering TV and Internet.

3. **Banks :** In financial, market basket analysis can be used to analyze credit card purchases of customers to build profiles for fraud detection purposes and cross-selling opportunities.
4. **Insurance :** Market basket analysis can be used to build profiles to detect medical insurance claim fraud. By building profiles of claims, you are able to then use the profiles to determine if more than one claim belongs to a particular claimee within a specified period of time.
5. **Medical :** Market basket analysis can be used for com or bid conditions and symptom analysis, with which a profile of illness can be better identified. It can also be used to reveal biologically relevant associations between different genes or between environmental effects and gene expression.

Example 4.5.1

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Use the above data and group them using k-means clustering algorithm. Show calculation of centroids. **SPPU : Aug.-18 (In Sem), Marks 6**

Solution :

Step 1 : Input - Dataset, Clustering variables and maximum number of Clusters. In this dataset, only two variables - height and weight - are considered for clustering.

Step 2 : Initialize cluster centroid. In this example, value of K is considered as 2. Cluster centroids are initialized with first 2 observations.

Cluster	Initial centroid	
	Height	Weight
K_1	185	72
K_2	170	56

Step 3 : Calculate euclidean distance

Euclidean is one of the distance measures used on K means algorithm. Euclidean distance between of a observation and initial cluster centroids 1 and 2 is calculated. Based on euclidean distance each observation is assigned to one of the clusters - based on minimum distance.

$$\text{Euclidean Distance} = \sqrt{(X_H - H_1)} + (X_W - W_1)^2$$

Where, X_H : Observation value of variable height.

H_1 : Observation value of variable height.

X_W : Centroid value of cluster 1 for variable weight.

W_1 : Centroid value of cluster 1 for variable weight.

First two observations

Height	Weight
185	72
170	56

Now initial cluster centroids are :

Cluster	Updated Centroid	
	Height	Weight
K_1	185	72
K_2	170	56

Euclidean distance calculation from each of the clusters is calculated.

Euclidian distance from cluster 1	Euclidian distance from cluster 2	Assignment
$(185-185)^2 + (72-72)^2 = 0$	$(185-170)^2 + (72-56)^2 = 21.93$	1
$(170-185)^2 + (56-72)^2 = 21.93$	$(170-170)^2 + (56-56)^2 = 0$	2

Step 4 : Move on to next observation and calculate euclidean distance

Height	Weight
168	60

Euclidean distance from cluster 1	Euclidean distance from cluster 2	Assignment
$(168-185)^2+(60-72)^2$ = 20.808	$(168-185)^2+(60-72)^2$ = 4.472	2

Since distance is minimum from cluster 2, so the observation is assigned to cluster 2. Now revise cluster centroid - mean value height and weight as cluster centroids. Addition is only to cluster 2, so centroid of cluster 2 will be updated.

Updated cluster centroids

Cluster	Updated centroid	
	Height	Weight
K = 1	185	72
K = 2	(170 + 168) / 2 = 169	(56 + 60) / 2 = 58

Step 5 : Calculate euclidean distance for the next observation, assign next observation based on minimum euclidean distance and update the cluster centroids.

Height	Weight
179	68

Euclidean distance calculation and assignment

Euclidain distance from cluster 1	Euclidain distance from cluster 2	Assignment
7.211103	14.14214	1

Update cluster centroid

Cluster	Updated centroid	
	Height	Weight
K = 1	182	70.6667
K = 2	169	58

Continue the steps until all observations are assigned

Cluster Centroids

Cluster	Updated centroid	
	Height	Weight
K = 1	182.8	72
K = 2	169	58

Review Questions

1. *What is market basket analysis ? Explain Apriori algorithm with example.* **SPPU : Aug.-18 (In Sem), Marks 6**
2. *Explain Apriori association rule mining algorithm.* **SPPU : Dec.-18 (End Sem), Marks 7**
3. *Write an Apriori algorithm.* **SPPU : May-19 (End Sem), Marks 5**
4. *Define following terms with example : Confidence and Lift.* **SPPU : May-19 (End Sem), Marks 5**
5. *Explain Apriori association rule mining algorithm.* **SPPU : Oct.-19 (In Sem), Marks 5**
6. *How association rules are helpful in developing business strategy ?* **SPPU : Oct.-19 (In Sem), Dec.-19 (End Sem), Marks 5**
7. *Define following terms : i) Confidence ii) Support* **SPPU : Oct.-19 (In Sem), Marks 5**
8. *How to improve 'Apriori's efficiency ?* **SPPU : Dec.-19 (End Sem), Marks 5**

4.6 Frequent Item Set Generation

SPPU : Aug.-18

- Frequent item set mining is a method for market basket analysis. It aims at finding regularities in the shopping behaviour of customers of on-line shop, super-marks etc.
  1. A set of items is referred to as an itemset. An itemset is an unordered set of distinct items. An itemset that contains k items is a k - itemset.
  2. The set {computer, antivirus software} is a 2 - itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency, support count, or count of the itemset.
  3. Frequent itemsets that cannot be extended with any item without making them infrequent are called maximal frequent itemsets. Exact support counts of the subsets cannot be directly derived from support of the maximal frequent itemset.

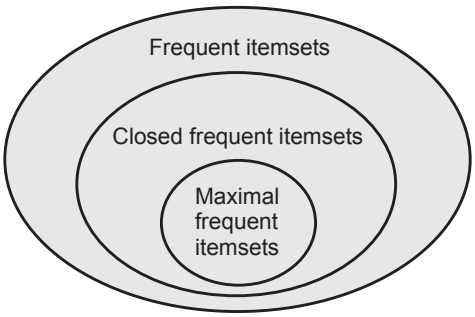


Fig. 4.6.1

Closed itemsets :

- An alternative approach is to try to retain some of the support information in the compacted representation.
- A closed itemset is an itemset whose all immediate supersets have different support count.
- A closed frequent itemset is a closed itemset that satisfies the minimum support threshold.
- Maximal frequent itemsets are closed by definition.
- An itemset X is closed in a data set S if there exists no proper super-itemset Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is bothclosed and frequent in S.
- An itemset is closed if none of its immediate supersets has the same support as the itemset.
- **Closed itemset example 1 :**

TID	Items
1	{A, B}
2	{B, C, D}
3	{A, B, C, D}
4	{A, B, D}
5	{A, B, C, D}

Itemset	Support
{A}	4
{B}	5
{C}	3
(D)	4
{A, B}	4
{A, C}	2
{A, D}	3
{B, C}	3
{B, D}	4
{C, D}	3

Itemset	Support
{A, B, C}	2
{A, B, D}	3
{A, C, D}	2
{B, C, D}	3
{A, B, C, D}	2

- Closed itemset are :

  {B}, {A, B}, {B, D}, {A, B, D}, {B, C, D}, {A, B, C, D}

- **Closed itemset example 2 :**

TID	Items
100	a, c, d, e, f
200	a, b, e
300	c, e, f
400	a, c, d, f
500	c, e, f

Total frequent itemsets : 20

{a}, {c}, {d}, {e}, {f}, {a, c}, {a, d}, {a, e}, {a, f}, {c, d}, {c, e}, {c, f}, {d, f}, {e, f},

{a, c, d}, {a, c, f}, {a d f}, {c, d, f}, {c e f}, {a, c, d, f}

Closed frequent itemssets :

{a, c, d, f}, {c, e, f}, {a, e}, {c, f}, {a}, {e}

4.6.1 The Apriori Algorithm

- The Apriori algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
- Innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item. It based on minimum support threshold.
- Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items.

  Let $D = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I.
- A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \phi$. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.
- The Apriori algorithm is used for mining frequent itemsets and devising association rules from a transactional database. The parameters "support" and "confidence" are used.
- Support refers to items' frequency of occurrence; confidence is a conditional probability.

- To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called Apriori property which helps by reducing the search space.
- Major components of Apriori algorithm are **Support, Confidence and Lift**
- The following are the main steps of the Apriori algorithm :
  1. Calculate the support of item sets (of size k = 1) in the transactional database. This is called generating the candidate set.
  2. Prune the candidate set by eliminating items with a support less than the given threshold.
  3. Join the frequent itemsets to form sets of size k + 1, and repeat the above sets until no more itemsets can be formed. This will happen when the set(s) formed have a support less than the given support.

Pseudocode of Apriori algorithm

L_1 = {frequent items};

for (k = 2; L_{k-1} ! = ϕ; k++) **do begin**

C_k= Candidates generated from L_{k-1} (that is : cartesian product $L_{k-1} \times L_{k-1}$ and eliminating any

k-1 size itemset that is not frequent);

foreach transaction t in database **do**

increment the count of all candidates in

C_k that are contained in t

L_k = candidates in C_k with min_sup

end

return $\bigcup_k L_k$;

4.6.2 Limitations of Apriori Algorithm

1. Needs several iterations of the data.
2. Uses a uniform minimum support threshold.
3. Difficulties to find rarely occuring events.
4. Some competing alternative approaches focus on partition and sampling.

Example 4.6.1 *Generate frequent itemsets and generate association rules based on it using apriori algorithm. Minimum support is 50 % and minimum confidence is 70 %.*

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

Solution : Apriori algorithm :

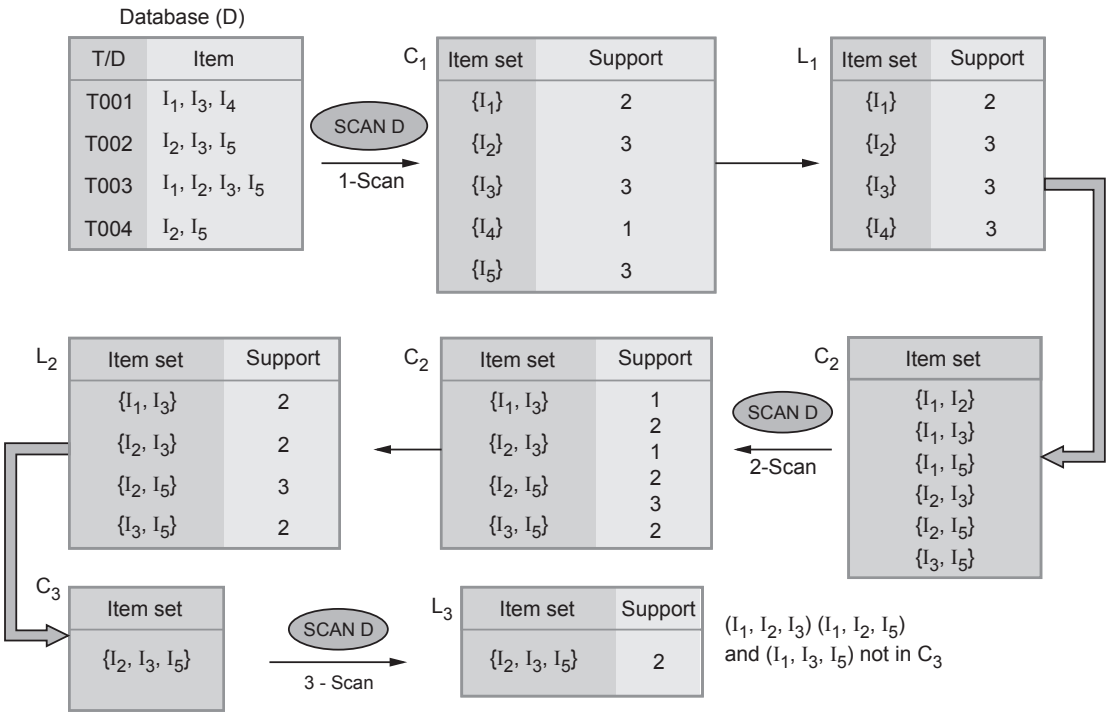


Fig. 4.6.2

Example 4.6.2 *Using Apriori algorithm, generate frequent item sets (min_sup > = 33.3 %) for the following transaction database.*

Trans_id	Itemlist
T_1	{A, B, D, K}
T_2	{A, B, C, D, E}
T_3	{A, B, C, E}
T_4	{B, D}
T_5	{A, C}
T_6	{B, D}

Solution :

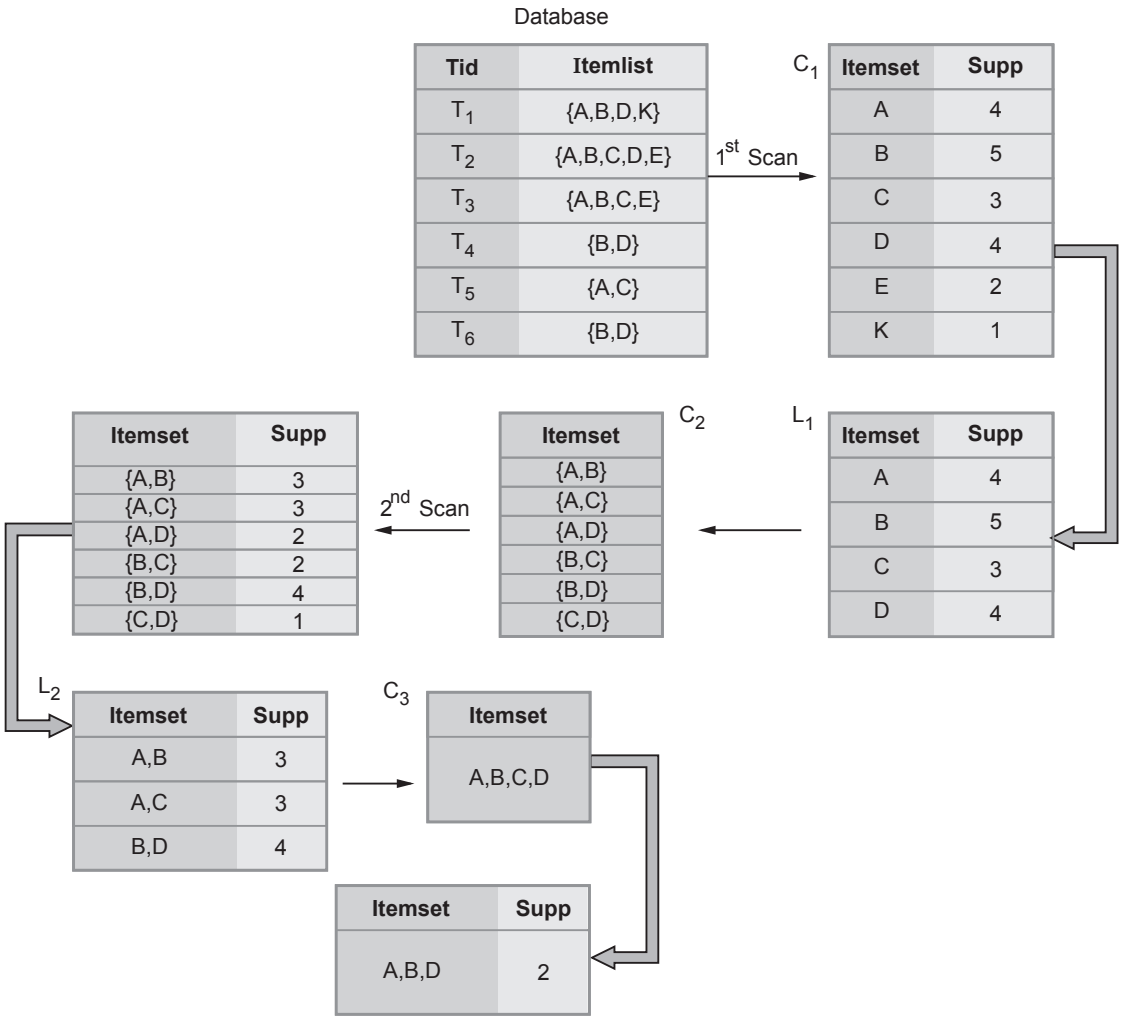


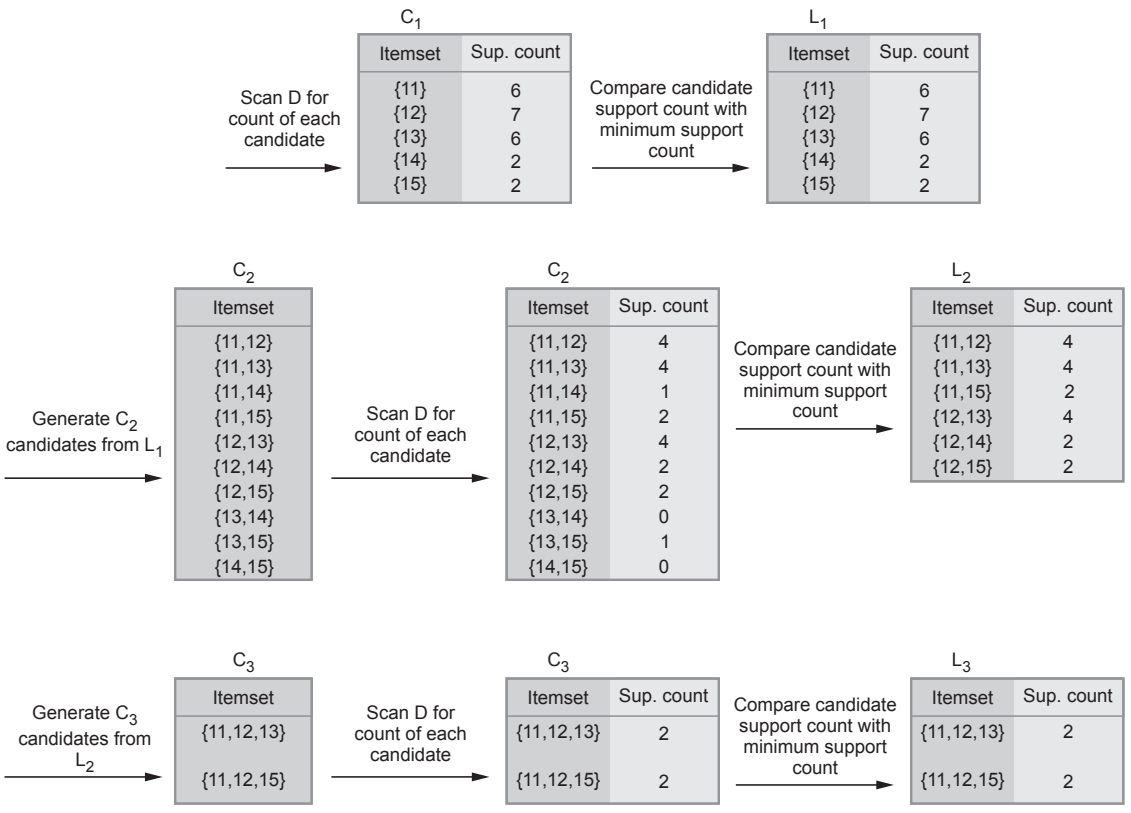
Fig. 4.6.3

Example 4.6.3 *State the Apriori property. Generate candidate itemsets, frequent itemsets and association rules using Apriori algorithm on the following data set with minimum support count is 2.*

	TID	List of items_IDs
1.	T100	11, 12, 15
2.	T200	12, 14
3.	T300	12, 13
4.	T400	11, 12, 14
5.	T500	11, 13
6.	T600	12, 13
7.	T700	11, 13
8.	T800	11, 12, 13, 15
9.	T900	11, 12, 13

SPPU : Aug.-18 (In Sem), Marks 6

Solution :



Example 4.6.4 *Consider the following set of frequent 3-itemsets : {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}. Assume that there are only five items in the data set.*

i) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori algorithm.

ii) List all candidate 4-itemsets that survive the candidate pruning setup of the Apriori algorithm.

Solution : Supports for 1 - itemssets :

Item	Support
1	5
2	5
3	6
4	4
5	4

i) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori :

{1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5,}, {2, 3, 4, 5} {2, 3, 4, 6}

ii) List all candidate 4-itemsets that survive the candidate pruning setp of the Apriori algorithm.

{1, 2, 3, 4}

Example 4.6.5 *A database has five transactions. Let minimum support is 60 %.*

TID	Items
1	*Butter, Milk*
2	*Butter, Dates, Balloon, Eggs*
3	*Milk, Dates, Balloon, Cake*
4	*Butter, Milk, Dates, Balloon*
5	*Butter, Milk, Dates, Cake*

Find all the frequent item sets using Apriori algorithm. Show each step.

Solution : Database is scanned once to generate frequent 1 - itemsets. To do this, use absolute support, where duplicate values are counted only once per TID.

Itemset	Support	Support %
(Butter)	4	80 %
{Milk}	4	80 %
{Dates}	4	80 %
{Balloon}	3	60 %
{Eggs}	1	20 %
{Cake)	2	40 %

- The total number of TID is 5, so minimum support of 60 % is equivalent to 3/5. Thus itemsets with 1 or 2 support counts are eliminated.

Itemset	Support	Support %
(Butter}	4	80 %
{Milk}	4	80 %
{Dates}	4	80 %
{Balloon}	3	60 %

- Now, database is scanned second time to generate frequent 2 - itemsets. Using absolute support, each combination is counted per TID, and combinations that are below support value of 3 are eliminated.

Itemset	Support	Support %
{Butter, Milk}	3	60 %
{Butter, Dates}	3	60 %
{Butter, Balloon}	2	40 %
{Butter, Cake}	1	20 %
{Butter, Eggs}	1	20 %
{Milk, Dates}	3	60 %
{Milk, Balloon}	2	40 %
{Milk, Cake}	2	40 %
{Dates, Balloon}	3	60 %
{Dates, Eggs}	1	20 %
{Dates, Cake}	2	40 %
{Balloon, Cake}	1	20 %
{Balloon, Eggs}	1	20 %
{Eggs, Cake)	0	0

2 - itemset results, consolidated :

Itemset	Support	Support %
(Butter, Milk}	3	60 %
{Milk, Dates}	3	60 %
{Dates, Dates}	3	60 %
{Balloon, Balloon}	3	60 %

To generate frequent 3 - itemsets :

Itemset	Support	Support %
(Butter, Milk, Dates}	2	40 %
{Milk, Dates, Balloon}	2	40 %

4.6.3 Challenges of Frequent Pattern Mining

Challenges :

1. Multiple scans of transaction database
2. Huge number of candidates
3. Tedious workload of support counting for candidates

- Improving Apriori : general ideas
  1. Reduce number of transaction database scans
  2. Shrink number of candidates
  3. Facilitate support counting of candidates

4.6.4 Improving Apriori Efficiency

- Apriori algorithm is a classical algorithm of association rule mining and widely used for generating frequent item sets. This classical algorithm is inefficient due to so many scans of database. And if the database is large, it will take too much time to scan the database. To overcome these limitations, researchers have made a lot of improvements to the Apriori.
- Techniques for improving efficiency of Apriori algorithm are as follows :
  1. **Hash based technique :** It is used to reduce the size of the candidate k-itemsets (C_k) for k > 1. These techniques work by creating a dictionary (hash table) that stores the candidate item sets as keys, and the number of appearances as the value. Initialization start with zero and Increment the counter for each item set that you see in the data.

2. **Transaction reduction :** In this approach, the number of transactions to be scanned is greatly reduced when comparing to the original Apriori algorithm by reducing the number of similar transactions in the database and this results in reduction of time.
3. **Partitioning :** The partitioning algorithm divides the transactional dataset D into "n" non-overlapping partitions, D_1, D_2, ..., D_n. The algorithm reduces the number of datasets process to two phases. During the first phase, the algorithm finds all item sets in each partition. Those local frequent item sets are collected into the global candidate item sets. During the second phase, these global item sets are counted to determine if they are large across the entire dataset. The partitioning algorithm improves the performance of finding frequent item sets and also provide several advantages. Small partitions might be fit into main memory than large one. Because the size of each partition is small, the algorithm might reduce the size of candidate item sets.
4. **Sampling :** The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D.

4.7 Mining Frequent Itemset without Candidate Generation

- FP-Growth Algorithm was introduced by Han, Pei and Yin in 2000 to eliminate the candidate generation of Apriori algorithm.
- FP-growth algorithm using a root-like data structure and divide and conquer strategy to find candidate, this makes the FP-Growth algorithm as an efficient algorithm to find rules.
- FP growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree. This tree structure will maintain the association between the itemsets.
- A FP-tree is a compact data structure that represents the data set in tree form. Each transaction is read and then mapped onto a path in the FP-tree. This is done until all transactions have been read. Different transactions that have common subsets allow the tree to remain compact because their paths overlap.
- The construction of a FP-tree is subdivided into three major steps :
  1. Scan the data set to determine the support count of each item, discard the infrequent items and sort the frequent items in decreasing order.
  2. Scan the data set one transaction at a time to create the FP-tree. For each transaction :
     - If it is a unique transaction form a new path and set the counter for each node to 1.
     - If it shares a common prefix itemset then increment the common itemset node counters and create new nodes if needed.

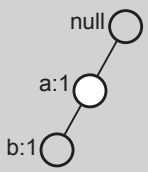
3. Continue this until each transaction has been mapped unto the tree.

- FP - growth algorithm can reduce memory and time used to find association rules because the FP - growth algorithm only needs to scan the database two times to find rules candidates.
- **FP - tree construction example :**

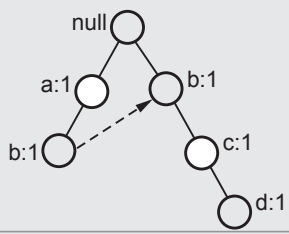
Transaction data set :

TID	Items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

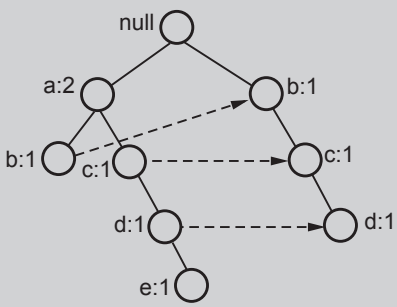
i) After reading TID = 1



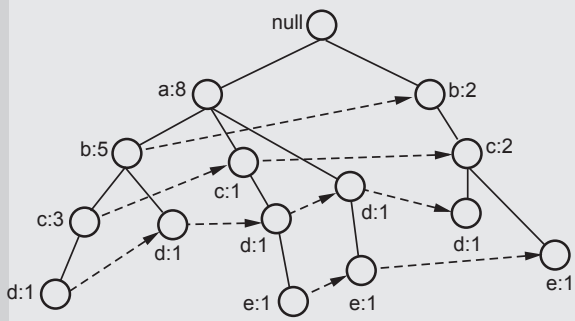
ii) After reading TID = 2



iii) After reading TID = 3



iv) After reading TID = 10



4.7.1 Advantages and Disadvantages of FP - Growth

Advantages :

1. This algorithm needs to scan the database only twice when compared to Apriori which scans the transactions for each iteration.
2. The pairing of items is not done in this algorithm and this makes it faster.
3. The database is stored in a compact version in memory.
4. It is efficient and scalable for mining both long and short frequent patterns.

Disadvantages :

1. FP Tree is more cumbersome and difficult to build than Apriori.
2. It may be expensive.
3. When the database is large, the algorithm may not fit in the shared memory.

4.7.2 Difference between FP - Growth and Apriori Algorithm

FP - growth	Apriori algorithm
FP - Growth algorithm using a root - like data structure and divide and conquer strategy to find candidate.	Apriori algorithm using a Brute-force strategyto find data patterns by scanning the database repeatedly.
FP - growth algorithm works better with a small dataset.	Apriori algorithm works better with a big dataset.
There is no candidate generation. no new candidate	Apriori algorithm uses candidate generation.
FP growth generates pattern by constructing a FP tree.	Apriori generates pattern by pairing the items into singletons, pairs and triplets.
Scan the database only two times.	Multiple scan for generating candidate sets.
It is tree based algorithm.	It is array based algorithm.
FP Growth uses a depth-first search.	Apriori uses a breadth-first search.

4.8 Regression

SPPU : Aug.-18, Oct.-19

- Regression is a data mining function that predicts a number. Profit, sales, mortgage rates, house values, square footage, temperature or distance could all be predicted using regression techniques.
- For example, a regression model could be used to predict the values of a data warehouse based on web-marketing, number of data entries, size and other factors.

- A regression task begins with a data set in which the target values are known. Regression analysis is a good choice when all of the predictor variables are continuously valued as well.
- For an input x, if the output is continuous, this is called a regresssion problem. For example, based on historical information of demand for tooth paste in your supermarket, you are asked to predict the demand for the next month.
- Regression is concerned with the prediction of continuous quantities. Linear regression is the oldest and most widely used preditive model in field of machine learning. The goal is to minimize the sum of the squared errors to fit a striaght line to a set of data points.
- For regression tasks, the typical accuracy metrics are Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). These metrics measure the distance between the the predicted numeric target and the actual numeric answer.

Regression Line

- **Least squares :** The least squares regression line is the line that makes the sum of squared residuals as small as possible. Linear means "straight line".
- **Regression line** is the line which gives the best estimate of one variable from the value of any other given variable.
- **The regression line** gives the average relationship between the two variables in mathematical form.
- For two variables X and Y, there are always two lines of regression.
- **Regression line of X on Y :** Gives the best estimate for the value of X for any specific given values of Y :

$$\mathbf{X = a + b\ Y}$$

where, a = X - intercept

b = Slope of the line

X = Dependent variable

Y = Independent variable

- **Regression line of Y on X :** Gives the best estimate for the value of Y for any specific given values of X :

$$\mathbf{Y = a + bx}$$

where a = Y - intercept

b = Slope of the line

Y = Dependent variable

x = Independent variable

4.8.1 Linear Regression

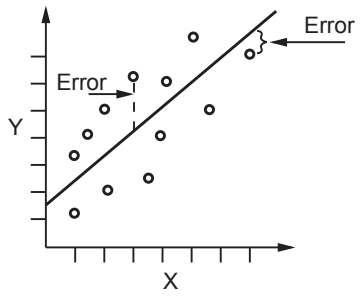


Fig. 4.8.1 Linear regression

- The simplest form of regression to visualize is linear regression with a single predictor. A linear regression technique can be used if the relationship between X and Y can be approximated with a straight line.
- Linear regression with a single predictor can be expressed with the equation :
  $y = \theta_2 x + \theta_1 + e$
- The regression parameters in simple linear regression are the slope of the line (θ_2), the angle between a data point and the regression line and the y intercept (θ_1) the point where x crosses the y axis (X = 0).
- Model **'Y'**, is a linear function of '**X**'. The value of 'Y' increases or decreases in linear manner according to which the value of 'X' also changes.

