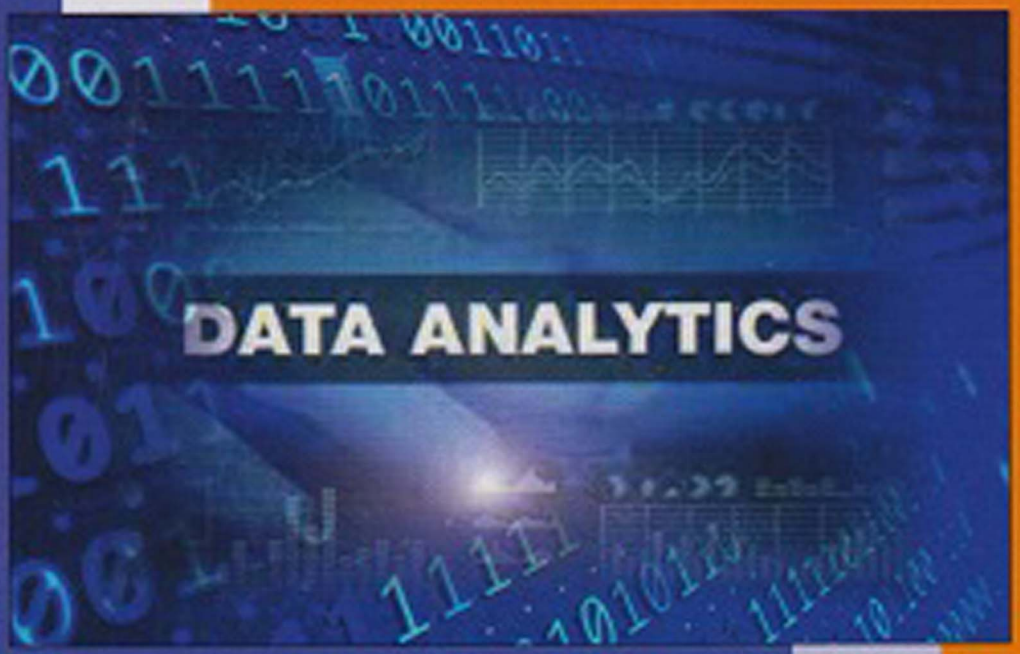
T. Y. B. Sc.
COMPUTER SCIENCE
SEMESTER-VI

NEW SYLLABUS
CBCS PATTERN

DATA ANALYTICS

Dr. Ms. MANISHA BHARAMBE
Dr. Mrs. HARSHA PATIL



NIRALI PRAKASHAN
ADVANCEMENT OF KNOWLEDGE

SPPU New Syllabus

A Book Of

DATA ANALYTICS

For T.Y.B.Sc. Computer Science : Semester – VI

[Course Code CS 364 : Credits – 2]

CBCS Pattern

As Per New Syllabus

Dr. Ms. Manisha Bharambe

M.Sc. (Comp. Sci.), M.Phil. Ph.D. (Comp. Sci.)

Vice Principal, Associate Professor, Department of Computer Science

MES's Abasaheb Garware College

Pune

Dr. Mrs. Harsha Patil

M.C.A., M.Phil. Ph.D. (Comp. Sci.)

Asst. Professor, AEF's Ashoka Center for Business and Computer Studies

Nashik

Price ₹ 230.00

NIRALI PRAKASHAN

ADVANCEMENT OF KNOWLEDGE

N5949

DATA ANALYTICS **ISBN 978-93-5451-317-6**

First Edition : February 2022

Published By :
NIRALI PRAKASHAN
Abhyudaya Pragati, 1312, Shivaji Nagar,
Off J.M. Road, Pune – 411005
Tel - (020) 25512336/37/39
Email : niralipune@pragationline.com

Polyplate

Printed By :
YOGIRAJ PRINTERS AND BINDERS
Survey No. 10/1A, Ghule Industrial Estate
Nanded Gaon Road
Nanded, Pune - 411041

DISTRIBUTION CENTRES

PUNE

Nirali Prakashan
(For orders outside Pune)
S. No. 28/27, Dhayari Narhe Road, Near Asian College
Pune 411041, Maharashtra
Tel : (020) 24690204; Mobile : 9657703143
Email : bookorder@pragationline.com

Nirali Prakashan
(For orders within Pune)
119, Budhwar Peth, Jogeshwari Mandir Lane
Pune 411002, Maharashtra
Tel : (020) 2445 2044; Mobile : 9657703145
Email : niralilocal@pragationline.com

MUMBAI

Nirali Prakashan
Rasdhara Co-op. Hsg. Society Ltd., 'D' Wing Ground Floor, 385 S.V.P. Road
Girgaum, Mumbai 400004, Maharashtra
Mobile : 7045821020, Tel : (022) 2385 6339 / 2386 9976
Email : niralimumbai@pragationline.com

DISTRIBUTION BRANCHES

DELHI

Nirali Prakashan
Room No. 2 Ground Floor
4575/15 Omkar Tower, Agarwal Road
Darya Ganj, New Delhi 110002
Mobile : 9555778814/9818561840
Email : delhi@niralibooks.com

BENGALURU

Nirali Prakashan
Maitri Ground Floor, Jaya Apartments,
No. 99, 6th Cross, 6th Main,
Malleswaram, Bengaluru 560003
Karnataka; Mob : 9686821074
Email : bengaluru@niralibooks.com

NAGPUR

Nirali Prakashan
Above Maratha Mandir, Shop No. 3,
First Floor, Rani Jhanshi Square,
Sitabuldi Nagpur 440012 (MAH)
Tel : (0712) 254 7129
Email : nagpur@niralibooks.com

KOLHAPUR

Nirali Prakashan
438/2, Bhosale Plaza, Ground Floor
Khasbag, Opp. Balgopal Talim
Kolhapur 416 012, Maharashtra
Mob : 9850046155
Email : kolhapur@niralibooks.com

JALGAON

Nirali Prakashan
34, V. V. Golani Market, Navi Peth,
Jalgaon 425001, Maharashtra
Tel : (0257) 222 0395
Mob : 94234 91860
Email : jalgaon@niralibooks.com

SOLAPUR

Nirali Prakashan
R-158/2, Avanti Nagar, Near Golden
Gate, Pune Naka Chowk
Solapur 413001, Maharashtra
Mobile 9890918687
Email : solapur@niralibooks.com

marketing@pragationline.com | www.pragationline.com

Also find us on www.facebook.com/niralibooks

Preface ...

We take an opportunity to present this Text Book on **"Data Analytics"** to the students of Third Year B.Sc. (Computer Science) Semester-VI as per the New Syllabus, June 2021.

The book has its own unique features. It brings out the subject in a very simple and lucid manner for easy and comprehensive understanding of the basic concepts. The book covers theory of Introduction to Data Analytics, Overview of Machine Learning, Mining Frequent Patterns, Associations and Correlations, Social Media and Text Analytics.

A special word of thank to Shri. Dineshbhai Furia, and Mr. Jignesh Furia for showing full faith in us to write this text book. We also thank to Mr. Amar Salunkhe and Mr. Akbar Shaikh of M/s Nirali Prakashan for their excellent co-operation.

We also thank Mrs. Yojana Despande, Mr. Ravindra Walodare, Mr. Sachin Shinde, Mr. Ashok Bodke, Mr. Moshin Sayyed and Mr. Nitin Thorat.

Although every care has been taken to check mistakes and misprints, any errors, omission and suggestions from teachers and students for the improvement of this text book shall be most welcome.

Authors

■■■

Syllabus ...

1. Introduction to Data Analytics (6 Lectures)

- Concept of Data Analytics
- Data Analysis vs Data Analytics
- Types of Analytics
  - Diagnostic Analytics
  - Predictive Analytics
  - Prescriptive Analytics
  - Exploratory Analysis
  - Mechanistic Analysis
- Mathematical Models:
  - Concept
- Model Evaluation:
- Metrics for Evaluating Classifiers:
  - Class Imbalance:
    - AUC, ROC (Receiver-Operator Characteristic) Curves
    - Evaluating Value Prediction Models

2. Machine Learning Overview (6 Lectures)

- Introduction to Machine Learning, Deep Learning, Artificial intelligence
- Applications for Machine Learning in Data Science
- The Modeling Process
  - Engineering Features and Selecting a Model
  - Training the Model
  - Validating the Model
  - Predicting New Observations
- Types of Machine Learning
  - Supervised Learning
  - Unsupervised Learning
  - Semi-supervised Learning
  - Ensemble Techniques
- Regression Models
  - Linear Regression
  - Polynomial Regression
  - Logistic Regression
- Concept of Classification, Clustering and Reinforcement Learning

3. Mining Frequent Patterns, Associations and Correlations (12 Lectures)

- What kind of Patterns can be Mined
- Class/Concept Description:
  - Characterization and Discrimination
  - Mining Frequent Patterns
  - Associations and Correlations
  - Classification and Regression for Predictive Analysis
  - Cluster Analysis
  - Outlier Analysis
- Mining Frequent Patterns: Market Basket Analysis
- Frequent Itemsets, Closed Itemsets and Association Rules
- Frequent Itemset Mining Methods
- Apriori Algorithm
- Generating Association Rules from Frequent Itemsets
- Improving Efficiency of Apriori Algorithm
- Frequent Pattern Growth (FP-Growth) Algorithm

4. Social Media and Text Analytics (12 Lectures)

- Overview of Social Media Analytics
  - Social Media Analytics Process
  - Seven Layers of Social Media Analytics
  - Accessing Social Media Data
- Key Social Media Analytics Methods
- Social Network Analysis
  - Link Prediction
  - Community Detection
  - Influence Maximization
  - Expert Finding
  - Prediction of Trust and Distrust among Individuals
- Introduction to Natural Language Processing
- Text Analytics:
  - Tokenization
  - Bag of Words
  - Word weighting: TF-IDF
  - n-Grams
  - Stop Words
  - Stemming and Lemmatization
  - Synonyms and Parts of Speech Tagging
  - Sentiment Analysis
- Document or Text Summarization
- Trend Analytics
- Challenges to Social Media Analytics

■■■

Contents ...

■■■

CHAPTER 1

Introduction to Data Analytics

Objectives...

- To understand Concept of Data Analytics
- To learn Types of Data Analytics
- To study different Types of Data Analytics

1.0 INTRODUCTION

- An important phase of technological innovation associated with the rise and rapid development of computer technology came into existence only a few decades ago.
- The technological innovation brought about a revolution in the way people work, first in the field of science and then in many others, from technology to business, as well as in day-to-day life.
- In today's data driven world the massive amount of data collected/generated/ produced at remarkable speed and high volume at every day. Data allows to makes better predictions about the future.
- For processing and analyzing need of this massive/huge amount of a new discipline is formed known as data science. The objective/goal of data science is to extract information from data sources.
- Data science is a collection of techniques used to extract value from data. Data science has become an essential tool for any organization that collects stores and processes data as part of its operations.
- Data science is the task of scrutinizing and processing raw data to reach a meaningful conclusion. Data science techniques rely on finding useful patterns, connections and relationships within data.
- Data science applies an ever-changing and vast collection of techniques and technology from mathematics, statistics, Machine Learning (ML) and Artificial Intelligence (AI) to decompose complex problems into smaller tasks to deliver insight and knowledge.
- Analytics is the systematic computational analysis of data. Analytics is the discovery, interpretation, and communication of meaningful patterns in data.

- Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.
- Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

1.1 CONCEPT OF DATA ANALYTICS

- Advancement in data science has created opportunities to sort, manage and analyze large or massive amounts of data more effectively and efficiently.
- Data science is closely related to the fields of data mining and machine learning, but it is broader in scope. Today, data science drives decision making in nearly all parts of modern societies.
- The term data comprises facts, observations and raw information. Data itself have little meaning if it is not processed. The processed data in meaningful form known as information.
- Analytics is used for the discovery, interpretation, and communication of meaningful patterns and/or insights in data. The term analytics is used to refer to any data-driven decision-making.
- Data analytics may analyze many varieties of data to provide views into patterns and insights that are not humanly possible.
- Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions from that information.
- Data analytics is used in many industries to allow companies and organizations to make better business decisions, and in the sciences to verify or disprove existing models or theories.

1.1.1 Definition of Data Analytics

- Data and information are increasing rapidly; the growth rate of the information is so high that the information available to us in the near future is going to unpredictable.
- So, there is need of technique like data analytics which operates at high-speed and efficiently on huge/massive amount data and helps organizations for making better decisions.
- Data analytics is defined as, a science of extracting meaningful, valuable information from raw data.
- Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.
- The goal of data analytics is to get actionable insights from raw data resulting better decisions.

- Data analytics and all associated strategies and techniques are essential when it comes to identifying different patterns, finding anomalies and relationships in large chunks/set of data and making the data or information collected more meaningful and more understandable.

1.1.2 Roles in Data Analytics

- Various roles in data analytics are explained below:
  1. **Data Analyst:** Data analyst is an individual, who performs mining of huge amount of data, models the data, looks for patterns, relationship, trends and so on. He/she comes up with visualization and reporting for analyzing the data for decision making and problem-solving process. The main role of a data analyst is to extract data and interpret the information attained from the data for analyzing the outcome of a given problem.
  2. **Data Scientist:** A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc. Data scientists mainly deal with large and complex data that can be of high dimension, and carry out appropriate machine learning and visualization tools to convert the complex data into easily interpretable meaningful information. The primary task of a data scientist is to use machine learning and deep learning-based techniques to make an in-depth analysis of input data.
  3. **Data Architect:** They are provides the support of various tools and platforms that are required by data engineers to carry out various tests with precision. Data architects should he well equipped with knowledge of data modeling and data warehousing. The main task of data architects is to design and implement database systems, data models, and components of data architecture.
  4. **Data Engineer:** A data engineer works with massive amount of data and responsible for building and maintaining the data architecture of a data science project. Data engineer also works for the creation of data set processes used in modeling, mining, acquisition and verification. Data engineers have a demanding role in data analytics as they help in assuring that data are made available in a form that can be easily used for analysis and interpretation.
  5. **Analytics Manager:** They are involved in the overall management of the various data analytics operations as discussed in this section. For each of the stakeholders of data analytics that have been mentioned in this section, the analytics manager deals with the team leader of each group and monitors and manages the work of each team.

1.1.3 Lifecycle of Data Analytics

- The data analytics lifecycle is a process that consists of six basic stages/phases (data discovery, data preparation, model planning, model building, communication results, and operationalize) as shown in Fig. 1.1 that define how information is created, gathered, processed used and analyzed for organizational goals.
- Fig. 1.1 shows the six phases of the data analytics lifecycle that is followed one phase after another to complete one cycle.
- The lifecycle of the data analytics provides a framework for the best performances of each phase from the creation of the project until its completion.
- Data Analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about the information.

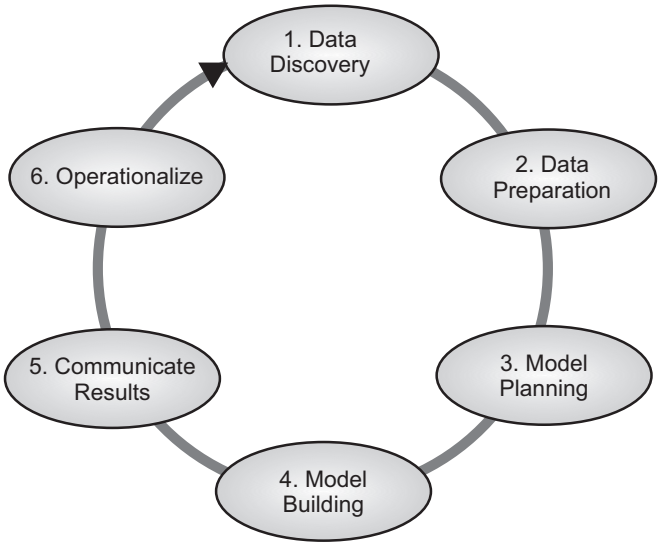


Fig. 1.1: Lifecycle of Data Analytics

- Various phases in data analytics lifecycle are explained below:

Phase 1 - Data Discovery:

- Data discovery is the 1st phase to set project's objectives and find ways to achieve a complete data analytics lifecycle.
- Data discovery phase defining the purpose of data and how to achieve it by the end of the data analytics lifecycle.
- Data discovery phase consists of identifying critical objectives a business is trying to discover by mapping out the data.

Phase 2 - Data Preparation:

- In the 2nd phase after the data discovery phase, data is prepared by transforming it from a legacy system into a data analytics form by using the sandbox platform, (a scalable platform commonly used by the data scientists for data preprocessing).

- Data preparation phase of the data analytics lifecycle involves data preparation, which includes the steps to explore, preprocess and condition data prior to modeling and analysis.
- The data preparation and processing phase involves collecting, processing and conditioning data before moving to the model building process.
- An analytics sandbox is a platform that allows us to store and process large amounts of data.
- Data are loaded in the sandbox in three ways namely, ETL (Extract, Transform and Load), ELT (Extract, Load, and Transform) and ETLT.

Phase 3 - Model Planning:

- The 3rd phase of the lifecycle is model planning, where the data analytics team members makes proper planning of the methods to be adapted and the various workflow to be followed during the next phase of model building.
- Model planning is a phase where the data analytics team members have to analyze the quality of data and find a suitable model for the project.

Phase 4 - Model Building:

- In this phase the team works on developing datasets for training and testing as well as for production purposes.
- This phase is based on the planning made in the previous phase, the execution of the model is carried out by the team.
- Model building is the process where team has to deploy the planned model in a real-time environment. It allows analysts to solidify their decision-making process by gain in-depth analytical information.
- The environment needed for the execution of the model is decided and prepared so that if a more robust environment is required, it is accordingly applied.

Phase 5 - Communicate Results:

- The 5th phase of the life cycle of data analytics checks the results of the project to find whether it is a success or failure.
- The result is scrutinized by the entire team along with its stakeholders to draw inferences on the key findings and summarize the entire work done.
- In communicate results phase, the business/organizational values are quantified and an elaborate narrative on the key findings is prepared.

Phase 6 - Operationalize:

- In 6th phase, the team delivers final reports is prepared by the team along with the briefings, source code and related technical documents.
- Operationalize phase also involves running the pilot project to implement the model and test it in a real-time environment.

- As data analytics help build models that lead to better decision making, it, in turn, adds values to individuals, customers, business sectors and other organizations.
- As soon the team prepares a detailed report including the key findings, documents, and briefings, the data analytics life cycle almost comes close to the end.
- The next step remains the measure the effectiveness of the analysis before submitting the final reports to the stakeholders.

1.1.4 Data Analytics Framework

- In data analytics, the framework allows to move through data analysis in an organized/structured way.
- Data analytics provides us with a process to follow as we scrutinize the data to identify and solve problems.
- Data analytics is the framework deals with technical aspects of managing data and analytics tools. It answers the following questions:
  1. What are the infrastructure requirements today and in 5-10 years?
  2. Should we build an on premise cloud infrastructure or store data in an off premise private virtual cloud?
  3. What are the infrastructure components for data storage and archiving?
  4. Which systems of record will be supported and designated as analytics platforms?
  5. What analytics tools will we support and what will our analytics tools library consist of?
  6. What technologies and vendor solutions will be supported as enterprise analytics systems to provide infrastructure and analytics tools?
  7. What is the analytics capability roadmap?
  8. What solutions do we intend to deploy in the next five years and in what priority?
- Fig. 1.2 shows the four layer framework of data analytics consists of a data management layer an analytics engine layer and a presentation layer.
- The four layers in data analytics framework is explained below:
  1. **Data Connection Layer:** In this layer, data analysts set up data ingestion pipelines and data connectors to access data. They might apply methods to identify metadata (data about data) in all source data repositories. Building this layer starts with making an inventory of where the data is created and stored. The data analysts might implement Extract, Transfer and Load (ETL) software tools to extract data from their source. Other data exchange standards such as X.12 might be used to transfer data to the data management layer. In number of architectures, the enterprise data warehouse may be connected to data sources through data gateways, data harvesters and connectors using APIs. Products offered by Amazon AWS, Microsoft Data Factory and Talend or similar systems are used as data connector tools.

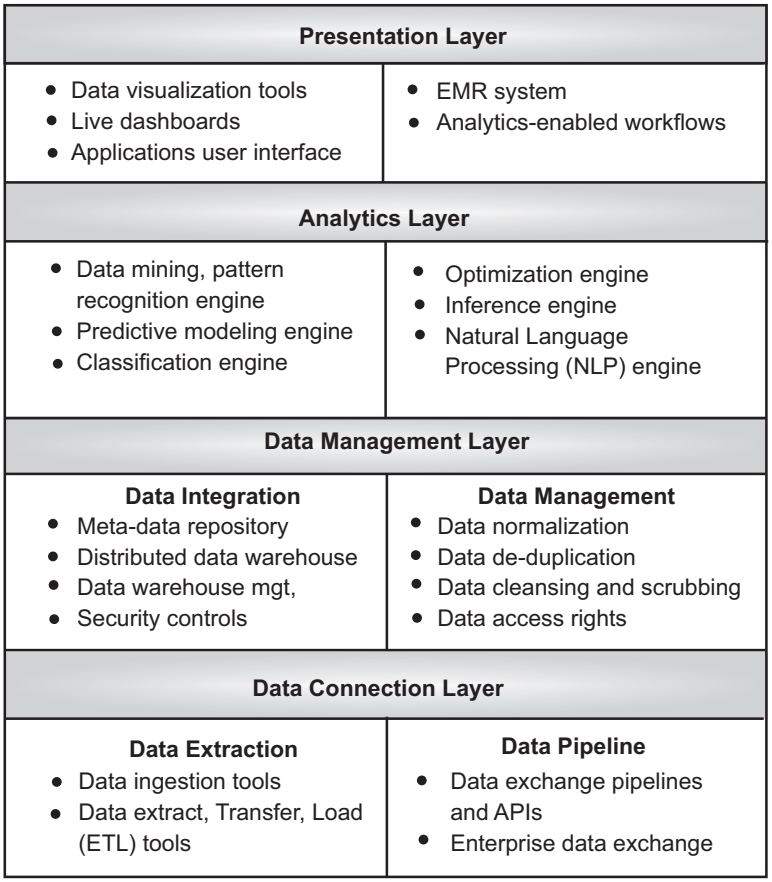


Fig. 1.2: Four Layer Framework of Data Analytics

2. **Data Management Layer:** Once, the data has been extracted, data scientists must perform a number of functions that are grouped under the data management layer. The data may need to be normalized and stored in certain database architectures to improve data query and access by the analytics layer. We'll cover taxonomies of database tools including SQL, NoSQL, Hadoop, Spark and other architecture in the upcoming sections.
3. **Analytics Layer:** In analytics layer, a data scientist uses a number of engines to implement the analytical functions. Depending on the task at hand, a data scientist may use one or multiple engines to build an analytics application. A more complete layer would include engines for optimization, machine learning, natural language processing, predictive modeling, pattern recognition, classification, inferencing and semantic analysis.
4. **Presentation Layer:** The presentation layer includes tools for building dashboards, applications and user-facing applications that display the results of analytics engines. Data scientists often mash up several data visualization widgets,

web parts and dashboards (sometimes called Mash boards) on the screen to display the results using info-graphic reports. These dashboards are active and display data dynamically as the underlying analytics models continuously update the results for dashboards.