Nonlinear Regression :

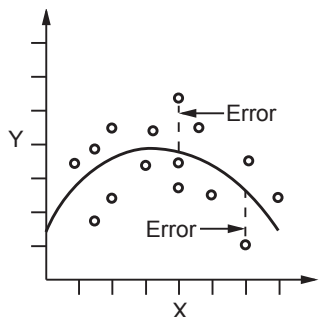


Fig. 4.8.2 Nonlinear regression

- Often the relationship between x and y cannot be approximated with a straight line. In this case, a nonlinear regression technique may be used.
- Alternatively, the data could be preprocessed to make the relationship linear. In Fig. 4.8.2 shows nonlinear regression.
- The X and Y have a nonlinear relationship.
- If data does not show a linear dependence we can get a more accurate model using a nonlinear regression model.
- For example : $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$
- Generalized linear model is foundation on which linear regression can be applied to modeling categorical response variables.

Advantages :

a. Training a linear regression model is usually much faster than methods such as neural networks.

b. Linear regression models are simple and require minimum memory to implement.

c. By examining the magnitude and sign of the regression coefficients you can infer how predictor variables affect the target outcome.

- There are two **important shortcomings** of linear regression :
  1. **Predictive ability :** The linear regression fit often has low bias but high variance. Recall that expected test error is a combination of these two quantities. Prediction accuracy can sometimes be improved by sacrificing some small amount of bias in order to decrease the variance.
  2. **Interpretative ability :** Linear regression freely assigns a coefficient to each predictor variable. When the number of variables p is large, we may sometimes seek, for the sake of interpretation, a smaller set of important variables.

Example 4.8.1 *Define linear and nonlinear regression using figures. Calculate the value of Y for X = 100 based on linear regression prediction method.*

X	Y
4	390
9	580
10	650
14	730
4	410
7	530
12	600
22	790
1	350
3	400
8	590
11	640
5	450
6	520
10	690
11	690
16	770
13	700
13	730
10	640

Solution :

X	Y	X.Y	X^2
4	390	1560	16
9	580	5220	81
10	650	6500	100
14	730	10220	196
4	410	1640	16
7	530	3710	49
12	600	7200	144
22	790	17380	484
1	350	350	1
3	400	1200	9
8	590	4720	56
11	640	7040	121
TT= 105	**TT= 6660**	**TT= 66740**	**TT= 1273**

N = Total number of samples = 12

$f = aA + b$

but linear equation calculate slope and interception prediction.

$$f = a_0 + a_1 N$$

$$\overline{X} = \frac{x}{N} = \frac{105}{12} = 8.75$$

$$\overline{Y} = \frac{y}{N} = \frac{6660}{12} = 555$$

$$X_1 = \frac{\begin{array}{l}(4-8.75)(390-555)+(9-8.75)(580-555)+(10-8.75)(650-555)\\+(14-8.75)(730-555)+(4-8.75)(410-555)+(7-8.75)(530-555)\\+(12-8.75)(600-555)+(22-8.75)(790-555)+(1-8.75)(350-555)\\+(3-8.75)(400-555)+(8-8.75)(590-555)+(11-8.75)(640-555)\end{array}}{\begin{array}{l}(4-8.75)^2+(9-8.75)^2+(10-8.75)^2+(14-8.75)^2+(4-8.75)^2\\+(7-8.75)^2+(12-8.75)^2+(10-8.75)^2+(14-8.75)^2+(4-8.75)^2\\+(7-8.75)^2+(11-8.75)^2\end{array}}$$

$$X_1 = \frac{8465}{362.25} = 23.36$$

$$X_0 = \overline{Y} - X \cdot x = 555 - (8.75)(23.36) = 350.6$$

$$F = 350.6 + 23.36 \times 12 = 630.92$$

4.8.2 Logistic Regression

- Logistic regression is a form of regression analysis in which the outcome variable is binary or dichotomous. A statistical method used to model dichotomous or binary outcomes using predictor variables.
- **Logistic component :** Instead of modeling the outcome, Y, directly, the method models the log odds (Y) using the logistic function.
- **Regression component :** Methods used to quantify association between an outcome and predictor variables. It could be used to build predictive models as a function of predictors.
- **In simple** logistic regression, logistic regression with 1 predictor variable.

Logistic Regression :

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- With logistic regression, the response variable is an indicator of some characteristic, that is, a 0/1 variable. Logistic regression is used to determine whether other measurements are related to the presence of some characteristic, for example, whether certain blood measures are predictive of having a disease.
- If analysis of covariance can be said to be a t test adjusted for other variables, then logistic regression can be thought of as a chi-square test for homogeneity of proportions adjusted for other variables. While the response variable in a logistic regression is a 0/1 variable, the logistic regression equation, which is a linear equation, does not predict the 0/1 variable itself.

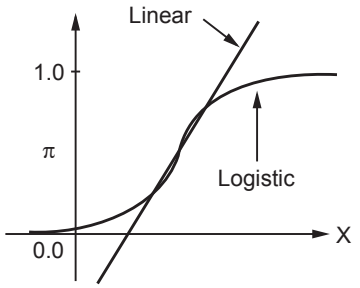


Fig. 4.8.3

- Fig. 4.8.3 shows Sigmoid curve for logistic regression.
- The linear and logistic probability models are :

Linear Regression :

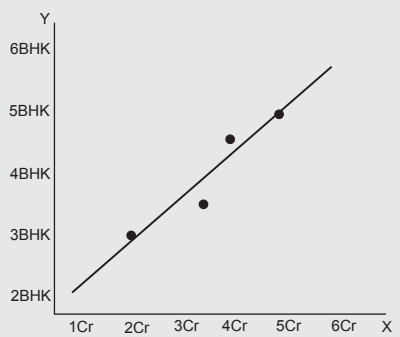
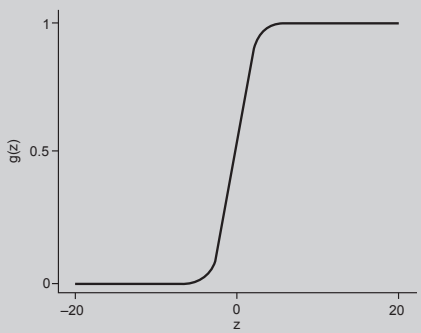
$$p = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_k X_k$$

Logistic Regression :

$$\ln[p/(1-p)] = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

- The linear model assumes that the probability p is a linear function of the regressors, while the logistic model assumes that the natural log of the odds $p/(1-p)$ is a linear function of the regressors.
- The major advantage of the linear model is its interpretability. In the linear model, if a 1 is 0.05, that means that a one-unit increase in X1 is associated with a 5 % point increase in the probability that Y is 1.
- The logistic model is less interpretable. In the logistic model, if b1 is 0.05, that means that a one - unit increase in X1 is associated with a 0.05 increase in the log odds that Y is 1. And what does that mean ? I've never met anyone with any intuition for log odds.

4.8.3 Difference between Linear and Logistic Regression

Sr. No.	Linear regression	Logistic Regression
1.	Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
2.	Target is an interval variable.	Target is discrete variable.
3.	Solve regression problem	Solve classification problem
4.	Example : Relationship between number of hours worked with your salary	Example : whether they are male or female
5.	Example : What is temperature ?	Example : Will it rain or not ?
6.		

Review Questions

1. *Explain logistic regression. Explain use cases of logistic regression.*
   SPPU : Aug.-18 (In Sem), Marks 4
2. *What is regression ? Explain any one type of regression in detail.*
   SPPU : Aug.-18 (In Sem), Marks 4
3. *Explain linear regression in detail.* **SPPU : Oct.-19 (In Sem), Marks 5**

4.9 Classification

SPPU : Dec.-18, May-19

- Classification predicts categorical labels (classes), prediction models continuous-valued functions. Classification is considered to be supervised learning.
- Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. **Prediction** means models continuous-valued functions, i.e., predicts unknown or missing values.
- Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation, such as generalizing the data to higher level concepts or normalizing data.
- Fig. 4.9.1 shows the classification.

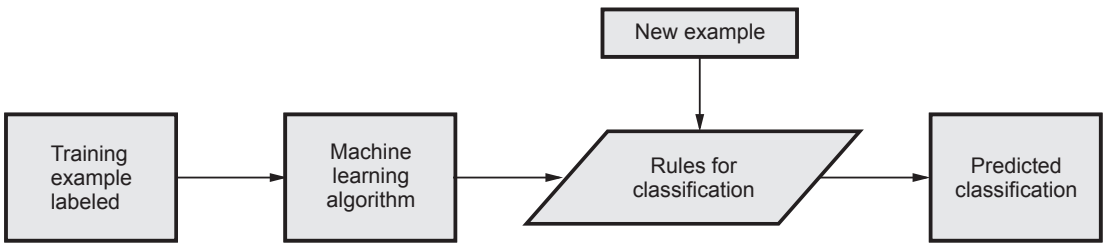


Fig. 4.9.1 Classification

Aim : To predict categorical class labels for new samples

Input : Training set of samples, each with a class label

Output : Classifier is based on the training set and the class labels

- Prediction is similar to classification. It constructs a model and uses the model to predict unknown or missing value.
- Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.
- Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process.
- Numeric prediction is the task of predicting continuous values for given input. For example, we may wish to predict the salary of college employee with 15 years of work experience, or the potential sales of a new product given its price.
- Some of the classification methods like back - propagation, support vector machines, and k - nearest - neighbor classifiers can be used for prediction.

4.9.1 Naïve Bayes

- Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. It is highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.
- A Naive Bayes Classifier is a program which predicts a class value given a set of attributes. For each known class value,
  1. Calculate probabilities for each attribute, conditional on the class value.
  2. Use the product rule to obtain a joint conditional probability for the attributes.
  3. Use Bayes rule to derive conditional probabilities for the class variable.
- Once this has been done for all class values, output the class with the highest probability.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.
- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

4.9.2 Naive Bayes Classifiers

- Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. It is highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.
- A Naive Bayes classifier is a program which predicts a class value given a set of attributes.
- For each known class value,
  1. Calculate probabilities for each attribute, conditional on the class value.
  2. Use the product rule to obtain a joint conditional probability for the attributes.
  3. Use Bayes rule to derive conditional probabilities for the class variable.
- Once this has been done for all class values, output the class with the highest probability.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

Conditional Probability

- Let A and B be two events such that P(A) > 0. We denote P(B | A) the probability of B given that A has occurred. Since A is known to have occurred, it becomes the new sample space replacing the original S. From this, the definition is,

$$\mathbf{P(B/A) \equiv \frac{P(A \cap B)}{P(A)}}$$

OR

$$\mathbf{P(A \cap B) = P(A)\ P(B/A)}$$

- The notation P(B | A) is read **"the probability of event B given event A"**. It is the probability of an event B given the occurrence of the event A.
- We say that, the probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs given that A has occurred. We call **P(B | A) the conditional probability of B given A**, i.e., the probability that B will occur given that A has occurred.
- Similarly, the conditional probability of an event A, given B by,

$$P(A/B) \equiv \frac{P(A \cap B)}{P(B)}$$

- The probability P(A | B) simply reflects the fact that the probability of an event A may depend on a second event B. If A and B are mutually exclusive $A \cap B = \phi$ and P(A | B) = 0.
- Another way to look at the conditional probability formula is :

$$P(\text{Second}/\text{First}) = \frac{P\ (\text{First choice and second choice})}{P\ (\text{First choice})}$$

- Conditional probability is a defined quantity and cannot be proven.
- The key to solving conditional probability problems is to :
  1. Define the events.
  2. Express the given information and question in probability notation.
  3. Apply the formula.

Joint Probability

- A joint probability is a probability that measures the likelihood that two or more events will happen concurrently.

- If there are two independent events A and B, the probability that A and B will occur is found by multiplying the two probabilities. Thus for two events A and B, the special rule of multiplication shown symbolically is :

  $P(A \text{ and } B) = P(A)\, P(B)$.

- The general rule of multiplication is used to find the joint probability that two events will occur. Symbolically, the general rule of multiplication is,

  $P(A \text{ and } B) = P(A)\, P(B|A)$.

- The probability $P(A \cap B)$ is called the joint probability for two events A and B which intersect in the sample space. Venn diagram will readily shows that

  $$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

Equivalently :

$$P(A \cap B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$$

- The probability of the union of two events never exceeds the sum of the event probabilities.
- A tree diagram is very useful for portraying conditional and joint probabilities. A tree diagram portrays outcomes that are mutually exclusive.

Bayes Theorem

- Bayes' theorem is a method to revise the probability of an event given additional information. Bayes's theorem calculates a conditional probability called a posterior or revised probability.
- Bayes' theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, $P(A|B)$ denotes the conditional probability of A occurring, given that B occurs. The two conditional probabilities $P(A|B)$ and $P(B|A)$ are in general different.
- Bayes theorem gives a relation between $P(A|B)$ and $P(B|A)$. An important application of Bayes' theorem is that it gives a rule how to update or revise the strengths of evidence-based beliefs in light of new evidence a posteriori.
- A **prior probability** is an initial probability value originally obtained before any additional information is obtained.
- A **posterior probability** is a probability value that has been revised by using additional information that is later obtained.

- Suppose that $B_1, B_2, B_3 \ldots B_n$ partition the outcomes of an experiment and that A is another event. For any number, k, with $1 \leq k \leq n$, we have the formula :

$$P(B_k/A) = \frac{P(A/B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A/B_i) \cdot P(B_i)}$$

Example 4.9.1 *At a certain university, 4 % of men are over 6 feet tall and 1 % of women are over 6 feet tall. The total student population is divided in the ratio 3 : 2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman ?*

Solution : Let us assume following :

M = {Student is Male},

F = {Student is Female},

T = {Student is over 6 feet tall}.

Given data : P(M) = 2/5,

P(F) = 3/5,

P(T | M) = 4/100

P(T | F) = 1/100.

We require to find P(F | T) ?

Using Bayes' Theorem we have :

$$P(F/T) = \frac{P(T/F)\,P(F)}{P(T/F)\,P(F) + P(T/M)\,P(M)}$$

$$= \frac{\frac{1}{100} \times \frac{3}{5}}{\frac{1}{100} \times \frac{3}{5} + \frac{4}{100} \times \frac{2}{5}} = \frac{\frac{3}{500}}{\frac{3}{500} + \frac{8}{500}}$$

$$P(F/T) = \frac{3}{11}$$

Example 4.9.2 *Bag contains 5 red balls and 2 white balls. Two balls are drawn successivly without replacement. Draw the probability tree for this.*

Solution : Let R_1 = for the event of getting a red ball on the first draw, W_2 for getting a white ball on the second draw, and so forth. Here's the probability tree.

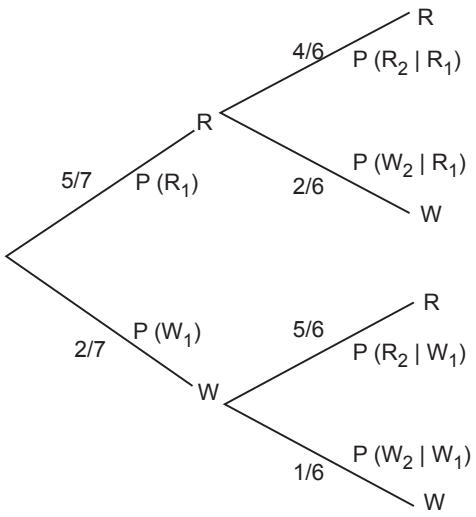


Fig. 4.9.2

Example 4.9.3 *The proportions of bike owner at the petrol pump in the city using regular, extra unleaded and premium petrol are 40 %, 35 % and 25 %, respectively. The respective propertions of filling their tanks are 30 %, 50 % and 60 %. If a randomly chosen motorist filled his/her tank, what is the probability that he / she used regular petrol ?*

Solution : Let assume that

R = Regular, E = Extra unleaded, P = Premium and F = Full tank

Given data : P(R) = 0.4, P(E) = 0.35, P(P) = 0.25, P(F | R) = 0.3, P(F | E) = 0.5, P(F | P) = 0.6, **P(R | F) = ?**

By using Bayes formula :

$$P(R \mid F) = \frac{P(F \mid R)\, P(R)}{P(F \mid R)\, P(R) + P(F \mid E)\, P(E) + P(F \mid P)\, P(P)}$$

$$= \frac{0.3 \times 0.4}{0.3 \times 0.4 + 0.5 \times 0.35 + 0.6 \times 0.25} = \frac{0.12}{0.12 + 0.175 + 0.15} = \frac{0.12}{0.445}$$

$$P(R \mid F) = 0.26966$$

Review Questions

1. *Explain Bayes 'theorem. Explain Naive Bayes' classifier.* **SPPU : Dec.-18 (End Sem), Marks 8**
2. *What is classification ? List the different classifiers.* **SPPU : Dec.-18 (End Sem), Marks 3**
3. *How Naive Baye's classification works ? Give its applications.* **SPPU : May-19 (End Sem), Marks 8**

4.10 Decision Trees

SPPU : Dec.-18, 19, May-19

- A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value).
- A decision tree or a classification tree is a tree in which each internal node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature.
- Each leaf of the tree is labeled with a class or a probability distribution over the classes.
- A decision tree is a simple representation for classifying examples.
- In this method a set of training examples is broken down into smaller and smaller subsets while at the same time an associated decision tree get incrementally developed.
- At the end of the learning process, a decision tree covering the training set is returned.
- The key idea is to use a decision tree to partition the data space into cluster regions and empty regions.
- Decision tree consists of
  1. **Nodes :** Test for the value of a certain attribute.
  2. **Edges :** Correspond to the outcome of a test and connect to the next node or leaf.
  3. **Leaves :** Terminal nodes that predict the outcome.
- A tree is defined as a set of logical conditions on attributes; a leaf represents the subset of instances corresponding to the conjunction of conditions along its branch or path back to the root.
- A simple approach to ranking is to estimate the probability of an instance's membership in a class and assign that probability as the instance's rank. A decision tree can easily be used to estimate these probabilities.
- There are several steps involved in the building of a decision tree.
  1. **Splitting :** The process of partitioning the data set into subsets. Splits are formed on a particular variable and in a particular location. For each split, two determinations are made : the predictor variable used for the split, called the splitting variable, and the set of values for the predictor variable, called the split point.

2. **Pruning :** The shortening of branches of the tree. Pruning is the process of reducing the size of the tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch.
3. **Tree selection :** The process of finding the smallest tree that fits the data. Usually this is the tree that yields the lowest cross-validated error.

- The Fig. 4.10.1 shows an example of a decision tree to determine what kind of contact lens a person may wear.

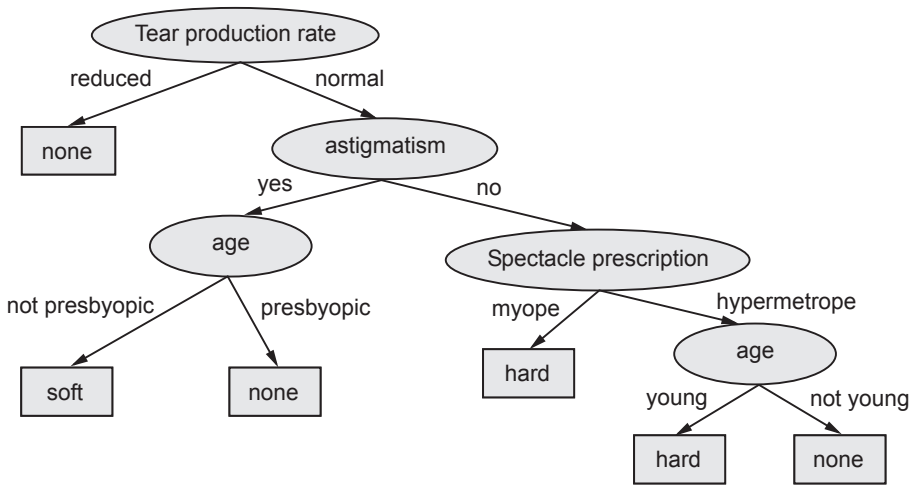


Fig. 4.10.1

- The choices (classes) are none, soft and hard. The attributes that we can obtain from the person are their tear production rate (reduced or normal), whether they have astigmatism (yes/no), their age category (presbyopic or not, young or not), their spectacle prescription (myopia or hypermetropia).
- The decision tree learning algorithm recursively learns the tree as follows :
  1. Assign all training instances to the root of the tree. Set current node to root node.
  2. For each attribute
     a) Partition all data instances at the node by the value of the attribute.
     b) Compute the information gain ratio from the partitioning.
  3. Identify feature that results in the greatest information gain ratio. Set this feature to be the splitting criterion at the current node.
     a) If the best information gain ratio is 0, tag the current node as a leaf and return.
  4. Partition all instances according to attribute value of the best feature.

5. Denote each partition as a child node of the current node.
6. For each child node :
   a) If the child node is "pure" (has instances from only one class) tag it as a leaf and return.
   b) If not set the child node as the current node and go to step 2.

4.10.1 Advantages and Disadvantages of Decision Tree

Advantages :

1. Rules are simple and easy to understand.
2. Decision trees can handle both nominal and numerical attributes.
3. Decision trees are capable of handling datasets that may have errors.
4. Decision trees are capable of handling datasets that may have missing values.
5. Decision trees are considered to be a nonparametric method.
6. Decision trees are self-explanatory.

Disadvantages :

1. Most of the algorithms require that the target attribute will have only discrete values.
2. Some problem are difficult to solve like XOR.
3. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
4. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.

4.10.2 Decision Tree Induction

1. Classification and Regression Tree (CART)

- **Classification Tree :** When decision or target variable is categorical, the decision tree is classification decision tree.
- **Regression tree :** When the decision or target variable is continuous variable, the decision tree is called regression decision tree.
- CART algorithm can be used for building both classification and regression decision trees. The impurity measure used in building decision tree in CART is Gini Index. The decision tree built by CART algorithm is always a binary decision tree.

- Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.
- Gini index, entropy and towing rule are some of the frequency used impurity measures.
- Gini Index for a given node t :

$$GINI(t) = p(j \mid t)\,(1 - p(j \mid t)) - p(j \mid t)^2$$

- Maximum of $\frac{1-1}{n_c}$ when records are equally distributed among all classes = Maximal impurity.
- Minimum of 0 when all records belong to one class = Complete purity.
- Entropy at a given node by :

$$\text{Entropy } (t) = \sum_j p(j / t) \log p(j \mid t)$$

- Maximum ($\log n_c$) when records are equally distributed among all classes (maximal impurity).
- Minimum (0.0) when all records belongs to one class (maximal purity).
- Entropy is the only function that satisfies all of the following three properties :
  1. When node is pure, measure should be zero.
  2. When impurity is maximal (i.e. all classes equally likely), measure should be maximal.
  3. Measure should obey multistage property.
- When a node p is split into k partitions (children), then quality of the split is computed as a weighted sum :

$$GIN_{Isplit} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) = \sum_j p(j / t)^2$$

where n_i = Number of records at child i,

and n = Number of records at node P.

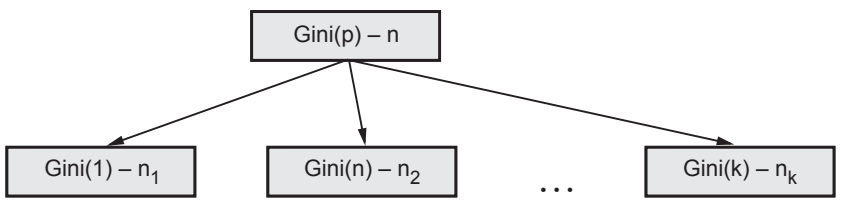


Fig. 4.10.2

- A problem with all impurity measures is that they depend only on the number of (training) patterns of different classes on either side of the hyperplane. Thus, if we change the class regions without changing the effective areas of class regions on either side of a hyperplane, the impurity measure of the hyperplane will not change.
- Thus the impurity measures do not really capture the geometric structure of class distributions. Also, all the algorithms need to optimize on some average of impurity of the child nodes and often it is not clear what kind of average is proper.

2. Information gain

- Entropy measures the impurity of a collection. Information gain is defined in terms of entropy.
- Information gain tells us how important a given attribute of the feature vectors is,
- Information gain of attribute A is the reduction in entropy caused by partitioning the set of examples S.

$$\text{Gain } (S, A) \equiv \text{Entropy } (S) - \sum_{v \in \text{values}} \frac{|S_v|}{S} \text{Entropy } (S_v)$$

where values (A) is the set of all possible values for attribute A and Sv is the subset of S for which attribute A has value v.

Pruning by information gain :

- The simplest technique is to prune out portions of the tree that result in the least information gain.
- This procedure does not require any additional data, and only bases the pruning on the information that is already computed when the tree is being built from training data.
- The process of information gain based pruning required us to identify "twigs", nodes whose children are all leaves.
- "Pruning" a twig removes all of the leaves which are the children of the twig and makes the twig a leaf. The Fig. 4.10.3 illustrates this.
  (Refer Fig. 4.10.3 on next page).
- The algorithm for pruning is as follows :
  1. Catalog all twigs in the tree.
  2. Count the total number of leaves in the tree.

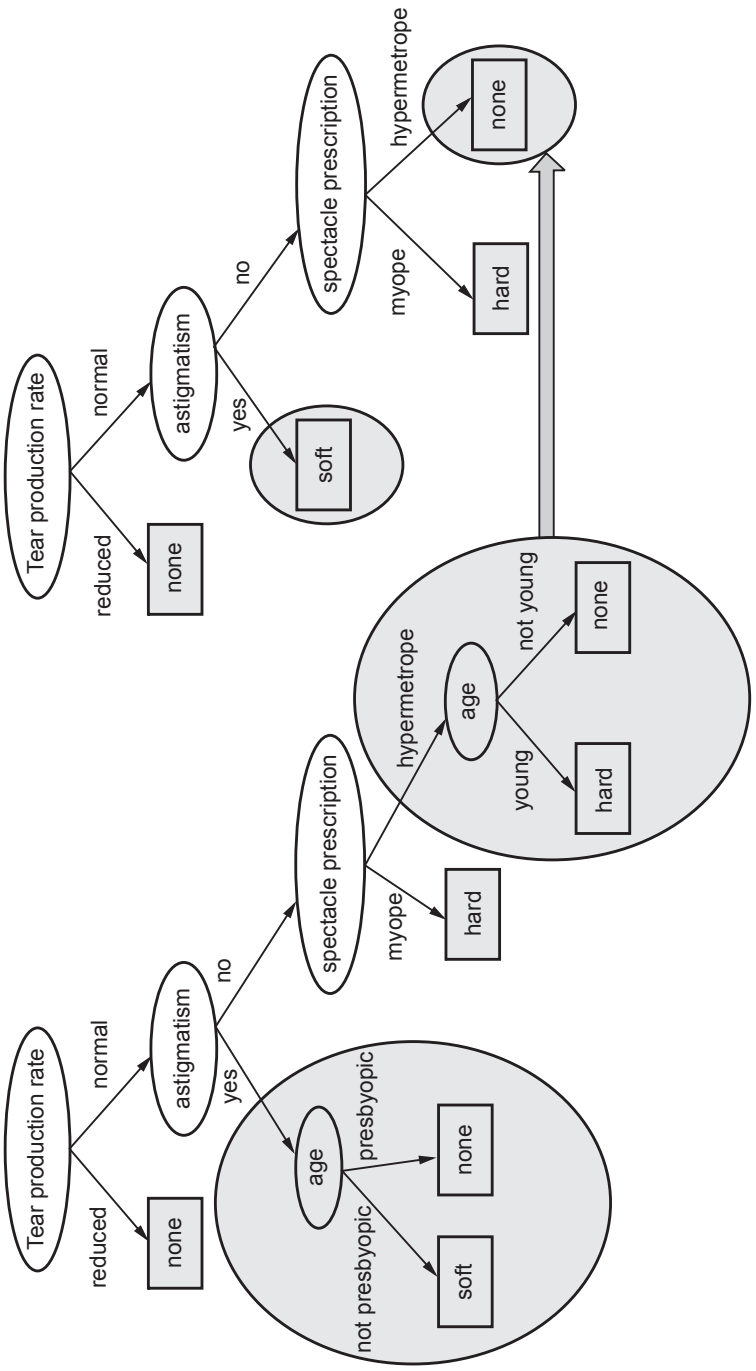

Tear production rate
normal
reduced
none
astigmatism
yes
no
soft
spectacle prescription
myope
hypermetrope
hard
none
Tear production rate
normal
reduced
none
astigmatism
yes
no
age
presbyopic
not presbyopic
none
soft
spectacle prescription
myope
hypermetrope
hard
age
not young
young
none
hard

Fig. 4.10.3

3. While the number of leaves in the tree exceeds the desired number :
   a) Find the twig with the least information gain
   b) Remove all child nodes of the twig.
   c) Relabel twig as a leaf.
   d) Update the leaf count

Example 4.10.1 *If S is a collection of 14 examples with 9 YES and 5 NO examples then calculate entropy.*

Solution :

$$\text{Entropy (S)} = \sum - p(I) \log_2 p(I)$$

Where p(I) is the proportion of S belonging to class I.

$\sum$ is over c.

$$\text{Entropy (S)} = -\left(\frac{9}{14}\right) \log 2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log\ 2 \left(\frac{5}{14}\right)$$

$$= -0.940$$

4.10.3 Tree Pruning

- If the classifier fits the training instances too closely, it may fit noisy instances and that reduces its usefulness. This phenomenon is called overfitting.
- Pruning simplifies a classifier by merging disjuncts that are adjacent in instance space. This can improve the classier's performance by eliminating error-prone components.
- Pruning of the decision tree is done by replacing a whole sub-tree by a leaf node. The replacement takes place if a decision rule establishes that the expected error rate in the sub-tree is greater than in the single leaf.
- For example :

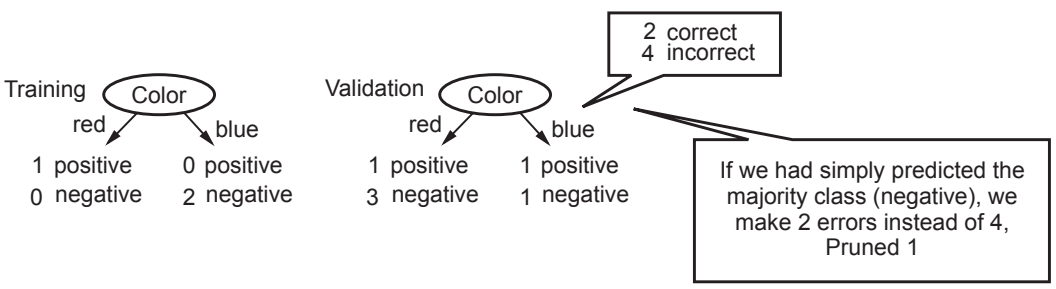


Fig. 4.10.4

4.10.4 ID3 Algorithm

- ID3 stands for Iterative Dichotomiser 3. This algorithm used to generate a decision tree. ID3 uses Entropy function and Information gain as metrics.
- The ID3 follows the Occam's razor principle. It attempts to create the smallest possible decision tree.
- The calculation for information gain is the most difficult part of this algorithm.
- ID3 performs a search whereby the search states are decision trees and the operator involves adding a node to an existing tree. It uses information gain to measure the attribute to put in each node, and performs a greedy search using this measure of worth.
- The algorithm goes as follows : Given a set of examples (S), categorised in categories c_i, then :
  1. Choose the root node to be the attribute, A, which scores the highest for information gain relative to S.
  2. For each value v that A can possibly take, draw a branch from the node.
  3. For each branch from A corresponding to value v, calculate S_v. Then :
     i. If S_v is empty, choose the category c default which contains the most examples from S, and put this as the leaf node category which ends that branch.
     ii. If S_v contains only examples from a category c, then put c as the leaf node category which ends that branch.
     iii. Otherwise, remove A from the set of attributes which can be put into nodes. Then put a new node in the decision tree, where the new attribute being tested in the node is the one which scores highest for information gain relative to S_v.

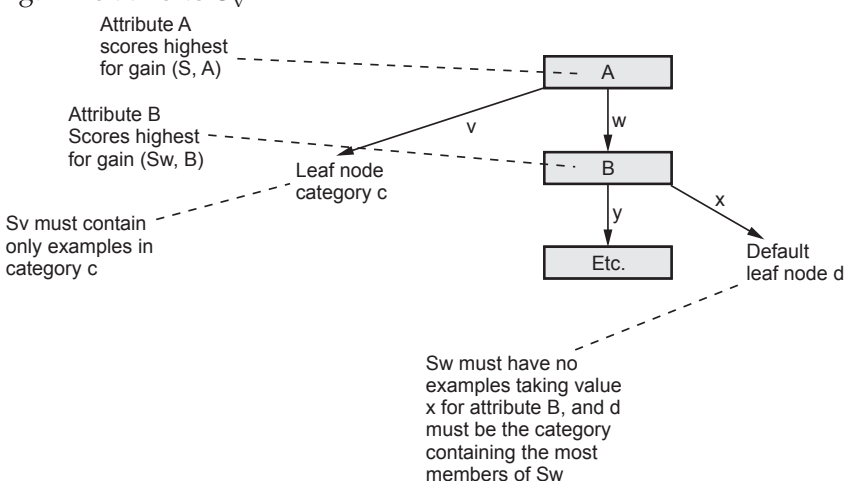


Fig. 4.10.5

- The ID3 algorithm is a classic data mining algorithm for classifying instances. The input is a set of training data for building a decision tree.
- By applying the ID3 algorithm, a decision tree is created. To create the decision tree, we have to choose a target attribute.
- Take all unused attributes and calculates their entropies. Chooses attribute that has the lowest entropy is minimum or when information gain is maximum. Makes a node containing that attribute.

Capabilities and limitations of ID3 :

1. Hypothesis space is a complete space of all discrete valued functions.
2. Cannot determine how many alternative trees are consistent with training data.
3. ID3 in its pure form performs no backtracking.
4. ID3 uses all training examples at each step to make statistically based decisions regarding how to refine its current hypothesis.

Example : Consider the following table :

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Calculate entropy and gain.

$$\text{Entropy (S)} = P_{\text{cinema}} \log_2(P_{\text{cinema}}) - P_{\text{tennis}} \log_2(P_{\text{tennis}}) - P_{\text{shopping}} \log_2(P_{\text{shopping}}) - P_{\text{stay\_in}} \log_2(P_{\text{stay-in}})$$

$$= -(6/10) * \log_2(6/10) - (2/10) * \log_2(2/10) - (1/10) * \log_2(1/10) - (1/10) * \log_2(1/10)$$

$$= -(6/10) * -0.737 - (2/10) * -2.322 - (1/10) * -3.322 - (1/10) * -3.322$$

$$= 0.4422 + 0.4644 + 0.3322 + 0.3322 = 1.571$$

and we need to determine the best of :

$$\text{Gain}(S, \text{weather}) = 1.571 - (|S_{sun}| / 10) * \text{Entropy}(S_{sun}) - (S_{wind}) / 10) * \text{Entropy}(S_{wind}) - (|S_{rain}| / 10) * \text{Entropy}(S_{rain})$$

$$= 1.571 - (0.3) * \text{Entropy}(S_{sun}) - (0.4) * \text{Entropy}(S_{wind}) - (0.3) * \text{Entropy}(S_{rain})$$

$$= 1.571 - (0.3) * (0.918) - (0.4) * (0.81125) - (0.3) * (0.918) = 0.70$$

$$\text{Gain}(S, \text{parents}) = 1571 - (|S_{yes}|) / 10) * \text{Entropy}(S_{yes}) - (S_{no} / 10) * \text{Entropy}(S_{no})$$

$$= 1.571 - (0.5) * 0 - (0.5) * 1.922 = 1.571 - 0.961 = 0.61$$

$$\text{Gain}(S, \text{money}) = 1571 - (|S_{rich}|) / 10) * \text{Entropy}(S_{rich}) - (S_{poor} / 10) * \text{Entropy}(S_{poor})$$

$$= 1.571 - (0.7) * (1.842) - (0.3) * 0 = 1.571 - 1.2894 = 0.2816$$

- This means that the first node in the decision tree will be the weather attribute. From the weather node, we draw a branch for the values that weather can take : sunny, windy and rainy :

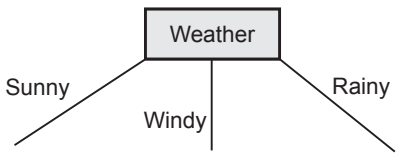


Fig. 4.10.6

- Now we look at the first branch. S_{sunny} = {W1, W2, W10}. This is not empty, so we do not put a default categorization leaf node here.
- The categorisations of W1, W2 and W10 are Cinema, Tennis and Tennis respectively. As these are not all the same, we cannot put a categorisation leaf node here. Hence we put an attribute node here, which we will leave blank for the time being.
- Looking at the second branch, S_{windy} = {W3, W7, W8, W9}. Again, this is not empty, and they do not all belong to the same class, so we put an attribute node here, left blank for now. The same situation happens with the third branch, hence our amended tree looks like this :

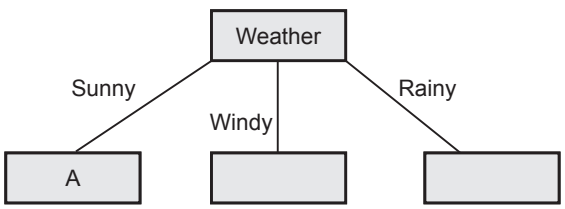


Fig. 4.10.7

- "In effect, we are interested only in this part of the table :

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

Hence we can calculate :

$$\text{Gain}(S_{\text{sunny}}, \text{parents}) = 0.918 - (|S_{\text{yes}}| / |S|) * \text{Entropy}(S_{\text{yes}}) - (|S_{\text{no}}| / |S|) * \text{Entropy}(S_{\text{no}})$$

$$= 0.918 - (1/3)* 0 - (2/3) * 0 = 0.918$$

$$\text{Gain}(S_{\text{sunny}}, \text{money}) = 0.918 - (|S_{\text{rich}}| / |S|) * \text{Entropy}(S_{\text{rich}}) - (|S_{\text{poor}}| / |S|) * \text{Entropy}(S_{\text{poor}})$$

$$= 0.918 - (3/3)* 0.918 - (0/3)*0 = 0.918 - 0.918 = 0$$

Rwview Questions

1. *What is decision tree ? Explain how decision tree is constructed using ID3 algorithm.*
   SPPU : Dec.-18 (End Sem), Marks 8
2. *Explain the following with their significance : i) Entropy ii) Information gain iii) Gain ratio*
   SPPU : Dec.-18 (End Sem), Marks 6
3. *Explain following decision tree algorithms : i) ID3 Algorithm ii) C4.5 iii) CART.*
   SPPU : May-19 (End Sem), Dec.-19 (End Sem),Marks 9
4. *What is decision tree ? Explain various terms used in decision tree ?*
   SPPU : Dec.-19 (End Sem), Marks 8

4.11 Introduction to Scikit-learn

- Scikit-learn is probably the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- In scikit-learn, an estimator for classification is a Python object that implements the methods fit(X, y) and predict(T).
- An example of an estimator is the class sklearn.svm. SVC, which implements support vector classification. The estimator's constructor takes as arguments the model's parameters.
- Scikit-learn comes loaded with a lot of features. Here are a few of them to help you understand :

1. **Supervised learning algorithms :** Think of any supervised learning algorithm you might have heard about and there is a very high chance that it is part of scikit-learn.
2. **Cross-validation :** There are various methods to check the accuracy of supervised models on unseen data.
3. **Unsupervised learning algorithms :** Again there is a large spread of algorithms in the offering - starting from clustering, factor analysis, principal component analysis to unsupervised neural networks.
4. **Various toy datasets :** This came in handy while learning scikit-learn. I had learnt SAS using various academic datasets (e.g. IRIS dataset, Boston House prices dataset). Having them handy while learning a new library helped a lot.
5. **Feature extraction :** Useful for extracting features from images and text (e.g. Bag of words).

4.11.1 Creating Training and Test Sets

- Machine learning is about learning some properties of a data set and applying them to new data. This is why a common practice in machine learning to evaluate an algorithm is to split the data at hand in two sets, one that we call a training set on which we learn data properties and one that we call a testing set, on which we test these properties.
- In training data, data are assign the labels. In test data, data labels are unknown but not given. The training data consist of a set of training examples.
- The real aim of supervised learning is to do well on test data that is not known during learning. Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy.
- The training error is the mean error over the training sample. The test error is the expected prediction error over an independent test sample.
- Problem is that training error is not a good estimator for test error. Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to over fitting and poor generalization.
- **Training set :** A set of examples used for learning, where the target value is known.
- **Test set :** It is used only to assess the performances of a classifier. It is never used during the training process so that the error on the test set provides an unbiased estimate of the generalization error.
- Training data is the knowledge about the data source which we use to construct the classifier.

4.11.2 Managing Categorical Data

- That categorical data is defined as variables with a finite set of label values. That most machine learning algorithms require numerical input and output variables. That an integer and one hot encoding is used to convert categorical data to integer data.
- Categorical data is very common in business datasets. For example, users are typically described by country, gender, age group etc., products are often described by product type, manufacturer, seller etc., and so on.
- Several regression and binary classification algorithms are available in scikit-learn. A simple way to extend these algorithms to the multi-class classification case is to use the so-called one-vs-all scheme.
- At learning time, this simply consists in learning one regressor or binary classifier per class. In doing so, one needs to convert multi-class labels to binary labels. LabelBinarizer makes this process easy with the transform method.
- At prediction time, one assigns the class for which the corresponding model gave the greatest confidence. LabelBinarizer makes this easy with the inverse_transform method.
- Example :

```
>>> from sklearn import preprocessing
>>> lb = preprocessing.LabelBinarizer()
>>> lb.fit([1, 2, 6, 4, 2])
LabelBinarizer(neg_label=0, pos_label=1, sparse_output=False)
>>> lb.classes_
array([1, 2, 4, 6])
>>> lb.transform([1, 6])
array([[1, 0, 0, 0],
       [0, 0, 0, 1]])
```

4.11.3 Managing Missing Features

- Sometimes a dataset can contain missing features, so there are a few options that can be taken into account :
  1. Removing the whole line.
  2. Creating sub-model to predict those features.
  3. Using an automatic strategy to input them according to the other known values.
- In real-world samples, it is not uncommon that there are missing one or more values such as the blank spaces in our data table.
- Quite a few computational tools, however, are unable to handle such missing values and might produce unpredictable results. So, before we proceed with further analyses, it is critical that we take care of those missing values.

- Mean imputation replaces missing values with the mean value of that feature/variable. Mean imputation is one of the most 'naive' imputation methods because unlike more complex methods like k-nearest neighbors imputation, it does not use the information we have about an observation to estimate a value for it.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import imputer
# Create an empty dataset
df = pd.DataFrame()
# Create two variables called x0 and x1. Make the first value of x1 a missing value
df['x0'] = [0.3051,0.4949,0.6974,0.3769,0.2231,0.341,0.4436,0.5897,0.6308,0.5]
df['x1'] = [np.nan,0.2654,0.2615,0.5846,0.4615,0.8308,0.4962,0.3269,0.5346,0.6731]
# View the dataset
df
```

	x0	x1
0	0.3051	NaN
1	0.4949	0.2654
2	0.6974	0.2615
3	0.3769	0.5846
4	0.2231	0.4615
5	0.3410	0.8308
6	0.4436	0.4962
7	0.5897	0.3269
8	0.6308	0.5346
9	0.5000	0.6731

Fit Imputer

```
# Create an imputer object that looks for 'Nan' values, then replaces them with the
  mean value of the feature by columns (axis=0)
  mean_imputer = Imputer(missing_values='NaN', strategy='mean', axis=0)
# Train the imputor on the df dataset
  mean_imputer = mean_imputer.fit(df)
```

Apply Imputer

```
# Apply the imputer to the df dataset
imputed_df = mean_imputer.transform(df.values)
```

View Data

```
# View the data
imputed_df
array([[ 0.3051    ,  0.49273333],
       [ 0.4949    ,  0.2654    ],
```

```
[ 0.6974    ,  0.2615    ],
[ 0.3769    ,  0.5846    ],
[ 0.2231    ,  0.4615    ],
[ 0.341     ,  0.8308    ],
[ 0.4436    ,  0.4962    ],
[ 0.5897    ,  0.3269    ],
[ 0.6308    ,  0.5346    ],
[ 0.5       ,  0.6731    ]])
```

Notice that 0.49273333 is the imputed value, replacing the np.NaN value.

4.11.4 Data Scaling and Normalization

- Generic dataset is made up of different values which can be drawn from different distributions, having different scales. Machine learning algorithm is not naturally able to distinguish among these various situations. For this reason it is always preferable to standardize datasets before processing them.
- Standardization : To transform data so that it has zero mean and unit variance. Also called scaling.
- Use function **sklearn.preprocessing.scale()**
- Parameters :

  X : Data to be scaled

  with_mean : Boolean. Whether to center the data (make zero mean)

  with_std : Boolean (whether to make unit standard deviation
- Normalization : To transform data so that it is scaled to the [0,1] range.
- Use function **sklearn.preprocessing.normalize()**
- Parameters :

  X : Data to be normalized

  norm : which norm to use : l1 or l2

  axis : whether to normalize by row or column
- Normalizing in scikit-learn refers to rescaling each observation (row) to have a length of 1.
- This preprocessing can be useful for sparse datasets (lots of zeros) with attributes of varying scales when using algorithms that weight input values such as neural networks and algorithms that use distance measures such as K-Nearest Neighbors.
- You can normalize data in Python with scikit-learn using the Normalizer class.
- Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using the transform method.

- Standardization of a dataset is a common requirement for many machine learning estimators : they might behave badly if the individual features do not more or less look like standard normally distributed data.
- **Examples :**

```
>>>Hide the prompts and output>>> from sklearn.preprocessing import
StandardScaler
>>> data = [[0, 0], [0, 0], [1, 1], [1, 1]]
>>> scaler = StandardScaler()
>>> print(scaler.fit(data))
StandardScaler(copy=True, with_mean=True, with_std=True)
>>> print(scaler.mean_)
[0.5 0.5]
>>> print(scaler.transform(data))
[ [-1. -1.]
[-1. -1.]
[ 1.  1.]
[ 1.  1.]]
>>> print(scaler.transform([[2, 2]]))
[[3. 3.]]
```

4.11.5 Feature Selection and Filtering

- Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.
- Feature selection is a critical step in the feature construction process. In text categorization problems, some words simply do not appear very often. Perhaps the word "groovy" appears in exactly one training document, which is positive. Is it really worth keeping this word around as a feature ? It's a dangerous endeavor because it's hard to tell with just one training example if it is really correlated with the positive class, or is it just noise. You could hope that your learning algorithm is smart enough to figure it out. Or you could just remove it.
- There are three general classes of feature selection algorithms : Filter methods, wrapper methods and embedded methods.
- The role of feature selection in machine learning is :
  1. To reduce the dimensionality of feature space
  2. To speed up a learning algorithm
  3. To improve the predictive accuracy of a classification algorithm
  4. To improve the comprehensibility of the learning results

- Features Selection Algorithms are as follows :
  1. **Instancebased approaches :** There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.
  2. **Nondeterministic approaches :** Genetic algorithms and simulated annealing are also used in feature selection.
  3. **Exhaustive complete approaches :** Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.
- Two examples of feature selection that use the classes SelectKBest and SelectPercentile.
- The classes in the sklearn.feature_selection module can be used for feature selection/dimensionality reduction on sample sets, either to improve estimators' accuracy scores or to boost their performance on very high-dimensional datasets.
- VarianceThreshold is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples.
- As an example, suppose that we have a dataset with boolean features, and we want to remove all features that are either one or zero (on or off) in more than 80% of the samples. Boolean features are Bernoulli random variables, and the variance of such variables is given by **Var[X] = p(1 - p)**.

```
>>> from sklearn.feature_selection import VarianceThreshold
>>> X = [[0, 0, 1], [0, 1, 0], [1, 0, 0], [0, 1, 1], [0, 1, 0], [0, 1, 1]]
>>> sel = VarianceThreshold(threshold=(.8 * (1 - .8)))
>>> sel.fit_transform(X)
array([[0, 1],
       [1, 0],
       [0, 0],
       [1, 1],
       [1, 0],
       [1, 1]])
```

- Univariate feature selection works by selecting the best features based on univariate statistical tests. It can be seen as a preprocessing step to an estimator. Scikit-learn exposes feature selection routines as objects that implement the transform method :
  1. SelectKBest removes all but the highest scoring features.
  2. SelectPercentile removes all but a user-specified highest scoring percentage of features.

3. Using common univariate statistical tests for each feature : false positive rate SelectFpr, false discovery rate SelectFdr, or family wise error SelectFwe.
4. GenericUnivariateSelect allows to perform univariate feature selection with a configurable strategy. This allows to select the best univariate selection strategy with hyper-parameter search estimator.

4.11.6 Mat plotlib

- Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.
- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- Matplotlib is a plotting library for the Python programming language. It allows to make quality charts in few lines of code. Most of the other python plotting library are build on top of Matplotlib.
- The library is currently limited to 2D output, but it still provides you with the means to express graphically the data patterns.
- Matplotlib plots can be saved as image files using the plt.savefig() function.
- The .savefig() method requires a filename be specified as the first argument. This filename can be a full path. It can also include a particular file extension if desired. If no extension is provided, the configuration value of savefig.format is used instead.
- The .savefig() also has a number of useful optional arguments :
  1. dpi can be used to set the resolution of the file to a numeric value.
  2. transparent can be set to True, which causes the background of the chart to be transparent.
  3. bbox_inches can be set to alter the size of the bounding box (whitespace) around the output image. In most cases, if no bounding box is desired using bbox_inches = 'tight is ideal.
  4. If bbox_inches is set to 'tight', then the pad_inches option sepcifies the amount of padding around the image.
- Line styles help differentiate graphs by drawing the lines in various ways. Following line style is used by Matplotlib.

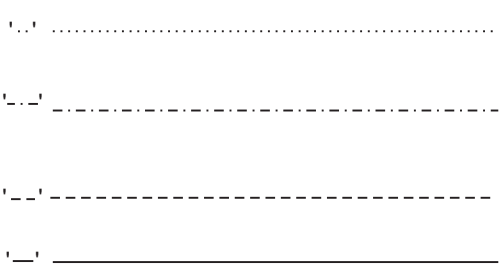


Fig. 4.11.1

- Matplotlib has an additional paramter to control the colour and style of the plot.

```
plt.plot(xa, ya 'g')
```

- This will make the line green. You can use any colour of red, green, blue, cyan, mangenta, yellow, white or black just by using the first character of the colour name in lower case (use"k" for black, as "b" means blue).
- You can also alter the lifestyle, for example two dashes -- makes a dashed line. This can be used added to the colour selector, like this :

```
plt.plot(xa,ya, 'r--')
```

- You can use "_" for a solid line (the default), "_" for dash-dot lines, or ":" for a dotted line. Here is an example.

```
from matplotlib import pyplot as plt
import numpy as np

xa = np.linspace(0, 5, 20)

ya = xa **2
plt.plot (xa, ya, 'g')

ya = 3*xa
plt.plot (xa, ya, 'r--')

plt.show()
```

Output

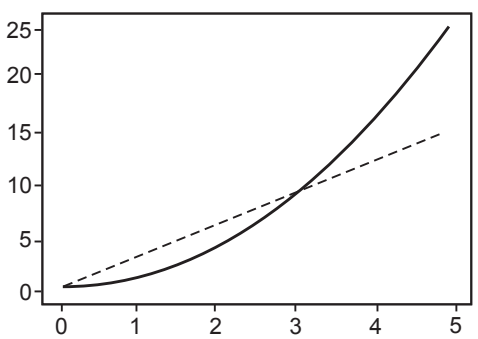


Fig. 4.11.2

- MatPlotLib colors are as follows :

Color	Character
Black	'k'
Yellow	'y'
Cyan	'c'
Blue	'b'
Green	'g'
Red	'r'
White	'w'
Magenta	'm'

4.12 Regression and Classification using Scikit-learn

- In scikit-learn, an estimator for classification is a Python object that implements the methods fit(X, y) and predict(T).
- If the prediction task is to classify the observations in a set of finite labels, in other words to "name" the objects observed, the task is said to be a classification task. On the other hand, if the goal is to predict a continuous target variable, it is said to be a regression task. When doing classification in scikit-learn, y is a vector of integers or strings.
- scikit-learn comes with a few standard datasets, for instance the iris and digits datasets for classification and the diabetes dataset for regression.
- In the following, we start a Python interpreter from our shell and then load the iris and digits datasets. Our notational convention is that $ denotes the shell prompt while >>> denotes the Python interpreter prompt :

```
$ python
>>> from sklearn import datasets
>>> iris = datasets.load_iris()
>>> digits = datasets.load_digits()
```

Classifying irises :

- The iris dataset is a classification task consisting in identifying 3 different types of irises (Setosa, Versicolour, and Virginica) from their petal and sepal length and width :

```
>>>
>>> import numpy as np
>>> from sklearn import datasets
>>> iris_X, iris_y = datasets.load_iris(return_X_y=True)
>>> np.unique(iris_y)
array([0, 1, 2])
```

```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from sklearn import datasets
from sklearn.decomposition import PCA

# import some data to play with
iris = datasets.load_iris()
X = iris.data[:, :2]  # we only take the first two features.
y = iris.target

x_min, x_max = X[:, 0].min() - 0.5, X[:, 0].max() + 0.5
y_min, y_max = X[:, 1].min() - 0.5, X[:, 1].max() + 0.5

plt.figure(2, figsize=(8, 6))
plt.clf()
```

Plot the training points :

```
plt.scatter(X[:, 0], X[:, 1], c=y, cmap=plt.cm.Set1, edgecolor="k")
plt.xlabel("Sepal length")
plt.ylabel("Sepal width")

plt.xlim(x_min, x_max)
plt.ylim(y_min, y_max)
plt.xticks(())
plt.yticks(())
# To getter a better understanding of interaction of the dimensions
# plot the first three PCA dimensions
fig = plt.figure(1, figsize=(8, 6))
ax = Axes3D(fig, elev=-150, azim=110)
X_reduced = PCA(n_components=3).fit_transform(iris.data)
ax.scatter(
        X_reduced[:, 0],
```