1.1.5 Advantages and Disadvantages of Data Analytics

- Data analytics represents the process of examining massive amount of data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions.

Advantages of Data Analytics:

1. **Improving Efficiency:** Data analytics can help analyze large amounts of data quickly and display it in a formulated manner to help achieve specific organizational goals. It encourages a culture of efficiency and teamwork by allowing the managers to share the insights from the analytics results to the employees. The improvement areas within an organization become evident and actions can be taken to increase the overall efficiency of the workplace thereby increasing productivity.
2. **Improving Quality of Products and Services:** Data analytics can help with enhancing the user experience by detecting and correcting errors or avoiding non-value-added tasks. For example, self-learning systems can use data to understand the way customers are interacting with the tools and make appropriate changes to improve user experience. In addition, data analytics can help with automated data cleansing and improving the quality of data and consecutively benefiting both customers and organizations.
3. **Witnessing the Opportunities:** The changing nature of technology the organizations want to keep pace with the latest trends. Here, Data Analytics offers refined sets of data that can help in observing the opportunities to avail.
4. **Helps an Organization make Better Decisions:** Data analytics can help with transforming the data that is available into valuable information for executives so that better decisions can be made.

Disadvantages of Data Analytics:

1. **Low Quality of Data:** One of the biggest limitations of data analytics is lack of access to quality data. It is possible that organizations already have access to a lot of data, but the question is do they have the right data that they need?
2. **Privacy Concerns:** Sometimes, data collection might breach/violate the privacy of the customers as their information such as purchases, online transactions and subscriptions are available to organizations whose services they are using.

1.2 DATA ANALYSIS vs DATA ANALYTICS

- The terms data analysis and data analytics are often used interchangeably and could be confusing.
- Data analytics is a broader term and includes data analysis as necessary subcomponent. Analytics defines the science behind the analysis.
- The science means understanding the cognitive processes an analyst uses to understand problems and explore data in meaningful ways.
- Data analysis is a process that refers to hands-on data exploration and evaluation. Data analysis looks backwards, providing marketers with a historical view of what has happened. Data analytics, on the other hand, models the future or predicts a result.
- The purpose of data analysis is to extract information that is not easily deducible but that, when understood, leads to the possibility of carrying out studies on the mechanisms of the systems that have produced them, thus allowing forecasting possible responses of these systems and their evolution in time.
- Analytics makes extensive use of mathematics and statistics and the use of descriptive techniques and predictive models to gain valuable knowledge from data.
- These insights from data are used to recommend action or to guide decision-making in a business context. Thus, analytics is not so much concerned with individual analysis or analysis steps, but with the entire methodology.
- Data analysis helps design a strong business plan for businesses, using its historical data that tell about what worked, what did not and what was expected from a product or service.
- On other hand, Data analytics helps organization in utilizing the potential of the past data and in turn identifying new opportunities that would help them plan future strategies.
- Data analysis helps to finding or extracting useful information for decision making. Data analytics helps in business growth by reducing risks, costs, and making the right decisions.
- Data Analytics is a wide area involving handling data with a lot of necessary tools to produce helpful decisions with useful predictions for a better output.
- While, Data analysis is actually a subset of data analytics which helps us to understand the data by questioning and to collect useful insights from the information already available.
- Data analytics is the process of exploring the data from the past to make appropriate decisions in the future by using valuable insights whereas, Data analysis helps in understanding the data and provides required insights from the past to understand what happened so far.

- Following table compares data analysis and data analytics:

Sr. No.	Data Analysis	Data Analytics
1.	The process of extracting information from raw data is called as data analysis.	The process of extracting meaningful valuable insights from raw data called as data analytics.
2.	Data analysis is a process involving the collection, manipulation and examination of data for getting insight from data.	Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed organizational decisions.
3.	Data analysis is a process of studying, refining, transforming, and training of the past data to gain useful information, suggest conclusions and make decisions.	Data analytics is the process of remodeling past data into actions through analysis and insights to help in organizational decision making and problem-solving.
4.	Data analysis looks backwards, with a historical view of what has happened.	Data analytics models the future or predicts a result.
5.	Data analysis is a subset of data analytics, which takes multiple data analysis processes to focus on why an event happened and what may happen in the future based on the previous data.	Data analytics is a multidisciplinary field with extensive use of computer skills, mathematics, statistics, the use of descriptive techniques and predictive models to gain valuable knowledge from data through analytics.
6.	Data analysis also makes decisions but less good than data analytics.	Data analytics is utilizing data, machine learning, statistical analysis and computer-based models to get better insight and make better decisions from the data.
7.	Data analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.	Data analytics help uncover the patterns from raw data and derive valuable insights from it.
8.	Data analysis is subset of data analytics.	Data analytics uses data analysis as subcomponent.
9.	Tools used for data analysis are Open Refine, Rapid Miner, KNIME, Google Fusion Tables, Node XL, Wolfram Alpha, Tableau Public, etc.	Tools used in data analytics are Python, Tableau Public, SAS, Apache Spark, Excel, etc.

1.3 TYPES OF DATA ANALYTICS

- Organizations from almost every sector are generating a large volume of data on a regular basis.
- Merely collecting large amounts of data will not serve any purpose and cannot be used directly for the profit of the company/organization.
- Organizations can extract very useful information from this data which can further support complex decision making hence, there is a need for data analytics.
- The art and science of refining data to fetch useful insight which further helps in decision making is known as Analytics.
- There are four types of data analytics as shown in Fig. 1.3 as explained below:
  1. **Descriptive Analytics:** What happened?
  2. **Diagnostic Analytics:** Why did it happen?
  3. **Predictive Analytics:** What will it happen?
  4. **Prescriptive Analytics:** How can we make it happen?

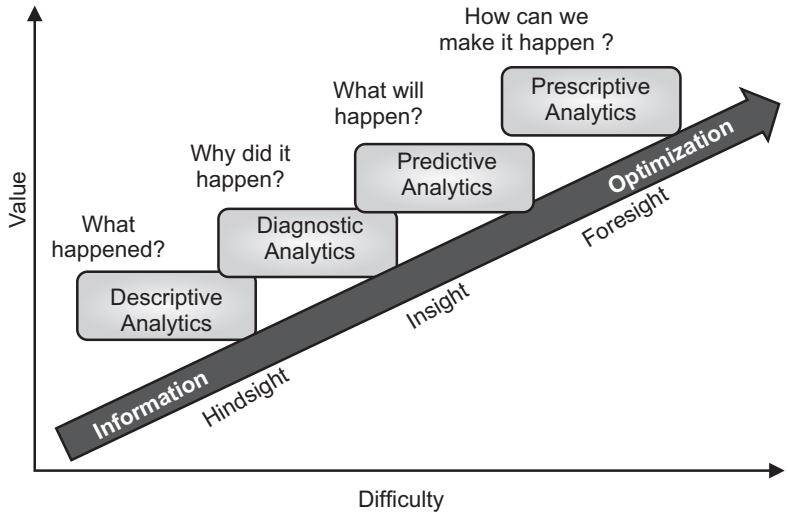


Fig. 1.3: Types of Data Analytics

1.3.1 Descriptive Analytics

- Descriptive analytics examines the raw data or content to answer question, what happened?, by analyzing valuable information found from the available past (historical) data.
- The goal of descriptive analytics is to provide insights into the past leading to the present, using descriptive statistics, interactive explorations of the data, and data mining.

- Descriptive analytics enables learning from the past and assessing how the past might influence future outcomes.
- Descriptive analytics is valuable as it enables associations to gain from past practices and helps them in seeing how they may impact future results.
- Descriptive analytics looks at data and analyzes past events for insight as to how to approach the future.
- It looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure.

Examples:

1. An organizations' records give a past review of their financials, operations, customers and stakeholders, sales and so on.
2. Using descriptive analysis, a data analyst will be able to generate the statistical results of the performance of the hockey players of team India. For generating such results, the data may need to be integrated from multiple data sources to gain meaningful insights through statistical analysis.

1.3.2 Diagnostic Analytics

- Diagnostic analytics is a form of analytics which examines data to answer the question, why did it happen?.
- It is kind of root cause analysis that focuses on the processes and causes, key factors and unseen patterns.
- The goal/objective of diagnostic analytics is to find the root cause of issues. It can be accomplished by techniques like data discovery, correlations, data mining and drill-down.
- Diagnostic analytics tries to gain a deeper understanding of the reasons behind the pattern of data found in the past. Here, business/organizational intelligence comes into play by digging down to find the root cause of the pattern or nature of data obtained.
- For example, with diagnostic analysis, a data analyst will be able to find why the performance of each player of the hockey team of India has risen (or degraded) in the recent past nine months.
- The main function of diagnostic analytics is to identify anomalies, drill into the analytics and determine the casual relationships.

Examples:

1. Some form of social media marketing campaign where the user is interested in retrieving the number of likes or reviews. Diagnostic analytics can help to filter out thousands of likes and reviews into a single view to see the progress of the campaign.
2. Drop in website traffic of an organization can lead to a decrease in the sales and thereby revenue will also be reduced. In this case, diagnostic analytics finds the

root cause initially, such as traffic has been reduced and from there, it will fine-tune the problem after finding the reasons for the downside in website traffic such as Software Engine Optimization (SEO), social marketing, email marketing and any other factors, which are not enabling the website to reach many people.

1.3.3 Predictive Analytics

- Predictive analysis, as the name suggests, deals with prediction of future based on the available current and past data.
- A predictive analysis uses past data to create a model that answer the question, what will happen?
- Prediction-based on historical data, build models and use them to forecast a future value. For example, demand for a particular package around holiday season.
- Predictive analytics is important to analyze the current/present data and make use of it to predict a solution for the future. For these future predictions in data analytics the predictive analytics is used.
- Predictive analytics makes predictions about future outcomes/result using historical data combined with statistical modeling, data mining techniques and machine learning.
- Organizations/firms employ predictive analytics to find patterns in this data to identify risks and opportunities.
- Using predictive analytics, users can prepare plans and implement corrective actions in a proactive manner in advance of the occurrence of an event.
- Predictive analytics is the use of data, machine learning techniques, and statistical algorithms to determine the likelihood of future results based on historical data.
- The primary goal of predictive analytics is to help you go beyond just what has happened and provide the best possible assessment of what is likely to happen in future.
- Predictive analytics extracts the information from the available datasets and this extraction helps us forecast the possibility that can happen in the future with risk analysis and mitigation.
- It is not guaranteed that all the predicted data can produce exact results; there may be a slight variation between the predicted and the future values.
- Based on the past events, a predictive analytics model forecasts what is likely to happen in future.
- Predictive analytics is critical to knowing about future events well in advance and implementing corrective actions.
- Predictive analysis uses techniques to include machine learning, statistical modeling and data mining.

- Predictive analytics turns data into valuable, actionable information and it uses data to determine the probable future outcome of an event or a likelihood of a situation occurring.

  Example: Using predictive analysis, a data analyst will be able to predict the performance of each player of the hockey team for the upcoming Olympics. Such prediction analysis can help the Indian Hockey Federation to decide on the players' selection for the upcoming Olympics.

1.3.4 Prescriptive Analytics

- Prescriptive analytics goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the decision maker the implications of each decision option.
- Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen.
- Further, prescriptive analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option.
- In practice, prescriptive analytics can continually and automatically process new data to improve prediction accuracy and provide better decision options.
- Predictive analytics is often associated with data science. To gain insights from the data, data scientists use deep learning and machine learning algorithms to find patterns and make predictions about future events.
- Predictive analysis is a type of data analysis, the insights gained from all the other three types of data analyzes are combined to determine the kind of action to be taken to solve a certain situation.
- Predictive analysis prescribes what steps are needed to be taken to avoid a future problem. It involves a high degree of responsibility, time and complicacy to reach to informed decision-making.

  Example: In the healthcare industry, we can use prescriptive analytics to manage the patient population by measuring the number of patients who are clinically obese.

Table 1.1: Comparisons between Data Analytics

Sr. No.	Analytics	Question	Methods	Context
1.	Predictive analytics	What will happen if? What is the pattern?	Predictive modeling. Statistical modeling.	It predicts the future possibilities that can happen in the firm/ organization.

contd. ...

2.	Descriptive analytics	What has happened? How many, when and where?	Canned reports. Ad hoc reports.	It describes the events that have occurs already in the past.
3.	Diagnostics analytics	Why did it happen? Where must we see?	Query and Drilldowns. Discovery alerts.	It justifies the reason for the occurrences of those events in the firm/ organization.
4.	Prescriptive analytics	What should we do about this? What will happen if we use this?	Optimization. Random testing.	It suggests the solutions to overcome the past events.

1.3.5 Exploratory Analytics

- Exploratory Data Analysis (EDA) is the most important aspect to any data analysis. Exploratory data analytics attempts to find hidden, unseen or previously unknown relationships.
- The EDA techniques are used to interactively discover and visualize trends, behaviors, and relationships in data. They also provide clues as to which variables might be good for building data analytics.
- The goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set.
- Exploratory analytics is an analytical approach that primarily focuses on identifying general patterns in the raw data to identify outliers and features that might not have been anticipated using other analytical types.
- EDA is an approach to analyzing datasets to summarize their main characteristics, often with visual/graphical methods.
- The purpose of exploratory data analysis is to:
  1. Check for missing data and other mistakes.
  2. Gain maximum insight into the data set and its underlying structure.
  3. Check assumptions associated with any model fitting or hypothesis test.
  4. Create a list of outliers or other anomalies.
  5. Find parameter estimates and their associated confidence intervals or margins of error.
- Exploratory data analysis is the process of analyzing and interpreting datasets while summarizing their particular characteristics with the help of data visualization methods.

- Exploratory data analysis alludes to the basic procedure of performing introductory examinations on information to find designs, to spot abnormalities, to test theories and to check presumptions with the assistance of insights from data and graphical portrayals.

  Example: If we want to go to a theater to watch a movie, the analysis and searching we do to select a movie to watch is nothing but our exploratory analytics.
- Some of the most common data science tools used to create an EDA includes Python and R programming.

1.3.6 Mechanistic Analytics

- Mechanistic analytics allow data scientists to understand clear alterations in variables which can result in changing of variables.
- The results of mechanistic data analytics are determined by equations in engineering and physical sciences.
- Mechanistic (or Algorithmic) analytics seeks to explain the relationship, influence and interdependence of variables such that changes to one variable are understood in the way they impact other variables.
- The goal of mechanistic data analytics is to understand exact changes in variables that lead to other changes in other variables.
- For example, we may want to know how the number of free doughnuts per employee per day affects employee productivity. Perhaps by giving them one extra doughnut we gain a 5% productivity boost, but two extra doughnuts could end up making them lazy.
- Regression analysis is a process for estimating the relationships among variables. It is a statistical tool used for the investigation of relationships between variables.
- Regression analysis is a method of predicting or estimating one variable knowing the value of the other variable.

1.4 MATHEMATICAL MODELS

- A variety of different tools have been developed to manipulate and interpret data. A model can be used for both predictive and explanatory applications.
- Modeling is the process of encapsulating information into a tool which can forecast and make predictions.
- Predictive models in mathematics are structured around some idea of what causes future events to happen.
- Extrapolating from current or recent trends and observations assumes a world view that the future will be like the past.
- More sophisticated models like the laws of physics, provide principled notions of causation; fundamental explanations of why things happen/occurs.
- Accurately evaluating or appraising the performance of a model can be surprisingly hard, but it is essential for knowing how to interpret the resulting predictions.

1.4.1 Concept

- In this section, we will see various ways of thinking about models to help shape the way we build them.

Occam's Razor:

- Occam's razor is a problem-solving principle arguing that simplicity is better than complexity.
- Named after 14th century logician and theologian William of Ockham, this theory has been helping many great thinkers for centuries.
- Occam's razor is the problem solving principle, which states that "entities should not be multiplied beyond necessity", sometimes inaccurately paraphrased as "the simplest explanation is usually the best one."
- In simple words, Occam's razor is the philosophical principle states that, the simplest explanation is the best explanation.
- Occam's notion of simpler generally refers to reducing the number of assumptions employed in developing the model.
- With respect to statistical modeling, Occam's razor tells or speaks to the need to minimize the parameter count of a model.
- Overfitting occurs when a mathematical model tries too hard to achieve accurate performance on its training data.
- It is the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably.
- Overfitting occurs or happens when there are so many parameters that the model can essentially memorize its training set, instead of generalizing appropriately to minimize the effects of error and outliers.
- Overfit models tend to perform extremely well on training data, but much less accurately on independent test data.
- An overfit model is a statistical model. An overfit model contains more parameters than can be justified by the data.
- Invoking Occam's razor requires that we have a meaningful way to evaluate how accurately our models are performing. Simplicity is not an absolute virtue, when it leads to poor performance.
- Deep learning is a powerful technique for building models with millions of parameters. Despite the danger of overfitting, these models perform extremely well on a variety of complex tasks.
- Occam would have been suspicious of such models, but come to accept those that have substantially more predictive power than the alternatives.

- Appreciate the inherent trade-off between accuracy and simplicity. It is almost always possible to improve the performance of any model by kludging-on extra parameters and rules to govern exceptions.
- Complexity has a cost, as explicitly captured in machine learning methods like LASSO/ridge regression. These techniques employ penalty functions to minimize the features used in the model.
- Underfitting occurs when a statistical model cannot adequately capture the underlying structure of the data.
- An under-fitted model is a model where some parameters or terms that would appear in a correctly specified model are missing.
- Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model will tend to have poor predictive performance.

Bias-Variance Trade-Offs:

- The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.
- Bias-variance trade-off is tension between the between model complexity and performance shows up in the statistical notion of the bias-variance trade-off:
  1. **Bias:** It is error from incorrect assumptions built into the model, such as restricting an interpolating function to be linear instead of a higher-order curve.
  2. **Variance:** It is error from sensitivity to fluctuations in the training set. If our training set contains sampling or measurement error, this noise introduces variance into the resulting model.
- Errors of bias produce or generate underfit models and they do not fit the training data as tightly as possible, were they allowed the freedom to do so.
- Underfitting occurs/happens when a statistical model cannot adequately capture the underlying structure of the data.
- Errors of variance result in overfit models (their quest for accuracy causes overfit models to mistake noise for signal and they adjust so well to the training data that noise leads them astray).
- Models that do much better on testing data than training data are overfit models. An underfitted model is a model where some parameters or terms that would appear in a correctly specified model are missing.

Nate Silver's Principles for Effective Modeling:

- Nate R. Silver is perhaps the most prominent public face of data science today. He outlines following principles for effective modeling:

  Principle #1 (Think Probabilistically): Forecasts which make concrete statements are less meaningful than those that are inherently probabilistic. The real world is an

uncertain place, and successful models recognize this uncertainty. There are always a range of possible outcomes that can occur with slight perturbations of reality, and this should be captured in the model.

Principle #2 (Change the Forecast in Response to New Information): Live models are much more interesting than dead ones. A model is live if it is continually updating predictions in response to new information. Fresh information should change the result of any forecast. Scientists should be open to changing opinions in response to new data and built the infrastructure that maintains a live model. Any live model should track and display its predictions over time, so the viewer can guess whether changes accurately reflected the impact of new information.

Principle #3 (Look for Consensus): Data should derive from as many different sources as possible to get the good forecast. Ideally, multiple models should be built, each trying to predict the same thing in different ways. We should have an opinion as to which model is the best, but be concerned when it substantially differs from the herd. Often third parties produce competing forecasts, which you can monitor and compare against.

Principle #4 (Employ Baysian Reasoning): The Bayes' theorem has several interpretations, but perhaps most clearly provides a way to calculate how probabilities change in response to new evidence. When stated as given below, it provides a way to calculate how the probability of event A changes in response to new evidence B.

$$P(A \mid B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Applying Bayes' theorem requires a prior probability P(A), the likelihood of event A before knowing the status of a particular event B. This might be the result of running a classifier to predict the status of A from other features or background knowledge about event frequencies in a population. Without a good estimate for this prior, it is very difficult to know how seriously to take the classifier.

1.4.2 Taxonomy of Models

- Mathematical models come in, well, many different models. Part of producing or developing a philosophy of modeling understands the available degrees of freedom in design and implementation.
- In this section, we will study at model types along several different dimensions, reviewing the primary technical issues which arise to distinguish each class.

First-Principle vs. Data-Driven Models:

- First-principle models are based on a belief of how the system under investigation really works. First-principle models might be a theoretical explanation, like Newton's laws of motion.

- The first-principle model might be a discrete event simulation. Data driven models are based on observed correlations between input parameters and outcome variables.
- The same basic model might be used to predict tomorrow's weather or the price of a given stock, differing only on the data it was trained on.
- Machine learning methods make it possible to build an effective and efficient model on a domain one knows nothing about, provided we are given a good enough training set.
- Ad hoc models are built using domain specific knowledge to guide their structure and design.
- These ad hoc models tend to be brittle in response to changing conditions, and difficult to apply to new tasks.
- In contrast, machine learning models for classification and regression are basic, because they employ no problem-specific ideas, only specific data.
- Retrain the models on fresh data and classification and regression adapt to changing conditions.
- Train them on a different data set and classification and regression can do something completely different. By this rubric, general models sound much better than ad hoc ones.
- The truth is that the best models are a mixture of both i.e., theory and data. It is important to understand the domain as deeply as possible, while using the best data or information we can in order to fit and evaluate the models.

Linear vs. Non-Linear Models:

- Linear models are governed by equations that weigh each feature variable by a coefficient reflecting its importance and sum up these values to produce a score.
- Powerful machine learning techniques, (for example, linear regression) can be used to identify the best possible coefficients to fit training data, yielding very effective models.
- But basically speaking, the world is not linear. Richer mathematical descriptions include higher order exponentials, logarithms and polynomials.
- These permit mathematical models that fit training data much more tightly than linear functions can.
- Generally speaking, it is much harder to find the best possible coefficients to fit non-linear models.
- But we do not have to find the best possible fit, ‘deep learning techniques, based on neural networks, offer excellent performance despite inherent difficulties in optimization’.

Blackbox vs. Descriptive Models:

- Black boxes are devices that do their task, but in some unknown manner. Stuff goes in and stuff comes out, but how the sausage is made is completely impenetrable to outsiders.

- By contrast, we prefer mathematical models that are descriptive, meaning these models provide some insight into why they are making their decisions.
- Generally, theory driven models are descriptive because they are explicit implementations of a particular well developed theory.
- If we believe the theory, we have a reason to trust the underlying model, and any resulting predictions. Certain Machine Learning (ML) models prove less opaque than others.
- Linear regression models are descriptive in nature. Because one can see exactly which variables receive the most weight, and measure how much they contribute to the resulting prediction.
- Decision tree models enable us to follow the exact decision path used to make a classification.
- But the unfortunate truth is that black box modeling techniques such as deep learning can be extremely effective.
- Generally, neural network models are completely opaque as to why they do what they do.