```
            X_reduced[:, 1],
            X_reduced[:, 2],
            c=y,
            cmap=plt.cm.Set1,
            edgecolor="k",
            s=40,
    )
    ax.set_title("First three PCA directions")
    ax.set_xlabel("1st eigenvector")
    ax.w_xaxis.set_ticklabels([])
    ax.set_ylabel("2nd eigenvector")
    ax.w_yaxis.set_ticklabels([])
    ax.set_zlabel("3rd eigenvector")
    ax.w_zaxis.set_ticklabels([])
    plt.show()
```

4.13 Multiple Choice Questions

Q.1 What are closed itemsets ?

a An itemset for which at least one proper super - itemset has same support.

b An itemset whose no proper super - itemset has same support.

c An itemset for which at least super - itemset has same confidence.

d An itemset whose no proper super - itemset has same confidence.

Q.2 What are maximal frequent itemsets ?

a A frequent itemset whose no super - itemset is frequent.

b A frequent itemset whose super - itemset is also frequent.

c A non - frequent itemset whose super - itemset is frequent.

d None of the above.

Q.3 What does FP growth algorithm do ?

a It mines all frequent patterns through pruning rules with lesser support.

b It mines all frequent patterns through pruning rules with higher support.

c It mines all frequent patterns by constructing a FP tree.

d All of the above.

Q.4 What is the relation between candidate and frequent itemsets ?

a A candidate itemset is always a frequent itemset.

b A frequent itemset must be a candidate itemset.

c No relation between the two.

d Both are same.

Q.5 Which of these is not a frequent pattern mining algorithm ?

a Apriori b FP growth

c Decision trees d Eclat

Q.6 What will happen if support is reduced ?

a Number of frequent itemsets remains same.

b Some itemsets will add to the current set of frequent itemsets.

c Some itemsets will become infrequent while others will become frequent.

d Can not say.

Q.7 What are maximal frequent itemsets ?

a A frequent itemset whose no super - itemset is frequent.

b A frequent itemset whose super - itemset is also frequent.

c A non - frequent itemset whose super - itemset is frequent.

d None of the above.

Q.8 The individual tuples making up the training set are referred to as _____ and are selected from the database under analysis.

a learning tuples b training tuples

c samples d database

Q.9 A _____ is a flowchart - like tree structure, where each internal node denotes a test on an attribute.

a decision tree b binary tree

c cluster d none of these

Q.10 ID3 stands for _____.

a Induction Decision tree b Iterative Database

c Iterative Dichotomiser d Iterative Decision tree

Q.11 ID3 uses _____ as its attribute selection measure.

a decision tree b Gini index

c information gain d attributes

Q.12 Attribute selection measures based on the _____ principle have the least bias toward multi - valued attribute.

a maximum description length

b minimum description length

c minimum distance length

d maximum distance length

Q.13 In theory, Bayesian classifiers have the _____ error rate in comparison to all other classifiers.

a equal b maximum

c minimum d zero

Q.14 A _____ is a flowchart - like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

a binary tree b decision tree

c classification d none

Q.15 In multiple regression, we use the _____ method to determine the best coefficients to attain good model fit.

a maximum likelihood b ordinary least square

c binary d decision tree

Q.16 In logistic regression, we use _____ method to determine the best coefficients and eventually a good model fit.

a maximum likelihood b ordinary least square

c binary d decision tree

Q.17 Linear regression gives a _____ output.

a constant b variable

c continuous d none

Q.18 Logistic regression assumes that the dependent variable follows a _____.

a binomial distribution b sigmoid function

c gradient descent d all of these

Q.19 Unlike multiple regression, logistic regression _____.

a does not have b weights

b does not have an indenpendent variable

c does not have a dependent variable

d does not have a score for the dependent variable

Q.20 Logistic regression is used when you want to _____

a predict a dichotomous variable from continuous or dichotomous variables

b predict a continuous variable from dichotomus variables

c predict any categorical variable from several other categorical variables.

d predict a continuous variable from dichotomous continuous variables

Answer Keys for Multiple Choice Questions :

Q.1	b	**Q.2**	a
Q.3	c	**Q.4**	b
Q.5	c	**Q.6**	b
Q.7	a	**Q.8**	b
Q.9	a	**Q.10**	c
Q.11	c	**Q.12**	b
Q.13	c	**Q.14**	b
Q.15	b	**Q.16**	a
Q.17	c	**Q.18**	b
Q.19	d	**Q.20**	a

❑❑❑

Notes

UNIT V

5 Big Data Analytics and Model Evaluation

Syllabus

Clustering Algorithms : *K-Means, Hierarchical Clustering, Time-series analysis.* ***Introduction to Text Analysis :*** *Text-preprocessing, Bag of words, TF-IDF and topics. Need and Introduction to social network analysis, Introduction to business analysis.* ***Model Evaluation and Selection :*** *Metrics for Evaluating Classifier Performance, Holdout Method and Random Sub sampling, Parameter Tuning and Optimization, Result Interpretation, Clustering and Time-series analysis using Scikit-learn, sklearn. metrics, Confusion matrix, AUC-ROC Curves, Elbow plot.*

Contents

5.1 Clustering Algorithms

SPPU : Aug.-18, Oct.-19, Dec.-18, 19

What is cluster analysis ?

- Given a set of objects, place them in groups such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups.
- Cluster analysis can be a powerful data - mining tool for any organisation that needs to identify discrete groups of customers, sales transactions or other types of behaviors and things. For example, insurance providers use cluster analysis to detect fraudulent claims, and banks use it for credit scoring.
- Cluster analysis uses mathematical models to discover groups of similar customers based on the smallest variations among customers within each group.
- Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.
- Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user understand the natural grouping or structure in a data set.
- Various types of clustering methods are partitioning methods, hierarchical clustering, Fuzzy clustering, Density - based clustering and Model - based clustering.
- Cluster analysis is process of grouping a set of data - objects into clusters.
- **Desirable properties of a clustering algorithm are as follows :**
  1. Scalability (in terms of both time and space)
  2. Ability to deal with different data types
  3. Minimal requirements for domain knowledge to determine input parameters.
  4. Interpretability and usability.
- Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important unsupervised learning problem.
- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Fig. 5.1.1 shows cluster.

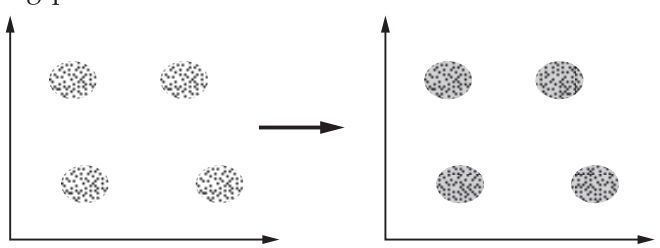
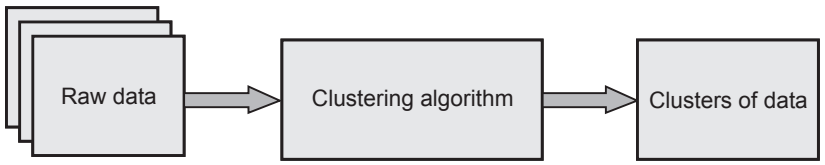
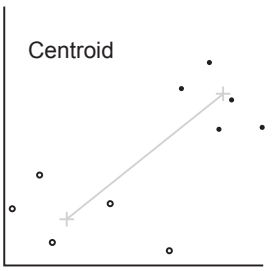


Fig. 5.1.1 Cluster

- In this case we easily identify the 4 clusters into which the data can be divided ; the similarity criterion is distance : Two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called **distance-based clustering**.



- Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.



- A clustering algorithm attempts to find natural groups of components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets.
- To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.
- **Cluster centroid :** The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Each cluster has a well defined centroid.
- **Distance :** The distance between two points is taken as a common metric to as see the similarity among the components of a population. The commonly used distance measure is the euclidean metric which defines the distance between two points $p = (p_1, p_2, ...)$ and $q = (q_1, q_2, ...)$ is given by :

$$d = \sum_{i=1}^{k} (p_i - q_i)^2$$

- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering ? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
- Clustering analysis helps construct meaningful partitioning of a large set of objects. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, etc.

- Clustering algorithms may be classified as listed below :
  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilistic clustering
- A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Examples of Clustering Applications

1. **Marketing :** Help marketers discover distinct groups in their customer bases and then use this knowledge to develop targeted marketing programs.
2. **Land use :** Identification of areas of similar land use in an earth observation database.
3. **Insurance :** Identifying groups of motor insurance policy holders with a high average claim cost.
4. **Urban planning :** Identifying groups of houses according to their house type, value, and geographical location.
5. **Seismology :** Observed earth quake epicenters should be clustered along continent faults.

5.1.1 Typical Requirements of Clustering in Data Mining

1. **Scalability :** Many clustering algorithms work well on small data sets.
2. **Ability to deal with different types of attributes :** Many algorithms are designed to cluster interval-based data.
3. **Discovery of clusters with arbitrary shape :** Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures.
4. **Minimal requirements for domain knowledge to determine input parameters :** Many clustering algorithms require users to input certain parameters in cluster analysis.
5. **Ability to deal with noisy data :** Most real - world databases contain outliers or missing, unknown or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
6. **Incremental clustering and insensitivity to the order of input records :** Some clustering algorithms cannot incorporate newly inserted data into existing clustering structures.

7. **High dimensionality :** A database or a data warehouse can contain several dimensions or attributes.
8. **Constraint-based clustering :** Real-world applications may need to perform clustering under various kinds of constraints.
9. **Interpretability and usability :** Users expect clustering results to be interpretable, comprehensible and usable.

5.1.2 Problems with Clustering

1. Current clustering techniques do not address all the requirements adequately ;
2. Dealing with large number of dimensions and large number of data items can be problematic because of time complexity ;
3. The effectiveness of the method depends on the definition of "distance" ;
4. If an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces ;
5. The result of the clustering algorithm can be interpreted in different ways.

5.1.3 Types of Clusters

- Type of clusters are as follows :
  a) Well - separated clusters
  b) Prototype - based clusters
  c) Contiguity - based clusters
  d) Density - based clusters

a) Well - separated clusters :

- A cluster is a set of points such that any point in a cluster is closer to every other point in the cluster than to any point not in the cluster.
- Fig. 5.1.2 shows well-separated cluster.

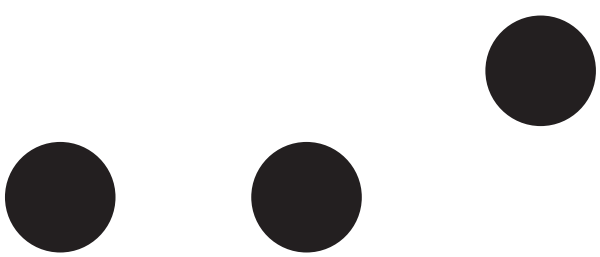


Fig. 5.1.2 Well-separated cluster

- Sometimes a threshold is used to specify that all the objects in a cluster must sufficiently close to one another. Definition of a cluster is satisfied only when the data contains natural clusters.

b) Prototype - based cluster

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the prototype or "center" of a cluster, than to the center of any other cluster. Prototype based clusters can also be referred to as "Center-Based" Clusters.
- The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster. Fig. 5.1.3 shows 4 center-based clusters.

Fig. 5.1.3 4 Center-based clusters

- If the data is numerical, the prototype of the cluster is often a centroid i.e., the average of all the points in the cluster.
- If the data has categorical attributes, the prototype of the cluster is often a medoid i.e., the most representative point of the cluster.
- Objects in the cluster are closer to the prototype of the cluster than to the prototype of any other cluster.
- K-Means and K-Medoids are the examples of Prototype-based Clustering algorithm.

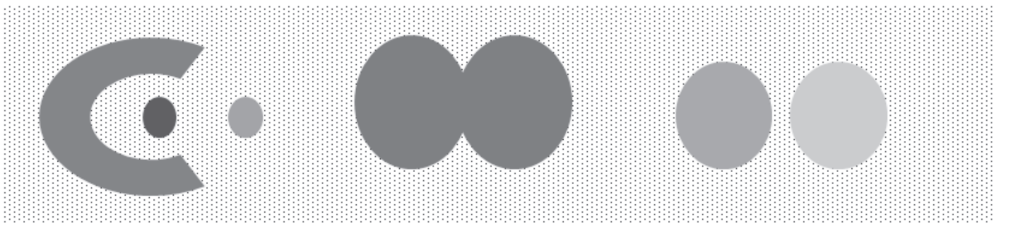
c) Contiguity - based clusters

- A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

Fig. 5.1.4 8 contiguous clusters

d) Density - based

- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.



5.1.4 Desired Features of Cluster Analysis

- Features are as follows :

1. **Scalability :** Data-mining problems can be large and therefore a cluster-analysis method should be able to deal with large problems gracefully. Ideally, performance should be linear with data-size.
2. **Only one scan of the dataset :** For large problems, data must be stored on disk, so cost of I/O disk can become significant in solving the problem.
3. **Ability to stop and resume :** For large dataset, cluster-analysis may require huge processor-time to complete the task. In such cases, the task should be able to be stopped and then resumed when convenient.
4. **Minimal input parameters :** The method should not expect too much guidance from the data-mining analyst.
5. **Robustness :** Most data obtained from a variety of sources has errors. Therefore, the method should be able to deal with noise, outlier and missing values gracefully.
6. **Ability to discover different cluster-shapes :** Clusters appear in different shapes and not all clusters are spherical. So the method should be able to discover cluster-shapes other than spherical.
7. **Different data types :** Many problems have a mixture of data types, for e.g. numerical, categorical and even textual. Therefore, the method should be able to deal with numerical, boolean and categorical data.
8. **Result independent of data input order :** The method should not be sensitive to data input-order.

5.1.5 K-Means

- K-Means clustering is heuristic method. Here each cluster is represented by the center of the cluster. "K" stands for number of clusters, it is typically a user input to the algorithm ; some criteria can be used to automatically estimate K.
- This method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart.
- Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.
- Given K, the K-means algorithm consists of four steps :
  1. Select initial centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.
- The x_1, ..., x_N are data points or vectors of observations. Each observation (vector x_i) will be assigned to one and only one cluster. The C(i) denotes cluster number for the observation. K-means minimizes within-cluster point scatter :

$$W(C) = \frac{1}{2} \sum_{K-1}^{K} \sum_{C(i)-K} \sum_{C(i)-K} ||x_i - x_j||^2$$

$$= \sum_{K-1}^{K} Nk \sum_{C(i)-K} ||x_i - m_K||^2$$

Where,

m_K is the mean vector of the K^{th} cluster.

N_K is the number of observations in K^{th} cluster.

K-means algorithm properties

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

The K-means algorithm process

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point,
   a. Calculate the distance from the data point to each cluster.
   b. If the data point is closest to its own cluster, leave it where it is.
   c. If the data point is not closest to its own cluster, move it into the closest cluster.
3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

- K-means algorithm is iterative in nature. It converges, however only a local minimum is obtained. It works only for numerical data. This method easy to implement.

Advantages of K-means algorithm :

1. Efficient in computation
2. Easy to implement

Weaknesses :

1. Applicable only when mean is defined.
2. Need to specify K, the number of clusters, in advance.
3. Trouble with noisy data and outliers.
4. Not suitable to discover clusters with non-convex shapes.

5.1.6 Hierarchical Clustering

- This method use distance matrix as clustering criteria. This method does not require the number of clusters K as an input, but needs a termination condition. Hierarchical clustering is a widely used data analysis tool.
- The idea is to build a binary tree of the data that successively merges similar groups of points. Visualizing this tree provides a useful summary of the data.

- Hierarchical clustering arranges items in a hierarchy with a tree like structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree - structured graph called a **dendrogram**.
- Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom - up (merging) or top - down (splitting) fashion.

Agglomerative hierarchical clustering

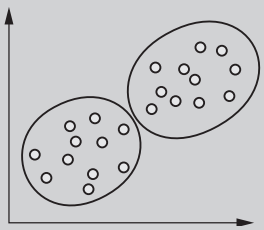
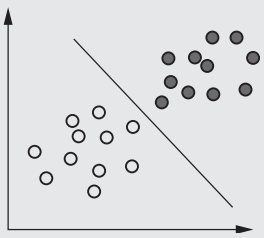
- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category.
- Initially, AGNES places each objects into a cluster of its own. The clusters are then merged step-by-step according to some criterion. For example, cluster C_1 and C_2 may be merged if an object in C_1 and object in C_2 form the minimum euclidean distance between any two objects from different clusters.
- In the agglomerative hierarchical approach, we start by defining each data point to be a cluster and combine existing clusters at each step. Here are four different methods for doing this :
  1. **Single linkage :** Smallest pairwise distance between elements from each cluster.
  2. **Complete linkage :** Largest distance between elements from each cluster.
  3. **Average linkage :** The average distance between elements from each cluster.
  4. **Centroid linkage :** Distance between cluster means.

Divisive hierarchical clustering

- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the clusters into smaller and smaller pieces, until each object form a cluster on its own or until it satisfies certain termination conditions, such as a desired number of cluster or the diameter of each cluster is within a certain threshold.

Agglomerative	Divisive hierarchical clustering
Initially each item in its own cluster.	Initially all items in one cluster.
Iteratively clusters are merged together.	Large clusters are successively divided.
Bottom up.	Top down.

5.1.7 Difference between Clustering vs Classification

Sr. No.	Clustering	Classification
1.	This function maps the data into one of several clusters which is the grouping of data items based on the similarities between them.	This model function classifies the data into one of several predefined categorical classes.
2.	Involved in unsupervised learning	Involved in supervised learning
3.	Training sample is not provided.	Training sample is provided.
4.	The number of cluster is not known before clustering, These are identified after the completion of clustering.	The number of classes is known before classification as there is predefined output based on input data.
5.	Data is not labeled.	Labeled data points.
6.	Asks how can I group this set of items ?	Asks what class does this item belong to ?
7.	Unknown number of classes	Known number of classes
8.	Used to understand data	Used to classify future observations
9.		

Review Questions

1. *What is clustering ? Explain k - means clustering algorithm with the use cases.* **SPPU : Aug.-18 (In Sem), Marks 6**
2. *Explain k-means clustering algorithm. What are its drawbacks ?* **SPPU : Dec.-18 (End Sem), Marks 7**
3. *Explain k-means algorithm.* **SPPU : Oct.-19 (In Sem), Marks 5**
4. *How k - means algorithm works ?* **SPPU : Dec.-19 (End Sem), Marks 5**

5.2 Time-Series Analysis

- Time series is a sequence of data points in chronological sequence, most often gathered in regular intervals. Time series analysis can be applied to any variable that changes over time and generally speaking, usually data points that are closer together are more similar than those further apart.

- Time series analysis is the way of studying the characteristics of the response variable with respect to time, as the independent variable. To estimate the target variable in the name of predicting or forecasting, use the time variable as the point of reference.
- Components of time series analysis are trend, seasonality, cyclical and irregularity.
  a) Trend : In which there is no fixed interval and any divergence within the given dataset is a continuous timeline. The trend would be negative or positive or null trend.

  b) Seasonality : In which regular or fixed interval shifts within the dataset in a continuous timeline. Would be bell curve or saw tooth.

  c) Cyclical : In which there is no fixed interval, uncertainty in movement and its pattern.

  d) Irregularity : Unexpected situations/events/scenarios and spikes in a short time span.
- Identifying seasonality in time series data is important for the development of a useful time series model.
- There are many tools that are useful for detecting seasonality in time series data :
  1. Background theory and knowledge of the data can provide insight into the presence and frequency of seasonality.
  2. Time series plots such as the seasonal subseries plot, the autocorrelation plot, or a spectral plot can help identify obvious seasonal trends in data.
  3. Statistical analysis and tests, such as the autocorrelation function, periodograms, or power spectrums can be used to identify the presence of seasonality.
- Time series data are data points collected over a period of time as a sequence of time gap. Time series data analysis means analyzing the available data to find out the pattern or trend in the data to predict some future values which will, in turn, help more effective and optimize business decisions.
- A forecast "error" is the difference between an observed value and its forecast.
- Time series analysis can be classified as :
  1. Parametric and Non-parametric
  2. Linear and Non-linear and
  3. Univariate and Multivariate
- Techniques used for time series analysis :
  1. ARIMA models
  2. Box-Jenkins multivariate models
  3. Holt winters exponential smoothing (single, double and triple)

5.2.1 ARIMA

- ARIMA stands for AutoRegressive Integrated Moving Average. It is a forecasting technique that projects the future values of a series based entirely on its own inertia.
- Its main application is in the area of short term forecasting requiring at least 40 historical data points. It works best when data exhibits a stable or consistent pattern over time with a minimum amount of outliers.
- Parameters of the ARIMA model :
  1. **p (lag order) :** Number of lag observations included in the model.
  2. **d (degree of differencing) :** Number of times that the raw observations are differenced.
  3. **q (order of moving average) :** Size of the moving average window.

Assumptions of ARIMA model :

1. **Data should be stationary :** By stationary it means that the properties of the series doesn't depend on the time when it is captured. A white noise series and series with cyclic behaviour can also be considered as stationary series.
2. **Data should be univariate :** ARIMA works on a single variable. Auto-regression is all about regression with the past values.

- **AR (Autoregression) :** A model that uses the dependent relationship between an observation and some number of lagged observations. 'p' is used as a notation to define this component, called the lag order.
- **I (Integrated) :** The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary. 'd' is used as a notation to define this component, called the degree of differencing.
- **MA (Moving Average) :** A model that uses the dependency between an observation and a residual error from a moving average model applied to lag observations. 'q' is used as a notation to define this component, called the order of moving average.

ARIMA forecasting equation :

- Let Y denote the original series and y denote the differenced (stationaries) series.

No difference (d = 0) : $y_t = Y_t$

First difference (d = 1) : $y_t = Y_t - Y_{t-1}$

Second difference (d = 2) :

$$y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$$

$$= Y_t - 2Y_{t-1} + Y_{t-2}$$

- Basic model for ARIMA is as follows :

1. ARIMA (1, 0, 0) : First-order autoregressive model-if the series is stationary and auto correlated, perhaps it can be predicted as a multiple of its own previous value, plus a constant.

$$y_t = \beta_0 + \beta_1 Y_{t-1}$$

2. ARIMA (2, 0, 0) : The series can be said to be dependent on its previous two values.

$$y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$$

3. ARIMA (0, 1, 0) : Random walk : If the series Y is not stationary, the simplest possible model for it is a random walk model, which can be considered as a limiting case of an AR(1) model in whict the auto-regressive coefficient is equal to 1, i.e. series with infinitely slow mean reversion.

$$y_t - y_{t-1} = \mu$$

4. ARIMA (1, 1, 0) : Differenced first-order autoregressive model : If the errors of a random walk model are auto correlated, perhaps the problem can be fixed by adding one lag of the depedent variable to the prediction equation i.e., by regressing the first difference of Y on itself lagged by one period. This would yield the following prediction equation.

$$y_t - y_{t-1} = \mu + \beta_1 (Y_{t-1} - Y_{t-2})$$

5. ARIMA (0,1,1) : Without constant moving average model and can expressed as -

$$y_t - y_{t-1} = \phi_1 (\alpha_{t-1} - \alpha_{t-2})$$

5.2.2 STL Approach

- STL is an acronym for "Seasonal and Trend decomposition using Loess.
- Seasonal trend decomposition using loess (STL) is a method for deomposing a time series into seasonal, trend and remainder components using local regression techniques.
- The goal of STL is to decompose a time series, y_i, $i = 1,\ldots,n$ into seasonal, trend and remainder components s_i, t_i, r_i respectively.

$$y_i = s_i + t_i + r_i$$

- STL works by iterating through smoothing of the seasonal and trend components. The seasonal component is defined based on the periodicitiy of the time series and how many observations there are per period.
- All smoothing operations done in STL. STL is a procedure for regular time series, so the design points of the smoothing operation are equally-spaced, and without loss of generality can thought of as 1, …, n.
- For time series smoothing, q is always an odd integer. Each smoothing operation in STL requires smoothing parameters to be specified, but may of them have nice defaults.

5.3 Introduction to Text Analysis

- Text Analysis (TA) aims to extract machine-readable information from unstructured text in order to enable data-driven approaches towards managing content. The purpose of Text Analysis is to create structured data out of free text content.
- Text analysis, also known as text mining, is the process of automatically classifying and extracting meaningful information from unstructured text.
- Text mining is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from textual document repositories.
- Text mining can be visualized as consisting of two phases : Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.
- A process of text mining involves a series of activities to be performed to mine the information. These activities are :

  1. Text Pre-processing : It involves a series of steps as shown in below :

  i. Text Clean-up : Text Clean-up means removing any unnecessary or unwanted information. Such as remove ads from web pages, normalize text converted from binary formats.

  ii. Tokenization : Tokenizing is simply achieved by splitting the text into white spaces.

  iii. Part of Speech Tagging : Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text. Taggers have to cope with unknown words (OOV problem) and ambiguous word-tag mappings.

  2. Text Transformation (Attribute Generation) : A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are : Bag of words and Vector Space.

3. **Feature Selection (Attribute Selection) :** Feature selection also is known as variable selection. It is the process of selecting a subset of important features for use in model creation. Redundant features are the one which provides no extra information. Irrelevant features provide no useful or relevant information in any context.
4. **Data Mining :** At this point, the text mining process merges with the traditional process. Classic data mining techniques are used in the structured database. Also, it resulted from the previous stages.
5. **Evaluate :** Evaluate the result, after evaluation, the result can be discarded.

- The five fundamental steps involved in text mining are :
  1. Gathering unstructured data from multiple data sources like plain text, web pages, pdf files, emails, and blogs, to name a few.
  2. Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing allows user to extract and retain the valuable information hidden within the data and to help identify the roots of specific words.
  3. Convert all the relevant information extracted from unstructured data into structured formats.
  4. Analyse the patterns within the data via Management Information System (MIS).
  5. Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organization.

5.3.1 Use of a Text Mining Tool

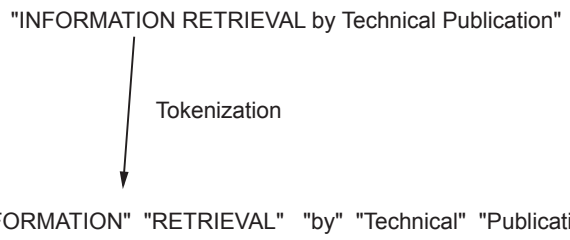
1. **Text analytics :** Involves extracting useful information and patterns from text. Most tools provide this feature.
2. **Text processing :** Involves transforming and manipulating unstructured text so that analysis methods can be applied to it.
3. **Classification/Categorization :** Many tools are used for classification and categorization of text/documents.
4. **Sentiment analysis :** Is used to identify subjective information from text. Many tools provide for sentiment analysis also called as opinion mining.
5. **Knowledge discovery :** Deals with identification of useful information from huge amount of text. Most tools provide for knowledge discovery and information retrieval features.

5.3.2 Text Pre - processing

- Text pre - processing is required to transform the text into an understandable format so that machine learning algorithms can be applied to it.
- As we know machine learning needs data in the numeric form. We basically used encoding technique to encode text into numeric vector. But before encoding we first need to clean the text data and this process to prepare or clean text data before encoding is called text pre-processing.
- The various text pre-processing steps are : Tokenization, Lower casing, Stop words removal, Stemming and Lemmatization.

5.3.2.1 Tokenization

- Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. The simplest form of analysis is to reduce different word forms into tokens.
- Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.
- Tokenization can be broadly classified into 3 types : Word, character, and subword (n-gram characters) tokenization.
- These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.



- Tokenization can be done to either separate words or sentences. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.
- Tokens are the building blocks of natural language, the most common way of processing the raw text happens at the token level.
- The major question of the tokenization phase is what are the correct tokens to use ? You chop on whitespace and throw away punctuation characters.
- Types of tokenization are white space, dictionary based, rule based, penn tree, spacy, subword etc.

5.3.2.2 Stemming

- Stemming is the process for reducing inflected words to their stem, base or root form, generally a written word form. Stemming allows a query term such as "orienteering" to match an occurrence of "orienteers", or "runs" to match "running".
- For example, "orienteering" and "orienteers" might reduce to the root form "orienteer"; "runs" and "running" might reduce to "run".
- In an IR system a stemmer may be applied at both indexing time and query time. During indexing each token is passed through the stemmer and the resulting root form is indexed.
- At query time, the query terms are passed through the same stemmer and matched against the index terms. Thus the query term "runs" would match an occurrence of "running" by way of their common root form "run".
- Stemming is the particular case of tokenization which reduces inflected forms to a single base form or stem. Stemming algorithms are basic string - handling algorithms, which depend on rules which identify affixes that can be stripped.
- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance :

  am, are, is $\Rightarrow$ be

  car, cars, car's, cars' $\Rightarrow$ car
- The result of this mapping of text will be something like : The boy's cars are different colors the boy car be differ color.

5.3.2.3 Stop Words

- The removal of high frequency words, 'stop' words or 'fluff' words is one way of implementing Luhn's upper cut-off. This is normally done by comparing the input text with a 'stop list' of words which are to be removed.
- Sometimes, some extremely common words that would appear to be of little value in helping select documents matching a user need are excluded from stop words the vocabulary entirely. These words are called stop words.
- Function words are words that have no well - defined meanings in and of themselves. Function words are usually among the most frequently occurring words in any language. At query time these stopwords are stripped from the query, and retrieval takes place on the basis of the remaining terms alone.
- Examples of a few stop words in English are "the", "a", "an", "so", "what".
- Removal of stop words definitely reduces the dataset size and thus reduces the training time due to the fewer number of tokens involved in the training.

- The phrase query "Capital of the India," which contains two stop words, is more precise than Capital AND "India".
- Stop word list is as follows :

  a an and are as at be by for from has he in

  is it its of on that the to was were will with
- With stopwords present in the index, the IR system can make a decision on a query - by - query basis. Ranking methods in a modern commercial search engine will incorporate many ranking features, including features based on term frequency and proximity.
- For features that do not consider proximity between query terms, stopwords may be eliminated. For features that do consider proximity between query terms, particularly to match their occurrence in phrases, it may be appropriate to retain stopwords.
- Good compression techniques means the space for including stopwords in a system is very small.
- Good query optimization techniques mean you pay little at query time for including stopwords.

5.3.2.4 Lemmatization

- Many complex or technical concepts and many organization and product names are multiword compounds or phrases. An information retrieval system uses phrases to index, retrieve, organize and describe documents.
- Phrases are identified that predict the presence of other phrases in documents. Documents are the indexed according to their included phrases.
- Most recent search engines support a double quotes syntax ("technical publications") for phrase queries, which has proven to be very easily understood and successfully used by users.
- One approach to handling phrases is to consider every pair of consecutive terms in a document as a phrase. For example the text ***Friends, Romans, Countrymen*** would generate the biwords :

  friends romans

  romans countrymen
- Each of these biwords is now a dictionary term. Two - word phrase query - processing is now immediate.
- Biword indexes are not the standard solution but can be part of a compound strategy.

5.3.3 Bag of Words

- Bag of words model helps convert the text into numerical representation such that the same can be used to train models using machine learning algorithms.
- The bag-of-words model is commonly used in methods of document classification where the occurrence of each word is used as a feature for training a classifier.
- Here are the key steps of fitting a bag-of-words model :
  1. Create a vocabulary index of words or tokens from the entire set of documents. The vocabulary indices can be created in alphabetical order.
  2. Construct the numerical feature vector for each document that represents how frequent each word appears in different documents. The feature vector representing each will be sparse in nature as the words in each document will represent only a small subset of words out of all words present in entire set of documents.
- Fig. 5.3.1 shows turning raw text into a bag of words representation.

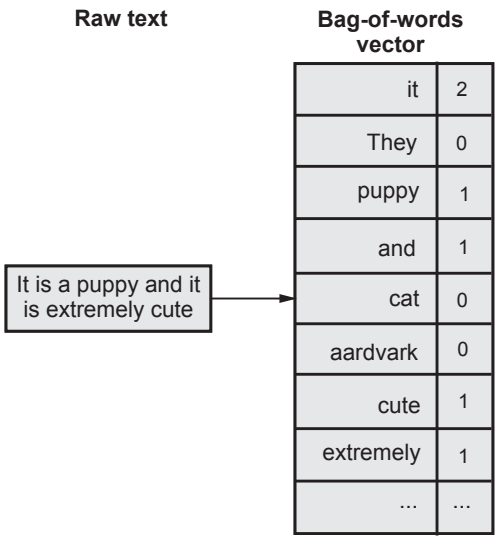


Fig. 5.3.1 Turning raw text into a bag of words representation

- Bag of words simply refers to a matrix in which the rows are documents and the columns are words. The values matching a document with a word in the matrix, could be a count of word occurrences within the document or use tf-idf.
- Classifiers are used to train the bag of words and a special kind of algorithm used to break words down into categories.
- Traditionally, text documents are represented in NLP as a bag-of-words. This means that each document is represented as a fixed-length vector with length equal to the vocabulary size.

- Each dimension of this vector corresponds to the count or occurrence of a word in a document. Being able to reduce variable-length documents to fixed-length vectors makes them more amenable for use with a large variety of Machine Learning (ML) models and tasks.

5.3.4 TF-IDF Weighting

- Term Frequency (TF) : Frequency of occurrence of query keyword in document.
- Inverse Document Frequency (IDF) : How many documents the query keyword occurs in.
- Inverse Document Frequency (IDF) is a popular measure of a word's importance. It's defined as the logarithm of the ratio of number of documents in a collection to the number of documents containing the given word. This means rare words have high IDF and common words have low IDF.
- Term frequency is a measure of the importance of terms *i* in document *j.*
- Inverse document frequency is a measure of the *general* importance of the term.
- High term frequency for "apple" means that apple is an important word in a specific document. But high document frequency (low inverse document frequency) for "apple", given a particular set of documents, means that apple is not all that important overall, since it is in all of the documents.
- The weight increases as the number of documents in which the term appears decreases. High value indicates that the word occurs more often in this document than average.
- The term frequency $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.
- A document with tf = 10 occurrences of the term is more relevant than a document with tf = 1 occurrence of the term. But not 10 times more relevant. Relevance does not increase proportionally with term frequency.
- The document frequency is the number of documents in the collection that the term occurs in. We define the idf weight of term t as follows :

  $$\text{idf weight } (idf_t) = \log 10 \frac{N}{df_t}$$

  here N is the number of documents in the collection
- The tf-idf weight of a term is the product of its tf weight and its idf weight

  $$W_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

Stop lists and Stemming :

- **Stoplists :** This is a list of words that we should ignore when processing documents, since they give no useful information about content.
- **Stemming :** This is the process of treating a set of words like "fights, fighting, fighter, …" as all instances of the same term - in this case the stem is "fight".

5.4 Need and Introduction to Social Network Analysis

- Social Network Analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The term "social network" has been introduced by Barnes in 1954.
- SNA is the study of social relations among a set of actors. The methods of data collection in network analysis are aimed at collecting relational data in a reliable manner. Data collection is typically carried out using standard questionnaires and observation techniques that aim to ensure the correctness and completeness of network data.
- Social network analysis is based on an assumption of the importance of relationships among interacting units. The social network perspective encompasses theories, models, and applications that are expressed in terms of relational concepts or processes.
- The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships.
- The advantage of social network analysis is that, unlike many other methods, it focuses on interaction. Network analysis allows us to examine how the configuration of networks influences how individuals and groups, organizations, or systems function.
- **Features of social network analysis :** Structural intuition, systematic relational data, graphic representation and mathematical or computational models.

Fundamental concepts in network analysis :

Following terminology is used in social network analysis.

1. Actor 2. Relational tie 3. Dyad 4. Triad
5. Subgroup 6. Group 7. Relation

a) Actor : Actor is discrete individual, corporate, or collective social units. Examples : people in a group, departments within in a corporation, public service agency in a city, nation - states in the world system.

b) Relational tie : Actors are linked to another by social ties. A tie establishes a linkage between a pair of actors.

c) Dyad : It is a tie between two actors and consists of a pair of actors and the tie(s) between them.

d) Triad : Triples of actors and associated ties. A subset of three actors and the tie(s) among them.

e) Subgroup of actors is defined as any subset of actors, and all ties among them.

f) Group : Group is the collection of all actors on which ties are to be measured.

g) Relation : It is the collection of ties of a specific kind among members of a group. Example : The set of friendship among pairs of children in a classroom.

- Network can be categorized by the nature of the sets of actors and the properties of the ties among them. The number of modes in a network refers to the number of distinct kinds of social entities in the network.
- One-mode networks are a single set of actors. Two-mode networks are focus on two sets of actors, or one set of actors and one set of events.

Principles of social network analysis

1. Actors and their actions are viewed as interdependent rather than independent, autonomous units.
2. Relational ties (linkages) between actors are channels for transfer or "flow" of resources (either material or nonmaterial).
3. Network models focusing on individuals view the network structure environment as providing opportunities for or constraints on individual action.
4. Network models conceptualize structure (social, economic, political, and so forth) as lasting patterns of relations among actors.

5.4.1 Development of Social Network Analysis

- A social network is a group of collaborating, and/or competing individuals or entities that are related to each other. It may be presented as a graph, or a multi-graph; each participant in the collaboration or competition is called an actor and depicted as a node in the graph theory.
- Valued relations between actors are depicted as links, or ties, either directed or undirected, between the corresponding nodes.
- Actors can be persons, organizations, or groups - any set of related entities. As such, SNA may be used on different levels, ranging from individuals, web pages, families, small groups, to large organizations, parties, and even to nations.

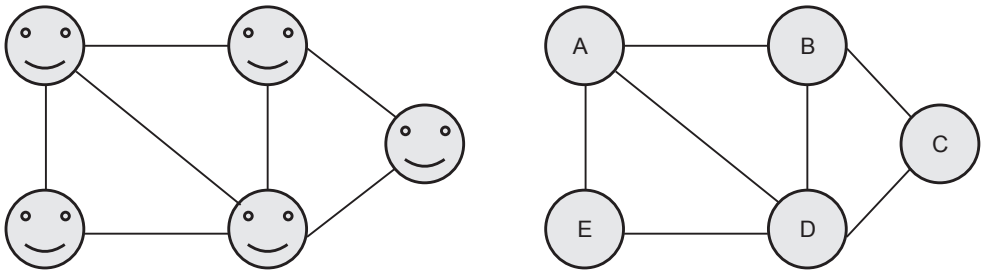
- In general, a social network consists of actors (e.g., persons, organizations) and some form of relation among them. The network structure is usually modeled as a graph, in which vertices represent actors, and edges represent ties, i.e., the existence of a relation between two actors.
- The vocabulary, models and methods of network analysis also expand continuously through applications that require to handle ever more complex data sets.
- An example of this process are the advances in dealing with longitudinal data. New probabilistic models are capable of modelling the evolution of social networks and answering questions regarding the dynamics of communities. Formalizing an increasing set of concepts in terms of networks also contributes to both developing and testing theories in more theoretical branches of sociology.
- The purpose of social network analysis is to identify important actors, crucial links, roles, dense groups, and so on, in order to answer substantive questions about structure.
- Analysis methods are divided into four main categories according to the level or subject of interest : vertex, dyad, group, and network level.
- Available analysis methods include actor-level centrality indices, e.g. closeness, betweenness, and pagerank, cohesive subgroups like cliques, k-cliques, and k-clans, centrality and connectedness.
- These levels break further down into measures of the same objective, e.g., connectedness or cohesiveness. Analysis methods are accessible using the analysis tab in the control area.

Key concepts and measures in network analysis

- Social network analysis has developed a set of concepts and methods specific to the analysis of social networks.
- Several analytic tendencies distinguish social network analysis :
  1. There is no assumption that groups are the building blocks of society : The approach is open to studying less - bounded social systems, from nonlocal communities to links among websites.
  2. Rather than treating individuals (persons, organizations, states) as discrete units of analysis, it focuses on how the structure of ties affects individuals and their relationships.
  3. In contrast to analyses that assume that socialization into norms determines behavior, network analysis looks to see the extent to which the structure and composition of ties affect norms.

5.4.2 Global Structure of Networks

- Social network can be represented as a graph G = (V, E)
  where V = The finite set of vertices
  E = Finite set of edges such
- The most network analysis methods work on an abstract, graph based representation of real world networks. It is shown in Fig. 5.4.1.



	A	B	C	D	E
A	0	1	0	1	1
B	1	0	1	1	0
C	0	1	0	1	0
D	1	1	1	0	1
E	1	0	0	1	0

Fig. 5.4.1 Graph based representation of real world networks

- When representing a network as a graph, all of the connections are pair-wise and hence represented by ties known as edges.
- Networks can be described using a mixture of local, global, and intermediate-scale perspectives. Accordingly, one of the key uses of network theory is the identification of summary statistics for large networks in order to develop a framework for analyzing and comparing complex structures.
- SNA can produce maps like the one featured below, and provide statistical measures of relationships between actors. In SNA maps, the nodes represent the different actors in the network, and the lines represent the relationships between the various actors.

- The size of the node often represents the relative importance of that actor in the network, and the thickness of the connecting line denotes the strength of the relationship.
- Clustering for a single vertex can be measured by the actual number of the edges between the neighbors of a vertex divided by the possible number of edges between the neighbors.
- When taken the average over all vertices, we get to the measure known as clustering coefficient. The clustering coefficient of a tree is zero, which is easy to see if we consider that there are no triangles of edges (triads) in the graph. In a tree, it would never be the case that our friends are friends with each other.
- The coordination degree measures the ability of the vertices in a graph to interchange information. There are several ways in which we can model this magnitude. One of the easiest is to consider the coordination degree to be exponentially related with the distance between the vertices.
- To define the total co-ordination degree of a vertex "i" in a graph as the sum of all the coordination degrees between that particular vertex and the rest :

$$\Gamma_i = \sum_{j=1}^{N} \gamma_{ij}$$

Where N is the order of the graph.

- Graph density (D) is defined as the total number of observed lines in a graph divided by the total number of possible lines in the same graph. Density ranges from 0 to 1.

$$\text{Density (D)} = \frac{\text{Number of lines (L)}}{(\text{Number of points (Number of points} - 1)) / 2} = \frac{2L}{g(g-1)}$$

5.4.3 Random Graphs with Arbitrary Degree Distributions

- A random graph is simple to define. One takes some number N of nodes or "vertices" and places connections or "edges" between them, such that each pair of vertices i, j has a connecting edge with independent probability p.
- Random graph can be generated by taking a set of vertices with no edges connecting them. Subsequently, edges are added by picking pairs of nodes with equal probability.

- Consider a vertex in a random graph. It is connected with equal probability p with each of the N – 1 other vertices in the graph, and hence the probability p_k that it has degree exactly k is given by the binomial distribution :

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

- A large random graph has a Poisson degree distribution. This degree distribution makes the random graph a poor approximation to the real-world networks.

5.4.4 Macro-Structure of Social Networks

- Network visualizations based on topographic or physical principles can be helpful in understanding the group structure of social networks and pinpoint hubs that naturally tend to gravitate toward the center of the visualization.
- Clustering a graph into subgroups allows us to visualize the connectivity at a group level.
- Core-Periphery structure is one where nodes can be divided in two distinct subgroups : Nodes in the core are densely connected with each other and the nodes on the periphery, while peripheral nodes are not connected with each other, only nodes in the core.
- By computing a network's core-periphery structure, one attempts to determine which nodes are part of a densely connected core and which are part of a sparsely connected periphery.
- Core nodes should also be reasonably well-connected to peripheral nodes, but the latter are not well-connected to a core or to each other.
- Node belongs to a core if and only if it is well-connected both to other core nodes and to peripheral nodes. A core structure in a network is thus not merely densely connected but also tends to be 'central' to the network.
- From network theory has it defined the dual relationships between nodes in the network, so that if an agent has a feature no other, for example, if it is good then it is not bad, is a bipartition graph in which each element of a subset is additional to another concept indeed implies that binding of the n subgroups partitions make the whole graph. So CPS, involves dividing the nodes of the network into two groups.
- Fig. 5.4.2 shows core-periphery structure that would be perfect without the edge between nodes.

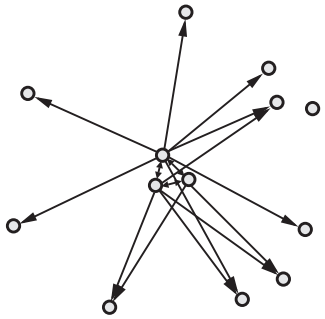


Fig. 5.4.2 (a) : Core - periphery structure

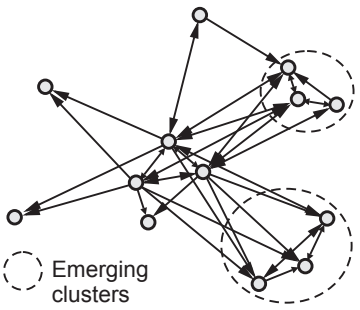


Fig. 5.4.2 (b) Cluster structure

1. Affiliation network :

- **Affiliation networks** contain information about the relationships between two sets of nodes : A set of subjects and a set of affiliations. An affiliation network can be formally represented as a bipartite graph, also known as a two-mode network.
- Affiliation networks are **two mode networks** that allow one to study the dual perspectives of the actors and the events. They look at collections or subsets of actors or subsets rather than ties between pairs of actors. Connections among members of one of the modes as based on linkages established through the second mode.
- An affiliation network is a network in which actors are joined together by common membership of groups or clubs of some kind.
- A distinctive feature of affiliation networks is **duality** i.e. events can be described as collections of individuals affiliated with them and actors can be described as collections of events with which they are affiliated.
- Based on two-mode matrix data, affiliation networks consist of sets of relations connecting actors and events, rather than direct ties between pairs of actors as in one-mode data. Familiar affiliation networks include persons belonging to associations, social movement activists participating in protest events, firms creating strategic alliances, and nations signing treaties.
- The representation of two-mode data should facilitate the visualization of three kinds of patterning :
  a. The actor-event structure
  b. The actor-actor structure
  c. The event-event structure

- Many ways to represent affiliation networks :
  1. Affiliation network matrix
  2. Bipartite graph
  3. Hypergraph
  4. Simplicial complex

Benefits of affiliations network

1. Affiliations of actors with events provide a direct linkage between actors through memberships in events, or between events through common memberships.
2. Affiliations provide conditions that facilitate the formation of pairwise ties between actors.
3. Affiliations enable us to model the relationships between actors and events as a whole system.

2. Bipartite graph :

- Nodes are partitions into two subsets and all lines are between pairs of nodes belonging to different subsets. Fig 5.4.3 shows bipartite network. As there are g actors and h events, there are g + h nodes.

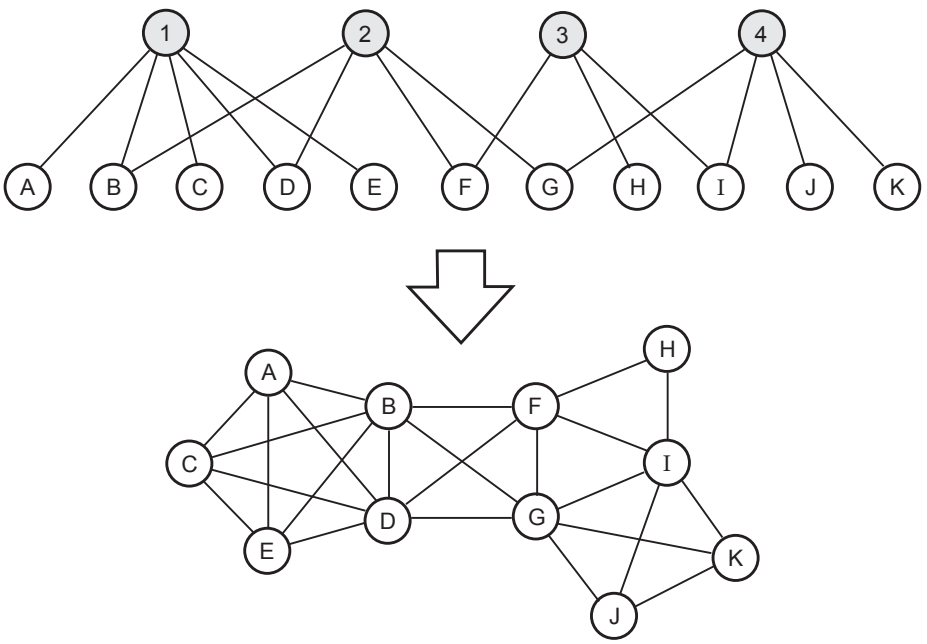


Fig. 5.4.3 Bipartite graph

- The lines on the graph represent the relation "is affiliated with" from the perspective of the actor and the relation "has as a member" from the perspective of the event.
- No two actors are adjacent and no two events are adjacent. If pairs of actors are reachable, it is only via paths containing one or more events. Similarly, if pairs of events are reachable, it is only via paths containing one or more actors.

Advantages

1. They highlight the connectivity in the network, as well as the indirect chains of connection.
2. Data is not lost and we always know which individuals attended which events.

Disadvantage

1. They can be unwieldy when used to depict larger affiliation networks.

5.4.5 Application of Social Network Analysis

- Social Network Analysis (SNA) is an important and valuable tool for knowledge extraction from massive and un-structured data. Social network provides a powerful abstraction of the structure and dynamics of diverse kinds of inter-personal connection and interaction.
- Facebook is a social networking service and website that connects people with other people, and share data between people. A user can create a personal profile, add other users as friends, exchange data, create and join common interest communities.
- Twitter is a social net-working and microblogging service. The users of Twitter can exchange text-based posts called tweets. A tweet is a maximum 140 characters long but can be augmented by pictures or audio recording. The main concept of Twitter was to build a social network formed by friends and followers. Friends are people who you follow, followers are those who follow you.
- The role of social networks in labor markets deserves attention for at least two reasons : First, because of the central role networks play in disseminating information about job openings they place a critical role in determining whether labor markets function efficiently; and second, because network structure ends up having implications for things like human capital investment as well as inequality.
- Social Network Analysis (SNA) primarily focuses on applying analytic techniques to the relationships between individuals and groups, and investigating how those relationships can be used to infer additional information about the individuals and groups.

- SNA is used in a variety of domains. For example, business consultants use SNA to identify the effective relationships between workers that enable work to get done; these relationships often differ from connections seen in an organizational chart.

5.5 Introduction to Business Analysis

- Business Analysis is a discipline and practice of defining business needs and recommending solutions to business problems. Business analysis deals with the current state of each company, desired future state, stakeholders' needs, processes, software and more.
- A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms.
- The adoption of a business intelligence system tends to promote a scientific and rational approach to the management of enterprises and complex organizations.
- Obstacle to business intelligence in an organization are as follows :

1. Lack of BI strategy : Organizations should proactively define the problems they trying to solve. Only then they will be able to identify the right Business Intelligence solution that will suit their requirements.

2. Business intelligence :

- When You Don't Know How to Code. Now a days, executives find it difficult to access the right data at right time. And even if they do find what they're looking for, data formats are typically so complex and unstructured it's hard to find out meaningful and relevant data.
- Now unless they are using Excel extensively, they probably would not get much satisfaction (or value) from their BI system.
- A good practice would be to replace Excel Sheets with intuitive dashboards to make data more engaging, meaningful and eventually very powerful.

3. Lack of training and execution :

- Many a times, companies might have well-articulated requirements, a sound BI strategy, and a good tool solution, but lack technical skills like designing, building, maintaining, and supporting BI applications.
- This results in BI applications to run slowly, break frequently, deliver uncertain results and eventually leading to rising cost of using the BI solution. The causes of lack of execution often are multiple and varied, as are its remedies.

4. Lack of BI impact (Low utilization) :

- Management might always wonder why there is no change in business results attributable to BI and might feel that business value of BI investments not captured. This indicates that the organization is not utilizing the BI solution at par with global standards and best practices.

5. Business intelligence with unstructured data :

- Most of the times data is unstructured for BI to analyze. This lead to a problem when users need to perform simple BI processes.
- Businesses may invest in big data analytics but cannot complete the tasks in time. They may result to people spending hours on cleaning and structuring the data first and then using the BI solution.

6. Installation and deployment :

- A painful BI solution installation and deployment would be difficult to maintain. Even an unplanned and rushed deployment would be unsuccessful so often.
- Doing this may leave users void with time to understand the system and develop the skills using the solution effectively.

5.6 Model Evaluation and Selection

SPPU : Dec.-18

- A binary classification rule is a method that assigns a class to an object, on the basis of its description.
- The performance of a binary classifier can be assessed by tabulating its predictions on a test set with known labels in a contingency table or confusion matrix, with actual classes in rows and predicted classes in columns.
- Measures of performance need to satisfy several criteria :
  1. They must coherently capture the aspect of performance of interest ;
  2. They must be intuitive enough to become widely used, so that the same measures are consistently reported by researchers, enabling community - wide conclusions to be drawn ;
  3. They must be computationally tractable, to match the rapid growth in the scale of modern data collection.
  4. They must be simple to report as a single number for each method - dataset combination.
- Performance metrics for binary classification are designed to capture trade-offs between four fundamental population quantities : True positives, false positives, true negatives and false negatives.

- The evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, named confusion matrix.
- Confusion matrix is also called a contingency table.
  1. **False positives :** Examples predicted as positive, which are from the negative class.
  2. **False negatives :** Examples predicted as negative, whose true class is positive.
  3. **True positives :** Examples correctly predicted as pertraining to the positive class.
  4. **True negatives :** Examples correctly predicted as belonging to the negative class.

$$\text{Accuracy rate} = \frac{|\text{True negatives}| + |\text{True positives}|}{|\text{False negatives}|+|\text{True positive}|+|\text{True negatives}|+|\text{True positives}|}$$

- The complement of accuracy rate is the error rate, which evaluates a classifier by its percentage of incorrect predictions.

$$\text{Error rate} = \frac{|\text{False negatives}| + |\text{False positives}|}{|\text{False negatives}|+|\text{False positive}|+|\text{True negatives}|+|\text{True positives}|}$$

$$\text{Error rate} = 1 - (\text{Accuracy rate})$$

- The recall and specifity measures evaluate the effectiveness of a classifier for each class in the binary problem. The recall is also known as sensitivity or true positive rate. Recall is the proportion of examples belonging to the positive class which were correctly predicted as positive.
- The **specificity** is a statistical measure of how well a binary classification test correctly identifies the negatives cases.

$$\text{Recall (R)} = \frac{|\text{True positive}|}{|\text{True positive}|+|\text{False negative}|}$$

$$\text{Specificity} = \frac{|\text{True negatives}|}{|\text{False positives}|+|\text{True negative}|}$$

- True Positive Rate (TPR) is also called sensivity, hit rate, and recall.

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}}$$

- A statistical measure of how well a binary classification test correctly identifies a conditon. Probability of correctly labeling members of the target class.

- No single measure tells the whole story. A classifier with 90 % accuracy can be useless if 90 percent of the population does not have cancer and the 10 % that do are misclassified by the classier. Use of multiple measures recommended.
- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predictive model.

5.6.1 Issues Regarding Classification and Prediction

Preparing the data for classification and prediction

Following pre-processing steps may be applied to the data to help improve the accuracy, efficiency, and scalability of the classification or prediction process.

a) **Data cleaning :** Preprocess data in order to reduce noise and handle missing values.

b) **Relevance analysis (Feature selection) :** Remove the irrelevant or redundant attributes.

c) **Data transformation :** Generalize and/or normalize data

Feature wise comparison between classification and prediction :

1. **Accuracy :** Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.
2. **Speed :** This refers to the computational cost in generating and using the classifier or predictor.
3. **Robustness :** It refers to the ability of classifier or predictor to make correct predictions from given noisy data.
4. **Scalability :** Scalability refers to the ability to construct the classifier or predictor efficiently ; given large amount of data.
5. **Interpretability :** It refers to what extent the classifier or predictor understands.

5.6.2 Holdout Method

- The data is split into two different datasets labelled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split. This technique is called the hold-out validation technique.
- Suppose we have a database with house prices as the dependent variable and two independent variables showing the square footage of the house and the number of rooms.

- Now, imagine this dataset has 30 rows. The whole idea is that you build a model that can predict house prices accurately.
- To 'train' your model, or see how well it performs, we randomly subset 20 of those rows and fit the model.
- The second step is to predict the values of those 10 rows that we excluded and measure how well our predictions were.
- As a rule of thumb, experts suggest to randomly sample 80 % of the data into the training set and 20 % into the test set.

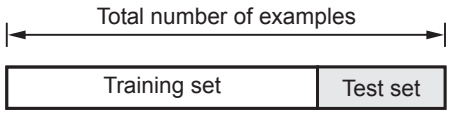


Fig. 5.6.1

- Training set : Used to train the classifier.
- The holdout method has two, basic drawbacks :
  1. It requires extra dataset
  2. It is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split.

5.6.3 Random Subsampling

- Random subsampling performs K data splits of the entire sample. For each data split, a fixed number of observations is chosen without replacement from the sample and kept aside as the test data.
- The prediction model is fitted to the training data from scratch for each of the K splits and an estimate of the prediction error is obtained from each test set.
- Let the estimated PE in the i^{th} test set be denoted by E_i. The true error estimate is obtained as the average of the separate estimates E_i.

$$\frac{1}{K}\sum_{i=1}^{K} E_i$$

1. Cross-Validation

- Cross-validation is a technique for evaluating estimating performance by training several machine learning models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, i.e., failing to generalize a pattern.
- In general, machine learning involves deriving models from data, with the aim of achieving some kind of desired behaviour, e.g., prediction or classification.

- Fig. 5.6.2 shows cross-validation.

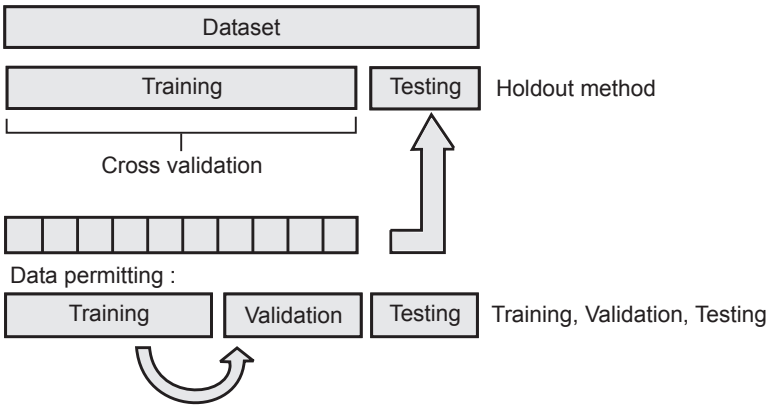


Fig. 5.6.2 Cross validation

- But this generic task is broken down into a number of special cases. When training is done, the data that was removed can be used to test the performance of the learned model on "new" data. This is the basic idea for a whole class of model evaluation methods called **cross validation**.
- Types of cross validation methods are holdout, K-fold and Leave-one-out.
- The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximate fits a function using the training set only.
- The K-fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times.
- Each time, one of the k subsets is used as the test set and the other k - 1 subsets are put together to form a training set. Then the average error across all k trials is computed.
- Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set.
- That means that N separate times, the function approximate is trained on all the data except for one point and a prediction is made for that point.
- Cross-validation ensures non-overlapping test sets.

K-fold cross-validation :

- In this technique, k – 1 folds are used for training and the remaining one is used for testing as shown in Fig. 5.6.3.

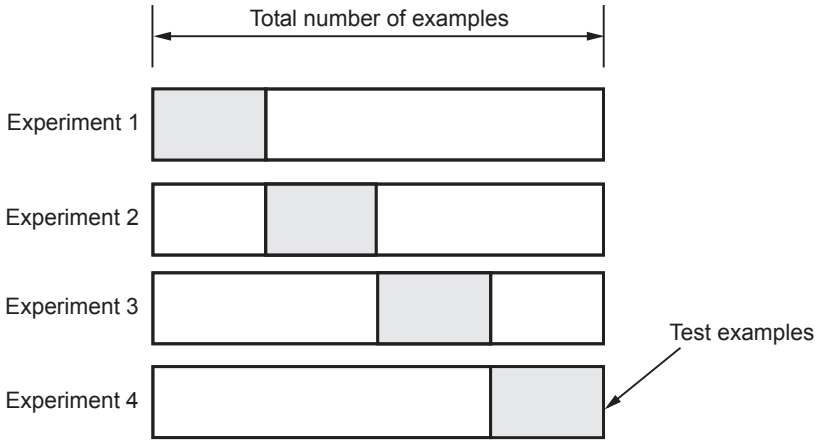


Fig. 5.6.3 K-fold cross validation

- The advantage is that entire data is used for training and testing. The error rate of the model is average of the error rate of each iteration.
- This technique can also be called a form the repeated hold-out method. The error rate could be improved by using stratification technique.

Review Question

1. *Explain any three of classification performance measures.*

SPPU : Dec.-18 (End Sem), Marks 6

5.7 Clustering and Time-series Analysis using Scikit-learn

- Time series data is widely used to analyse different trends and seasonalities of products over time by various industries. Sktime is a unified python framework/library providing API for machine learning with time series data and sklearn compatible tools to analyse, visualize, tune and validate multiple time series learning models such as time series forecasting, time series regression and classification.
- Time series are a stream of data that are created by making measures of something such as sales, temperature, stocks, etc. in fixed frequency. They have to be indexed in time order and usually used in weather forecasting, econometrics, earthquake prediction, signal processing, etc.
- Clustering of unlabeled data can be performed with the module sklearn.cluster.

5.7.1 Scikit-learn

- Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

- It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.
- Scikit-learn is a library, i.e. a collection of classes and functions that users import into Python programs. Using scikit-learn therefore requires basic Python programming knowledge
- The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes :
  1. NumPy : Base n-dimensional array package
  2. SciPy : Fundamental library for scientific computing
  3. Matplotlib : Comprehensive 2D/3D plotting
  4. IPython : Enhanced interactive console
  5. Sympy : Symbolic mathematics
  6. Pandas : Data structures and analysis
- Scikit-learn is the package for machine learning and data science experimentation favoured by most data scientists. It contains a wide range of well-established learning algorithms, error functions, and testing procedures.

5.7.2 Understanding Classes in Scikit-learn

- Scikit-learn takes a highly object-oriented approach to machine learning models. Every major Scikit-learn class inherits from ***sklearn.base.BaseEstimator***.
- Scikit-learn features some base classes on which all the algorithms are built. Apart from *BaseEstimator*, the class from which all other classes inherit, there are four class types covering all the basic machine learning functionalities :
  1. Classifying
  2. Regressing
  3. Grouping by clusters
  4. Transforming data
- Scikit-learn takes a highly object-oriented approach to machine learning models. Every major Scikit-learn class inherits from ***sklearn.base.BaseEstimator***.
- All objects within scikit-learn share a uniform common basic API consisting of three complementary interfaces : an **estimator** interface for building and fitting models, a **predictor** interface for making predictions and a **transformer** interface for converting data.

1. Estimators

- The estimator interface is at the core of the library. It defines instantiation mechanisms of objects and exposes a fit method for learning a model from training data.

- All supervised and unsupervised learning algorithms (e.g., for classification, regression or clustering) are offered as objects implementing this interface. Machine learning tasks like feature extraction, feature selection or dimensionality reduction are also provided as estimators.

2. Predictors

- The predictor interface extends the notion of an estimator by adding a predict method that takes an array X test and produces predictions for X test, based on the learned parameters of the estimator.
- In the case of supervised learning estimators, this method typically returns the predicted labels or values computed by the model.

3. Transformers

- Since it is common to modify or filter data before feeding it to a learning algorithm, some estimators in the library implement a transformer interface which defines a transform method.
- It takes as input some new data X test and yields as output a transformed version of X test.
- Preprocessing, feature selection, feature extraction and dimensionality reduction algorithms are all provided as transformers within the library.

5.8 Confusion Matrix

SPPU : May-19

- A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The numbers displayed give the frequency of each data point.
- The confusion matrix for binary classification shown below :

	Predicted class	
True class	**Positive**	**Negative**
Positive	True negative	False negative
Negative	False positive	True negative

- A confusion matrix contains information about actual and predicted classification done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Confusion matrix is also called a contingency table.
  1. **False positives :** Examples predicted as positive, which are from the negative class.

2. **False negatives :** Examples predicted as negative, whose true class is positive.
3. **True positives :** Examples correctly predicted as pertraining to the positive class.
4. **True negatives :** Examples correctly predicted as belonging to the negative class.

- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predictive model.
- Accuracy (ACC) measures the fraction of correct predictions. Precision measures the fraction of actual positives among those examples that are predicted as positive. Recall measures how many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall.

5.8.1 ROC Curve

- Receiver Operating Characteristics (ROC) graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates noisy channel. Recent years have seen an increase in the use of ROC graphs in the machine learning community.
- An ROC plot plots true positive rate on the Y-axis against false positive rate on the X-axis ; a single contingency table corresponds to a single point in an ROC plot.
- The performance of a ranker can be assessed by drawing piecewise linear curve in an ROC plot, known as an ROC curve. The curve starts in (0, 0), finishes in (1, 1) and is monotonically non-decreasing in both axes.
- A useful technique for organizing classifiers and visualizing their performance. Especially useful for domains with skewed class distribution and unequal classification error costs.
- It allows to create ROC curve and a complete sensitivity / specificity report. The ROC curve is a fundamental tool for diagnostic test evaulation.
- In a ROC curve the true positive rate (sensitivity) is plotted in function of the false positive rate (100 -specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity / specificity pair corresoponding to a particular decision threshold. The area under the ROC curve is a measure of how well a parameter two segments.

- Each point on an ROC curve connecting two segments corresponds to the true and false positive rates achieved on the same test set by the classifier obtained from the ranker by splitting the ranking between those two segments.
- An ROC curve is convex if the slopes are monotonically non-increasing when moving along the curve from (0, 0) to (1, 1). A concavity in an ROC curve, i.e. in an ROC curve, i.e. two or more adjacent segment with increasing slopes, indicates a locally worse than random ranking. In this case, we would get better ranking performance by joining the segments involved in concavity, thus creating a coarser classifier.

Review Question

1. *Explain the term confusion matrix.* **SPPU : May-19 (End Sem), Marks 4**

5.9 Elbow Plot

- The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k.
- If k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.
- It involves running the algorithm multiple times over a loop, with an increasing number of cluster choice and then plotting a clustering score as a function of the number of clusters.
- The Elbow and Silhouette methods are the two state-of-the-art methods used to identify the correct cluster number in the dataset.
- The Elbow method is the oldest method to distinguish the potential optimal cluster number for the analyzed dataset, whose basic idea is to specify K = 2 as the initial optimal cluster number K, and then keeps increasing K by step 1 to the maximal specified for the estimated potential optimal cluster number, and finally distinguish the potential optimal cluster number K corresponding to the plateau.
- The optimal cluster number K is distinguished by the fact that before reaching K, the cost rapidly decreases to the called cost peak value, and after exceeding K, it continues to increase with the called cost peak value almost unchanged, as shown in Fig. 5.9.1 (a) with an explicit elbow point.

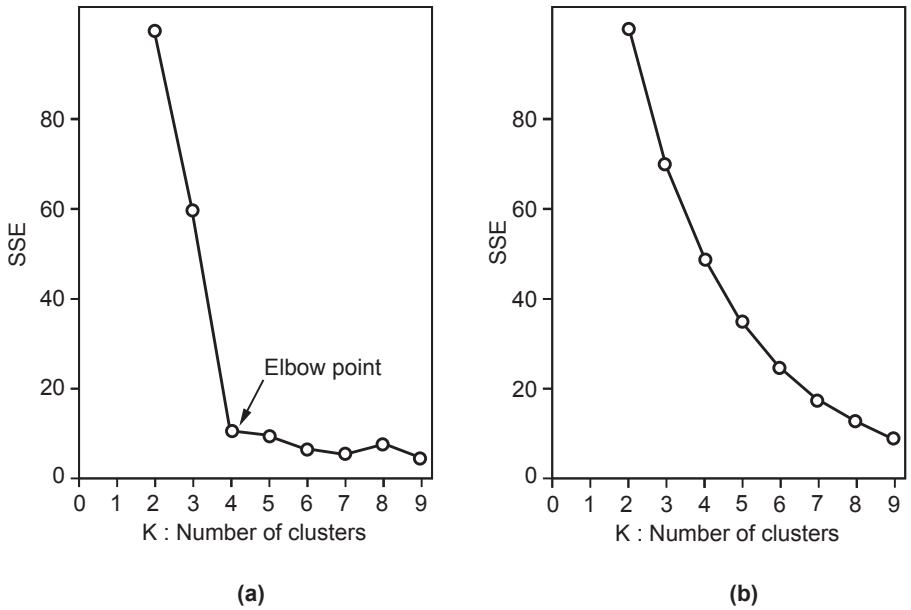


Fig. 5.9.1 Elbow point

- Meanwhile, the optimal cluster number corresponding to the elbow point depends on the manmade selection. There is, however, a problem with the Elbow method in that the elbow point cannot be unambiguously distinguished by the experienced analysts when the plotted curve is fairly smooth, as shown in Fig. 5.9.1 (b) with an ambiguous elbow point.
- To select the best K, we need to plot the mean in-cluster distance for each K. As K increases from 1, before reaching the optimal K, the decrease speed is relatively fast because the number of centers are too low from the very beginning and each new center will incur a large decrease in the mean distance.
- But after the optimal K, the decrease is slow since the correct cluster structure is already discovered and any newly added center will appear in a certain cluster already formed. That will not decrease the mean in-cluster distance too much. The entire curve looks like an L shape and the best K lies in the turning point or the elbow of the L shape.

5.10 Multiple Choice Questions

Q.1 A _____ is a a flowchart-like tree structure, where each internal node denotes a test on an attribute.

a desicion tree b binary tree

c cluster d none of these

Q.2 A node without further branches is called as _____.

a internal node
b root node
c lead node
d binary node

Q.3 _____ occurs when the gap between the training error and test error is too large.

a Underfitting
b Overfitting
c Overloaded
d Purning

Q.4 What is the approach of basic algorithm for decision tree induction ?

a Greedy
b Top down
c Procedural
d Step by Procedural

Q.5 What are two steps of tree pruning work ?

a Pessimistic pruning and Optimistic pruning
b Postpruning and Prepruning
c Cost complxity pruning and time complexity pruning
d None of the options

Q.6 How will you counter over-fitting in decision tree ?

a By pruning the longer rules
b By creating new rules
c Both by pruning the longer rules and by creating new rules.
d None of the above

Q.7 Which of the following is not a forecasting technique ?

a Judgemental
b Time series
c Time horizon
d Associative

Q.8 Which of the following is not true for forecasting ?

a Forcasts are rarely perfect.
b The underlying casual system will remain same in the future.
c Forecast for group of items is accurate than individual item.
d Short range forcasts are less accurate than long range forecasts.

Q.9 ETL stand for _____.

a Extract Transform and Load

b Exact Transfer and Language

c Extract Transmission and Language

d None

Q.10 ARIMA is a _____ model that uses time series data to either better understand the data set or to predict future trends.

a statistical analysis b analytical

c descriptive d all of these

Answer Keys for Multiple Choice Questions :

Q.1	a	**Q.2**	c	**Q.3**	b	**Q.4**	a	**Q.5**	b
Q.6	a	**Q.7**	c	**Q.8**	d	**Q.9**	a	**Q.10**	a

❑❑❑

UNIT VI

6 Data Visualization and Hadoop

Syllabus

Introduction to Data Visualization, Challenges to Big data visualization, Types of data visualization, Data Visualization Techniques, Visualizing Big Data, Tools used in Data Visualization, Hadoop ecosystem, Map Reduce, Pig, Hive, Analytical techniques used in Big data visualization. ***Data Visualization using Python :*** *Line plot, Scatter plot, Histogram, Density plot, Box- plot.*

Contents

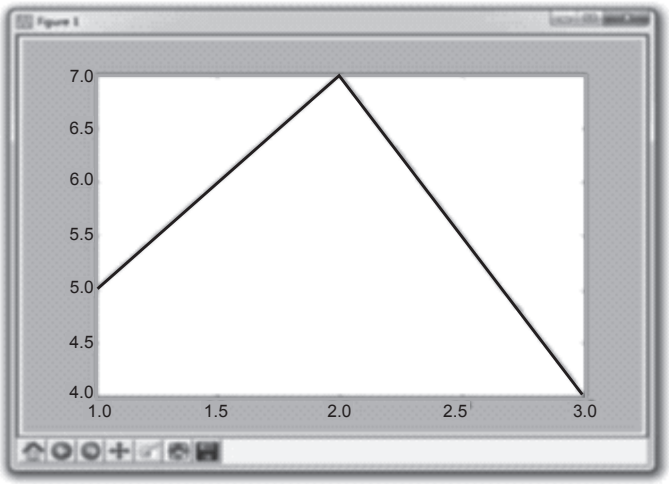
6.1 Introduction to Data Visualization

SPPU : Dec.-18, May-19

- Data visualization is the presentation of quantitative information in a graphical form. In other words, data visualizations turtan large and small datasets into visuals that are easier for the human brain to understand and process.
- Good data visualizations are created when communication, data science and design collide. Data visualizations done right offer key insights into complicated datasets in ways that are meaningful and intuitive.
- Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers and new insights about the information represented in the data.
- In order to craft a good data visualization, we need to start with clean data that is well sourced and complete. Once data is ready to visualize, we need to pick the right chart. This can be tricky, but there are many resources available to help us to choose the right type of chart for data.
- A graph is simply a visual representation of numeric data. MatplotLib supports a large number of graph and chart types.
- MatplotLib is popular Python package used to build plots. MatplotLib can also be used to make 3D plots plots and animations.
- Line plots can be created in Python with MatplotLib's pyplot library. To build a line plot, first import MatplotLib. It is a standard convention to import MatplotLib's pyplot library as plt.
- To define a plot, we need some values, the matplotlib. pyplot module and an idea of what you want to display.