Flat vs. Hierarchical Models:

- Interesting problems often exist on several different levels, each of which may require independent sub-models.
- Predicting the future price for a particular stock really should involve sub-models for analyzing such separate issues as given below:
  - the general state of the economy,
  - the company's balance sheet, and
  - the performance of other companies in its industrial sector.
- Imposing a hierarchical structure on a model permits it to be built and evaluated in a logical and transparent way, instead of as a black box.
- Certain sub-problems lend themselves to theory-based models, first-principle models, which can then be used as features in a general data driven model.
- Explicitly hierarchical models are descriptive in nature (one can trace a final decision back to the appropriate top-level sub-problem, and report how strongly it contributed to making the observed result).
- The first step to build a hierarchical model is explicitly decomposing the problem into sub-problems. Basically, these represent mechanisms governing the underlying process being modeled.
- Deep learning models in mathematics can be thought of as being both at and hierarchical, at the same time.
- Deep learning models are typically trained on large sets of unwashed data, so there is no explicit definition of sub-problems to guide the sub-process.

- Looked at as a whole, the network does only one thing. But because they are built from multiple nested layers (the deep in deep learning), these deep learning models presume that there are complex features there to be learned from the lower level inputs.

Stochastic vs. Deterministic Models:

- Demanding a single deterministic prediction from a mathematical model can be a fool's errand. The world is a complex and critical place of many realities, with events that generally would not unfold in exactly the same way if time could be run over again.
- Good forecasting models incorporate such thinking and produce probability distributions over all possible events.
- Stochastic (meaning “randomly determined") modeling techniques that explicitly build some notion of probability into the model include logistic regression and Monte Carlo simulation.
- It is important that the model observe the basic properties of probabilities, including:
  - **Each probability is a value between 0 and 1:** Scores that are not constrained to be in 0 and 1 range do not directly estimate probabilities. The solution is often to put the values through a logit() function to turn them into probabilities in a principled way.
  - **Rare events do not have probability zero:** Any event i.e., possible must have a greater than zero probability of occurrence. Discounting is a way of evaluating the likelihood of unseen but possible events. Probabilities are a measure of humility about the accuracy of our model and the uncertainty of a complex world. Models must be honest in what they do and don't know.
  - **That they must sum to 1:** Independently generating values between 0 and 1 does not mean that they together add up to a unit probability, over the full event space. The solution here is to scale these values so that they do, by dividing each by the partition function. Alternately, rethink the model to understand why they didn't add up in the first place.

1.4.3 Baseline Models

- To assess the complexity of the task involves building baseline models (the simplest reasonable models that produce answers we can compare against).
- More sophisticated models should do better than baseline models, but verifying that they really do and, if so by how much, puts its performance into the proper context.
- Certain forecasting tasks are inherently harder than others. A simple baseline (yes) has proven very accurate in predicting whether the sun will rise tomorrow.
- By contrast, we could get rich predicting whether the stock market will go up or down 51% of the time. Only after we decisively beat our baselines can our models really be deemed effective.

- There are two common tasks for data science models namely, classification and value prediction.

Baseline Models for Classification:

- In classification tasks, we are given a small set of possible labels for any given item, like (man or woman), (spam or not spam) or (car or truck).
- We seek a system that will generate or produce a label accurately describing a particular instance of a person, e-mail or vehicle.
- Representative baseline models for classification include:
  1. **Uniform or Random Selection among Labels:** If we have absolutely no prior distribution on the objects, we might as well make an arbitrary selection using the broken watch method. Comparing the stock market prediction model against random coin flips will go a long way to showing how hard the problem is.
  2. **The most common Label appearing in the Training Data:** A large training dataset usually provides some notion of a prior distribution on the classes. Selecting the most frequent label is better than selecting them uniformly or randomly. This is the theory behind the sun-will-rise-tomorrow baseline model.
  3. **The most Accurate Single-feature Model:** Powerful classification baseline models strive to exploit all the useful features present in a given data set. But it is valuable to know what the best single feature can do. Occam's razor deems the simplest and easiest model to be best. Only when the complicated model beats all single-factor models does it start to be interesting.
  4. **Somebody else's Model:** Often we are not the first person to attempt a particular task. Our firm/organization may have a legacy model that we are charged with updating or revising. One of two things can happen/occur when we compare the model against someone else's work: either we beat them or we don't. If we beat them, we now have something worth bragging about. If we don't, it is a chance to learn and improve. Why didn't we win? The fact that we lost gives we certainty that your model can be improved, at least to the level of the other guy's model.
  5. **Clairvoyance:** There are circumstances when even the best possible classification baseline model cannot theoretically reach 100% accuracy.

Baseline Models for Value Prediction:

- In baseline value prediction models problems, we are given a collection of feature value pairs (f_i, v_i) to use to train a function F such that $F(v_i) = v_i$.
- Baseline models for value prediction problems follow from similar techniques to what were proposed for classification, as follows:
  1. **Mean or Median:** Just ignore the features, so we can always output the consensus value of the target. This proves that, to be quite an informative baseline, because if we cannot substantially beat always guessing the mean, either we have the wrong features or is working on a hopeless task.

2. **Linear Regression:** Linear regression suffices to understand that this powerful but simple-to-use technique builds the best possible linear function for value prediction problems. This baseline enables us to better judge the performance of non-linear models. If they do not perform substantially better than the linear classifier, they are probably not worth the effort.
3. **Value of the Previous Point in Time:** Time series forecasting is a common task/job, where we are charged with predicting the value $f(t_n, x)$ at time t_n given feature set x and the observed values $f'(t_i)$ for $1 \leq i < n$. But today's weather is a good guess for whether it will rain tomorrow. Similarly, the value of the previous observed value $f'(t_n - 1)$ is a reasonable forecast for time $f(t_n)$. It is often surprisingly difficult to beat this baseline models in practice.

1.5 MODEL EVALUATION

- The informal sniff test is perhaps the most important criteria for evaluating a model. To really know what is happening, we need to do a sniff test.
- The personal sniff test involves looking carefully at a few example instances where the model got it right and a few where it got it wrong.
- The formal evaluations that will be detailed below reduce the performance of a model down to a few summary statistics, aggregated over many instances.
- But number of sins in a model can be hidden when we only interact with these aggregate scores.
- We have no way of knowing whether there are bugs in the implementation or data normalization, resulting in poorer performance than it should have.
- Perhaps we intermingled the training and test data, yielding much better scores on the test-bed than we deserve.
- Another important issue is the degree of surprise at the evaluated accuracy of the model. Is it performing better or worse than we expected? How accurate do we think we would be at the given task, if we had to use human judgment?
- A related question is establishing a sense of how valuable it would be if the model performed just a little better.
- An NLP (Natural Language Processing) task that classifies words correctly with 95% accuracy makes a mistake roughly once every two to three sentences.
- Is this good enough? The better its current performance is, the harder it will be to make further improvements.

1.5.1 Metrics for Evaluating Classifiers

- A matrix (plural matrices) refers to a set of numbers (or object) arranged in rows and column. The advantage of matrices is that it can often simplify representing larger amount of data or relationship.

- Evaluating a classifier means measuring how accurately our predicted labels match the gold standard labels in the evaluation set.
- For the common case of two distinct labels or classes (binary classification), we typically call the smaller and more interesting of the two classes as positive and the larger/other class as negative.
- In a spam classification problem, the spam would typically be positive and the ham (non-spam) would be negative.
- This labeling aims to ensure that identifying the positives is at least as hard as identifying the negatives, although often the test instances are selected so that the classes are of equal cardinality.
- There are four possible results of what the classification model could do on any given instance, which defines the confusion matrix or contingency table shown in Fig. 1.4.
- A confusion matrix contains information about actual and predicted classifications done by a classifier. Performance of such systems is commonly evaluated using the data in the matrix.
- A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.
- The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. A confusion matrix also known as an error matrix.
- A confusion matrix is a technique for summarizing the performance of a classification algorithm.
- A confusion matrix is nothing but a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, “False Negatives (FN)”.

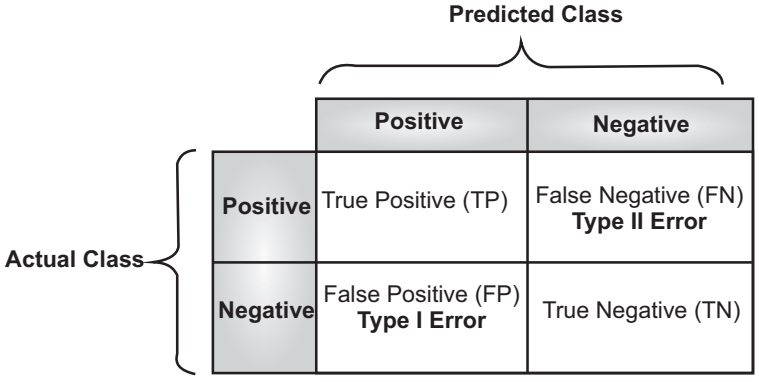


Fig. 1.4: Confusion Matrix for Binary Classifiers, Defining different Classes of Correct and Erroneous Predictions

- The rows in a confusion matrix represent actual class while the columns represent predicted class.

- The explanation of the terms associated with confusion matrix are as follows:
  - **True Positives (TP):** Here our classifier labels a positive item as positive, resulting in a win for the classifier.
  - **True Negatives (TN):** Here the classifier correctly determines that a member of the negative class deserves a negative label. Another win.
  - **False Positives (FP):** The classifier mistakenly calls a negative item as a positive, resulting in a “Type I" classification error.
  - **False Negatives (FN):** The classifier mistakenly declares a positive item as negative, resulting in a “Type II" classification error.
- Fig. 1.5 shows where these result classes fall in separating two distributions (men and women), where the decision variable is height as measured in centimeters.
- The classifier under evaluation labels everyone of height ≥ 168 centimeters as male. The purple regions represent the intersection of both male and female.
- The four possible results in the confusion matrix reflect which instances were classified correctly (TP and TN) and which ones were not (FN and FP).

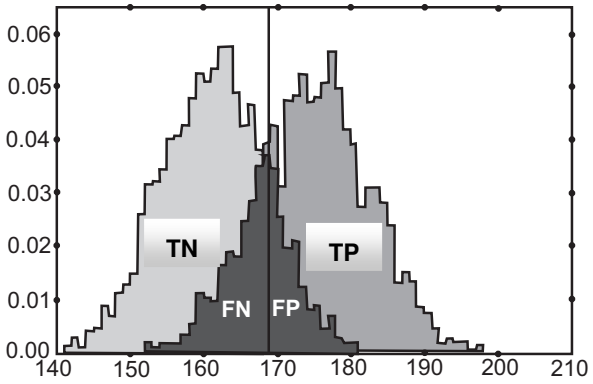


Fig. 1.5

Example: Consider Test set of 1100 images, from these images 1000 are non cat images and 100 are cat images. The following fig shows the confusion matrix with TP=90, FN=10, TN= 940 and FP=60

		Actual Class	
		Cat	Non-Cat
Predicted Class	Cat	90	60
	Non-Cat	10	940

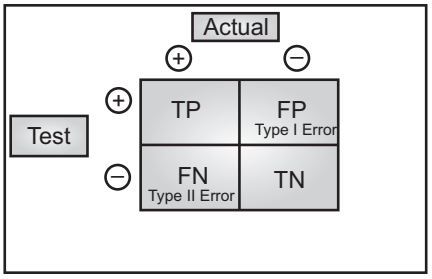


Fig. 1.6

True Positive: We predicted positive and it's true. In the Fig. 1.6, we predicted that 90 images are cat images.

True Negative: We predicted negative and it's true. In the Fig. 1.6, we predicted that 940 images are non cat.

False Positive (Type I Error): We predicted positive and it's false. In the Fig. 1.6, we predicted that 60 images are cat images but actually not.

False Negative (Type II Error): We predicted negative and it's false. In the Fig. 1.6, we predicted that 10 images are non cat but actually yes.

Statistic Measures for Classifier:

1. **Accuracy:** The accuracy of classifier, the ratio of the number of correct predictions over total predictions. We can calculate accuracy by confusion matrix with the help of following formula:

   $$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

2. **Precision:** The precision measures the number of positive values predicted by the classifier that are actually positive. We can calculate precision by confusion matrix with the help of following formula:

   $$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Recall or Sensitivity:** Recall determines the proportion of the positives values that were accurately predicted. Sensitivity or Recall means out of all actual positives, how many did we predict as positive. We can calculate recall by confusion matrix with the help of following formula:

   $$\text{Recall} = \frac{TP}{TP + FN}$$

4. **F-score:** The F-score (or sometimes F1-score) is such a combination, returning the harmonic mean of precision and recall. We can calculate recall by confusion matrix with the help of following formula:

   $$F = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

- For above example, consider following the confusion matrix:

  [[90 60]

  [10 940]]

  TP = 90, FN = 10, TN = 940 and FP = 60

- For above binary classifier,

  TP + TN = 90 + 940 = 1030 and

  TP + FP + FN + TN = 90 + 60+ 10 + 940 = 1100

Hence, Accuracy = 1030 / 1100 = 0.9364.

- For the above binary classifier,
  TP = 90 and
  TP + FP = 90 + 60= 150
  Hence, Precision = 90 / 150 = 0.6.
- For above built binary classifier,
  TP = 90 and
  TP + FN = 90 + 10 = 100
  Hence, Recall = 90 / 100 = 0.9.
- The F1 score takes into account the precision and recall of the model. The F1 score ranges from 0 to 1.
- The F1 score computes the harmonic mean of precision and recall, giving a higher weight to the low values.

1.5.2 Class Imbalance

- TheROC curve is a graphical plot that shows the performance of a binary classifier. The curve is created by plotting the True positive rate (sensitivity) against the False positive rate (1-specificity) at various thresholds.
- The value of AUC_{ROC} can vary between 1.0 and 0, where 1 indicates a perfect classifier with an ideal separation of the two classes and an AUC_{ROC} of 0.5 represents worthless classifier.
- The AUC_{ROC} is insensitive to class imbalance; if the majority labels of the data are positive or negative, a classifier which always outputs 1 or 0, respectively, will have a 0.5 score although it will achieve a very high accuracy.

1.5.2.1 ROC and AUC Curves

- When we need to visualize the performance of the binary classifier, we use the AUC (Area Under the Curve) and ROC (Receiver Operating Characteristics) curve.
- A Receiver Operating Characteristic Curve (or ROC curve) is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The Receiver Operating Characteristic (ROC) curve provides a visual/graphical representation of our complete space of options in putting together a classifier.
- Each point on ROC curve represents a particular classifier threshold, defined by its false positive and false negative rates.
- These rates are in turn defined by the count of errors divided by the total number of positives in the evaluation data, and perhaps multiplied by one hundred to turn into percentages.

- Consider what happens as we sweep our threshold from left to right over these distributions.
- Every time we pass over another example, we either increase the number of true positives (if this example was positive) or false positives (if this example was in fact a negative).
- At the very left, we achieve true/false positive rates of 0%, since the classifier labeled nothing as positive at that cutoff.
- Moving as far to the right as possible, all examples will be labeled positively, and hence both rates become 100%.
- Each threshold in between defines a possible classifier, and the sweep defines a staircase curve in true/false positive rate space taking us from (0%, 0%) to (100%, 100%).
- An ROC curve is the most commonly used way to visualize the performance of a binary classifier and AUC is (arguably) the best way to summarize its performance in a single number.
- The area under the ROC curve (AUC) is often used as a statistic measuring the quality of scoring function defining the classifier.
- The best possible ROC curve has an area of 100% $\times$ 100 % $\rightarrow$ 1, while the monkey's triangle has an area of 1/2. The closer the area is to 1, the better the classification function is.
- The Area Under the Curve (AUC) is another evaluation metric that we can use for classification models.
- The 45 degree line is the baseline for which the AUC is 0.5. The perfect model will have an AUC of 1.0. The closer the AUC to 1.0, the better the predictions.

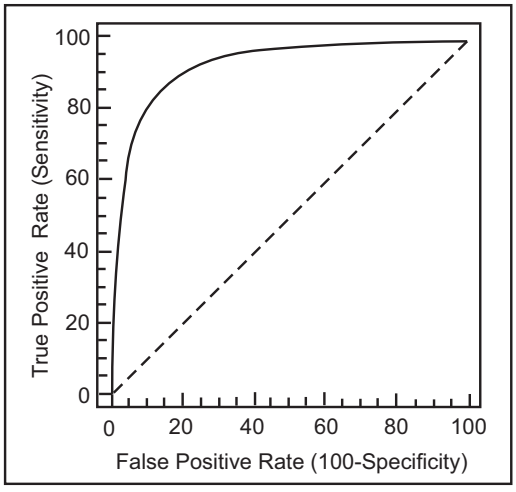


Fig. 1.7: ROC Curve

1.5.3 Evaluating Value Prediction Models

- Value prediction problems can be thought of as classification tasks, but over an infinite number of classes.
- However, there are more direct ways to evaluate regression systems, based on the distance between the predicted and actual values.
- For numerical values, error is a function of the difference between a forecast y' = f(x) and the actual result y.
- Measuring the performance of a value prediction system involves following two decisions:
  1. fixing the specific individual error function, and
  2. selecting the statistic to best represent the full error distribution.
- The choices for the individual error function include following types of errors:
  - **Absolute Error:** The value $\Delta = y' - y$ has the virtue of being simple and symmetric, so the sign can distinguish the case where $y' > y$ from $y > y'$. The problem comes in aggregating these values into a summary statistic. Do offsetting errors like − 1 and 1 mean that the system is perfect? Typically the absolute value of the error is taken to obliterate the sign.
  - **Relative Error:** The absolute magnitude of error is meaningless without a sense of the units involved. An absolute error of 1.2 in a person's predicted height is good if it is measured in millimeters, but terrible if measured in miles.

    Normalizing the error by the magnitude of the observation produces a unit-less quantity, which can be sensibly interpreted as a fraction or (multiplied by 100%) as a percentage: $\in = (y - y')/y$. Absolute error weighs instances with larger values of y as more important than smaller ones, a bias corrected when computing relative errors.
  - **Squared Error:** The value Δ^2 $(y' - y)^2$ is always positive and hence these values can be meaningfully summed. Large errors values contribute disproportionately to the total when squaring: Δ^2 for $\Delta = 2$ is four times larger than Δ^2 for $\Delta = 1$. Thus, outliers can easily come to dominate the error statistic in a large ensemble.
- It is a very better idea to plot a histogram of the absolute error distribution for any value predictor, as there is much we can learn from it.
- The distribution should be symmetric and centered around zero. It should be bell-shaped, meaning small errors are more common than big errors. And extreme outliers should be rare.
- If any of the conditions are wrong, there is likely a simple way to improve the forecasting procedure.
- For example, if it is not centered around on zero, adding a constant offset to all forecasts will improve the consensus results.

- Fig. 1.8 shows the absolute error distributions from two models for predicting the year of authorship of documents from their word usage distribution.
- On the left, we see the error distribution for the monkey, randomly guessing a year from 1800 to 2005. What do we see? The error distribution is broad and bad, as we might have expected, but also asymmetric.
- Far more documents produced positive errors than negative ones. Why? The test corpus apparently contained more modern documents than older ones, so is more often positive than negative.

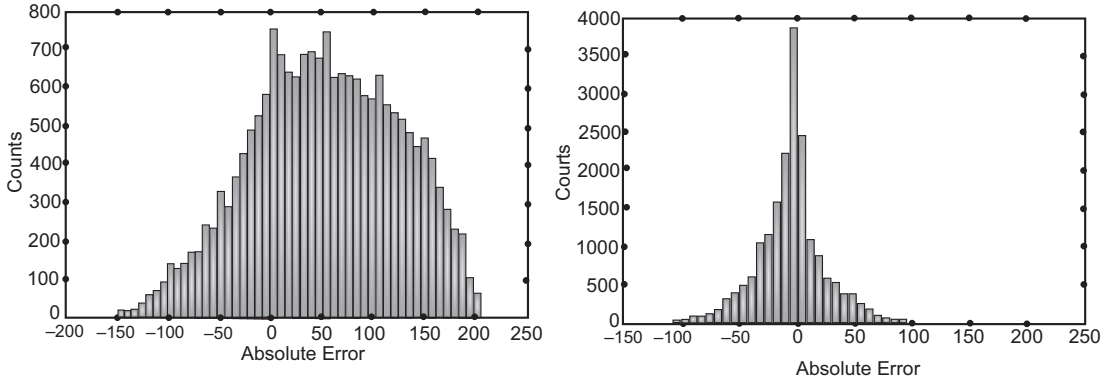


Fig. 1.8: Error Distribution Histograms for Random (Left) and Naive Bayes Classifiers Predicting the Year of Authorship for Documents (Right)

- In contrast, Fig. 1.8 (right) presents the error distribution for our naïve Bayes classifier for document dating. This looks much better: there is a sharp peak around zero and much narrower tails.
- But the longer tail now resides to the left of zero, telling us that we are still calling a distressing number of very old documents modern. We need to examine some of these instances, to figure out why that is the case.
- We need a summary statistic reducing such error distributions to a single number, in order to compare the performance of different value prediction models.
- A commonly-used statistic is Mean Squared Error (MSE), which is computed as follows:

$$\text{MSE}(Y, Y') = \frac{1}{n} \sum_{i=1}^{n} (y'_i - y_i)^2$$

- Because it weighs each term quadratically, outliers have a disproportionate effect. Thus, median squared error might be a more informative statistic for noisy instances.
- Root Mean Squared (RMSD) error is simply the square root of mean squared error:

$$\text{RMSD}(\Theta) = \sqrt{\text{MSE}(Y, Y')}.$$

PRACTICE QUESTIONS

Q. I Multiple Choice Questions:

1. Which is a collection of techniques used to extract value from data.?
   (a) Data science (b) Data analysis
   (c) Data analytics (d) Exploratory analytics
2. Which is the science of examining raw data with the purpose of drawing conclusions about that information?
   (a) Data science (b) Data analysis
   (c) Data analytics (d) Exploratory analytics
3. Which is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information or insights?
   (a) Data science (b) Data analysis
   (c) Data analytics (d) Exploratory analytics
4. The types of data analytics includes,
   (a) Descriptive Analytics (what happened?)
   (b) Diagnostic Analytics (why did it happen?)
   (c) Predictive Analytics (what will it happen?)
   (d) All of the mentioned
5. Which analytics deals with prediction of future based on the available current and past data?
   (a) Descriptive Analytics (b) Diagnostic Analytics
   (c) Mechanistic analytics (d) Predictive Analytics
6. Which is data analytics attempts to find hidden, unseen or previously unknown relationships?
   (a) Exploratory Analytics (b) Diagnostic Analytics
   (c) Mechanistic Analytics (d) Predictive Analytics
7. Which analytics allow data scientists to understand clear alterations in variables which can result in changing of variables?
   (a) Exploratory Analytics (b) Diagnostic Analytics
   (c) Mechanistic Analytics (d) Predictive Analytics
8. Which is the philosophical principle states that, the simplest explanation is the best explanation?
   (a) Occam's Analysis (b) Occam's mazor
   (c) Occam's Analytics (d) Occam's razor
9. Which is the process where team has to deploy the planned model in a real-time environment?
   (a) Model analytics (b) Model building
   (c) Model analysis (d) Model science

10. Which analytics looks at data and analyzes past events for insight as to how to approach the future?
    (a) Descriptive Analytics (b) Diagnostic Analytics
    (c) Mechanistic Analytics (d) Predictive Analytics
11. Which is analytics is kind of root cause analysis that focuses on the processes and causes, key factors and unseen patterns?
    (a) Descriptive Analytics (b) Diagnostic Analytics
    (c) Mechanistic Analytics (d) Predictive Analytics
12. Which is a graphical plot that illustrates the performance of a binary classifier?
    (a) ROC curve (b) COR curve
    (c) ETL curve (d) None of the mentioned
13. First-principle models can employ the full weight of classical mathematics such as,
    (a) calculus (b) algebra
    (c) geometry (d) All of the mentioned
14. Which analysis uses past data to create a model that answer the question, what will happen?
    (a) descriptive (b) diagnostic
    (c) predictive (d) Predictive

Answers

1. (a)	2. (c)	3. (b)	4. (d)	5. (d)	6. (a)	7. (c)	8. (d)	9. (b)	10. (a)
11. (b)	12. (a)	13. (d)	14. (c)						

Q. II Fill in the Blanks:

1. The purpose of _____ data analysis is to check for missing data and other mistakes.
2. Data _____ is a broad term capturing the endeavor of analyzing data into information into knowledge.
3. _____ is used for the discovery, interpretation, and communication of meaningful patterns and/or insights in data.
4. Data analytics is defined as, a science of _____ meaningful, valuable information from raw data.
5. A data _____ works with massive amount of data and responsible for building and maintaining the data architecture of a data science project.
6. An analytics _____ is a part of data lake architecture that allows you to store and process large amounts of data.
7. the _____ -layer framework of data analytics consists of a data management layer an analytics engine layer and a presentation layer.
8. EDA is an approach to analyzing datasets to _____ their main characteristics, often with visual/graphical methods.

9. Mechanistic analytics allow data scientists to understand clear _____ in variables which can result in changing of variables.
10. The _____ –variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.
11. Errors of bias produce _____ models.
12. _____ analysis is a process for estimating the relationships among variables. It is a statistical tool used for the investigation of relationships between variables.
13. _____ is the process of encapsulating information into a tool which can forecast and make predictions.
14. _____ models are structured around some idea of what causes future events to happen.
15. Occam's razor is a problem-solving principle arguing that _____ is better than complexity.
16. _____ models are built using domain-specific knowledge to guide their structure and design.
17. _____ models are governed by equations that weigh each feature variable by a coefficient reflecting its importance, and sum up these values to produce a score.
18. _____ models for value prediction problems follow from similar techniques to what were proposed for classification.
19. A _____ matrix contains information about actual and predicted classifications done by a classifier.
20. The _____ of classifier, the ratio of the number of correct predictions over total predictions.
21. The _____ is another evaluation metric that we can use for classification models.
22. Predictive analytics makes predictions about future outcomes using _____ data combined with statistical modeling, data mining techniques and machine learning.

Answers

1. exploratory	2. science	3. Analytics	4. extracting
5. engineer	6. sandbox	7. four	8. summarize
9. alterations	10. bias	11. underfit	12. Regression
13. Modeling	14. Predictive	15. simplicity	16. Ad hoc
17. Linear	18. Baseline	19. confusion	20. accuracy
21. AUC	22. historical		

Q. III State True or False:

1. Data science is a collection of techniques used to extracting insights from large datasets.

2. Recall determines the proportion of the positives values that were accurately predicted.
3. Analytics defines the science behind the analysis.
4. Data analytics is the process of exploring the data from the past to make appropriate decisions in the future by using valuable insights.
5. The art and science of refining data to fetch useful insight which further helps in decision making is known as analysis.
6. Descriptive analytics enables learning from the past and assessing how the past might influence future outcomes.
7. Overfitting occurs when a model tries too hard to achieve accurate performance on its training data.
8. Exploratory data analytics attempts to find hidden, unseen, or previously unknown relationships.
9. Predictive analysis, as the name suggests, deals with prediction of future based on the available current and past data.
10. We need a summary statistic reducing such error distributions to a single number, in order to compare the performance of different value prediction models.
11. Predictive analytics is often associated with data science.
12. The accuracy is such a combination, returning the harmonic mean of precision and recall.
13. Underfitting occurs when a statistical model cannot adequately capture the underlying structure of the data.
14. A confusion matrix is also known as error matrix.
15. An ROC curve is the most commonly used way to visualize the performance of a binary classifier.

Answers

1. (T)	2. (T)	3. (T)	4. (T)	5. (F)	6. (T)	7. (T)	8. (T)	9. (T)	10. (T)
11. (T)	12. (F)	13. (T)	14. (T)	15. (T)	16. (T)				

Q. IV Answer the following Questions:

(A) Short Answer Questions:

1. What is data science?
2. Define the term analytics.
3. Enlist types of data analytics.
4. Define data analysis.
5. Define mathematical model.
6. What is the purpose of diagnostic analytics?
7. Define class imbalance.

8. Differentiate between predictive analytics and prescriptive analytics. Any two points.
9. Define exploratory analysis.
10. Define linear model.
11. What is model evaluation?
12. Define predictive analytics.
13. What is the purpose of AUC and AOC curves?
14. Define baseline model.
15. Define descriptive analytics.
16. Define the terms metric and classifier.

(B) Long Answer Questions:

1. Define data science. What is its purpose? Explain in detail.
2. What is data analytics? Enlist its different roles. Also state its advantages and disadvantages.
3. With the help of diagram describe lifecycle of data analytics.
4. Explain four layers in data analytics framework diagrammatically.
5. Differentiate between data analysis and data analytics.
6. What are the types of data analytics? Describe two of them in detail.
7. What is prescriptive analytics? Explain in detail.
8. What exploratory analytics? What is its purpose? Explain with example.
9. Write a short note on: Mechanistic analytics.
10. What is mathematical model? List its types. Explain two of them in detail.
11. What is linear and non linear model? Compare them.
12. What is baseline model? Enlist two of them in detail.
13. How to evaluate a model? Describe in detail.
14. Write a short note on: Metrics for evaluating classifiers.
15. What is confusion matrix? How to use it in data analytics? Explain diagrammatically.
16. Define accuracy, precision, recall and f-score.
17. What is ROC curve? How to implement it? Explain with example.
18. What is class imbalance? Describe in detail.
19. Write a short note on: Evaluating value prediction models.

■■■

CHAPTER 2

Machine Learning Overview

Objectives...

- To understand Concept of Machine Learning
- To learn Deep Learning and Artificial Intelligence
- To learn Connect of Classification and Regression
- To study different Types of Machine Learning

2.0 INTRODUCTION

- Machine learning is buzzwords in today's technical and data driven world. Learning is the process of converting experience into expertise or knowledge.
- Machine Learning (ML) is a field of computer science that studies algorithms and techniques for automating solutions to complex problems that are hard to program using conventional programming methods.
- The conventional/traditional programming method consists of following two distinct steps. Given a specification for the program (i.e., what the program is supposed to do and not and how).
  - 1^{st} step is to create a detailed design for the program i.e., a fixed set of steps/stages used to solving the problem.
  - 2^{nd} step is to implement the detailed design as a program in a computer language.
- Though data science includes machine learning as one of its fundamental areas of study, machine learning in itself is a vast research area of study that requires good skills and experience to expertise.
- The basic idea of machine learning is to allow machines (computers) to independently learn from the wealth of data that is fed as input into the machine.
- To master in machine learning, a learner needs to have an in-depth knowledge of computer fundamentals, programming skills, data modeling and evaluation skills, probability and statistics.
- With the advancement of new technology, machines are being trained to behave like a human in decision-making capability.

- In doing so, it is necessary to automate decisions that can be inferred by the machines with the interaction with the environment and understanding from past knowledge.
- The field of machine learning deals with all those algorithms that help machines to get self-trained in this process.
- Machine learning techniques are broadly categorized into supervised machine learning, unsupervised machine learning, and reinforcement learning.
  1. **Supervised Machine Learning** sometimes described as "learn from the past to predict the future". Supervised machine learning is a field of learning where the machine learns with the help of a supervisor and instructor.
  2. **Unsupervised Machine Learning** the machine learns without any supervision. The goal of unsupervised learning is to model the structure of the data to learn more about the data.
  3. **Reinforcement Machine Learning** happens with an interaction with the environment. If we are assumed to be a program and with every encounter with the environment the program eventually starts learning then the process is called reinforcement learning.
- Today, Deep Learning (DL) is a fast-growing field of research and its applications run the gamut of structured and unstructured data (text, voice, images, video and so on).
- Deep learning (DL) is a subset of ML and ML is a subset of AI. Nowadays, AI and DL are the latest technologies that are doing much more.
- They are supporting humans in complex and creative problem-solving by analyzing vast amounts of data and identifying trends that were previously impossible to detect.

2.1 INTRODUCTION TO MACHINE LEARNING, DEEP LEARNING AND ARTIFICIAL INTELLIGENCE

- The recent extraordinary growth of artificial intelligence and its applications has been paralleled by a surge of interest in machine learning.
- Machine learning a field concerned with the developing computational theories of learning processes and building learning machines.
- Because the ability to learn is clearly fundamental to any intelligent behavior, the concerns and goals of machine learning are central to the progress of Artificial Intelligence (AI).
- Machine Learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. ML is seen as a part of artificial intelligence.
- Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

- Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans. The ultimate goal of AI is to make machines as intelligent as humans.
- Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.
- Deep Learning (DL) in short is part of the family of machine learning methods which are themselves a subset of the broader field of Artificial Intelligence (AI).
- Artificial Intelligence (AI) is any code, algorithm or technique that enables a computer to mimic human cognitive behavior or intelligence.
- Machine Learning (ML) is a subset of AI that uses statistical methods to enable machines to learn and improve with experience.
- Deep Learning is a subset of ML, which makes the computation of multi-layer neural networks feasible.
- Machine Learning is seen as shallow learning while Deep Learning is seen as hierarchical learning with abstraction.
- Artificial Intelligence is one of the most popular trends of recent times. Machine learning and deep learning constitute artificial intelligence.
- Fig. 2.1 shows the relationship of Artificial Intelligence, Machine Learning and Deep Learning.

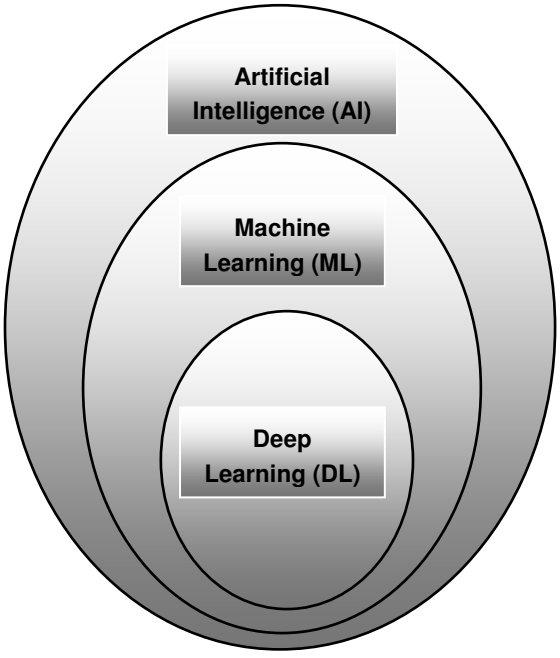
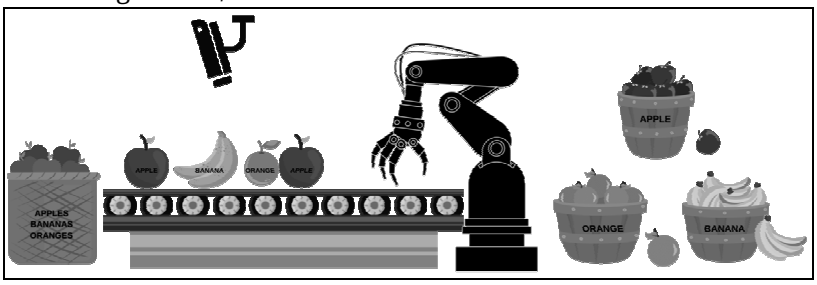
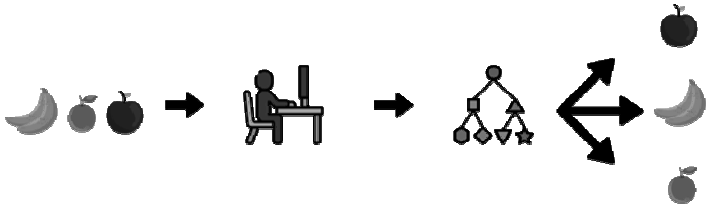
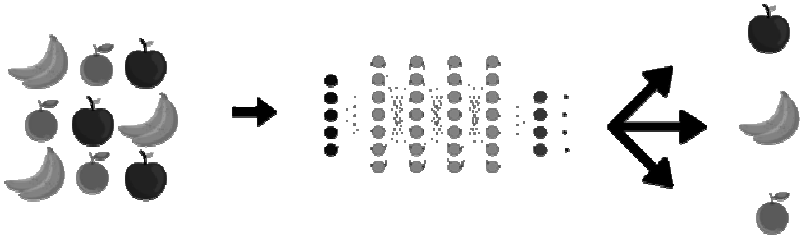


Fig. 2.1: Relation between AI, ML and DL

Comparison between AI, ML and DL:

Artificial Intelligence (AI)	▪ AI is the study of pattern recognition and mimicking human behavior. ▪ AI powered computers has started simulating the human brain work style, sensation, actions, interaction, perception and cognitive abilities. Though all of them are at low/basic level. ▪ AI is a subset of Data Science that was created with the aim of creating machines that are intelligent. ▪ Artificial Intelligence (AI) refers to the simulation of human intelligence processes by machines, including learning, reasoning and so on. In simple words, AI represents simulated intelligence in machines. ▪ Making machines intelligent may or may not need high computational power as it depends on the nature of the task that needs to be learnt by the machine or in better terms, needs to be automated. ▪ The objective of AI is to building machines that are capable to think likes humans. ▪ AI is a vast field containing several algorithms that may not necessarily be easy to understand based solely on logical thinking. ▪ In other words, AI means ability of machines to imitate human intelligence. ▪ Example, an AI-based algorithm is created that segregates the fruits using decision logic within a rule-based engine. For example, if an apple is on the convey or belt, a scanner would scan the label, informing the AI algorithm that the fruit is indeed an apple. Then the apple would be routed to the apple fruit tray via sorting rollers/arms.  ▪ **Application:** AI is usually adopted to solve customer service issues, inform people about the latest news along with giving them live traffic updates and weather forecast.

Machine Learning (ML)	▪ Machine learning is generally considered to be a subset of AI. ▪ ML is the practice of getting machines to make decisions without being programmed. ▪ Machine Learning (ML) algorithms incorporate intelligence into to machines by automatically learning from data. ▪ The objective of ML is to learn through data to solve the problem. ▪ ML requires low computational power as its algorithms can works on normal low performance machines. ▪ ML is a way to archive AI. ▪ In ML understanding how certain results are given by these algorithms is very easy as they work on a set of defined rules. ▪ Example, An ML-based algorithm is now proposed to solve the problem of fruit sorting by enhancing the AI-based approach when labels are not present. To create a ML model, a definition of what each fruit looks like is required: this is termed feature extraction. To do this, features and attributes that characterize each fruit are used to create a blueprint. Features such as sizes, colors, shapes, etc., are extracted and used to train the algorithm to classify the fruits accordingly. For example, once the ML algorithm has seen what a banana looks like many times i.e., has been trained, when a new fruit is presented, it can then compare the attributes against the learned features to classify the fruit.  **Input** **Feature Extraction** **Classification** **Output** ▪ **Application:** ML is used to power recommendation engines that provide suggestions based on past customers' behaviors.

Deep Learning (DL)	▪ DL is a subset of ML. DL is based on an neural network which is nothing but a mimic of the working of the human brain. ▪ The objective of DL is to build neural networks that automatically discover patterns for feature predictions. ▪ DL algorithms are dependent on high performance computational power because it solve high complex problems. ▪ DL is essentially a set of techniques that help to parameterize deep neural network structures i.e. neural networks numerous layers and parameters. ▪ In DL is very difficult to understand why the predicted values are the way they are. We can derive the reason behind the predictions mathematically but that's very complex, complicated and expensive. ▪ DL refers to a particular approach used for creating and training neural networks that are considered highly promising decision-making nodes. ▪ Example, a DL based algorithm is now proposed to solve the problem of sorting any fruit by totally removing the need for defining what each fruit looks like. By providing the DL model with lots of images of the fruits, it will build up a pattern of what each fruit looks like. The images will be processed through different layers of neural network within the DL model. Then each network layer will define specific features of the images, like the shape of the fruits, size of the fruits, color of the fruits, etc.  **Input** **Future Extraction + Classification** **Output** ▪ **Application:** DL is used to develop highly automated systems such as self-driving cars. Through their sensors and onboard analytics, these cars can reorganize obstacle and facilitate situational awareness.

2.1.1 Machine Learning

- With advances in computer technology, we currently have the ability to store and process large/massive amounts of data, as well as to access it from physically distant locations over a computer network.
- The data or information is increasing day by day, but the real challenge is to make sense of all the data.
- Businesses and organizations are trying to deal with it by building intelligent systems using the concepts and methodologies from Data science, Data Mining and Machine learning.
- Among them, machine learning is the most exciting field of computer science. It would not be wrong if we call machine learning the application and science of algorithms that provides sense to the data.
- Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data.
- Machine Learning (ML) is that field of computer science with the help of which computer systems can provide sense to data in much the same way as human beings do.
- In simple words, ML is a type of artificial intelligence that extracts patterns out of raw data by using an algorithm or method.
- The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.
- The primary aim of ML is to allow the computers learn automatically without human intervention.
- The idea behind machine learning is that instead of hard-coding the logic, data is fed into the machines and make machines themselves learn from the data by identifying patterns from the data.

2.1.1.1 Definition and Model of Machine Learning

- Machine learning is a set of methods that computers use to make and improve predicts or behaviors based on data. For example, to predict the value of a farmhouse, the computers would learn patterns from past farmhouse sales.
- The formal definition of ML by Mitchell, “a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”
- The above definition is basically focusing on three parameters, also the main components of any learning algorithm, namely Task (T), Performance (P) and experience (E).

- ML is a field of AI consisting of learning algorithms that Improve their performance (P), At executing some task (T) and Over time with experience (E). Based on these parameters, the Fig. 2.2 represents a machine learning model.

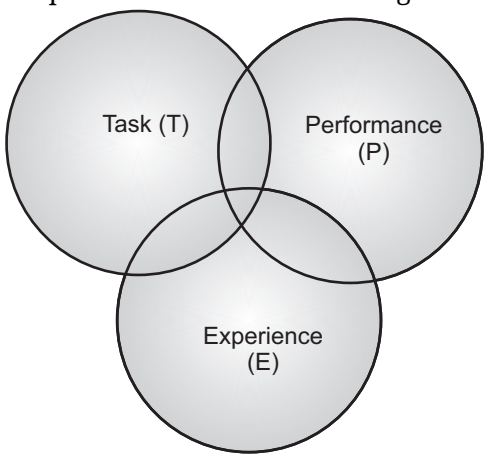


Fig. 2.2: Machine Learning (ML) Model

- Machine learning is an application of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- A computer program is said to learn from Experience E with respect to some Task T and some Performance measure P, if its performance on T, as measured by P, improves with experience E.
- For example, Smart Homes, where
  - **T:** Estimate the desired Temperature.
  - **E:** Learning from Temperature dataset.
  - **P:** Accuracy of the desired Temperature.
- Let us discuss the parameters in ML in detail:

1. Task (T):

  - From the perspective of problem, we may define the task T as the real-world problem to be solved.
  - The problem can be anything like finding best house price in a specific location or to find best marketing strategy etc.
  - On the other hand, if we talk about machine learning, the definition of task is different because it is difficult to solve ML based tasks by conventional programming approach.
  - A task T is said to be a ML based task when it is based on the process and the system must follow for operating on data points.
  - The examples of ML based tasks are Classification, Regression, Structured annotation, Clustering, Transcription etc.

2. **Experience (E):**
   - As name suggests, it is the knowledge gained from data points provided to the algorithm or model.
   - Once, provided with the dataset, the model will run iteratively and will learn some inherent pattern. The learning thus acquired is called Experience (E).
   - Making an analogy with human learning, we can think of this situation as in which a human being is learning or gaining some experience from various attributes like situation, relationships etc.
   - Supervised, unsupervised and reinforcement learning are some ways to learn or gain experience. The experience gained by out ML model or algorithm will be used to solve the Task (T).
3. **Performance (P):**
   - An ML algorithm is supposed to perform task and gain experience with the passage of time.
   - The measure which tells whether ML algorithm is performing as per expectation or not is its performance (P).
   - The P is basically a quantitative metric that tells how a model is performing the Task (T) using its Experience (E).
   - There are many metrics that help to understand the ML performance, such as accuracy score, F1 score, confusion matrix, precision, recall, sensitivity etc.

Defining the Learning Task

Improve on task T, with respect to Performance metric P,
based on Experience E

Sr. No.	Task (T)	Performance (P)	Experience (E)
1.	Playing Checkers	% of games won against an arbitrary opponent.	Playing practice games against itself.
2.	Recognizing hand written words.	% of words correctly classified.	Data-based of human-labeled images of handwritten words.
3.	Driving on four-lane highways using vision sensors.	Average distance travelled before a human-judged error.	A sequence of images and steering commands recorded while observing a human driver.
4.	Categorize email messages as spam or legitimate.	% of email messages correctly classified.	Database of emails, some with human- given labels.

2.1.1.2 Need for Machine Learning

- The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly.

When to apply Machine Learning (ML):

1. Human expertise is absent (For example, Navigating on Mars).
2. Humans are unable to explain their expertise (For examples, Vision language, Speech language and so on).
3. Solution changes with time (For examples, Tracking, Temperature controller and so on).
4. The problem size is to vast for limited reasoning capabilities (For examples, Matching ads to Facebook pages, Calculating Webpage ranks and so on).

- As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.
- Human beings, at this moment, are the most intelligent and advanced species on earth because they can think, evaluate and solve complex problems.
- On the other side, AI is still in its initial stage and haven't surpassed human intelligence in many aspects.
- Then the question is that what is the need to make machine learn? The most suitable reason for doing this is, "to make decisions, based on data, with efficiency and scale".
- Lately, organizations are investing heavily in the latest technologies like Artificial Intelligence, Machine Learning and Deep Learning to get the key information from data to perform several real-world tasks and solve problems.
- We can call it data-driven decisions taken by machines, particularly to automate the process.
- These data-driven decisions can be used, instead of using programming logic, in the problems that cannot be programmed inherently.
- The fact is that we can't do without human intelligence, but other aspect is that we all need to solve real-world problems with efficiency at a huge scale. That is why the need for machine learning arises.
- Recent progress in machine learning has been driven by the development of new learning algorithms and innovative researches, backed by the ongoing explosion in online and offline data.
- Also, the availability of low cost computation plays an important role. Here, are the few driving forces that justify the need of machine learning:
  1. **Diversity of Data:** Data is being generated from different channels and its nature and format are different.