```
import MatplotLib.pyplot as plt
plt.plot([1, 2,3], [5,7,4])
plt.show( )
```

- The plt.plot will "draw" this plot in the background, but we need to bring it to the screen when we're ready, after graphing everything we intend to.
- plt.show() : With that, the graph should pop up. If not, sometimes it can pop under or we may have gotten an error. Graph should look like :



- This window is matplotlib window, which allows us to see our graph, as well as interact with it and navigate it.
- **Three principal drivers of this technology :**
  1. **Visual :** Data are represented in a graphic/visual format.
  2. **Insight :** Data visualization, helps manager to understand data immediately and provides advice and suggestions on the possible actions he may take.
  3. **Sharing :** Advice and suggestions on the possible actions can be easily shared across the company which will lead to a consequent.
- For fully document graph, we usually have to resort to labels, annotations and legends. Each of these elements has a different purpose, as follows :
  1. **Label :** Make it easy for the viewer to know the name or kind of data illustrated.
  2. **Annotation :** Help extend the viewer's knowledge of the data, rather than simply identify it.
  3. **Legend :** Provides cues to make identification of the data group easier.
- Benefits of data visualization.
  1. Constructing ways in absorbing information. Data visualization enables users to receive vast amounts of information regarding operational and business conditions
  2. Visualize relationships and patterns in businesses
  3. More collaboration and sharing of information
  4. More self-service functions for the end users.

- Big data visualization is important because :
  1. It provides clear knowledge about patterns of data.
  2. Detects hidden structures in data
  3. Identify areas that need to be improved
  4. It helps us to understand which products to place where
  5. Clarify factors which influence human behaviour.

6.1.1 Challenges to Big Data Visualization

- Big data analytics plays a key role through reducing the data size and complexity in big data applications. Visualization is an important approach to helping big data get a complete view of data and discover data values.
- Scalability and dynamics are two major challenges in visual analytics.
- Volume : The methods are developed to work with an immense number of datasets and enable to derive meaning from large volumes of data.
- Variety : The methods are developed to combine as many data sources as needed.
- Velocity : With the methods, businesses can replace batch processing with real-time stream processing.
- Value : The methods not only enable users to create attractive info graphics and heat maps, but also create business value by gaining insights from big data.
- Visualization of big data with diversity and heterogeneity (structured, semi-structured and unstructured) is a big problem. Speed is the desired factor for big data analysis.
- There are also following problems for big data visualization :
  1. **Visual noise :** Most of the objects in the dataset are too relative to each other. Users cannot divide them as separate objects on the screen.
  2. **Information loss :** Reduction of visible data sets can be used, but leads to information loss.
  3. **Large image perception :** Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.
  4. **High rate of image change :** Users observe data and cannot react to the number of data changes or its intensity on display.
  5. **High performance requirements :** It can be hardly noticed in static visualization because of lower visualization speed requirements--high performance requirements.

- Following problems are encountered during visualizing big data :
  a) Scalability and dynamics are two major challenges in visual analytics.
  b) Visualization of big data with diversity and heterogeneity (structured, semi-structured and unstructured) is a big problem. Speed is the desired factor for big data analysis. Designing a new visualization tool with efficient indexing is not easy in big data.
  c) Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability.
  d) Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees and other metadata. Big data often has unstructured formats.
  e) Due to bandwidth limitations and power requirements, visualization should move closer to the data to extract meaningful information efficiently. Visualization software should be run in an in situ manner.
  f) Visual noise : It is messy to represent the whole array of data being studied on the screen. This problem comes when most of the objects share too much of relativeness, and that's the only reason why viewers cannot view them as separate objects.
  g) High - performance requirements : The graphical analysis does not stop at just static picture representation, so the above issues turn out to be more critical in unique perception.
  h) Large image perception : This problem occurs due to the human perceptions which different for different entities. In spite of the higher level of graphical data visualizations, it has its own limitations when compared with the table representation.
  i) High rate of image change : This issue turns into the hugest in checking assignments, when a man who analyses the information just can't respond to the quantity of information changes or its power on display.

Review Questions

1. *What are the challenges in big data visualization ?* **SPPU : Dec.-18 (End Sem), Marks 8**
2. *What is data visualization ? Describe any four data visualization techniques* **SPPU : May-19 (End Sem), Marks 8**
3. *Why is it difficult to visualize big data ? Also explain analytical techniques used in big data visualization* **SPPU : May-19 (End Sem), Marks 9**

6.2 Types of Data Visualization

SPPU : Dec.-18, 19

- Various types of data visualization are as follows :
  1. Multidimensional : 2D Area
  2. Temporal
  3. Hierarchical
  4. Network

Sr. No.	Types	Descriptions
1.	Multidimensional : 2D Area	**1. Cartogram :** It distorts map space to express information such as travel time or population of the alternate variable. It mainly consists of two main types : Area based and distance-based cartograms. **2. Choropleth :** It is used to represent the statistical measurement such as population density rate or website visitors count per city. **3. Dot distortion map :** It uses a dot symbol to represent a feature on the map, depending on the visual scatter for displaying spatial patterns.
2.	Temporal	**1. Pie chart :** The circle is divided into sectors to represent numeric proportions. The length of the arc and angle length of the sector is proportional to the particular quantity it represents. **2. Histogram :** In a histogram, the data are grouped into ranges (e.g. 10 - 19, 20 - 29) and then plotted as connected bars. Each bar represents a range of data. The width of each bar is proportional to the width of each category and the height is proportional to the frequency or percentage of that category. **3. Scatter plot :** It displays collection of all the points for the set of data limited only for two values.
3.	Hierarchical	**1. Dendrogram :** It is nothing but a tree diagram used to represent clusters generated by hierarchical clustering. **2. Ring chart :** It is a multi-level pie chart which is represented by the nested circles. **3. Tree diagram :** It represents the data or the hierarchy in the graph form, which can be visualized from left to right or top to bottom.
4.	Network	**1. Alluvial diagram :** It is a flow diagram which visualizes over time changes in network structure. **2. Node link diagram :** In this representation, nodes are visualized as dots whereas links are represented as line segments to display the data connection. **3. Matrix :** It shows relation between two to four groups of information and gives information regarding the same.

Review Questions

1. *What is data visualization ? Explain any four data visualization techniques*

SPPU : Dec.-18 (End Sem), Marks 9

2. *Explain data visualization with respect to 1-D, 2-D and 3-D data.*

SPPU : Dec.-19 (End Sem), Marks 9

6.3 Data Visualization Techniques

SPPU : Dec.-18, 19

- Whenever collection of data is started and the range of data increases rapidly, an efficient and convenient technique for representing data is needed.
- Higher authorities do not have enough time to go through whole reports regarding the progress of their firm or organization, so it is required for presenting the data in such a manner that enables readers to interpret the important data with minimum effort and time.
- Data visualization techniques are helping you avoid overloading the working memory.
- Techniques for data presentation are broadly classified in two ways :
  1. **Non graphical techniques :** Tabular form, case form
  2. **Graphical techniques :** Pie chart, bar chart, line graphs, geometrical diagrams.

6.3.1 Line Graph

- It is also called **stick graphs**. It gives relationships between variables.
- Line graphs are usually used to show time series data - that is how one or more variables vary over a continuous period of time. They can also be used to compare two different variables over time.
- Typical examples of the types of data that can be presented using line graphs are monthly rainfall and annual unemployment rates.
- Line graphs are particularly useful for identifying patterns and trends in the data such as seasonal effects, large changes and turning points. Fig. 6.3.1 show line graph.
- As well as time series data, line graphs can also be appropriate for displaying data that are measured over other continuous variables such as distance.
- For example, a line graph could be used to show how pollution levels vary with increasing distance from a source, or how the level of a chemical varies with depth of soil.

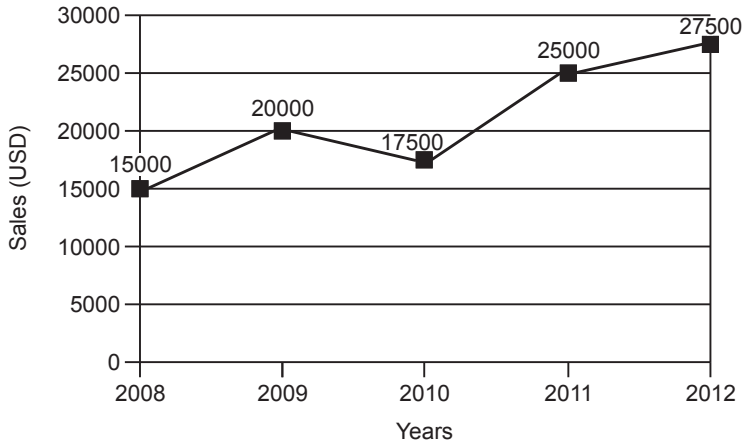


Fig. 6.3.1 Line graph

- In a line graph the x-axis represents the continuous variable (for example year or distance from the initial measurement) whilst the y-axis has a scale and indicates the measurement.
- Several data series can be plotted on the same line chart and this is particularly useful for analysing and comparing the trends in different datasets.
- Line graph is often used to visualize the rate of change of a quantity. It is more useful when the given data has peaks and valleys. Line graphs are very simple to draw and quite convenient to interpret.

6.3.2 Pie Chart

- A type of graph in which a circle is divided into sectors that each represent a proportion of the whole. Each sector shows the relative size of each value.
- A pie chart displays data, information and statistics in an easy to read "pie slice" format with varying slice sizes telling how much of one data element exists.
- Pie chart is also known as circle graph. The bigger the slice, the more of that particular data was gathered. The main use of a pie chart is to show comparisons. Fig. 6.3.2 shows pie chart.

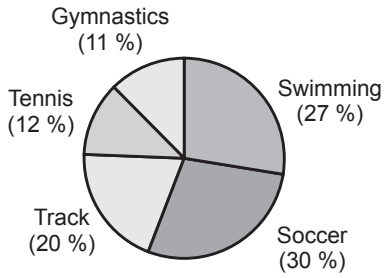


Fig. 6.3.2 Pie chart

- Various applications of pie charts can be found in business, school and at home. For business pie charts can be used to show the success or failure of certain products or services.

- At school, pie chart applications include showing how much time is allotted to each subject. At home pie charts can be useful to see expenditure of monthly income in different needs.
- Reading of pie chart is as easy as figuring out which slice of an actual pie is the biggest.
- Pie charts can be drawn using the function pie() in the pyplot module. The below python code example draws a pie chart using the pie() function. By default the pie() fucntion of pyplot arranges the pies or wedges in a pie chart in counter clockwise direction.

```
# import the pyplot library
import matplotlib.pyplot as plotter

# The slice names of a student distribution pie chart
pieLabels = 'Rakshita', 'Ritesh', 'Rupali', 'Rutu', 'Rushi', 'Radhika'

# marks data
marksShare = [59.69, 16, 9.94, 7.79, 5.68, 0.54]
figureObject, axesObject = plotter.subplots()

# Draw the pie chart
axesObject.pie(marksShare, labels=pieLabels, autopct='%1.2f', startangle=90)

# Aspect ratio - equal means pie is a circle
axesObject.axis('equal')
plotter.show()
```

- The essential part of a pie chart is the values. You could create a basic pie chart using just the values as input.
- Limitations of pie chart :
  a) It is difficult to tell the difference between estimates of similar size.
  b) Error bars or confidence limits cannot be shown on pie graph.
  c) Legends and labels on pie graphs are hard to align and read.
  d) The human visual system is more efficient at perceiving and discriminating between lines and line lengths rather than two-dimensional areas and angles.
  e) Pie graphs simply don't work when comparing data.

6.3.3 Venn Diagram

- Venn diagram is a diagram that visually displays all the possible logical relationships between collections of sets. Each set is typically represented with a circle.

- Venn diagram shows the similarities and differences of two or more data sets by using overlapping circles. The overlapping areas show the similarities and the non-overlapping areas show the differences.
- Venn diagrams may also be called primary diagrams, set diagrams, or logic diagrams.
- Venn diagrams can be useful tools for analysis or support the decision-making process. Although Venn diagrams can have unlimited circles (each circle representing a data set), they usually have just two or three overlapping circles.
- By the size of the circle, we can show the importance of an organization or projects. The bigger a circle is, the more important is a project.
- Overlapping circles represent interacting organizations. There is also the possibility of a subset. This means that a small circle is placed within a larger circle.
- The small circle stands for a component in a big organization or project which is symbolized by a big circle.
- Example : There are a total of 55 books, 23 available in hard copy, 20 available on Kindle, and 12 books available in both formats. Fig. 6.3.3 shows venn diagram of this data.

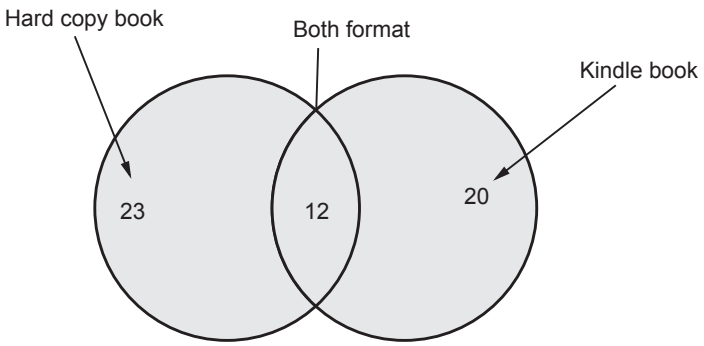


Fig. 6.3.3 : Venn diagram

- Venn diagrams can become much more complex with more data sets and are often shaded to help better visualize the relationships between data sets.

Advantages of Venn Diagram

1. Easy way to show similarities and differences amongst systems
2. Works without much technical equipment
3. A tool which is easy to understand and to use
4. Clearly orientated towards output
5. To solve complex mathematical problem.

Disadvantages of Venn Diagram

1. Venn diagram is often a snapshot of a group interaction and negotiations
2. Growing complexity if more than four circles are drawn
3. If the Venn diagrams are done by groups, the views of weaker actors are likely to be submerged.

6.3.4 Scatter Diagram

- **Scatter diagram** is also called scatter plot, X-Y graph. The scatter plot is the model of data visualization depicting two sets of unconnected dots as parameter values.
- Scatter plots which use horizontal and vertical axes to plot data points and display how much one variable is affected by another. The position of each dot on the horizontal and vertical axis indicates values for an individual data point.
- Fig. 6.3.4 shows scatter plots of two variables.

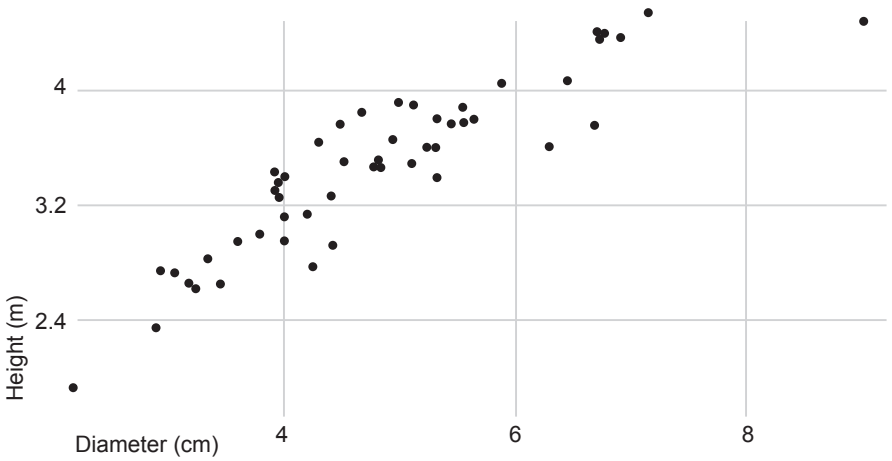


Fig. 6.3.4 : Scatter plot

- The example scatter plot above shows the diameters and heights for a sample of fictional trees. Each dot represents a single tree; each point's horizontal position indicates that tree's diameter (in centimeters) and the vertical position indicates that tree's height (in meters).
- From the plot, we can see a generally tight positive correlation between a tree's diameter and its height. We can also observe an outlier point, a tree that has a much larger diameter than the others.
- While working with statistical data it is often observed that there are connections between sets of data.

- A scatter diagram is a tool for analyzing relationships between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis.
- The pattern of their intersecting points can graphically show relationship patterns. Commonly a scatter diagram is used to prove or disprove cause-and-effect relationships.
- Scatter plot's primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole. Identification of correlational relationships are common with scatter plots.
- A scatter plot can also be useful for identifying other patterns in data. We can divide data points into groups based on how closely sets of points cluster together. Scatter plots can also show if there are any unexpected gaps in the data and if there are any outlier points. This can be useful if we want to segment the data into different parts, like in the development of user personas.

Merits :

a) Scatter diagrams are easy to draw.

b) It can be easily understood and interpreted.

c) Shows both positive and negative types of graphical correlation.

Demerits :

a) You cannot use scatter diagrams to show the relation of more than two variables.

b) Interpretation can be subjective.

Review Questions

1. *Explain how data visualization is done or visually represented, if data is 1-D, if data 2-D and data is 3-Dimentional ?* **SPPU : Dec.-18 (End Sem), Marks 6**
2. *Explain analytical techniques used in big data visualization.* **SPPU : Dec.-18 (End Sem), Marks 3**
3. *Why it is difficult to visualize big data ?* **SPPU : Dec.-19 (End Sem), Marks 8**

6.4 Visualizing Big Data

SPPU : Dec.-18, 19, May-19

- Big data visualization is the process of displaying data in charts, graphs, maps and other visual forms.
- There are various analytical techniques used in big data processing in order to extract, collect, store, process and analyze the huge amount of data coming very fast with the different variety.

1. Machine Learning :

- A machine learning algorithm then takes these examples and produces a program that does the job. The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers. If we do it right, the program works for new cases as well as the ones we trained it on.
- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Another goal is to develop computational models of the human learning process and perform computer simulations.
- The goal of machine learning is to build computer systems that can adapt and learn from their experience.
- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instructions.
- For example, The addition of four numbers is carried out by giving four number as input to the algorithm and output is the sum of all four numbers. For the same task, there may be various algorithms. It is interesting to find the most efficient one, requiring the least number of instructions or memory or both.
- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behaviour of a system for any set of input values, after an initial training phase.
- In contrast to supervised learning, unsupervised or self-organized learning does not require an external teacher. During the training session, the neural network receives a number of different input patterns, discovers significant features in these patterns and learns how to classify input data into appropriate categories.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real-time. Unsupervised learning is frequently employed for data clustering, feature extraction etc.

- Reinforcement learning : This is an advanced machine learning technique. This is based on probability theory where mapping can be done based on input received and changes based on the environment around it.
- Deep learning : This is also advanced machine learning technique which has multiple processing layers so as to produce non-linear response based on input data. There are so many small processors called as **neuron working** parallel in data processing.
- Predictive analytics : This technique refers to prediction based on past experience and it uses both data mining and machine learning.
- Association rule learning : This is used to identify interesting relations between different attributes from large datasets

Review Questions

1. *Explain big data visualization tools in short (any four tools).*

SPPU : Dec.-18 (End Sem), Marks 8

2. *Explain various tools to visualize big data. (Any four)*

SPPU : May-19, Dec.-19 (End Sem), Marks 8

6.5 Tools used in Data Visualization

SPPU : May-19, Dec.-19

- Traditional data visualization tools are often inadequate to handle big data. Methods for interactive visualization of big datawere presented.
- First, design space of scalable visual summaries that use data reduction approaches was described to visualize a variety of data types.
- Methods were then developed for interactive querying among binned plots through a combination of multivariate data tiles and parallel query processing.
- Lot of big data visualization tools run on the Hadoop platform. The common modules in Hadoop are : Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop Map Reduce.
- They analyze big data efficiently, but lack adequate visualization. Some software with the functions of visualization and interaction for visualizing data has been developed.

6.5.1 Pentaho

- Pentaho tightly couple's data integration with full business analytics to solve data integration challenges while providing business analytics in a single, seamless platform.

- Pentaho's Java-based data integration engine integrates with the MapRHadoop cache for automatic deployment as a MapReduce task across every data node in a Hadoop cluster, making use of the massively parallel processing and high availability of Hadoop.
- Pentaho's open-source heritage drives our continued innovation in a modern, integrated, embeddable platform built for the future of analytics, including diverse and big data requirements.
- Within a single platform it provides visual tools to extract and prepare our data plus the visualizations and analytics that will change the way we run our business.
- Pentaho's modern, simplified and interactive approach empowers business users to access, discover and blend all types and sizes of data. With a spectrum of increasingly advanced analytics, from basic reports to predictive modeling, users can analyze and visualize data across multiple dimensions, all while minimizing dependence on IT.
- The business analytics platform is a web application that allows users to publish and manage reports within an enterprise business intelligence system.
- The business analytics platform offers many capabilities, including the management and execution of Pentaho reports. By combining Pentaho reporting and Pentaho's business analytics platform, information technologists may utilize Pentaho reporting in their environment without writing any code.

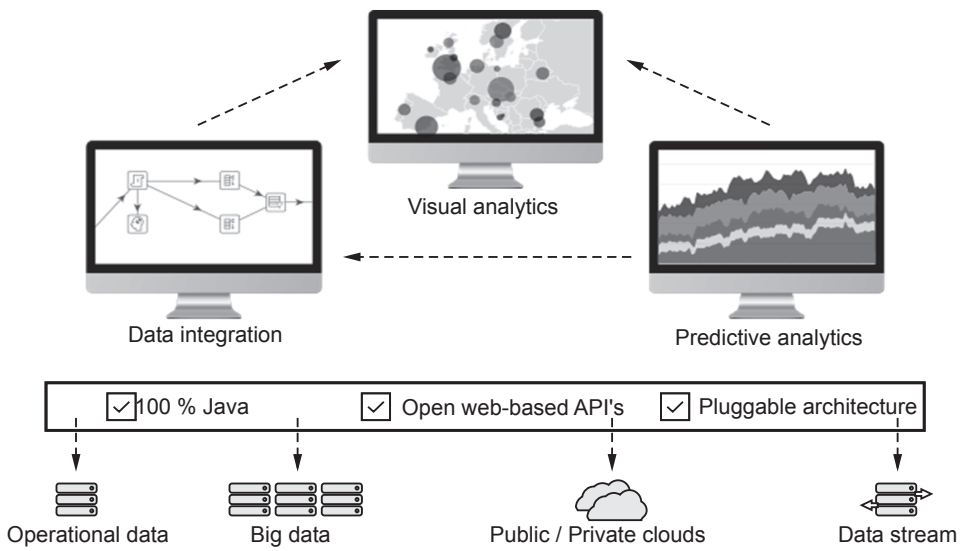


Fig. 6.5.1

- In addition to the publishing and execution of reports, the open source business analytics platform allows for scheduling, background execution, security and much more.

Advantages

1. Pentaho is an intuitive platform, where IT as well as business people can access and visualize data easily.
2. Easy access to data from diverse sources ranging from Excel to Hadoop.
3. Reporting is fast due to in-memory caching techniques.
4. Detailed visualisation and easy to understand infographics, with drilling and filters available. Seamless integration with third party applications, such as Google Maps.
5. The devices supported covers almost every platform : Android, iPhone, iPad, Mac, Web-based, Windows.

Disadvantages

1. All the products in Pentaho suite are inconsistent in the manner in which they work.
2. The metadata layer is cumbersome to use and understand. The documentation also is of little help at times.
3. There is no system of perpetual licensing. The usage rights have to be bought every year, at the same price.
4. Advanced analytics and corresponding data visualisation need more improvement, when compared with the same in Tableau.

6.5.2 Datameer

- Datameer's flipside provides simple, highly accessible, visual data profiling that lets users easily spot outliers in data, quickly and early in the analytics process. Datameer runs natively on Hadoop.
- Datameer, an end-to-end big data analytics platform, is built on Apache Hadoop to perform integration, analysis and visualization of massive volumes of both structured and unstructured data. It can be rapidly integrated with any data sources such as new and existing data sources to deliver an easy-to-use, cost-effective and sophisticated solution for big data analytics.
- It simplifies data extraction, data transformation, data loading and real-time data retrieval. It helps gain actionable insights from complex organizational data through data preparation and analytics.

- Fig. 6.5.2 shows all Datameer functionality occurs across three major components.

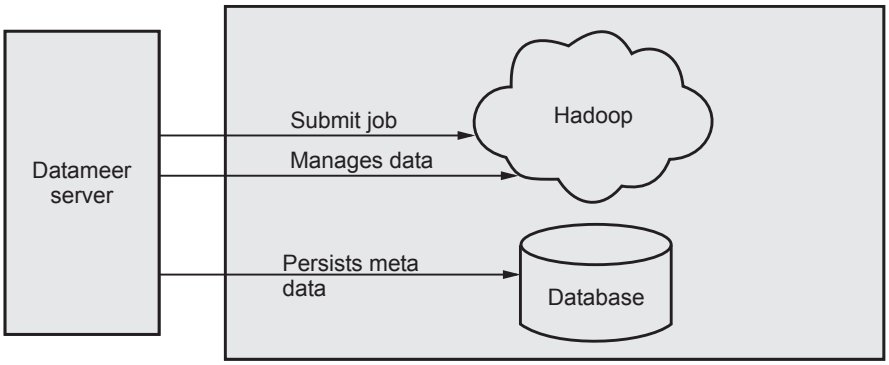


Fig. 6.5.2 Datameer functionality

- The Datameer server : Server is also called conductor. This server orchestrates all work and manages the configuration of all jobs performed on the Hadoop cluster. It also hosts the web app that lets users interact via the software's web UI. All processing done during the design of a workbook in real time on the Datameer server. Datameer provides real-time feedback during the design phase using intelligent previews generated by our smart sampling technology.
- Database for metadata storage : Datameer uses a database to store all metadata.
- Hadoop cluster : The Hadoop cluster provides persistent storage for all data, pre-views and other job artifacts, as well as a big data processing framework for executing long-running operations.
- Fundamental to the design of Datameer software is the fact that all resource-intensive processes are submitted to Hadoop clusters. This approach allows Datameer to scale up and scale out easily by distributing work across the entire Hadoop cluster.

6.5.3 JasperReport

- JasperReports is a powerful open source reporting package, but generating reports with data from multiple sources is hard and often impossible without the enterprise version.
- Fig. 6.5.3 shows JasperReport.

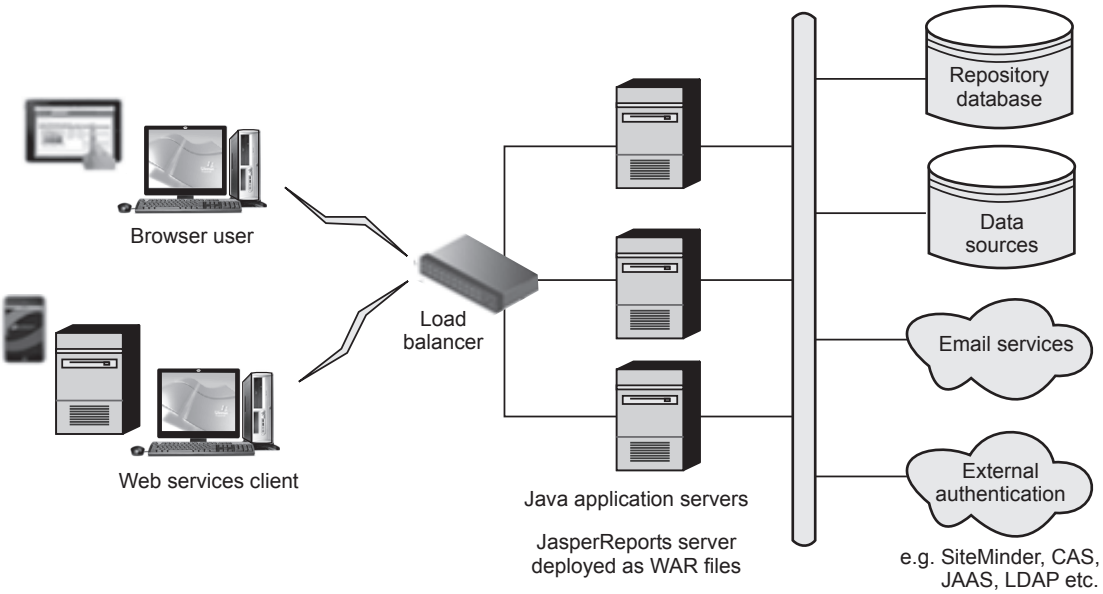


Fig. 6.5.3 JasperReport

- JasperReports is an open source java reporting engine. JasperReports is a Java class library and it is meant for those Java developers who need to add reporting capabilities to their applications.
- The main purpose of JasperReports is to create page oriented, ready to print documents in a simple and flexible manner.
- JasperReports Server is a stand-alone and embeddable reporting server.
- It provides reporting and analytics that can be embedded into a web or mobile application as well as operate as a central information hub for the enterprise by delivering mission critical information on a real-time or scheduled basis to the browser, mobile device, or email inbox in a variety of file formats.
- JasperReports Server is optimized to share, secure and centrally manage Jaspersoft reports and analytic views.
- Data sources are structured data containers. While generating the report, JasperReports engine obtains data from the datasources. Data can be obtained from the databases, XML files, arrays of objects and collection of objects.
- JasperReports has a feature <style> which helps to control text properties in a report template. This element is a collection of style settings declared at the report level.

- Properties like foreground color, background color, whether the font is bold, italic, or normal, the font size, a border for the font and many other attributes are controlled by <style> element.

6.5.4 Dygraphs

- Dygraphs is an open-source JavaScript library that produces interactive, zoomable charts of time series. It is designed to display dense data sets and enable users to explore and interpret them.
- It can handle large data sets with millions of plot points. It works in all browsers and zooms down for mobile devices. The dygraphs package is an R interface to the dygraphs JavaScript charting library.
- This library can be used to develop interactive charts on the X and Y axis and to display powerful diagrams. Dygraphs.js can use five types of input : CSV data, URL, array, function, DataTable.
- Some of the features of dygraphs :
  1. Plots time series without using an external server or flash
  2. Works in Internet Explorer (using excanvas)
  3. Lightweight (69 kb) and responsive
  4. Displays values on mouseover, making interaction easily discoverable
  5. Supports error bands around data series
  6. Interactive zoom
  7. Displays annotations on the chart
  8. Adjustable averaging period
  9. Can intelligently chart fractions
  10. Customizable click-through actions
  11. Compatible with the Google Visualization API.
- The dygraphs package is available on CRAN now and can be installed with :

```
install.packages("dygraphs")
```

- Dygraphs work primarily with time series. If you have a DSS dataset with a "date" column, you'll need to convert your dataframe to a time series or XTS object.
- For example, the following will create a time-series of revenue by order_ts

```
library(xts)
df <- dkuReadDataset("orders")
timeseries <- xts(df$revenue, order.by=as.Date(df$order_ts))
# You can then plot timeseries
dkuDisplayDygraph(dygraph(timeseries) %>% dyRangeSelector())
```

- It allows users to explore and interpret dense data sets. All the charts are inter-active : It can be used mouse over to highlight individual values, or click and drag to zoom. It is possible to change the number and hit enter to adjust the averaging period. Dygraphs handles huge data sets.

6.5.5 Tableau

- Tableau is one of the fastest evolving Business Intelligence (BI) and data visualization tools. Tableau server is a business intelligence application that provides browser-based analytics anyone can use. It's a rapid-fire alternative to the slow pace of traditional business intelligence software.
- A business intelligence and data visualization tool allowing users to make sense of their data through interactive charts, graphs and diagrams.
- Why use Tableau ?
  1. Traditional BI tools require complex installations
  2. Rapid results to useful information
  3. Easy to use for all skill levels
  4. Excellent migration path for excel users
  5. It can use many different sources of data.
- Tableau uses a visual query language. The tableau data engine is a breakthrough in-memory analytics database designed to overcome the limitations of existing databases and data silos.
- Capable of being run on ordinary computers, it leverages the complete memory hierarchy from disk to L1 cache. It shifts the curve between big data and fast analysis.
- Tableau allows the users to directly connect to databases, cubes and data warehouses etc. After analysing the data, the results can be shared live with just a few clicks. The dashboard can be published to share it live on web and mobile devices.
- Tableau is relatively new in the business intelligence market but its market share is growing on a daily basis. It is being nearly all industries, from transportation to healthcare used Tableau.
- Tableau software does not support expanded analytics such as box plots, network graphs, tree - maps, heat-maps, 3D-scatter plots, profile charts or data relationships tools which allow users to mine data for relationships like another data visualization software does.

- Tableau connects and extracts the data stored in various places. It can pull data from any platform imaginable. A simple database such as an excel, pdf, to a complex database like Oracle, a database in the cloud such as Amazon web services, Microsoft Azure SQL database, Google Cloud SQL and various other data sources can be extracted by Tableau.
- Tableau saves time when updating daily and weekly reports that currently reside in spreadsheets. That's because Tableau separates the data layer from the presentation layer and makes updating a spreadsheet data source a trivial append to the bottom of your source spreadsheet.
- Tableau is not an ETL engine for cleaning-up bad data, although it can be very helpful in identifying missing or erroneous data in existing data sources. Visualizing data via time series, bar charts, scatter plots or in maps highlights errors and outliers more effectively than grids of data in a spreadsheet.

6.5.6 1-D, 2-D and 3-D Data

- Every data set has a general structure. It is always characterised by a group of variables and the records the database contains. The first group consists of one-dimensional, two-dimensional, three-dimensional and high-dimensional data sets.
- The variable in one-dimensional data is usually time. An example is the log of interrupts in a processor.
- Two-dimensional data can often be found in statistics like the number of financial transactions in a certain period of time.
- Three dimensional data can be positioned in three-dimensional space or points on a surface whereas time varies. High-dimensional data contains all those sets of data that have more than three considered variables. Examples are locations in space that vary with time.
- **Two-dimensional data** can be visualized in different ways. A very common visualization form is the **scatter plot.** In a scatterplot the frame for the data presentation is a Cartesian coordinate system, in which the axes correspond to the two dimensions.
- Another important visualization technique for two-dimensional data is the line graph. The difference to scatter plots is that this time the relation between the dimension on the horizontal axis and the one on the vertical axis is definite.

- **Three-dimensional data :** The two-dimensional techniques can easily be extended to three dimensions. The third dimension is achieved in scatter plots and bar charts by adding a further axis, orthogonal to the other two.
- A scatter plot, more commonly called a graph of y versus x, shows the relationship of 2 variables and with the addition of colour can represent a 3rd variable. A scatterplot matrix of n variables is obtained by projection of the data onto n*(n–1) scatter plots, i.e., all possible combinations of scatter plots are drawn as illustrated in Fig. 6.5.4 which is an example for pressure, temperature and velocity data.

Pressure	PvT	PvV
TvP	Temperature	TvV
VvP	VvT	Velocity

Fig. 6.5.4

Review Question

1. *Explain data visualization tool - Tableau.* **SPPU : Dec.-19, May-19 (End Sem), Marks 8**

6.6 Hadoop Ecosystem

SPPU : Dec.-18, 19, May-19

- Hadoop ecosystem is neither a programming language nor a service, it is a platform or framework which solves big data problems.
- The Hadoop ecosystem refers to the various components of the Apache Hadoop software library, as well as to the accessories and tools provided by the Apache software foundation for these types of software projects and to the ways that they work together.
- Hadoop is a Java-based framework that is extremely popular for handling and analysing large sets of data. The idea of a Hadoop ecosystem involves the use of different parts of the core Hadoop set such as MapReduce, a framework for handling vast amounts of data and the Hadoop Distributed File System (HDFS), a sophisticated file-handling system. There is also YARN, a Hadoop resource manager.
- In addition to these core elements of Hadoop, Apache has also delivered other kinds of accessories or complementary tools for developers.
- Some of the most well-known tools of the Hadoop ecosystem include HDFS, Hive, Pig, YARN, MapReduce, Spark, HBase, Oozie, Sqoop, Zookeeper, etc.

- Fig. 6.6.1 shows Apache Hadoop ecosystem.

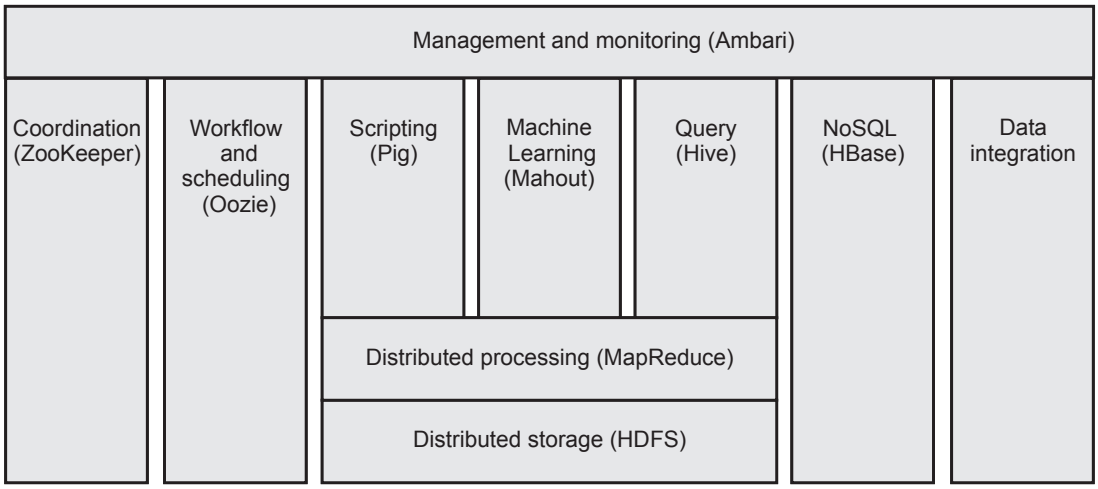


Fig. 6.6.1 : Apache Hadoop ecosystem

- Hadoop Distributed File System (HDFS), is one of the largest Apache projects and primary storage system of Hadoop. It employs a NameNode and DataNode architecture. It is a distributed file system able to store large files running over the cluster of commodity hardware.
- YARN stands for Yet Another Resource Negotiator. It is one of the core components in open source Apache Hadoop suitable for resource management. It is responsible for managing workloads, monitoring and security controls implementation.
- Hive is an ETL and data warehousing tool used to query or analyze large datasets stored within the Hadoop ecosystem. Hive has three main functions : Data summarization, query and analysis of unstructured and semi-structured data in Hadoop.
- Map-Reduce : It is the core component of processing in a Hadoop ecosystem as it provides the logic of processing. In other words, Map-Reduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside the Hadoop environment.
- Apache Pig is a high-level scripting language used to execute queries for larger datasets that are used within Hadoop.
- Apache Spark is a fast, in-memory data processing engine suitable for use in a wide range of circumstances. Spark can be deployed in several ways, it features Java, Python, Scala, and R programming languages and supports SQL, streaming

data, machine learning and graph processing, which can be used together in an application.

- Apache HBase is a Hadoop ecosystem component which is a distributed database that was designed to store structured data in tables that could have billions of rows and millions of columns. HBase is a scalable, distributed, and NoSQL database that is built on top of HDFS. HBase provide real-time access to read or write data in HDFS.

6.6.1 Hadoop Architecture

- Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It provides a software framework for distributed processing of large datasets in real-time applications.
- Hadoop manages to process and store vast amounts of data by using interconnected affordable commodity hardware. Hundreds or even thousands of low-cost dedicated servers working together to store and process data within a single ecosystem.
- Hadoop provides the basic platform for big data processing. The Hadoop architecture has mainly two parts : Hadoop Distributed File System (HDFS) and the MapReduce engine.
- Fig. 6.6.2 shows HDFS archtiecture

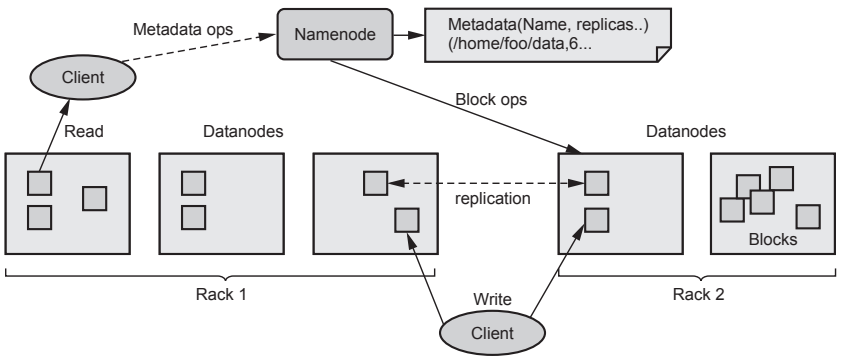


Fig. 6.6.2 Hadoop architecture

- Hadoop distributed file system is a block-structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines.

- Apache Hadoop HDFS architecture follows a master/slave architecture, where a cluster comprises of a single NameNode (Master node) and all the other nodes are DataNodes (Slave nodes).
- DataNodes process and store data blocks, while NameNodes manage the many DataNodes, maintain data block metadata and control client access.

1. NameNode and DataNode

- Namenode holds the meta data for the HDFS like Namespace information, block information etc. When in use, all this information is stored in main memory. But this information also stored in disk for persistence storage.
- Namenode manages the file system namespace. It keeps the directory tree of all files in the file system and metadata about files and directories.
- DataNode is a slave node in HDFS that stores the actual data as instructed by the NameNode. In brief, NameNode controls and manages a single or multiple data nodes.
- DataNode serves to read or write requests. It also creates, deletes and replicates blocks on the instructions from the NameNode.
- Fig. 6.6.3 shows Namenode. It shows how NameNode stores information on disk.

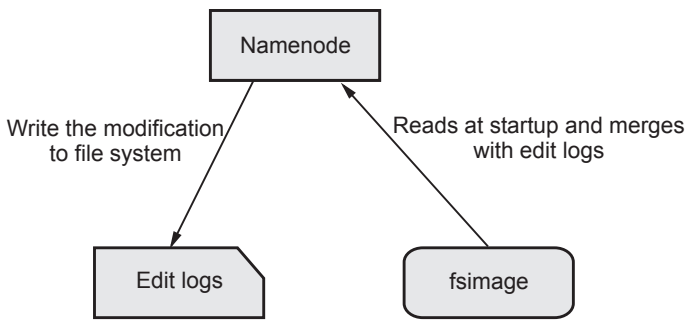


Fig. 6.6.3 Name node

- Two different files are :
  1. **fsimage :** It's the snapshot of the file system when name node started.
  2. **Edit logs :** It's the sequence of changes made to the file system after name node started.
- Only in the restart of namenode, edit logs are applied to fsimage to get the latest snapshot of the file system.

- But namenode restart are rare in production clusters which means edit logs can grow very large for the clusters where namenode runs for a long period of time.
- The following issues we will encounter in this situation :
  1. Editlog become very large, which will be challenging to manage it.
  2. Namenode restart takes long time because lot of changes to be merged.
  3. In the case of crash, we will lost huge amount of metadata since fsimage is very old.
- So to overcome this issues we need a mechanism which will help us reduce the edit log size which is manageable and have up to date fsimage, so that load on namenode reduces.
- Secondary Namenode helps to overcome the above issues by taking over responsibility of merging editlogs with fsimage from the namecode.
- Fig. 6.6.4 shows secondary Namenode.

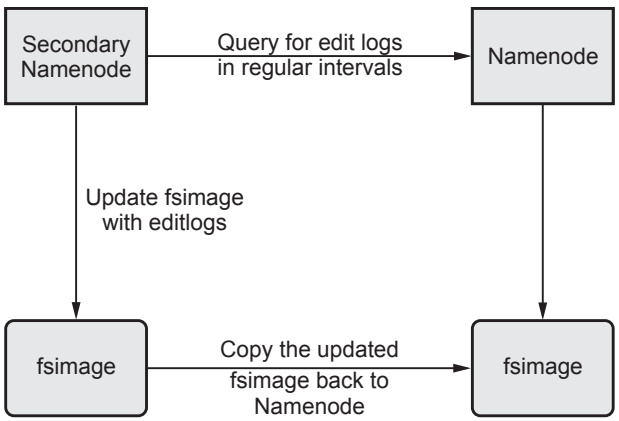


Fig. 6.6.4 Secondary Namenode

- Working of secondary Namenode :
  1. It gets the edit logs from the Namenode in regular intervals and applies of fsimage.
  2. Once it has new fsimage, it copies back to Namenode.
  3. Namenode will use this fsimage for the next restart, which will reduce the startup time.
- Secondary Namenode's whole purpose is to have a checkpoint in HDFS. Its just a helper node for Namecode. That's why it also known as checkpoint node inside the community.

Hadoop Distributed File System :

- Hadoop Distributed File System (HDFS) is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds of nodes.
- A block is the minimum amount of data that it can read or write. HDFS blocks are 128 MB by default and this is configurable. When a file is saved in HDFS, the file is broken into smaller chunks or "blocks".
- HDFS is a fault-tolerant and resilient system, meaning it prevents a failure in a node from affecting the overall system's health and allows for recovery from failure too. In order to achieve this, data stored in HDFS is automatically replicated across different nodes.
- HDFS supports a traditional hierarchical file organization. A user or an application can create directories and store files inside these directories. The file system namespace hierarchy is similar to most other existing file systems; one can create and remove files, move a file from one directory to another, or rename a file.
- Hadoop Distributed File System is a block-structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines.
- Apache Hadoop HDFS Architecture follows a Master/Slave Architecture, where a cluster comprises of a single NameNode (MasterNode) and all the other nodes are DataNodes (Slave nodes).
- HDFS can be deployed on a broad spectrum of machines that support Java. Though one can run several DataNodes on a single machine, but in the practical world, these DataNodes are spread across various machines

6.6.2 MapReduce

- MapReduce is a programming model and software framework first developed by Google. Intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.
- Characteristics of MapReduce :
  1. Very large scale data : peta, exa bytes
  2. Write once and read many data. It allows for parallelism without mutexes
  3. Map and reduce are the main operations : simple code
  4. All the maps should be completed before reduced operation starts

5. Map and reduce operations are typically performed by the same physical processor
6. Number of map tasks and reduced tasks are configurable.

- MapReduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside the Hadoop environment.
- In a MapReduce program, Map() and Reduce() are two functions.
  1. The Map function performs actions like filtering, grouping and sorting.
  2. While the reduce function aggregates and summarizes the result produced by the map function.
  3. The result generated by the Map function is a key value pair (K, V) which acts as the input for Reduce function.
- MapReduce works by breaking the processing into two phases :
  1. Map phase 2. Reduce phase
- Each phase has key-value pairs as input and output. In addition the programmer also specifies two functions: map function and reduce function.
- Map function takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
- Reduce function takes the output from the Map as an input and combines those data tuples based on the key and accordingly modifies the value of the key.
- Every Map/Reduce program must specify a Mapper and typically a Reducer. The Mapper has a map method that transforms input (key, value) pairs into any number of intermediate (key', value') pairs. The Reducer has a reduce method that transforms intermediate (key', value'*)aggregates into any number of output (key'', value'') pairs.
- Fig. 6.6.5 shows MapReduce logical data flow.

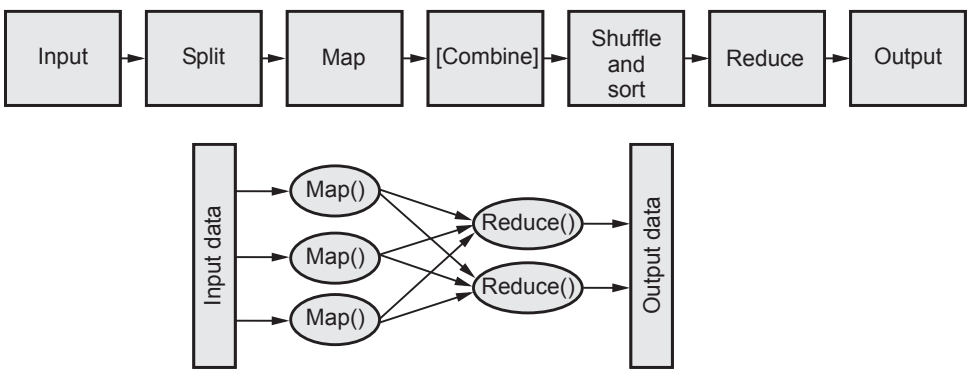


Fig. 6.6.5 Map-Reduce logical data flow

1. **Input :** This is the input data / file to be processed.
2. **Split :** Hadoop splits the incoming data into smaller pieces called "splits".
3. **Map :** In this step, MapReduce processes each split according to the logic defined in map() function. Each mapper works on each split at a time. Each mapper is treated as a task and multiple tasks are executed across different TaskTrackers and coordinated by the JobTracker.
4. **Combine :** This is an optional step and is used to improve the performance by reducing the amount of data transferred across the network. Combiner is the same as the reduce step and is used for aggregating the output of the map() function before it is passed to the subsequent steps.
5. **Shuffle and Sort :** In this step, outputs from all the mappers are shuffled, sorted to put them in order and grouped before sending them to the next step.
6. **Reduce :** This step is used to aggregate the outputs of mappers using the reduce() function. Output of reducer is sent to the next and final step. Each reducer is treated as a task and multiple tasks are executed across different TaskTrackers and coordinated by the JobTracker.
7. **Output :** Finally the output of reduce step is written to a file in HDFS.

- Map tasks write their output to the local disk, not to HDFS. Map output is intermediate output : It's processed by reducing tasks to produce the final output and once the job is complete, the map output can be thrown away. So, storing it in HDFS with replication would be overkill. If the node running the map task fails before the map output has been consumed by the reduced task then Hadoop will automatically rerun the map task on another node to re-create the map output.
- The map tasks partition their output, each creating one partition for each reduce task. There can be many keys and their associated values in each partition, but the records for any given key are all in a single partition. The partitioning can be controlled by a user-defined partitioning function, but normally the default partitioner, which buckets keys using a hash function, works very well.
- The partitions are sorted and transferred across the network to the node where the respective reduce task is running, where they are merged and passed to the user-defined reduce function.
- Consider an **ecommerce system** that receives a million requests every day to process payments. There may be several exceptions thrown during these requests such as "payment declined by a payment gateway," "out of inventory," and "invalid address."

- A developer wants to analyze last four days' logs to understand which exception is thrown how many times.

1. Map

- Let's assume that the Hadoop framework runs just four mappers. Mapper 1, Mapper 2, Mapper 3 and Mapper 4.
- The value input to the mapper is one record of the log file. The key could be a text string such as "file name + line number." The mapper, then, processes each record of the log file to produce key value pairs. Here, we will just use a filler for the value as '1.' The output from the mappers look like this :