2. **Capacity and Dimension:** The increases in the number of data sources and the globalization of diversification of businesses have led to the exponential growth of the data.
3. **Speed:** As data volume increases, so must the speed at which data is captured and transformed.
4. **Complexity:** With the increasing complexity of data, high data quality and security is required to enable data collection, transformation, and analysis to achieve expedient decision making.
5. **Applicability:** These aforementioned factors can compromise the applicability of the data to business process and performance improvement.

2.1.1.3 Advantages and Disadvantages of Machine Learning

Advantages of Machine Learning:

1. It is used in variety of applications such as banking and financial sector, healthcare, retail, publishing and social media, robot locomotion, game playing etc.
2. It has capabilities to handle multi-dimensional and multi-variety data in dynamic or uncertain environments.
3. It allows time cycle reduction and efficient utilization of resources.
4. Source programs such as RapidMiner help in increased usability of algorithms for various applications.
5. Due to Machine Learning, there are tools available to provide continuous quality improvement in large and complex process environments.
6. The process of automation of tasks is easily possible.

Disadvantages of Machine Learning:

1. Acquisition of relevant data is the major challenge. Based on different algorithms data need to be processed before providing as input to respective algorithms. This has significant impact on results to be achieved or obtained.
2. It is impossible to make immediate accurate predictions with a machine learning system.
3. Machine learning needs a lot of training data for future prediction.
4. Interpretation of results is also a major challenge to determine effectiveness of machine learning algorithms.
5. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.
6. ML models are the consumption of time especially for data acquisition, feature extraction and retrieval.
7. Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.
8. Complexity of the ML model makes it quite difficult to be deployed in real life.

2.1.1.4 Uses of Machine Learning

- Below are some most trending real-world uses of Machine Learning (ML):
  1. **Speech Recognition:** Speech recognition is a process of converting voice instructions into text and known as computer speech recognition. At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Cortana and Alexa are using speech recognition technology to follow the voice instructions.
  2. **Image Recognition:** It is one of the most common applications of machine learning used to identify objects, persons, places, digital images, etc. The popular example includes Google Lens uses image **recognition**. **Google Lens** identifies objects through a camera and Facebook provides us a feature of auto friend tagging suggestion.
  3. **Stock Market Trading:** Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's **long short term memory neural network** is used for the prediction of stock market trends.
  4. **Medical Sector:** Machine learning can never completely automate the medical field. However, it can help the doctors in diagnosis. In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models.
  5. **Automatic Language Translation:** Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.
  6. **Banking:** Various banking areas use ML. It is now used in customer service, mobile banking, etc. It can help to detect credit card frauds.
  7. **Online Fraud Detection:** Machine learning is used to track monetary frauds online. For example, Paypal is using ML to prevent money laundering. Traffic
  8. **Traffic Congestion Analysis and Predictions:** GPS navigation services monitor the user's location and velocities and use them to build a map of current traffic. This helps in preventing the traffic congestions. Machine learning in such scenarios helps to estimate the regions where congestion can be found based on previous records.
  9. **Product Recommendations:** If a user purchases or searches for a product online, he/she keeps on receiving emails for shopping suggestions and ads about that product. Based on previous user behavior, on a website/app, past purchases, items

liked or added to cart, brand preferences etc., the product recommendations are sent to the user. Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix etc., for product recommendation to the user.

2.1.2 Deep Learning

- The newest member of the data science is deep learning. Deep learning is a subset of machine learning.
- Deep learning is field in machine learning that has started gaining popularity with the computational power support by the hardware and multi-core processors' availability.
- Deep Learning (DL) belongs to a family of AI techniques whose models extract essential features from the data and find meaningful patterns in the dataset.
- Unlike traditional methods in which the developer of the model has to choose and encode features ahead of time, DL enables a model to learn features automatically.
- In this way, a DL model determines a representation of the data, making deep learning part of the larger field of representation learning.
- DL is a category of ML in which a network or model learns directly from dataset. The word "deep" denotes the number of layers in the network - the deeper network consists of many layers.
- A deep learning network is any network with three or more hidden layers between the input and output layers.
- DL is particularly suitable to identify modern smart applications such as face recognition, text translation, voice recognition, and advanced driver assistance systems including lane classification and traffic sign recognition.
- Deep learning is a new branch of the machine learning technique that enables computers to follow the human learning process, that is, to learn by example.
- In the deep learning technique, the model is trained by a large set of labeled data and neural network architectures that contain many hidden layers.
- The model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance.
- Deep learning is a key technology behind the current research in driverless cars, enabling them to distinguish a pedestrian from a lamppost, or to recognize the traffic symbols, etc.
- While deep learning was first proposed in the 1980s, it has only recently become a success, for following two reasons:
  1. Deep learning requires large amounts of labeled data, which nowadays has become available due to the large number of sensors and electronics that we use

in our daily lives, and the huge amount of data those sensors routinely generate. For example, driverless car development requires millions of images and thousands of hours of video.

2. Deep learning requires substantial computing power, including high-performance GPUs that have a parallel architecture, efficient for deep learning. When combined with cloud computing or distributed computing, this enables the training time for a deep learning network to be reduced from the usual weeks to hours or even less.

2.1.2.1 Definition of Deep Learning

- Deep learning is defined as, a class of machine learning algorithms thatuses multiple layers to progressively extract higher-level features from the raw input.
- For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces.
- Deep Learning (DL) is a machine learning technique that constructs Neural Networks (NNs) and Artificial Neural Networks (ANNs) to mimic the structure and function like a human brain.
- As DL is the subset of both AI and ML, all the characteristics of AI and ML are propagated to DL inherently. It deals with the large amount of data.
- Deep learning is the subfield of artificial intelligence that focuses on creating large neural network models that are capable of making accurate data-drive decisions.
- Deep learning is a branch of machine learning that uses neural networks with many layers.
- Following are the various deep learning tools available in the market today:
  1. **Tensor Flow** is one of the best frameworks is used for natural language processing, text classification and summarization, speech recognition and translation and more. It is flexible and has a comprehensive list of libraries and tools which lets us to build and deploy ML applications.
  2. **Microsoft Cognitive Toolkit** is the most effective for image, speech and text-based data.
  3. **Caffe** is the deep learning tools built for scale, Caffe helps machines to track speed, modularity and expression. It uses interfaces with C, C++, Python, MATLAB and is especially relevant for convolution neural networks.
  4. **Chainer** is a Python-based deep learning framework, Chainer provides automatic differentiation APIs based on the define-by-run approach (a.k.a. dynamic computational graphs). It can also build and train neural networks through high-level object-oriented APIs.
  5. **Deeplearning4j** is a JVM-based, industry-focused, commercially supported, distributed deep-learning framework. The most significant advantage of using Deeplearning4j is speed. It can skim through massive volumes of data in very little time.

2.1.2.2 How Deep Learning Works?

- The term “deep” in deep learning usually refers to the number of hidden layers in the neural network, which defines the underlying architecture of any deep learning framework.
- Deep learning gets its name from the fact that it involves going deep into several layers of network, which also includes a hidden layer.
- Deep learning is often used in data science as it is computationally very competent compared to traditional/conventional machine learning methods, which require human intervention before being machine trained.
- DL is based on neural networks, a conceptual model of the brain. The word deep comes from DL algorithms that are trained/run on deep neural networks.
- The central concept of deep learning is the automatic extraction of representation from data.
- Deep learning helps in analyzing a massive amount of data through a hierarchical learning process.
- The amount of data generated in the organizations is massive/huge, raw and unstructured for which deep learning approaches are used to generate meaningful results.
- Deep learning techniques have proven to outperform all other machine learning techniques especially in the field of image and speech recognition systems.
- Deep learning is a vast and rapidly emerging field of knowledge. A deep learning network obliges representation learning incorporating multiple levels of representation.
- In a simple sense, it could be understood as such that the higher levels of the network amplify input aspects that are relevant to classification ignoring the irrelevant features that are not significant to the classification process.
- The interesting fact to note is that these layers of features in the deep network are not designed by human engineers but are learned from data using general-purpose learning procedures.
- Most deep learning methods use neural network architectures, which is why deep learning models are often referred to as deep neural networks. The term “deep” usually refers to the number of hidden layers in the neural network.
- Deep Neural Networks (DNNs) are such types of networks where each layer can perform complex operations such as representation and abstraction that make sense of images, sound, and text.
- Deep learning models are trained by using large sets of labeled data and neural network architectures that learn features directly from the data without the need for manual feature extraction.

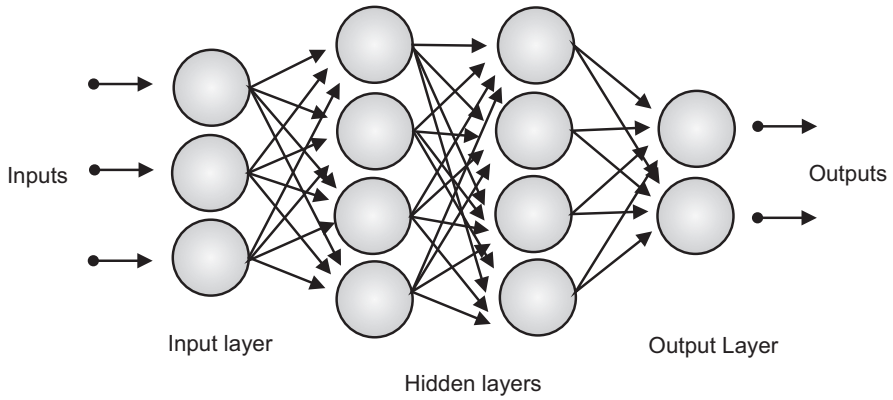


Fig. 2.3: Neural Networks (NNs) are organized in layers consisting of a set of interconnected nodes. Networks can have tens or hundreds of hidden layers

- The layers in Fig. 2.3 (Deep Learning Network) are explained below:
  1. **Output Layer:** The prime function of this layer is to respond to the information according to the learning of a particular task. These responses come from the units described above.
  2. **Input Layer:** The function of this layer is to import input as data into the network. On this data, the system starts learning, recognizing, and performing different types of processes.
  3. **Hidden Layer:** This layer lies between the input and the output layers. The function of this layer is to change the information into a format that can be used by the output progressively.

2.1.2.3 Advantages and Disadvantages of Deep Learning

Advantages of Deep Learning:

1. In DL the features are automatically deduced and optimally tuned for desired outcome.
2. In DL same neural network based approach can be applied to many different applications and data types.
3. The deep learning architecture is flexible to be adapted to new problems in be future.
4. DL provides maximum utilization of unstructured data to obtain insights from it.
5. DL has ability to execute feature engineering by itself. In this approach, an algorithm scans the data to identify features which correlate and then combine them to promote faster learning without being told to do so explicitly.
6. A deep learning model becomes able to perform thousands of routine, repetitive tasks within a relatively shorter period of time compared to what it would take for a human being.

Disadvantages of Deep Learning:

1. DL requires very large amount of data in order to perform better than other techniques.
2. DL is extremely expensive to train due to complex data models.
3. There is no standard theory in DL to guide users in selecting right deep learning tools as it requires knowledge, training method and other parameters.

2.1.2.4 Applications/Uses of Deep Learning

- Various applications of DL are explained below:
  1. **Natural Language Processing (NPL):** The deep learning techniques have led to improvements in translation and language modeling. Google Translate uses deep learning techniques to translate based on the semantics of an entire sentence instead of just memorizing phrase-to-phrase translations.
  2. **Automatic Speech Recognition:** Large-scale automatic speech recognition is the first and most convincing successful case of deep learning.
  3. **Image Recognition:** Deep learning-based image recognition has become "superhuman", producing more accurate results than human contestants. Deep learning-trained vehicles now interpret 360° camera views.
  4. **Medical Image Analysis:** Deep learning has been shown to produce competitive results in medical application such as cancer cell classification, lesion detection, organ segmentation and image enhancement.
  5. **Mobile Advertising:** Finding the appropriate mobile audience for mobile advertising is always challenging. Deep learning has been used to interpret large, many-dimensioned advertising datasets.
  6. **Financial Fraud Detection:** Deep learning is being successfully applied to financial fraud detection, tax evasion detection and anti-money laundering.
  7. **Military:** The United States (US) Department of Defense applied deep learning to train robots in new tasks through observation.

2.1.3 Artificial Intelligence

- In today's world, technology is growing very fast, and we are getting in touch with different new technologies day by day.
- The invention of computers or machines, their capability to perform various tasks went on growing exponentially.
- Humans have developed the power of computer systems in terms of their diverse working domains, their increasing speed, and reducing size with respect to time.
- Here, one of the booming technologies of computer science is Artificial Intelligence which is ready to create a new revolution in the world by making intelligent machines.

- Artificial intelligence (AI), also known as machine intelligence, is a branch of computer science that focuses on building and managing technology that can learn to autonomously make decisions and carry out actions on behalf of a human being.
- Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems.
- Artificial Intelligence is a way of making a computer, a computer-controlled robot, or a software think intelligently, in the similar manner the intelligent humans think.
- A goal of AI is to develop machines that behave as though they were intelligence.
- Artificial Intelligence exists when a machine can have human based skills such as learning, reasoning, and solving problems.
- The word Artificial Intelligence comprises of two words “Artificial” and “Intelligence”. Artificial refers to something which is made by human or non natural thing and Intelligence means ability to understand or think.
- Artificial intelligence is a science and technology based on disciplines such as Computer Science, Biology, Psychology, Linguistics, Mathematics, and Engineering.
- A major thrust of AI is in the development of computer functions associated with human intelligence, such as reasoning, learning, and problem solving.

2.1.3.1 Definition of Artificial Intelligence

- Artificial Intelligence (AI) is a wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence.
- According to the father of Artificial Intelligence, John McCarthy, “it is the science and engineering of making intelligent machines, especially intelligent computer programs”. **OR**
- Artificial Intelligence (AI) is a branch of computer science by which we can create intelligent machines which can behave like a human, think like humans, and able to make decisions."

2.1.3.2 Advantages and Disadvantages of Artificial Intelligence

Advantages of Artificial Intelligence:

1. **High Reliability:** AI machines are highly reliable and can perform the same action multiple times with high accuracy.
2. **High Accuracy with Less Errors:** AI machines or systems are prone to less errors and high accuracy as it takes decisions as per pre-experience or information.
3. **High-Speed:** AI systems can be of very high-speed and fast-decision making, because of that AI systems can beat a chess champion in the Chess game.
4. **Useful as a Public Utility:** AI can be very useful for public utilities such as a self-driving car which can make the journey safer.

5. **Useful for Risky Areas:** AI machines can be helpful in situations such as defusing a bomb, exploring the ocean floor, where to employ a human can be risky.
6. **Digital Assistant:** AI can be very useful to provide digital assistant to the users such as AI technology is currently used by various E-commerce websites to show the products as per customer requirement.

Disadvantages of Artificial Intelligence:

1. **High Cost:** The hardware and software requirement of AI is very costly as it requires lots of maintenance to meet current world requirements.
2. **No Original Creativity:** As humans are so creative and can imagine some new ideas but still AI machines cannot beat this power of human intelligence and cannot be creative and imaginative.
3. **Increase dependency on machines:** With the increment of technology, people are getting more dependent on devices and hence they are losing their mental capabilities.
4. **No feelings and emotions:** AI machines can be an outstanding performer, but still it does not have the feeling so it cannot make any kind of emotional attachment with human, and may sometime be harmful for users if the proper care is not taken.

2.1.3.3 Applications of Artificial Intelligence

- Applications of AI include:
  1. **Robotics:** Artificial Intelligence has a remarkable role in Robotics. With the help of AI, we can create intelligent robots.
  2. **Finance:** The finance industry is implementing automation, chatbot, adaptive intelligence, algorithm trading, and machine learning into financial processes.
  3. **Data Security:** The security of data is crucial for every organization and cyber-attacks are growing very rapidly in the digital world. AI can be used to make the data more safe and secure. AI uses AEG bot, AI2 Platform to determine software bug and cyber-attacks in a better way.
  4. **Natural Language Processing**: It is possible to interact with the computer that understands natural language spoken by humans.
  5. **Expert Systems**: There are some applications which integrate machine, software, and special information to impart reasoning and advising. They provide explanation and advice to the users.
  6. **Vision Systems**: These systems understand, interpret, and comprehend visual input on the computer. For example,
     - Doctors use clinical expert system to diagnose the patient.
     - Police use computer software that can recognize the face of criminal with the stored portrait made by forensic artist.

7. **Speech Recognition**: Some intelligent systems are capable of hearing and comprehending the language in terms of sentences and their meanings while a human talks to it. It can handle different accents, slang words, noise in the background, change in human's noise due to cold, etc.
8. **Social Media:** Social Media sites such as Facebook, Twitter etc. contain billions of user profiles, which need to be stored and managed in a very efficient way. AI can organize and manage massive amounts of data. AI can analyze lots of data to identify the latest trends, hashtag and requirement of different users.
9. **Travel:** AI is becoming highly demanding for travel industries with doing various travel related works such as from making travel arrangement to suggesting the hotels, flights, and best routes to the customers. Travel industries are using AI-powered chatbots which can make human-like interaction with customers for better and fast response.
10. **Agriculture:** Agriculture is an area which requires various resources, labor, money, and time for best result. Now a day's agriculture is becoming digital, and AI is emerging in this field. Agriculture is applying AI as agriculture robotics, solid and crop monitoring, predictive analysis. AI in agriculture can be very helpful for farmers.
11. **E-commerce:** AI is providing a competitive edge to the e-commerce industry, and it is becoming more demanding in the e-commerce business.

2.2 APPLICATIONS FOR MACHINE LEARNING IN DATA SCIENCE

- Machine Learning (ML) is a buzzword for today's technology, and it is growing very rapidly day by day.
- We are using machine learning in our daily life even without knowing it such as Google Maps, Google Assistant etc.
- Machine learning analyzes large chunks of data automatically. Machine learning basically automates the process of data analysis and makes data-informed predictions in real-time without any human intervention.
- A data model is built automatically and further trained to make real-time predictions. This is where the Machine Learning (ML) algorithms are used in the data science lifecycle.
- The three key machine learning algorithms in Data Science are Regression, Classification and Clustering.
- Regression and classification are of primary importance to a data scientist.

- The uses for regression and automatic classification are wide ranging, such as the following:
  1. Proactively identifying car parts that are likely to fail (regression).
  2. Finding oil fields, gold mines, or archeological sites based on existing sites (classification and regression).
  3. Predicting the number of eruptions of a volcano in a period (regression).
  4. Finding place names or persons in text (classification).
  5. Recognizing birds based on their whistle (classification).
  6. Face recognition or retina recognition, biometric (classification).
  7. Predicting which team will win the Champions League in soccer (classification).
  8. Identifying profitable customers (regression and classification).
  9. Identifying tumors and diseases (classification).
  10. Predicting the amount of money a person will spend on product X (regression).
  11. Predicting your company's yearly revenue (regression).
  12. Identifying people based on pictures or voice recordings (classification).
  13. Retail Marketing (clustering): Retail organizations often use clustering to identify groups of households that are similar to each other. For example, a retail organization may collect the following information on households:
      - Household income.
      - Household size.
      - Head of household Occupation.
      - Distance from nearest urban area.
      - These variables are used to identify the clusters of the families.

      They can then feed these variables into a clustering algorithm to perhaps identify the following clusters:
      - The organization can then send personalized advertisements or sales letters to each household based on how likely they are to respond to specific types of advertisements.
  14. Streaming services often use clustering analysis to identify viewers who have similar behavior. For example, a streaming service may collect the following data about individuals:
      - Minutes watched per day.
      - Total viewing sessions per week.
      - Number of unique shows viewed per month.

      Using these metrics, a streaming service can perform cluster analysis to identify high usage and low usage users so that they can know who they should spend most of their advertising dollars on, (clustering).

2.3 THE MODELING PROCESS

- A model is an abstraction of reality. A model is the representation of a relationship between variables in a dataset.
- A model describes how one or more variables in the data are related to other variables.
- Modeling is a process in which a representative abstraction is built from the observed dataset.
- For example, based on credit score, income level and requested loan amount, a model can be developed to determine the interest rate of a personal loan. For this task, previously known observational data such as credit score, income level, loan amount and interest rate are needed.
- The model serves following two purposes:
  - The model predicts the output (interest rate) based on the new and unseen set of input variables (credit score, income level and loan amount), and
  - the model can be used to understand the relationship between the output variable and all the input variables.
- The process of modeling comprises following four steps:
  1. Feature engineering and selecting a model.
  2. Training the model.
  3. Validating the model.
  4. Testing the model on new data.
- The first three steps in modeling are usually repeated because we most likely will not build an optimal model for the project on the first try.
- As such, we will be building several models and then select the one that performs the best on the testing data set (which is unseen data).
- In addition, the testing step of the modeling process is not always performed because, in some cases, the goal of the data science project is root cause analysis (basically explanation) instead of predictions.
- For example, the goal of the project might be to determine the reason why some species are going extinct (explanation) instead of predicting which species is most likely to go extent (a prediction).
- Another trick commonly used in machine learning is chaining. The way chaining works is that, when several models are chained, the output of one model is used as input by another model and so on.
- When chaining different models together, each of them needs to be trained independently of the others and only their results are combined together. In machine learning terms, this technique is known as ensemble learning.

- A model consists of constructs of information called features or predictors and a target or response variable. The model's goal is to predict the target variable, for example, tomorrow's high temperature.
- The variables that help us do this and are (usually) known to we are the features or predictor variables such as today's temperature, cloud movements, current wind speed, and so on.
- In practice the best models are those that accurately represent reality, preferably while staying concise and interpretable.
- To achieve this goal, feature engineering is the most important and arguably most interesting part of modeling.
- For instance, an important feature in a model that tried to explain the extinction of large land animals in the last 60,000 years in Australia turned out to be the population number and spread of humans.

2.3.1 Engineering Features and Selecting a Model

- In engineering features, we basically create feasible predictors for the model. With engineering features, we must come up with and create possible predictors for the model.
- Engineering features is the very most important and crucial steps in the modeling process because the model will be able to give predictions by combining these features and the accuracy of the predictions depends on how good the predictors are.
- Engineering features is important steps in the process because a model recombines these features to achieve its predictions.
- Often we may need to consult a machine learning expert to come up with meaningful features. It is very difficult for beginners to create good predictors.
- Once the engineering features have been created, we need to choose the respective model to use with it.

2.3.2 Training the Model

- With the right and accurate predictors in place and a modeling technique in mind, we can progress to model training. In model training we present the model data from which it can learn.
- After we have created the right predictors (engineering features) and have selected the appropriate modeling technique for the project, we can now proceed to train the model on a training data set.
- A training data set is a data sample that we select for the model to learn to perform actions from.
- Popular modeling techniques are available to be implemented in almost any programming language that we may choose, including Python.

- These techniques essentially allow us to train your model by using a simple set of lines of codes.
- Advanced data science techniques require the scientist to be capable of using heavy mathematical calculations and then implement modern data science techniques to use with these calculations.
- Once we have trained the model, the next thing to do is check whether it works as we intended it to.
- Training time is the number of hours to train the model also plays a vital role in determining the selection of the model. It is directly related to the accuracy of the obtained model.