```
Mapper 1 -> <Exception A, 1>, <Exception B, 1>, <Exception A, 1>, <Exception C, 1>,
<Exception A, 1>
Mapper 2 -> <Exception B, 1>, <Exception B, 1>, <Exception A, 1>, <Exception A, 1>
Mapper 3 -> <Exception A, 1>, <Exception C, 1>, <Exception A, 1>, <Exception B, 1>,
<Exception A, 1>
Mapper 4 -> <Exception B, 1>, <Exception C, 1>, <Exception C, 1>, <Exception A, 1>
```

- Assuming that there is a combiner running on each mapper - Combiner 1 … Combiner 4 - that calculates the count of each exception (which is the same function as the reducer), the input to Combiner 1 will be :

```
<Exception A, 1>, <Exception B, 1>, <Exception A, 1>, <Exception C, 1>,
<Exception A, 1>
```

2. Combine : The output of Combiner 1 will be :

```
<Exception A, 3>, <Exception B, 1>, <Exception C, 1>
```

- The output from the other combiners will be :

```
Combiner 2: <Exception A, 2> <Exception B, 2>
Combiner 3: <Exception A, 3> <Exception B, 1> <Exception C, 1>
Combiner 4: <Exception A, 1> <Exception B, 1> <Exception C, 2>
```

3. Partition : After this, the partitioner allocates the data from the combiners to the reducers. The data is also sorted for the reducer.

- The input to the reducers will be as below :

```
Reducer 1: <Exception A> {3,2,3,1}
Reducer 2: <Exception B> {1,2,1,1}
Reducer 3: <Exception C> {1,1,2}
```

- If there were no combiners involved, the input to the reducers will be as below :

```
Reducer 1: <Exception A> {1,1,1,1,1,1,1,1,1}
Reducer 2: <Exception B> {1,1,1,1,1}
Reducer 3: <Exception C> {1,1,1,1}
```

- Now, each reducer just calculates the total count of the exceptions as :

```
Reducer 1: <Exception A, 9>
Reducer 2: <Exception B, 5>
Reducer 3: <Exception C, 4>
```

- The data shows that exception A is thrown more often than others and requires more attention. When there are more than a few weeks or months of data to be processed together, the potential of the MapReduce program can be truly exploited.
- With HDFS, we are able to distribute the data so that data is stored on hundreds of nodes instead of a single large machine. Mapreduce provides the framework for highly parallel processing of data across clusters of commodity hardware.
  Fig. 6.6.6 shows MapReduce data processing.

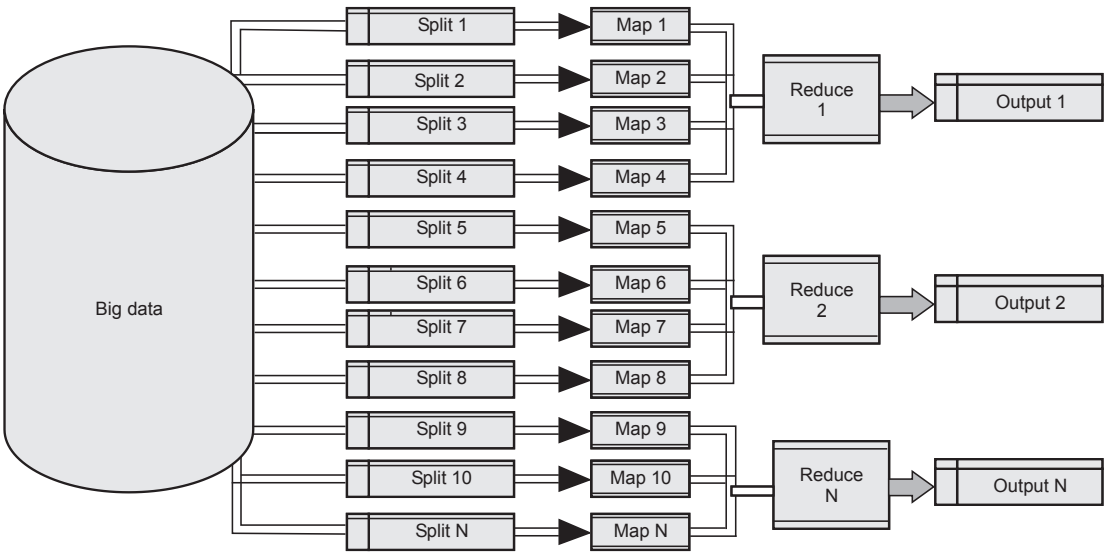


Fig. 6.6.6 Map-Reduce data processing

- It removes the complicated programming part from the programmers and moves into the framework. Programmers can write simple programs to make use of the parallel processing.
- The framework splits the data into smaller chunks that are processed in parallel on cluster of machines by programs called **mappers.**
- The output from the mappers is then consolidated by reducers into desired result. The share nothing architecture of mappers and reducers make them highly parallel.
- Data locality is achieved by mapreduce by working closely with HDFS. When you specify the file system as HDFS for mapreduce, it automatically schedules the mappers on the same node as where the block of data exists.

- Mapreduce can get the blocks from HDFS and process them. The final output from Mapreduce also can be stored in HDFS file system. However, the intermediate files between mappers and reducers are not stored in HDFS and are stored on the local file system of the mappers.

Function of job tracker :

- There is a single job tracker that runs on the master node. It is the driver for the mapreduce jobs. Its functions are :
  1. Accepts jobs from client and divides into tasks
  2. Schedules tasks on worker nodes called **task trackers**
  3. Keeps heartbeat info from task trackers on worker nodes
  4. Reschedules the task on alternate worker if a worker fails.

Function of task tracker :

- Task tracker runs on each worker node and there are as many task trackers as the worker nodes. If HDFS is also used, then data nodes of HDFS also become worker nodes for task tracker. The functions of a task tracker are :
  a. Takes assignments from job tracker
  b. Executes the tasks locally
  c. Each worker node has specific number of mapper and reducer tasks it can take at one time
  d. The tasks assigned are run in parallel
  e. Normally they can take more map jobs than reduce tasks
  f. Task tracker does a task attempt before executing task
  g. Task tracker may do multiple attempts before declaring a task as failed.
  h. Task tracker maintains a connection with the task attempt called **umbilical** protocol
  i. Task tracker sends a regular heartbeat signal to job tracker indicating its status including available map and reduce tasks
  j. Task tracker runs each task attempt in a separate JVM. So even if the task has bad code due to which it fails, it will not cause task tracker to abort.

- Hadoop configs are contained under /etc/hadoop/conf in CDH

Sr. No.	Name of File	Description
1.	hadoop-env.sh	• Used for environment-specific settings • It update the JAVA path to configure user JAVA_HOME. • It also specify JVM options for various Hadoop components
2.	core-site.xml	• System-level Hadoop configuration items, such as the HDFS URL, • It configure the Hadoop temporary directory and script locations for rack-aware Hadoop clusters
3.	hdfs-site.xml	• Used for HDFS settings such as File replication count, the block size, permissions
4.	mapred-site.xml	• Hadoop distributed file settings i.e. no. of reduce tasks, memory sizes
5.	Masters	• List of hosts that are Hadoop masters, i.e. secondary name nodes
6.	Slaves	• List of set of hosts that are going to act as slaves

- The default settings for above configuration are available at
  http://hadoop.apache.org/common/docs/r1.0.0/core-default.html

6.6.3 Pig

- Pig is an open-source high level data flow system. A high-level platform for creating MapReduce programs used in Hadoop. It translates into efficient sequences of one or more MapReduce jobs.
- Pig offers a high-level language to write data analysis programs which we call as Pig Latin. The salient property of pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.
- Pig makes use of both, the Hadoop Distributed File System as well as the MapReduce.

Features of Pig Hadoop :

1. In-built operators : Apache Pig provides a very good set of operators for performing several data operations like sort, join, filter, etc.
2. Ease of programming.
3. Automatic optimization : The tasks in Apache Pig are automatically optimized.
4. Handles all kinds of data : Apache Pig can analyze both structured and unstructured data and store the results in HDFS.

- Fig. 6.6.7 shows Pig architecture.

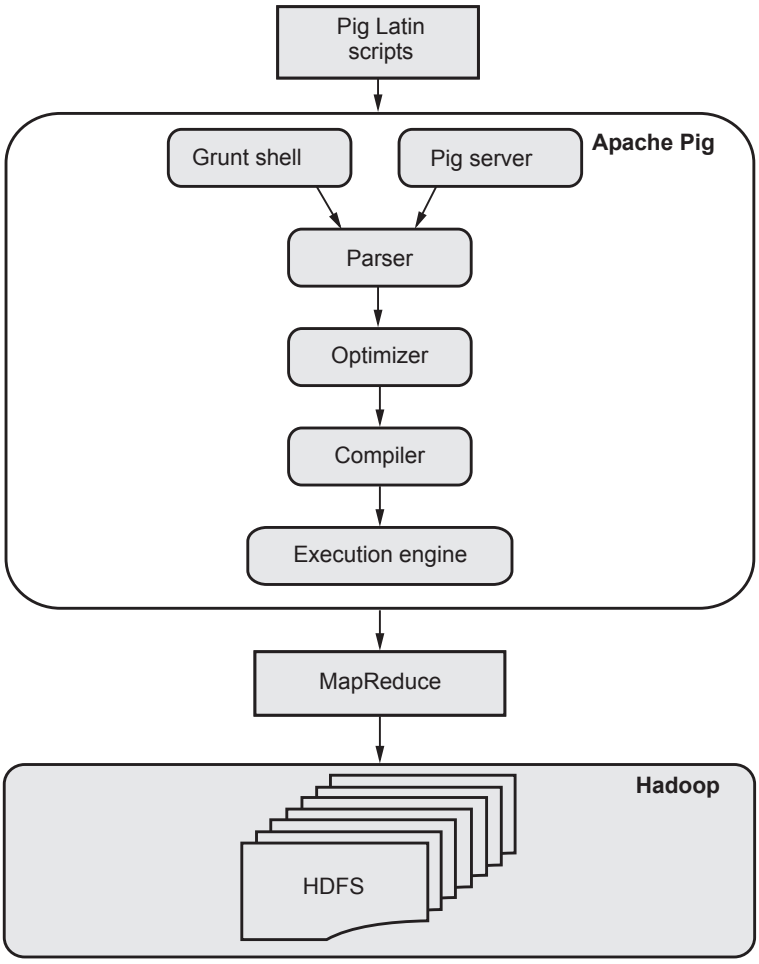


Fig. 6.6.7 Pig architecture

- Pig has two execution modes :

1. **Local mode :** To run pig in local mode, we need access to a single machine; all files are installed and run using local host and file system. Specify local mode using the -x flag (pig-x local).

2. **Mapreduce mode :** To run pig in mapreduce mode, we need access to a Hadoop cluster and HDFS installation. Mapreduce mode is the default mode; but don't need to, specify it using the -x flag

- Pig Hadoop framework has four main components :

1. **Parser :** When a Pig Latin script is sent to Hadoop Pig, it is first handled by the parser. The parser is responsible for checking the syntax of the script, along with

other miscellaneous checks. Parser gives an output in the form of a Directed Acyclic Graph (DAG) that contains Pig Latin statements, together with other logical operators represented as nodes.

2. **Optimizer :** After the output from the parser is retrieved, a logical plan for DAG is passed to a logical optimizer. The optimizer is responsible for carrying out the logical optimizations.
3. **Compiler :** The role of the compiler comes in when the output from the optimizer is received. The compiler compiles the logical plan sent by the optimize the logical plan is then converted into a series of MapReduce tasks or jobs.
4. **Execution Engine :** After the logical plan is converted to MapReduce jobs, these jobs are sent to Hadoop in a properly sorted order and these jobs are executed on Hadoop for yielding the desired result.

- Pig can run on two types of environments : The local environment in a single JVM or the distributed environment on a Hadoop cluster.
- Pig has variety of scalar data types and standard data processing options. Pig supports Map data; a map being a set of key-value pairs.
- Most pig operators take a relation as an input and give a relation as the output. It allows normal arithmetic operations and relational operations too.
- Pig's language layer currently consists of a textual language called **Pig Latin.** Pig Latin is a data flow language. This means it allows users to describe how data from one or more inputs should be read, processed and then stored to one or more outputs in parallel.
- These data flows can be simple linear flows, or complex workflows that include points where multiple inputs are joined and where data is split into multiple streams to be processed by different operators. To be mathematically precise, a Pig Latin script describes a directed acyclic graph (DAG), where the edges are data flows and the nodes are operators that process the data.
- The first step in a Pig program is to LOAD the data, which we want to manipulate from HDFS. Then run the data through a set of transformations. Finally, DUMP the data to the screen or STORE the results in a file somewhere.

Advantages of Pig :

1. Fast execution that works with MapReduce, Spark and Tez.
2. Its ability to process almost any amount of data, regardless of size.
3. A strong documentation process that helps new users learn Pig Latin.

4. Local and remote interoperability that lets professionals work from anywhere with a reliable connection.

Pig disadvantages :

1. Slow start-up and clean-up of MapReduce jobs
2. Not suitable for interactive OLAP analytics
3. Complex applications may require many user defined function.

6.6.4 Hive

- Apache Hive is an open source data warehouse software for reading, writing and managing large data set files that are stored directly in either the Apache Hadoop Distributed File System (HDFS) or other data storage systems such as Apache HBase.
- Data analysts often use Hive to analyze data, query large amounts of unstructured data and generate data summaries.
- Features of Hive :
  1. It stores schema in a database and processes data into HDFS.
  2. It is designed for OLAP.
  3. It provides SQL type language for querying called HiveQL or HQL.
  4. It is familiar, fast, scalable and extensible.
- Hive supports variety of storage formats : TEXTFILE for plaintext, SEQUENCEFILE for binary key-value pairs, RCFILE stores columns of a table in a record columnar format
- Hive table structure consists of rows and columns. The rows typically correspond to some record, transaction, or particular entity detail.
- The values of the corresponding columns represent the various attributes or characteristics for each row.
- Hadoop and its ecosystem are used to apply some structure to unstructured data. Therefore, if a table structure is an appropriate way to view the restructured data, Hive may be a good tool to use.
- Following are some Hive use cases :
  1. Exploratory or ad-hoc analysis of HDFS data : Data can be queried, transformed and exported to analytical tools.
  2. Extracts or data feeds to reporting systems, dashboards, or data repositories such as HBase.

3. Combining external structured data to data already residing in HDFS.

Advantages :

1. Simple querying for anyone already familiar with SQL.
2. Its ability to connect with a variety of relational databases, including Postgres and MySQL.
3. Simplifies working with large amounts of data.

Disadvantages :

1. Updating data is complicated
2. No real time access to data
3. High latency.

- **Program Example :** Write a code in JAVA for a simple Word Count application that counts the number of occurrences of each word in a given input set using the Hadoop Map-Reduce framework on local-standalone set-up.

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {
  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    public void map(Object key, Text value, Context context )
       throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
      }
    }
  }
  public static class IntSumReducer
       extends Reducer<Text,IntWritable,Text,IntWritable> {
```

```
  private IntWritable result = new IntWritable();
  public void reduce(Text key, Iterable<IntWritable> values,
                     Context context
                     ) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {
      sum += val.get();
    }
    result.set(sum);
    context.write(key, result);
  }
 }
 public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  Job job = Job.getInstance(conf, "word count");
  job.setJarByClass(WordCount.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setCombinerClass(IntSumReducer.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
  FileInputFormat.addInputPath(job, new Path(args[0]));
  FileOutputFormat.setOutputPath(job, new Path(args[1]));
  System.exit(job.waitForCompletion(true) ? 0 : 1);
 }
}
```

6.6.5 Difference between Pig and Hive

Sr. No.	Pig	Hive
1.	Pig used for data transformations and processing.	Hive used for warehousing and querying data.
2.	Pig works on structured, semi-structured and unstructured data.	Hive works only on structured data.
3.	Pig does not support web interface.	Hive support web interface.
4.	Pig is a scripting platform that runs on Hadoop clusters, designed to process and analyze large datasets. Pig uses a language called Pig Latin, which is similar to SQL.	Hive is a data warehouse system used to query and analyze large datasets stored in HDFS. Hive uses a query language called HiveQL, which is similar to SQL.
5.	Pig support Avro file format.	Hive does not support Avro file format.
6.	Creating schema is not required to store data in Pig.	Hive supports schema.

7.	Pig loads data quickly.	Hive takes time to load but executes quickly.
8.	Pig works on the client-side of the cluster.	Hive works on the server-side of the cluster.
9.	Used for programming.	Used for reporting.

6.6.6 HBase

- HBase is an open source, non-relational, distributed database modeled after Google's BigTable. HBase is an open source and sorted map data built on Hadoop. It is column oriented and horizontally scalable.
- It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop file system. It runs on top of Hadoop and HDFS, providing Big Table-like capabilities for Hadoop.
- HBase supports massively parallelized processing via MapReduce for using HBase as both source and sink.
- HBase supports an easy-to-use Java API for programmatic access. It also supports Thrift and REST for non-Java front-ends.
- HBase is a column oriented distributed database in Hadoop environment. It can store massive amounts of data from terabytes to petabytes. HBase is scalable, distributed big data storage on top of the Hadoop eco system.
- The HBase physical architecture consists of servers in a Master-Slave relationship. Typically, the HBase cluster has one Master node, called HMaster and multiple Region Servers called HRegionServer. Fig. 6.6.8 shows Hbase architecture.

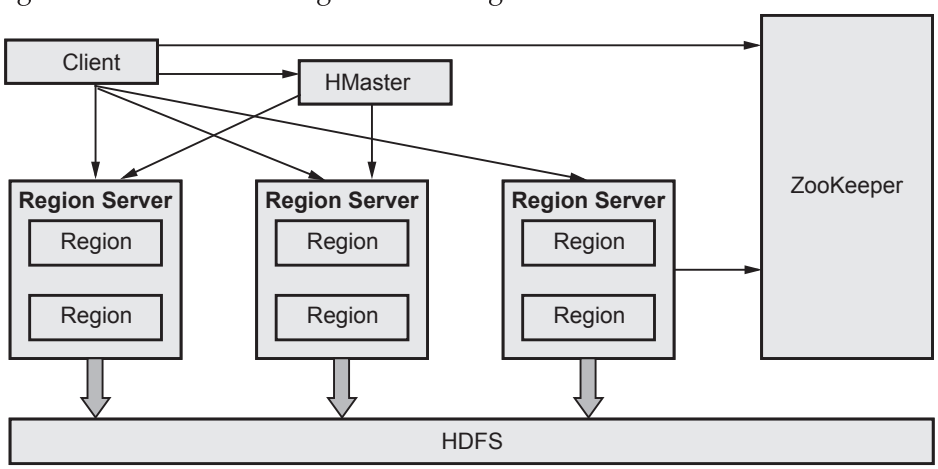


Fig. 6.6.8 Hbase architecture

- Zookeeper is a centralized monitoring server which maintains configuration information and provides distributed synchronization. If the client wants to communicate with regions servers, client has to approach Zookeeper.
- HMaster in the master server of Hbase and it coordinates the HBase cluster. HMaster is responsible for the administrative operations of the cluster.
- HRegions servers : It will perform the following functions in communication with HMaster and Zookeeper.
  1. Hosting and managing regions.
  2. Splitting regions automatically.
  3. Handling read and writes requests.
  4. Communicating with clients directly
- HRegions : For each column family, HRegions maintain a store. Main components of HRegions are Memstore and Hfile.
- Data model in HBase is designed to accommodate semi-structured data that could vary in field size, data type and columns.
- HBase is a column-oriented, non-relational database. This means that data is stored in individual columns and indexed by a unique row key. This architecture allows for rapid retrieval of individual rows and columns and efficient scans over individual columns within a table.
- Both data and requests are distributed across all servers in an HBase cluster, allowing user to query results on petabytes of data within milliseconds. HBase is most effectively used to store non-relational data, accessed via the HBase API.

6.6.7 Difference between HDFS and HBase

Sr. No.	HDFS	HBase
1.	HDFS is a distributed file system suitable for storing large files.	HBase is a database built on top of the HDFS.
2.	HDFS does not support fast individual record lookups.	HBase provides fast lookups for larger tables.
3.	It provides high latency batch processing; no concept of batch processing.	It provides low latency access to single rows from billions of records (Random access).
4.	It provides only sequential access of data.	HBase internally uses Hash tables and provides random access and it stores the data in indexed HDFS files for faster lookups.

5.	HDFS are suited for high latency operations.	HBase is suited for low latency operations.
6.	In HDFS, data are primarily accessed through Map Reduce jobs.	HBase provides access to single rows from billions of records.
7.	HDFS doesn't have the concept of random read and write operations.	HBase data is accessed through shell commands, client API in Java, REST, Avro or Thrift.

6.6.8 Mahout

- Mahout is an open source machine learning library from Apache written in java. It also supports a number of clustering algorithms like k-means, mean-shift and canopy.
- The primitive features of Apache Mahout include :
  1. The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment.
  2. Mahout uses the Apache Hadoop library to scale effectively in the cloud.
  3. Mahout offers the coder a ready-to-use framework for doing data mining tasks on large volumes of data.
  4. Mahout lets applications to analyze large sets of data effectively and in quick time
  5. Includes several MapReduce enabled clustering implementations such as k-means, fuzzy k-means, Canopy etc.
  6. Supports Distributed Naive Bayes and Complementary Naïve Bayes classification implementations.
  7. Comes with distributed fitness function capabilities for evolutionary programming.
  8. Includes matrix and vector libraries.
- Mahout is an open source machine learning library built on top of Hadoop to provide distributed analytics capabilities. Mahout incorporates a wide range of data mining techniques including collaborative filtering, classification and clustering algorithms.

Review Questions

1. *Explain MapReduce paradigm with example.* **SPPU : Dec.-18 (End Sem), Marks 6**
2. *Explain Hadoop distributed file system.* **SPPU : Dec.-18 (End Sem), Marks 5**
3. *Explain the Hadoop Ecosystem in detail with Pig, Hive, HBase and Mahout.* **SPPU : Dec.-18 (End Sem), Marks 8**
4. *Explain working of Apache Hadoop with HDFS and MapReduce.* **SPPU : Dec.-19 (End Sem), Marks 9**
5. *Explain following terms : i) Pig ii) Hive iii) HBase iv) Mahout* **SPPU : Dec.-19 (End Sem), Marks 8**
6. *What is Map-Reduce ? Explain working of Map-Reduce with example.* **SPPU : May-19 (End Sem), Marks 9**
7. *Explain HDFS with respect to NameNode, DataNodes, Secondary NameNode with example.* **SPPU : May-19 (End Sem), Marks 8**

6.7 Multiple Choice Questions

Q.1 3D scatter plots are used to plot data points on three axes in the attempt to show the relationship _____ variables.

a two three b three

c four d six

Q.2 _____ projection techniques help users find interesting projections of multidimensional data sets.

a Geometric b Pixel oriencted

c Circle segments d None

Q.3 List categorization of visualization methods.

a Pixel-oriented visulization techniques.

b Geometric visualization techniques

c Icon-based visulization projection technique

d All of these

Q.4 Line graph is also called _____ graph.

a X-Y b stick

c column d row

Q.5 Treemaps display hierarchical data using _____.

a rectangles b square

c triangle d circule

Q.6 Mahout is an open-source machine learning library from Apache written in _____.

a C b C++

c Python d Java

Q.7 HBase is a _____ , non-relational database.

a row-oriented b column-oriented

c horizontal d vertical

Q.8 Pig support _____ file format.

a mp3 b jpeg

c doc d Avro

Q.9 Pig works on the _____ of the cluster

a server-side b master node

c client-side d none

Q.10 Hive works on the _____ of the cluster.

a server-side b master node

c client-side d none

Q.11 MapReduce is a programming model and software framework first developed by _____.

a Microsoft b Amazon

c TCS d Google

Q.12 HDFS blocks are _____ MB by default and this is configurable.

a 32 b 64

c 128 d 256

Q.13 Apache Hadoop HDFS architecture follows a _____ architecture.

a client/server b master/slave

c peer to peer d all of these

Q.14 Hive is _____ and data warehousing tool used to query or analyze large datasets stored within the Hadoop ecosystem.

a STL b HDFS

c ETL d data mining

Q.15 Hadoop ecosystem include _____.

a Hive b Pig

c YARN d all of these

Answer Keys for Multiple Choice Questions :

Q.1	b	**Q.2**	a
Q.3	d	**Q.4**	b
Q.5	a	**Q.6**	d
Q.7	b	**Q.8**	d
Q.9	c	**Q.10**	a
Q.11	d	**Q.12**	c
Q.13	b	**Q.14**	c
Q.15	d		

❑❑❑

SOLVED MODEL QUESTION PAPER (In Sem)
Data Science and Big Data Analytics
T.E. (Computer) Semester - VI (As Per 2019 Pattern)

Time : 1 Hour] **[Maximum Marks : 30**

N. B. :

i) Attempt Q.1 or Q.2, Q.3 or Q.4.

ii) Neat diagrams must be drawn wherever necessary.

iii) Figures to the right side indicate full marks.

iv) Assume suitable data, if necessary.

Q.1 **a)** *What is data reduction ? Explain its strategies.* **(Refer section 1.10)** **[3]**

b) *What is big data ? Explain 3V's of big data.* **(Refer section 1.3)** **[4]**

c) *Explain data integration and transformation.* **(Refer section 1.9)** **[8]**

OR

Q.2 **a)** *What is data discretization ?* **(Refer section 1.11)** **[3]**

b) *Explain data science life cycle.* **(Refer section 1.5)** **[5]**

c) *What is data wrangling ? Explain process of data wrangling.* **(Refer section 1.7)** **[7]**

Q.3 **a)** *What is Bayes theorem ?* **(Refer section 2.4)** **[4]**

b) *Explain difference between null hypothesis and alternative hypothesis.* **(Refer section 2.5.3)** **[4]**

c) *What is Chi-square test ? List its characteristics ? Explain Chi -sqaure test for independence of attributes.* **(Refer section 2.7)** **[7]**

OR

Q.4 **a)** *What is need of statistics in data science and big data analytics ?* **(Refer section 2.1)** **[3]**

b) *Explain Wilcoxon Rank - sum test* **(Refer section 2.8.1)** **[5]**

c) *Explain various measures of central tendency.* **(Refer section 2.2)** **[7]**

SOLVED MODEL QUESTION PAPER (End Sem)

Data Science and Big Data Analytics

T.E. (Computer) Semester - VI (As Per 2019 Pattern)

Time : $2\frac{1}{2}$ Hours] **[Maximum Marks : 70**

N. B. :

i) Attempt Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.

ii) Neat diagrams must be drawn wherever necessary.

iii) Figures to the right side indicate full marks.

iv) Assume suitable data, if necessary.

Q.1 a) *What is big data ? Explain big data ecosystem.* **(Refer section 3.1)** **[8]**

b) *Explain data analytics life cycle.* **(Refer section 3.3)** **[10]**

OR

Q.2 a) *What is analytics sandbox ? Explain.* **(Refer section 3.2.4)** **[5]**

b) *Explain data analytics architecture with suitable diagram.* **(Refer section 3.1.4)** **[6]**

c) *What is data repository ? Explain advantages and disadvantages of data repository. Which are the factor responsible for data volume in big data.* **(Refer sections 3.2.1, 3.2.3 and 3.2.5)** **[7]**

Q.3 a) *What is regression ? Explain logistics regression. What is the difference between linear and logistics regression ?* **(Refer section 4.8)** **[8]**

b) *What is decision tree ? Explain how decision tree is constructed using ID3 algorithm.* **(Refer section 4.10)** **[9]**

OR

Q.4 a) *Generate frequent itemsets and generate association rules based on it using apriori algorithm. Minimum support is 50 % and minimum confidence is 70 %.* **(Refer example 4.6.1)** **[8]**

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

b) *What is data pre-processing ? How to remove duplicates ? How it handles missing data value ?* **(Refer section 4.3)** **[9]**

Q.5 **a)** *What is clustering ? Explain hierarchical clustering.* **(Refer section 5.1.6)** **[6]**

b) *What is confusion matrix ? Explain ROC curve.* **(Refer section 5.8)** **[6]**

c) *What is time series analysis ? Explain assumptions of ARIMA model.* **(Refer section 5.2)** **[6]**

OR

Q.6 **a)** *What is social network analysis ? How to develop social network analysis ?* **(Refer section 5.4)** **[6]**

b) *Explain random sampling.* **(Refer section 5.6.3)** **[6]**

c) *Explain various text pre-processing techniques.* **(Refer sections 5.3.2)** **[6]**

Q.7 **a)** *Explain the following data visualization techniques :* **(Refer section 6.3)** **[9]**

a. Venn diagram b. Line graph c. Pie chart

b) *Explain Hadoop ecosystem in details.* **(Refer section 6.6)** **[8]**

OR

Q.8 **a)** *What is data visualization ? Explain challenges to big data visualization.* **(Refer section 6.1)** **[8]**

b) *Explain the following data visualization tools :* **(Refer section 6.5)** **[9]**

a. Pentaho b. Datameter c. Tableau

❑❑❑

TEXT BOOKS FOR T.E. (COMP) SEM VI

Compulsory Subjects

1. Web Technology *(A. A. Puntambekar)*
2. Data Science and Big Data Analytics *(I. A. Dhotre, Dr. Kalpana V. Metre)*
3. Artificial Intelligence *(Anamitra Deshmukh-Nimbalkar, Dr. Vaishali P. Vikhe)*

Elective Subjects

4. Information Security *(I. A. Dhotre, Dr. Swati Nikam)*
5. Augmented and Virtual Reality *(Dr. Ninad More, Sunita Patil)*
6. Cloud Computing *(I. A. Dhotre)*
7. Software Modeling and Architecture *(A. A. Puntambekar)*