2.3.3 Validating the Model

- Validation is of the model is extremely important because it determines whether the model works in real-life conditions.
- Once, a model is trained, it's time to test whether it can be extrapolated to reality i.e., model validation.
- Data science has many modeling techniques and the question is which one is the right one to use. A good model has following two properties:
  1. it has good predictive power, and
  2. it generalizes well to data it hasn't seen.
- To achieve above properties we define an error measure (how wrong the model is) and a validation strategy.
- Two common error measures in machine learning are the classification error rate for classification problems and the mean squared error for regression problems.
  1. The **classification error rate** is the percentage of observations in the test data set that your model mislabeled; lower is better.
  2. The **mean squared error** measures how big the average error of your prediction is. Squaring the average error has following two consequences:
     - We can't cancel out a wrong prediction in one direction with a faulty prediction in the other direction. For example, overestimating future turnover for next month by 5,000 doesn't cancel out underestimating it by 5,000 for the following month.
     - As a second consequence of squaring, bigger errors get even more weight than they otherwise would. Small errors remain small or can even shrink (if <1), whereas big errors are enlarged and will definitely draw your attention.
- Number of validation strategies exist, including the following common ones:
  1. **Dividing the data into a training set with X% of the observations and keeping the rest as a holdout data set (a data set that's never used for model creation)** is the most common technique.

2. **K-folds cross validation** is the strategy divide the data set into k parts and use each part one time as a test data set while using the others as a training data set. This has the advantage that we use all the data available in the data set.
3. **Leave-1 out** is an approach in which the same as k-folds but with k=1. You always leave one observation out and train on the rest of the data. This is used only on small data sets, so it's more valuable to people evaluating laboratory experiments than to big data analysts.

- Regularization is another popular term in machine learning. When applying the term regularization, we incur a penalty for every extra variable used to construct the model.
  - With L_1 regularization we ask for a model with as few predictors as possible. This is important for the model's robustness: simple solutions tend to hold true in more situations.
  - The L_2 regularization aims to keep the variance between the coefficients of the predictors as small as possible.
- Overlapping variance between predictors in a model makes it hard to make out the actual impact of each predictor. Keeping their variance from overlapping will increase interpretability.
- To keep it simple, the regularization is mainly used to stop a model from using too many features and thus prevent over-fitting.

2.3.4 Predicting New Observations

- If we have performed the first three steps of the machine learning process successfully, then we will end up with a model that can generalize well to new data.
- Model scoring is defined as, using a machine learning model on unseen data. The process of applying the model to new data is called model scoring.
- Model scoring is done in following two steps:
  1. Preparing a data set with features that are defined by the model.
  2. Using the model on this prepared data set.

2.4 TYPES OF MACHINE LEARNING

- We can divide the different approaches to machine learning by the amount of human effort that's required to coordinate them and how they use labeled data.
- The labeled data is the data with a category or a real-value number assigned to it that represents the outcome of previous observations.
- Various types of machine learning include:
  1. **Supervised Learning** attempt to discern results and learn by trying to find patterns in a labeled data set. Human interaction is required to label the data.

2. **Unsupervised Learning** doesn't rely on labeled data and attempt to find patterns in a data set without human interaction.
3. **Semi-supervised Learning** needs labeled data and therefore human interaction, to find patterns in the data set, but they can still progress toward a result and learn even if passed unlabeled data as well.

2.4.1 Supervised Learning

- Supervised learning is a learning technique that can only be applied on labeled data. In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs.
- The purpose of supervised learning is for the algorithm to be able to "learn" by comparing its actual output with the "target" outputs to find errors, and modify the model accordingly.
- So, supervised learning or training is the process of providing the network with a series of sample inputs and comparing the output with the expected responses.
- In fact, the supervised learning is a typical case of pure inductive inference, where the free variables of the network are adjusted by knowing a priori the desired outputs for the investigated system.

Basic Idea/Concept of Supervised Learning:

- Supervised learning as the name indicates the presence of a supervisor as a teacher. In supervised learning the machines are trained using well labeled training data.
- Basically supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer.
- After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

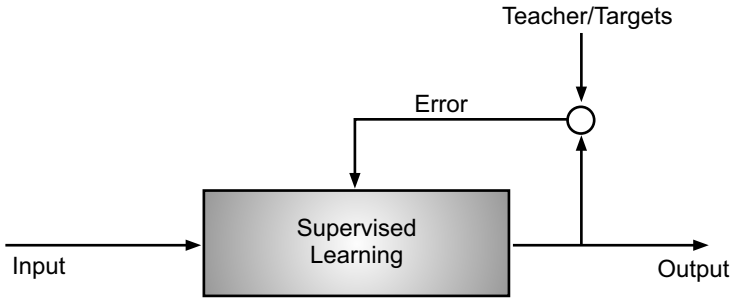


Fig. 2.4: Supervised Learning

- Based on the ML tasks, supervised learning algorithms can be divided into two classes namely, Classification and Regression.

1. Classification:

- The key objective of classification-based tasks is to predict categorical output labels or responses for the given input data. The output will be based on what the model has learned in training phase.
- As we know that the categorical output response means unordered and discrete values, hence each output response will belong to a specific class or category.
- Classification refers to process of predicting discrete output values for an input. For example, given an input predicting whether a student will pass or fail the exam.

2. Regression:

- The key objective of regression-based tasks is to predict output labels or responses which are continue numeric values, for the given input data. The output will be based on what the model has learned in its training phase.
- Basically, regression models use the input data features (independent variables) and their corresponding continuous numeric output values (dependent or outcome variables) to learn specific association between inputs and corresponding outputs.
- In regression problems the task of machine learning model is to predict a continuous value. For example, for given input, predict the marks obtained by a student on an exam etc.

How Supervised Learning Works?

- Supervised learning is the types of machine learning in which machines are trained using well labeled training data and on basis of that data, machines predict the output. The labeled data means some input data is already tagged with the correct output.
- In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.
- Supervised learning is a process of providing input data as well as correct output data to the machine learning model.
- The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).
- For example, we have x (input variables) and Y (output variable). Now, apply an algorithm to learn the mapping function from the input to output as: Y=f(x).
- In supervised learning, models are trained using labeled dataset, where the model learns about each type of data.
- Once the training process is completed, the model is tested on the basis of test data (a subset of the training set) and then it predicts the output.
- Fig. 2.5 shows working of supervised learning. Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle and Polygon.

- Now the first step is that we need to train the model for each shape.
  - If the given shape has four sides, and all the sides are equal, then it will be labeled as a Square.
  - If the given shape has three sides, then it will be labeled as a triangle.
  - If the given shape has six equal sides then it will be labeled as hexagon.
- Now, after training, we test our model using the test set, and the task of the model is to identify the shape.
- The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

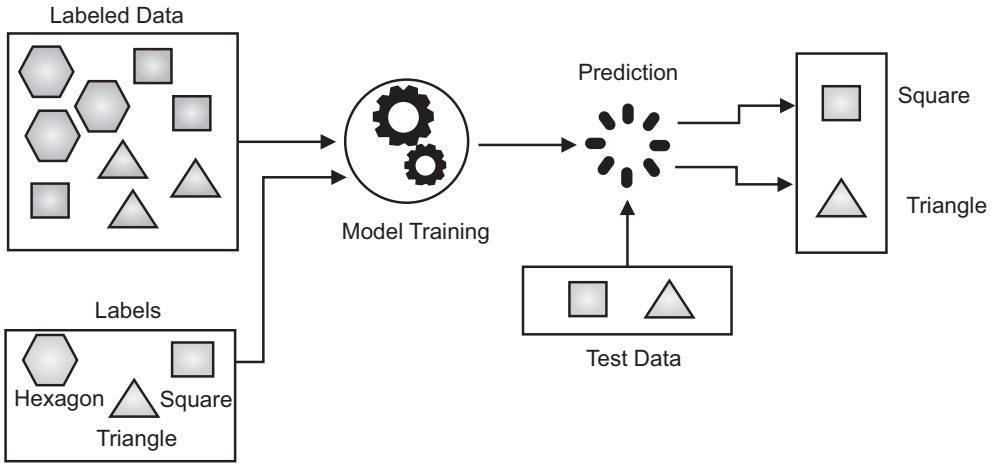


Fig. 2.5: Working of Supervised Learning

Advantages of Supervised learning:

1. Supervised learning model helps us to solve various real-world problems such as fraud detection.
2. With the help of supervised learning, the model can predict the output on the basis of prior experiences.
3. In supervised learning, we can have an exact idea about the classes of objects.

Disadvantages of supervised learning:

1. Supervised learning cannot predict the correct output if the test data is different from the training dataset.
2. Training required in supervised learning consumes lots of time.
3. Supervised learning models are not suitable for handling the complex tasks.

- There are several algorithms available for supervised learning. Some of the widely used algorithms of supervised learning are given below:
  1. k-Nearest Neighbours
  2. Decision Trees
  3. Naive Bayes

4. Logistic Regression
5. Support Vector Machines (SVMs)

k-Nearest-Neighbors (kNN):

- The k-NN algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems.
- The k-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- The k-NN algorithm stores all the available data and classifies a new data point based on the similarity; means when new data appears then it can be easily classified into a well suite category by using k-NN algorithm.
- The k-NN algorithm can be used for regression as well as for classification but mostly it is used for the classification problems.
- The k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation.
- The k-NN algorithm uses 'feature similarity' to predict the values of new data-points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

Need for k-NN Algorithm:

- Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x_1, so this data point will lie in which of these categories. To solve this type of problem, we need a k-NN algorithm.
- With the help of k-NN, we can easily identify the category or class of a particular dataset.

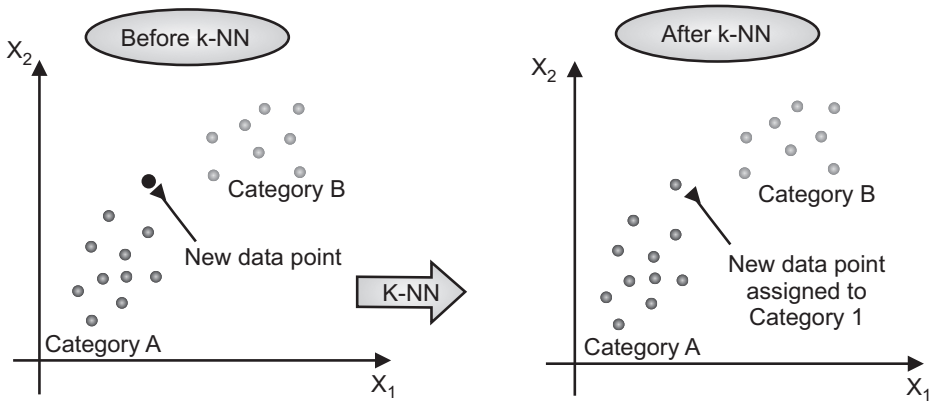


Fig. 2.6

- For example, consider dataset which can be plotted in Fig. 2.7.

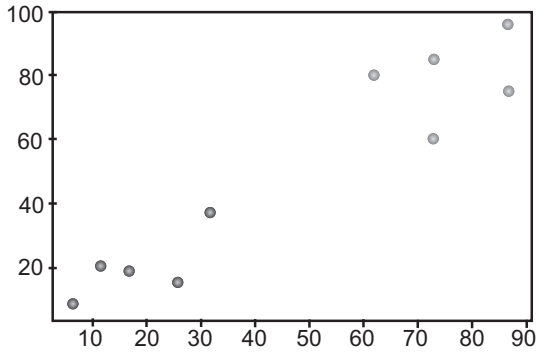


Fig. 2.7

- Now, we need to classify new data point with black dot (at point 60, 60) into gray or black class. We are assuming k = 3 i.e., it would find three nearest data points and shown in the Fig. 2.8.

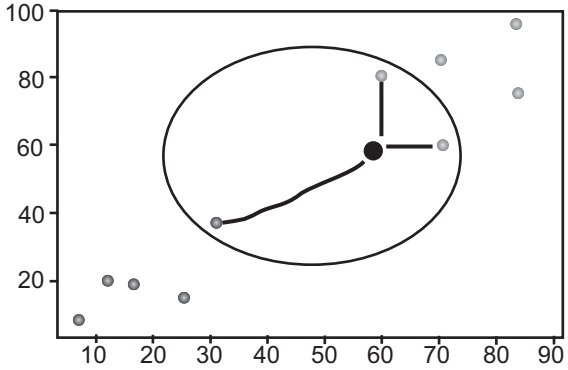


Fig. 2.8

- We can see in the Fig. 2.8 the three nearest neighbors of the data point with square black dot. Among those three, two of them lie in black class hence the black dot will also be assigned in black class.
- Take another example, we have a new data point and we need to put it in the required category (See Fig. 2.9).

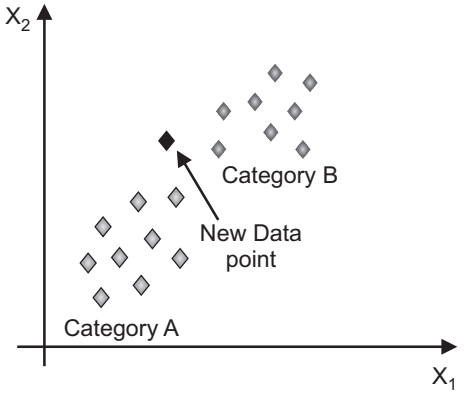


Fig. 2.9

- Firstly, we will choose the number of neighbors, so we will choose the k=5. As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

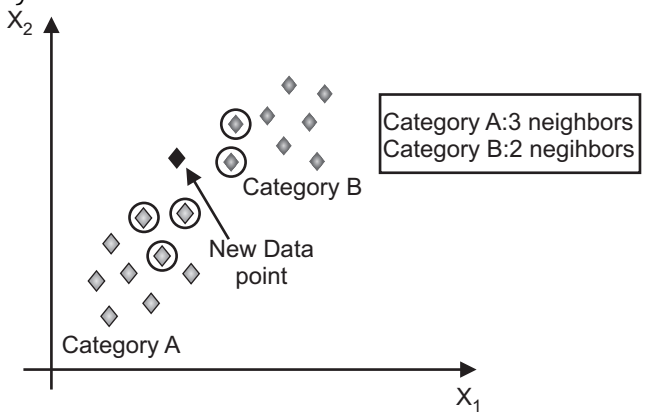


Fig. 2.10

Advantages of k-NN:

1. The k-NN algorithm is simple and easy to implement.
2. The k-NN a versatile algorithm as we can use it for classification as well as regression.
3. The k-NN is very useful for nonlinear data because there is no assumption about data in this algorithm.

Disadvantages of k-NN:

1. The k-NN algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.
2. The k-NN algorithm is computationally a bit expensive algorithm because it stores all the training data.
3. The k-NN algorithm requires high memory storage.

Decision Tree:

- In supervised learning the decisions are performed on the basis of features of the given dataset. The decision tree is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- Decision tree is a supervised learning technique that can be used for both classification and regression problems
- A decision tree is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- A decision tree builds classification or regression models in the form of a tree structure. The decisions are performed on the basis of features of the given dataset.

- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- Fig. 2.11 general structure of a decision tree. Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node. Decision nodes are used to make any decision and have multiple branches.

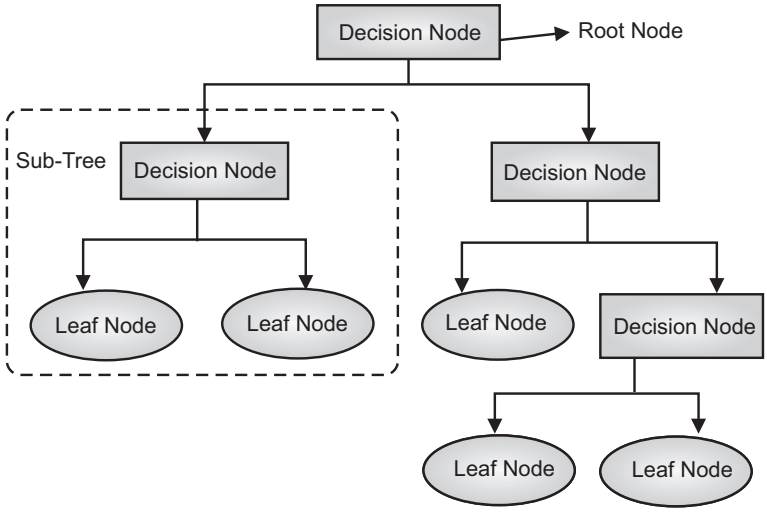


Fig. 2.11

How does the Decision Tree Work?

- In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree.
- This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
- For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.
- Fig. 2.12 shows an example of decision tree. Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute).
- The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node.
- Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer).

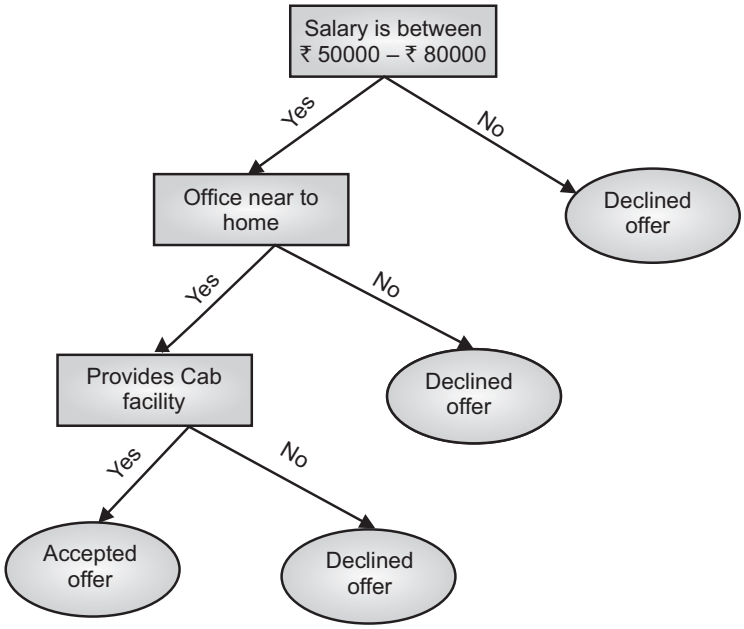


Fig. 2.12

Advantages of Decision Tree:

1. Decision trees are simple to understand and interpret.
2. Decision trees are able to handle both numerical and categorical data.
3. Decision trees works well with large datasets.
4. Decision trees are fast and accurate.

Disadvantages of Decision Tree:

1. A small change in the training data can result in a larger change in the tree and consequently the final predictions.
2. Decision trees performance is not good if there are lots of uncorrelated variables in the data set.
3. Decision trees are generally easy to use, but making them, particularly huge ones with numerous divisions or branches, is complex.

Support Vector Machine (SVM):

- Support Vector Machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression problems.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.
- The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N- the number of features) that distinctly classifies the data points.

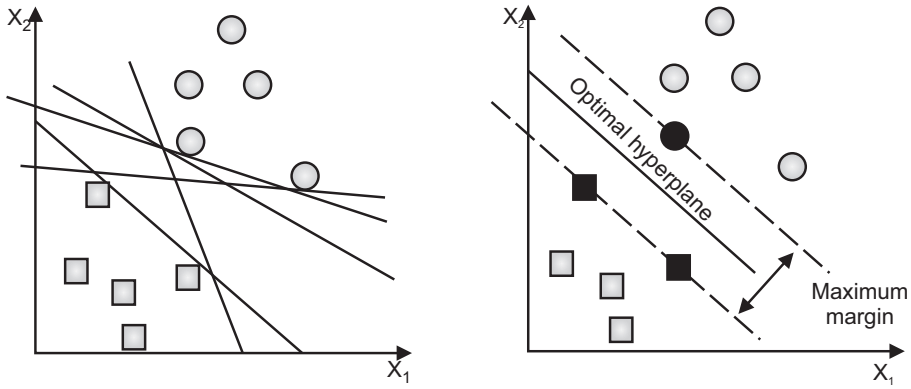


Fig. 2.13: Possible Hyperplanes

Working of SVM:

- An SVM model is basically a representation of different classes in a hyperplane in multidimensional space.
- The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized.
- The goal of SVM is to divide the datasets into classes to find a Maximum Marginal Hyperplane (MMH).

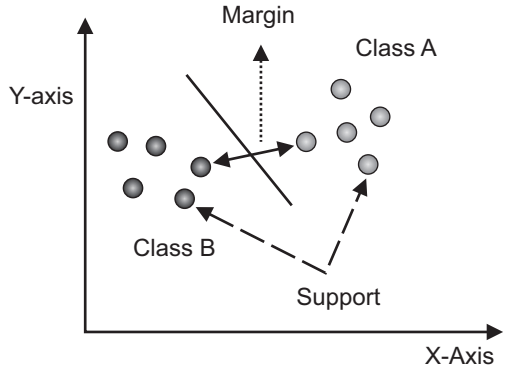


Fig. 2.14

- The important concepts in SVM are explained below:
  1. Data points that are closest to the hyperplane is called **support vectors**. Separating line will be defined with the help of these data points.
  2. **Hyperplane** is a decision plane or space which is divided between a set of objects having different classes.
  3. **Margin** may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

Advantages of SVM:

1. SVM offers great accuracy.
2. SVM work well with high dimensional space.
3. It is effective in cases where number of dimensions is greater than the number of samples.
4. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

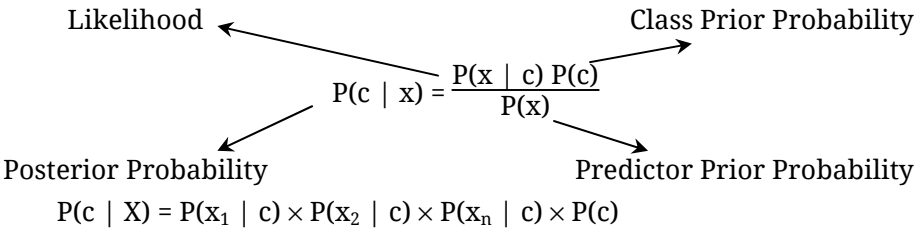
Disadvantages of SVM:

1. SVMs have high training time hence in practice not suitable for large datasets.
2. It also does not perform very well, when the data set has more noise i.e. target classes are overlapping.

Naïve Bayes:

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes' theorem and used for solving classification problems. Bayes' theorem is also known as Bayes' Rule or Bayes' law.
- The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as follows:
  1. **Naïve** is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.
  2. **Bayes** is called Bayes because it depends on the principle of Bayes' Theorem.
- Naïve Bayes classifier is one of the simple and most effective classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- Naïve Bayes is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors.
- In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Naive Bayes model is easy to build and particularly useful for very large data sets. Naive Bayes is used for creating classifiers. Suppose we want to sort out (classify) fruits of different kinds from a fruit basket.
- We may use features such as color, size and shape of a fruit, For example, any fruit that is red in color, is round in shape and is about 9 cm in diameter may be considered as Apple.
- So to train the model, we would use these features and test the probability that a given feature matches the desired constraints. The probabilities of different features are then combined to arrive at a probability that a given fruit is an Apple.

- Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times P(x_n \mid c) \times P(c)$$

Where,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Advantages of Naïve Bayes:

1. Naïve Bayes is fast and easy ML algorithms to predict a class of datasets.
2. Naïve Bayes will converge faster than discriminative models like logistic regression.
3. It can make probabilistic predictions and can handle continuous as well as discrete data.
4. Naïve Bayes requires less training data.
5. It is highly scalable in nature, or they scale linearly with the number of predictors and data points.
6. Naïve Bayes can be used for binary as well as multi-class classification problems both.

Disadvantages of Naïve Bayes:

1. If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction.
2. Naïve Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

2.4.2 Unsupervised Learning

- As the name suggests, it is opposite to supervised ML methods or algorithms which means in unsupervised machine learning algorithms we do not have any supervisor to provide any sort of guidance.
- Unsupervised learning is a machine learning technique in which models are not supervised using training dataset.

- Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.
- Unsupervised learning can be defined as, a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.
- In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data.
- As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable.
- The system itself must then decide which features it will use to group the input data.
- The training process extracts the statistical properties of the training set and groups of similar vectors into classes or clusters.
- Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next.

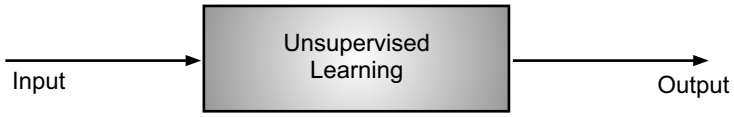


Fig. 2.15: Unsupervised Learning

- Unsupervised machine learning is the task of drawing inference of a function from data set containing input without labeled data or target value to describe hidden patterns from unlabeled data.
- The most common unsupervised machine learning method is hierarchical cluster analysis. It is used for exploratory data analysis in order to find the hidden patterns.
- The unsupervised learning algorithm can be further categorized into following two types of problems:
  1. **Clustering:** Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities. A clustering problem is where we want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
  2. **Association:** An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose

a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis. An association rule learning problem is where we want to discover rules that describe large portions of the data, such as people that buy X also tend to buy Y.

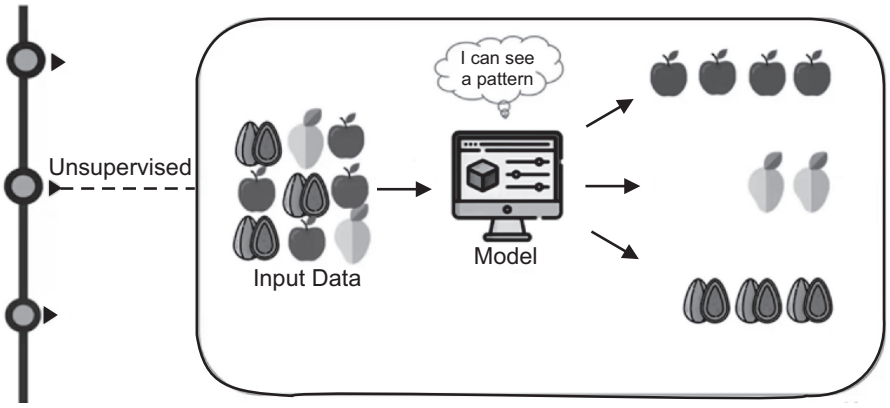


Fig. 2.16: Unsupervised Learning

Advantages of Unsupervised Learning:

1. Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.
2. Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.

Disadvantages of Unsupervised Learning:

1. The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.
2. Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.

- The popular unsupervised learning algorithms in clued k-means clustering, Apriori algorithms and Anomaly detection.

k-Means Clustering Algorithm:

- Clustering is one of the widely used unsupervised learning techniques in machine learning. The most common and simplest clustering algorithm is the k-means clustering.
- The k-means algorithm involves us to telling the algorithms how many possible cluster (or k) there are in the dataset.
- The algorithm then iteratively moves the k-centers and selects the datapoints that are closest to that centroid in the cluster.
- The k-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. The number of clusters identified from data by algorithm is represented by 'k' in k-means.

- In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum.
- It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.
- The k-means clustering is an unsupervised learning algorithm, which groups the unlabeled dataset into different clusters.
- Here k defines the number of pre-defined clusters that need to be created in the process, as if k=2, there will be two clusters, and for k=3, there will be three clusters, and so on.
- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.
- The k-means clustering algorithm mainly performs two tasks:
  1. Determines the best value for K center points or centroids by an iterative process.
  2. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.
- Hence, each cluster has data points with some commonalities, and it is away from other clusters.
- Fig. 2.17 shows the working of the k-means clustering algorithm.

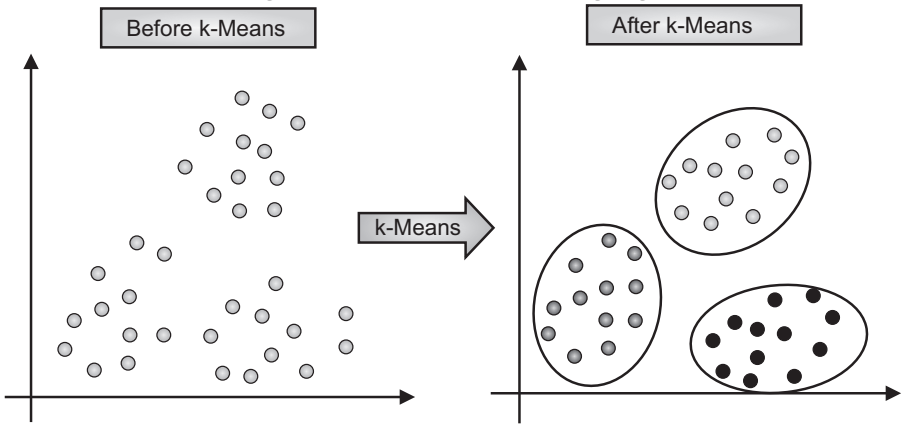


Fig. 2.17

How does the k-Means Algorithm Work?

- The k-means clustering is a simple and popular clustering algorithm that originated in signal processing.
- The goal of the k-means algorithm is to partition examples from a data set into k clusters.
- Each example is a numerical vector that allows the distance between vectors to be calculated as a Euclidean distance.
- The simple example below visualizes the partitioning of data into k = 2 clusters, where, the Euclidean distance between examples is smallest to the centroid (center) of the cluster, which indicates its membership.
- Fig. 2.18 shows simple example of k-means clustering algorithm.

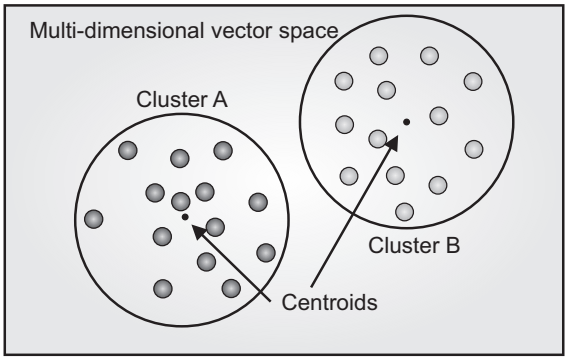
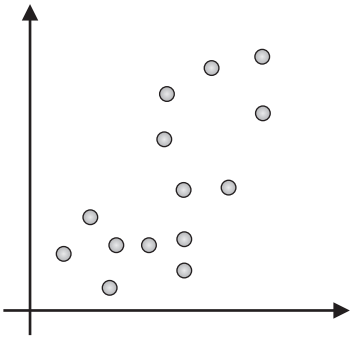


Fig. 2.18

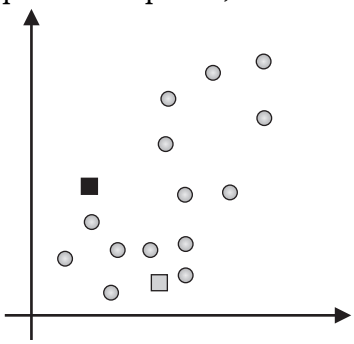
- The k-means algorithm is extremely simple and easy to understand and implement.
- We begin by randomly assigning each example from the data set into a cluster, calculate the centroid of the clusters as the mean of all member examples
- Then iterate the data set to determine whether an example is closer to the member cluster or the alternate cluster (given that k = 2).
- If the member is closer to the alternate cluster, the example is moved to the new cluster and its centroid recalculated. This process continues until no example moves to the alternate cluster.
- As illustrated, k-means partitions the example data set into k clusters without any understanding of the features within the example vectors (that is, without supervision).

Example for k-means Algorithm:

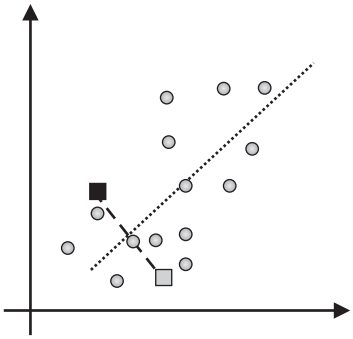
- Suppose we have two variables A1 and A2. The x-y axis scatter plot of these two variables is given below:



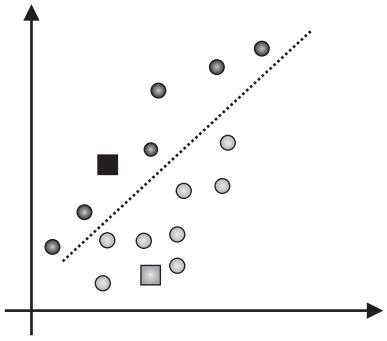
- Let us take number k of clusters, i.e., k=2, to identify the dataset and to put them into different clusters mean here we will try to group these datasets into two different clusters. We need to choose some random k points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as k points, which are not the part of our dataset.



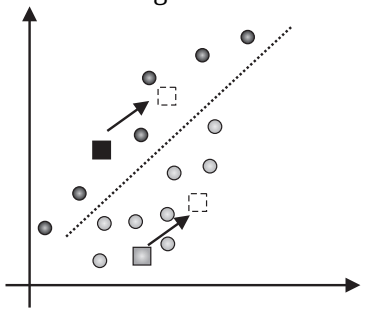
- Now we will assign each data point of the scatter plot to its closest k-point or centroid. We will compute it by applying some mathematics that we have studied to calculate the distance between two points. So, we will draw a median between both the centroids.



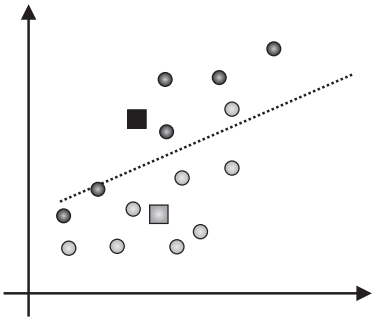
- From the above image, it is clear that points left side of the line is near to the k1 or black centroid, and points to the right of the line are close to the gray centroid. Let us color them as black and gray for clear visualization.



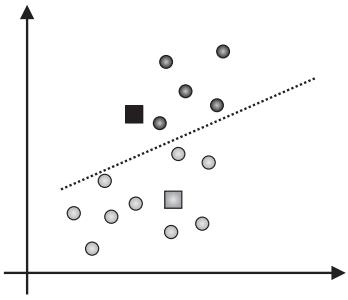
- As we need to find the closest cluster, so we will repeat the process by choosing a new centroid. To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as given below:



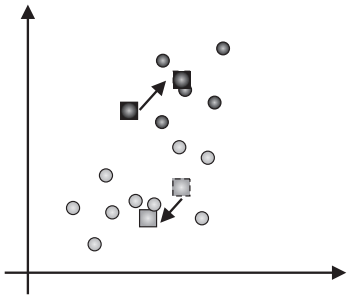
- Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line. The median will be given below:



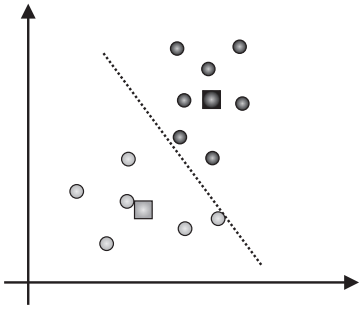
- From the above image, we can see, one gray point is on the left side of the line and two black points are right to the line. So, these three points will be assigned to new centroids.



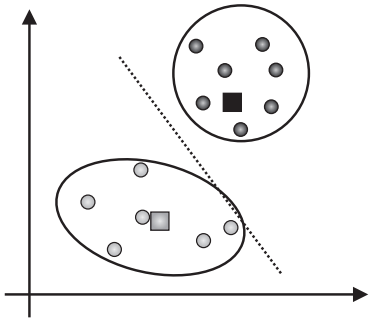
- As reassignment has taken place, so we will again go to the Point 4, which is finding new centroids or k-points. We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below:



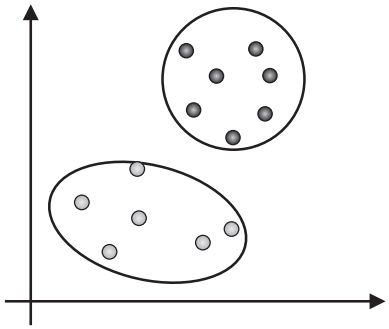
- As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:



- We can see in the above image; there are no dissimilar data points on either side of the line, which means our model is formed.



- As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be shown below:



Advantages of k-means Clustering Algorithm:

1. The k-means algorithm is simple easy to understand and to implement.
2. The k-means algorithm is the most popular clustering algorithm, because it provides easily interpretable clustering results.
3. The k-means algorithm is fast and efficient in terms of computational cost.

Disadvantages of k-means Clustering Algorithm:

1. The k-means algorithm is difficult to predict the number of clusters i.e. the value of k.
2. The k-means algorithm is not good in doing clustering job if the clusters have a complicated geometric shape.
3. The output of k-means algorithm is strongly impacted by initial inputs like number of clusters (value of k).
4. It is very sensitive to rescaling.
5. The k-means algorithms are slow and do not scale to a large number of data points.

Association Rule Mining:

- 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 analysis is primarily concerned about searching for frequent items in a given dataset.
- Another useful unsupervised ML method is Association which is used to analyze large dataset to find patterns which further represents the interesting relationships between various items.
- It is also termed as Association Rule Mining or Market basket analysis which is mainly used to analyze customer shopping patterns.
- Association rule mining discovers strong association or correlation relationships among data.
- Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.
- The main applications of association rule mining:
  1. **Basket Data Analysis** is to analyze the association of purchased items in a single basket or single purchase, for example, peanut butter and jelly are often bought together because a lot of people like to make PBandJ sandwiches.

2. **Cross Marketing** is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.
3. **Catalog Design** the selection of items in a business' catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related.

Apriori Algorithm:

- The Apriori algorithm was proposed by Agrawal and Srikant in 1994.
- The Apriori algorithm helps in building/creating the association rules from the frequent itemsets that remains after applying Apriori property during each iteration.

Anomaly Detection:

- Anomaly detection is an unsupervised ML method.
- Anomaly detection is used to find out the occurrences of rare events or observations that generally do not occur.

Difference between Supervised and Unsupervised Learning:

- Machine learning defines basically two types of learning namely, supervised and unsupervised learning. But both the techniques are used in different scenarios and with different datasets.
- Supervised learning is a machine learning method in which models are trained using labeled data. Unsupervised learning is another machine learning method in which patterns inferred from the unlabeled input data.
- Following table differences between Supervised learning and Unsupervised learning:

Sr. No.	Supervised Learning	Unsupervised Learning
1.	In supervised learning both input and output variables are provided on the basis of which the output could be predicted and probability of its correctness is higher.	In unsupervised learning only input variables are provided and no output variable are available due to which the outcome or resultant learning is dependent on one intellectual observation.
2.	As supervised learning is treated as highly accurate and trustworthy method so the accuracy and correctness is better as compare to unsupervised learning.	Unsupervised learning is comparatively less accurate and trustworthy method.
3.	Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.

contd. ...

4.	Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
5.	Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
6.	In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
7.	The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
8.	Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train the model.
9.	Supervised learning can be categorized in Classification and Regression problems.	Unsupervised learning can be classified in Clustering and Associations problems.
10.	Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
11.	Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.
12.	Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
13.	It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

2.4.3 Semi-supervised Learning

- Semi-supervised learning is an important category that lies between the Supervised and Unsupervised machine learning.
- Semi-supervised Learning algorithms or methods are neither fully supervised nor fully unsupervised. They basically fall between the two i.e. supervised and unsupervised learning methods.

- Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training.
- Semi-supervised learning falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data).
- Unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.
- In this algorithm, training data is a combination of both labeled and unlabeled data. However, labeled data exists with a very small amount while it consists of a huge amount of unlabeled data.
- Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labeled data. It is why label data is a comparatively, more expensive acquisition than unlabeled data.
- The semi-supervised machine learning uses a small amount of labeled data and a large amount of unlabeled data, which provides the benefits of both unsupervised and supervised learning while avoiding the challenges of finding a large amount of labeled data.
- Semi-supervised learning is applied in cases where it is expensive to acquire a fully labeled dataset while more practical to label a small subset.
- For example, it often requires skilled experts to label certain remote sensing images, and lots of field experiments to locate oil at a particular location, while acquiring unlabeled data is relatively easy.
- We can follow any of the following approaches for implementing semi-supervised learning methods:
    1. The first and simple approach is to build the supervised model based on small amount of labeled and annotated data and then build the unsupervised model by applying the same to the large amounts of unlabeled data to get more labeled samples. Now, train the model on them and repeat the process.
    2. The second approach needs some extra efforts. In this approach, we can first use the unsupervised methods to cluster similar data samples, annotate these groups and then use a combination of this information to train the model.

Basic Idea/Concept of Semi-supervised Learning:

- The idea behind semi-supervised learning is to learn from labeled and unlabeled data to improve the predictive power of the models.
- The notion is explained with a simple illustration in Fig. 2.19, which shows that when a large amount of unlabeled data is available.
- For example, HTML documents on the web, the expert can classify a few of them into known categories such as sports, news, entertainment, and so on.
- This small set of labeled data together with the large unlabeled dataset can then be used by semi-supervised learning techniques to learn models.

- Thus, using the knowledge of both labeled and unlabeled data, the model can classify unseen documents in the future.

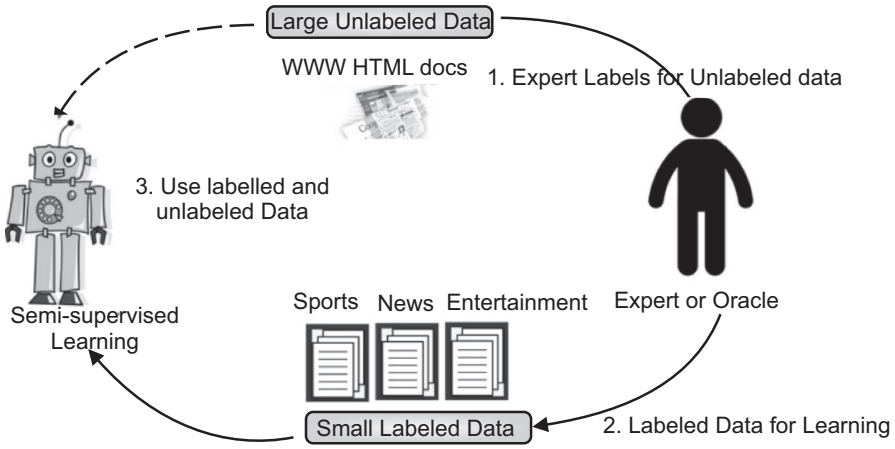


Fig. 2.19

- Take for example the plot in Fig. 2.20 In this case, the data has only two labeled observations; normally this is too few to make valid predictions.

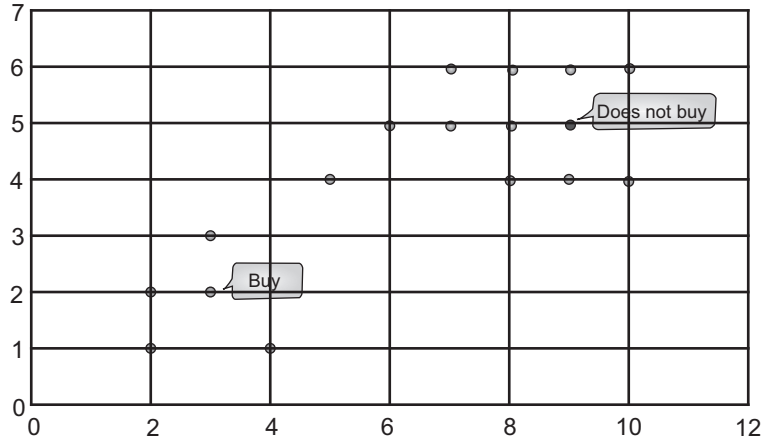


Fig. 2.20: Plot has only Two Labeled Observations - too few for Supervised Observations, but enough to Start with an Unsupervised or Semi-supervised Approach

- A common semi-supervised learning technique is label propagation. In this technique, we start with a labeled data set and give the same label to similar data points.
- This is similar to running a clustering algorithm over the data set and labeling each cluster based on the labels they contain.
- If we were to apply this approach to the data set in Fig. 2.20, we might end up with something like Fig. 2.21.
- The previous Fig. 2.20 shows that the data has only two labeled observations, far too few for supervised learning. The Fig. 2.21 shows how we can exploit the structure of the underlying data set to learn better classifiers than from the labeled data only.

- The data is split into two clusters by the clustering technique; we only have two labeled values, but if we're bold we can assume others within that cluster have that same label (buyer or non-buyer), as depicted here. This technique isn't flawless; it's better to get the actual labels if we can.

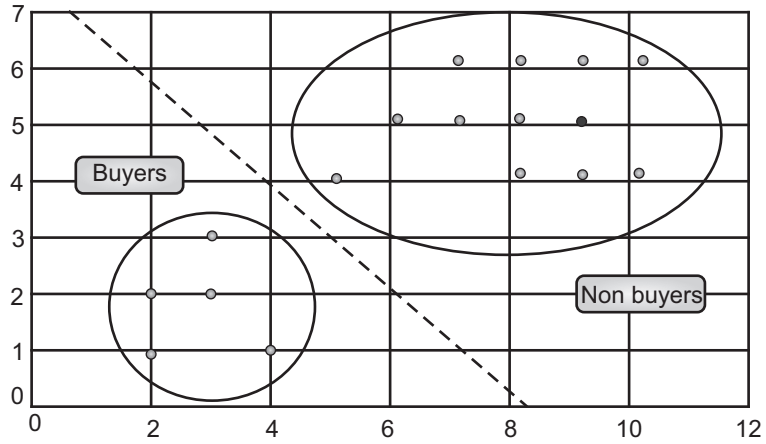


Fig. 2.21

- The special approach to semi-supervised learning worth mentioning here is active learning, in which the program points out the observations it wants to see labeled for its next round of learning based on some criteria we have specified.
- For example, we might set it to try and label the observations the algorithm is least certain about or we might use multiple models to make a prediction and select the points where the models disagree the most.

Advantages of Semi-supervised Machine Learning Algorithms:

1. It is easy to understand and simple to implement.
2. It reduces the amount of annotated data used.
3. It is a stable algorithm.
4. It has high efficiency.

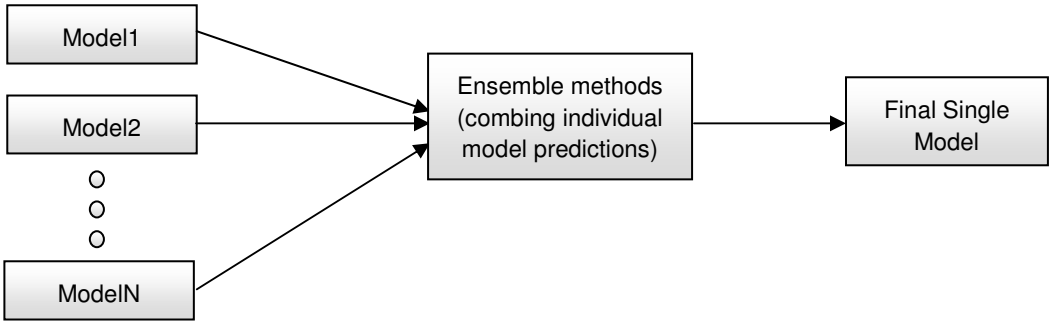
Disadvantages of Semi-supervised Machine Learning Algorithms:

1. Iteration results are not stable.
2. It is not applicable to network-level data.
3. It has low accuracy.

2.4.4 Ensemble Techniques

- In ML, ensemble methods from multiple machines learning models trend to have a better generalized performance than any single machine learning model.
- In supervised machine learning, the objective is to build a model that can explain the relationship or interaction between inputs and output.
- The model can be considered as a hypothesis that can map new input data to predicted output.

- Ensemble methods are the machine learning technique that combines several base models in order to produce one optimal predictive model.
- Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model.
- The following picture shows basic concept of ensemble methods and/or techniques.



- Ensemble methods are models composed of multiple weaker models that are independently trained and their prediction result are combined together in a single model which makes the predictions more accurate and better performance.
- Bagging, boosting and random forests are examples of ensemble methods/techniques.

1. Bagging Ensemble Technique:

- Bagging, (short form for bootstrap aggregating), is mainly applied in classification and regression.
- Bagging is classified into two types namely, bootstrapping and aggregation.
  - **Bootstrapping** is a sampling technique where samples are derived from the whole population (set) using the replacement procedure. The sampling with replacement method helps make the selection procedure randomized. The base learning algorithm is run on the samples to complete the procedure.
  - **Aggregation** in bagging is done to incorporate all possible outcomes of the prediction and randomize the outcome. Without aggregation, predictions will not be accurate because all outcomes are not put into consideration. Therefore, the aggregation is based on the probability bootstrapping procedures or on the basis of all outcomes of the predictive models.

2. Boosting Ensemble Technique:

- Boosting is an ensemble technique that learns from previous predictor mistakes to make better predictions in the future.
- The technique combines several weak base learners to form one strong learner, thus significantly improving the predictability of models.
- Boosting works by arranging weak learners in a sequence, such that weak learners learn from the next learner in the sequence to create better predictive models.
- Boosting takes many forms, including gradient boosting, Adaptive Boosting (AdaBoost), CatBoost and XGBoost (Extreme Gradient Boosting).

3. Random Forests Ensemble Technique:

- We now present another ensemble method called random forests. Imagine that each of the classifiers in the ensemble is a decision tree classifier so that the collection of classifiers is a "forest."
- The individual decision trees are generated using a random selection of attributes at each node to determine the split.
- More formally, each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.
- Some other ensemble techniques include **stacking ensemble technique** (uses output predictions from one model as input features to another model) and **blending ensemble technique** (the predictions are made only on the validation dataset). In stacking, we treat the result of individual predictions as next training data.
- During classification, each tree votes and the most popular class is returned. Random forests can be built using bagging in tandem with random attribute selection.
- **Example:** Suppose we are a movie director and we have created a short movie on a very important and interesting topic. Now, we want to take preliminary feedback (ratings) on the movie before making it public. What are the possible ways by which we can do that?
  1. We may ask one of our friends to rate the movie for us: Now it's entirely possible that the person we have chosen loves us very much and doesn't want to break our heart by providing a 1-star rating to the horrible work you have created.
  2. Another way could be by asking five colleagues of yours to rate the movie. Answer not be again appropriate.
  3. Now we can ask to 50 random people.
- The responses, in this case, would be more generalized and diversified since now we have people with different sets of skills. And as it turns out – this is a better approach to get honest ratings.
- With these examples, we can infer that a diverse group of people are likely to make better decisions as compared to individuals.
- Similar is true for a diverse set of models in comparison to single models. This diversification in Machine Learning (ML) is achieved by a technique called Ensemble Learning.
- Ensembling is nothing but the technique to combine several individual predictive models to come up with the final predictive model. Max voting, Averaging and Weighted averaging are the basic ensemble techniques/methods.
- Other ensemble techniques are:
  1. **Max Voting:** The max voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a 'vote'. The predictions which we get from the majority of the models are used as the final prediction.

2. **Averaging:** In this method, we take an average of predictions from all the models and use it to make the final prediction.
3. **Weighted Averaging:** All models are assigned different weights defining the importance of each model for prediction. For example, if two of our colleagues are critics, while others have no prior experience in this field, then the answers by these two friends are given more importance as compared to the other people.

2.5 REGRESSION MODELS

- Regression helps us to understand the relationship between various data points and helps us to find hidden patterns among the data.
- Regression is one of the most powerful and popular statistical tool or a learning technique that helps to discover the best relationship between a dependent variable and an independent variable.
- The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x.
- Regression analysis is a set of statistical processes for estimating the relationships among variables.
- Regression analysis is a set of statistical methods used to estimate relationships between a dependent variable (target) and one or more independent variables (predictor).
- More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.

2.5.1 Linear Regression

- Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.
- Linear regression is the most representative machine learning method to build models for value prediction and classification from training data.
- Linear regression maps an independent variable to a dependent variable by a linear equation. Many times an independent variable can have a deterministic mapping to a dependent variable.
- Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables.
- Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.

- If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression.
- The linear regression model provides a sloped straight line representing the relationship between the variables, (see Fig. 2.22).

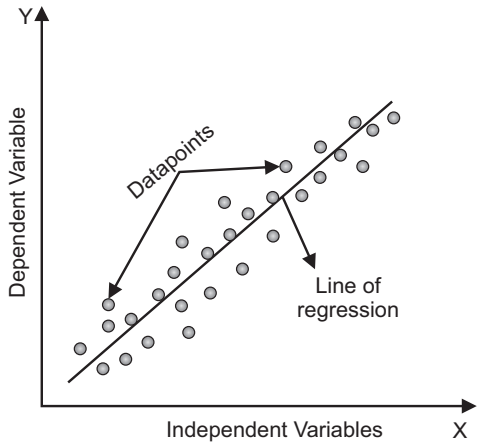


Fig. 2.22

- Mathematically the relationship can be represented with the help of following equation,

$$Y = aX + b \quad \ldots (2.1)$$

Here, Y = dependent variables (target variables), X = Independent variables (predictor variables), a and b are the linear coefficients.

- To quantify the number of errors the regression line has committed, all the errors have to be added up.
- However, if the sign of the errors are both positive and/or negative then certain errors might cancel out each other and therefore wouldn't be reflected in the overall error computation.
- It is therefore, necessary that the error is squared and one of the popular ways of computing the error could be the Root Mean Squared Error (RMSE). The lower the RMSE, the better is the regression model.
- Equation 2.1 could be modified in line with the definition of the linear regression model to incorporate the error term. The modified equation is represented in Equation 2.2 in which e is the error term:

$$Y = a_0 + a_1X + \varepsilon \quad \ldots (2.2)$$

Here,

Y = Dependent Variable (Target Variable)

X = Independent Variable (predictor Variable)

a_0 = intercept of the line (Gives an additional degree of freedom)

a_1 = Linear regression coefficient (scale factor to each input value).

ε = random error

- The values for X and Y variables are training datasets for Linear Regression model representation.
- Linear regression uses the relationship between the data-points to draw a straight line through all them.
- A linear line showing the relationship between the dependent and independent variables is called a regression line.
- A regression line can show following two types of relationship:
  1. **Positive Linear Relationship:** If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

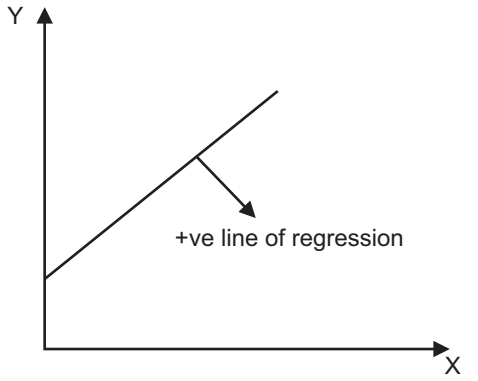


Fig. 2.23

  2. **Negative Linear Relationship:** If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

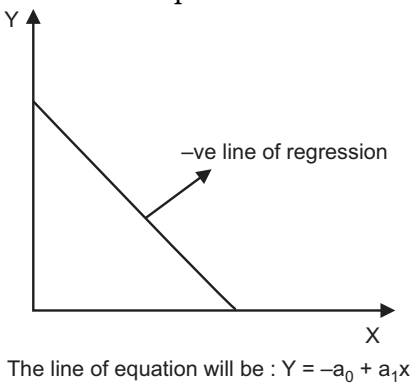


Fig. 2.24

- Linear regression can be further divided into two types of the algorithm:
  1. **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a linear regression algorithm is called simple linear regression.
  2. **Multiple Linear Regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a linear regression algorithm is called multiple linear regression.
- The relationship between variables in the linear regression model can be shown in the Fig. 2.25. Here, we are predicting the salary of an employee on the basis of the year of experience.

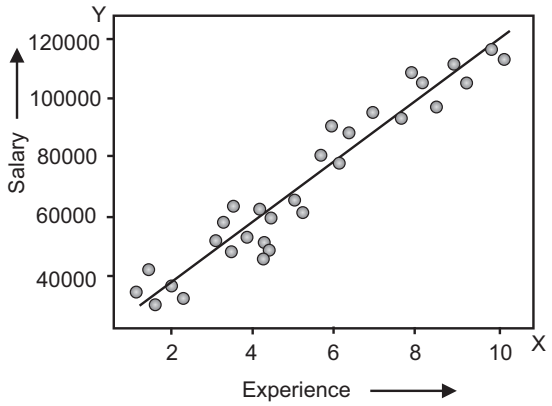


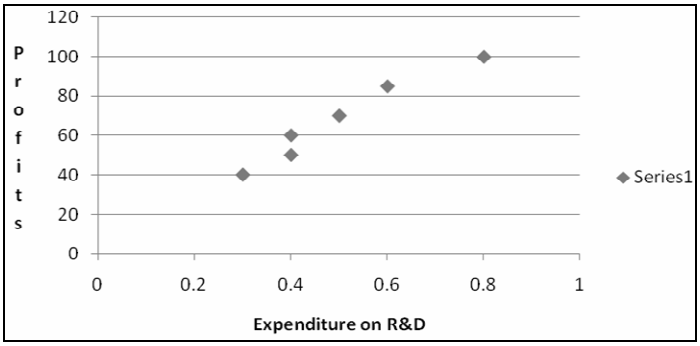
Fig. 2.25

Example on Fitting of Linear Regression: The following data gives expenditure on R&D and profit of a company

Profit	50	60	40	70	85	100
Expenditure on R&D	0.40	0.40	0.30	0.50	0.60	0.80

1. Find the regression equation of profit on R & D expenditure.
2. Estimate the profit when expenditure on R & D is budgeted at Rs 1 Crore.
3. Find the correlation coefficient.
4. What proportion of variability in profit is explained by variability in expenditure on R&D.

Solution:



From the above Scatter diagram, WE observed that profit and expenditure on R&D are highly positively correlated.

In the above example, let X denote expenditure on R&D and Y denote profit

Profit (Y)	**Expen-diture on R&D (X)**	**X^2**	**Y^2**	**XY**	$\hat{Y}$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
50	0.4	0.16	2500	20	55.3125	5.3125	28.22266
60	0.4	0.16	3600	24	55.3125	4.6875	21.97266
40	0.3	0.09	1600	12	43.125	-3.125	9.765625
70	0.5	0.25	4900	35	67.5	2.5	6.25
85	0.6	0.36	7225	51	79.6875	5.3125	28.22266
100	0.8	0.64	10000	80	104.0625	4.0625	16.50391
$\Sigma y = 405$	$\Sigma x = 3$	$\Sigma x^2 = 1.66$	29825	$\Sigma yx = 222$			110.9375

Step 1: Means,

$$\bar{x} = \frac{\Sigma x}{n} = 0.5; \quad \bar{y} = \frac{\Sigma y}{n} = 67.5$$

Step 2: Variances

$$\text{var}(X) = \frac{\Sigma x^2}{n} - \bar{x}^2 = 0.026667; \quad \text{Var}(Y) = \frac{\Sigma y^2}{n} - \bar{y}^2 = 414.5833$$

Step 3: Standard Deviations,

SD(X) = SQRT(V(X)) = 0.163299;

SD(Y) = SQRT(V(Y)) = 20.36132

Step 4: Covariance,

$$\text{cov}(X,Y) = \frac{\Sigma xy}{n} - \bar{x} \times \bar{y} = 3.25$$

Step 5: Karl Pearson's coefficient of correlation,

$$r(X,Y) = \frac{\text{Cov}(X, Y)}{SD(X) \times SD(Y)} = 0.977447$$

Step 6: Regression coefficient of Y on X,

$$b = \frac{\text{Cov}(X, Y)}{\text{Var}(x)} = 121.875$$

Step 7: Intercept

$$a = \bar{y} - (b \times \bar{x}) = 6.5625$$

Step 8: Hence, the equation of least square of line of the regression equation of profit on R & D expenditure is $Y = 6.5625 + 121.875 \times X$

Step 9: Total sum of squares = $\Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - (n \times \bar{Y})^2 = 2487.5$

Step10: Sum of squares due to error = = $\Sigma(Y - \hat{Y})^2 = 110.9375$

Step 11: Coefficient of determination = $r^2 = 1 - \frac{SSE}{TSS} = 0.955402$

2. To Estimate the profit when expenditure on R&D is budgeted at ₹ 1 Crore.
   For X = 1, the regression Estimate of the profit when expenditure on R&D is budgeted at ₹ 1 Crore is $\hat{Y} = 6.5625 + 121.875 \times 1 = 128.4375$
3. Find the correlation coefficient, r(X,Y) = 0.977447
   There exists high degree of positive correlation between the variables X and Y.
4. What proportion of variability in profit is explained by variability in expenditure on R&D?
   Since $r^2 = 0.955402$.
   It means 95.54% of variability in profit is explained by variability in expenditure on R&D.

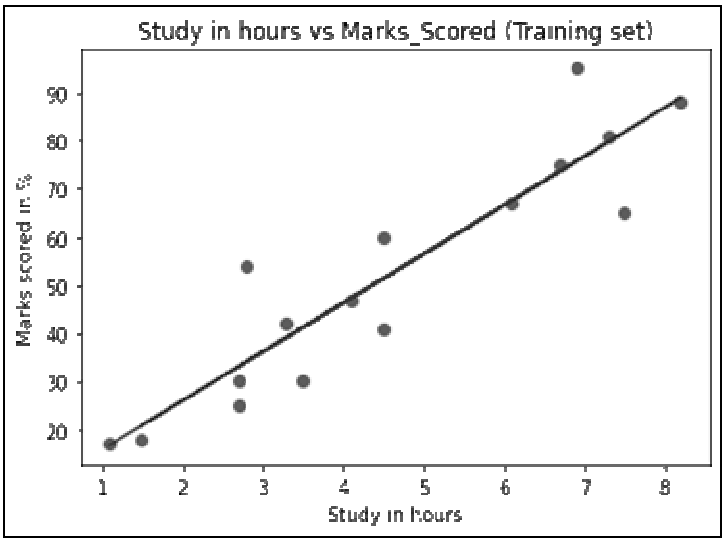
Program for Linear Regression:

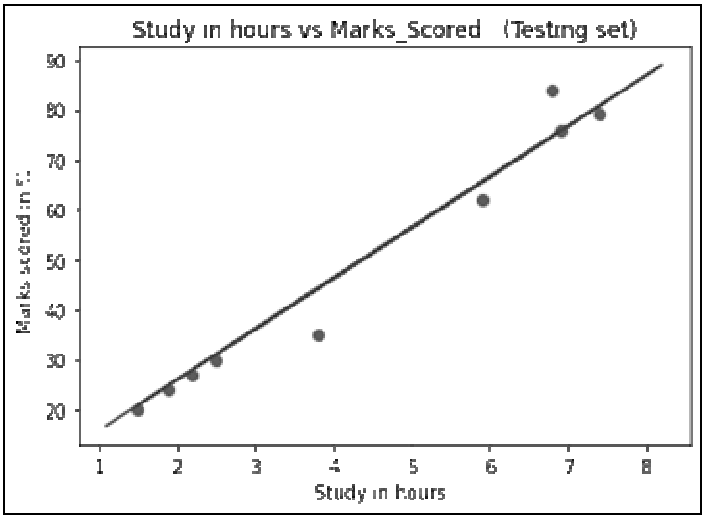
```
# importing the dataset
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from google.colab import files
data_to_load = files.upload()
import io
df = pd.read_csv(io.BytesIO(data_to_load['stsc.csv']))
#df = pd.read_csv('\b1.csv')
# data preprocessing
A = df.iloc[:, :-1].values  #independent variable array
B= df.iloc[:,1].values  #dependent variable vector
# splitting the dataset
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(A,B,test_size=1/3,
                                                    random_state=0)
# fitting the regression model
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train,y_train) #actually produces the linear
                                                    eqn for the data
# predicting the test set results
y_pred = regressor.predict(X_test)
y_pred
```

```
y_test
# graphical representation
#plot for the TRAIN data
plt.scatter(X_train, y_train, color='Green') # plotting the
                                                    observation line
plt.plot(X_train, regressor.predict(X_train), color='Red') # plotting
                                                    the regression line
plt.title("Study in hours vs Marks_Scored (Training set)") # stating
                                                  the title of the graph
plt.xlabel("Study in hours") # adding the name of x-axis
plt.ylabel("Marks scored in %") # adding the name of y-axis
plt.show() # specifies end of graph
#plot for the TEST data
plt.scatter(X_test, y_test, color='Green')
plt.plot(X_train, regressor.predict(X_train), color='Brown') # plotting
                                                    the regression line
plt.title("Study in hours vs Marks_Scored   (Testing set)")
plt.xlabel("Study in hours ")
plt.ylabel("Marks scored in % ")
plt.show()
```

Output:

stsc.csv(application/vnd.ms-excel)-214 bytes, last modified: 2/1/2022-100% done





2.5.2 Polynomial Regression

- Polynomial regression, like linear regression. It uses the relationship between the variables x and y to find the best way to draw a line through the data points.
- The dataset used in Polynomial regression for training is of non-linear nature. It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.
- Polynomial regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.
- The Polynomial regression equation is: $y = b_0 + b_1x_1 + b_2x_1^2 + ... b_nx_1^n$, where data points are arranged in a non-linear fashion, we need the Polynomial regression model.
- Fig. 2.26 shows the curve is near exponential due to the presence of the term x_1^2. The Fig. 2.26 shows that no way a linear line could fit all the data points.
- However, by transforming the linear line into a polynomial form, the curve is made to pass through all the points.

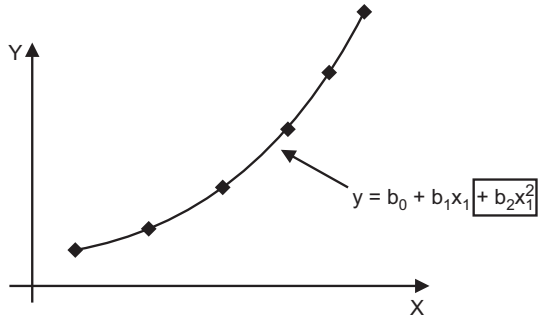
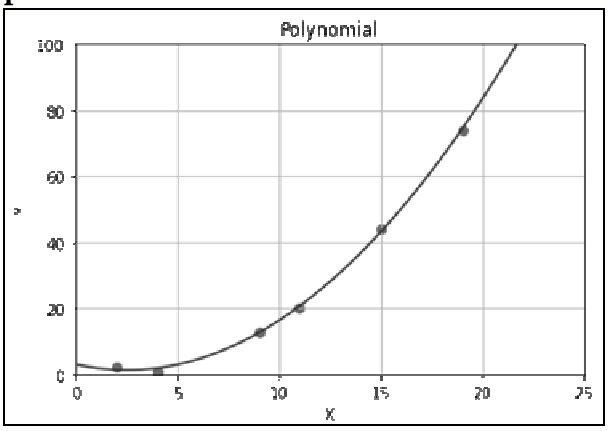


Fig. 2.26: Polynomial Regression

Program for Polynomial Regression:

```
import numpy as np
import matplotlib.pyplot as plt
X = [2, 4,9, 11, 15, 19]
Y = [2, 1, 13, 20, 44, 74]
# Train Algorithm (Polynomial)
degree = 2
poly_fit = np.poly1d(np.polyfit(X,Y, degree))
# Plot data
ab = np.linspace(0, 26, 100)
plt.plot(ab, poly_fit(ab), c='r',linestyle='-')
plt.title('Polynomial')
plt.xlabel('X')
plt.ylabel('Y')
plt.axis([0, 25, 0, 100])
plt.grid(True)
plt.scatter(X, Y)
plt.show()
# Predict price
print( poly_fit(14) )
```

Output:




```
36.73998103448068
```

2.5.3 Logistic Regression

- Logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it.
- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.

- The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.
- In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).
- Logistic regression is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic regression is much similar to the Linear Regression except that how they are used.
- Linear regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- The logistic regression primarily deals with binary output such as play versus not-play or success versus failure.
- In logistic regression, multiple independent variables are mapped to a single dependent variable. Popularly two types of logistic regression are found-binary and multinomial.
- Binary logistic regression is used when the dependent variable partitions the output class into two subsets and independent variables are found to be either categorical or continuous.
- However, if the dependent variable divides the target class into more than two subsets then the logistic regression is multinomial.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). A logistic curve is an S-shaped or sigmoid shape.

Sigmoid Function:

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- The sigmoid function maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.
- Fig. 2.27 shows assumptions logistic function.

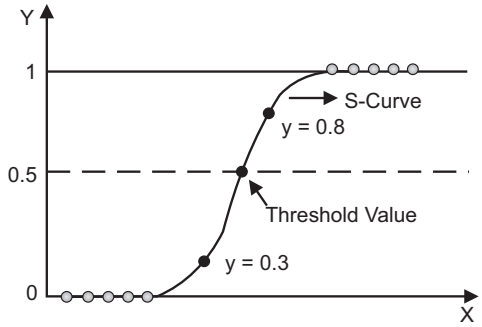


Fig. 2.27: Logistic Regression Assumptions for Logistic Regression

- Before diving into the implementation of logistic regression, we must be aware of the following assumptions about the same:
  - In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.
  - There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other.
  - We must include meaningful variables in our model.
  - We should choose a large sample size for logistic regression.
- Logistic regression can be divided into following types:
  1. **Binary or Binomial:** In such a kind of classification, a dependent variable will have only two possible types either 1 or 0. For example, these variables may represent pass or fail, success or failure, yes or no, win or loss etc.
  2. **Multinomial:** In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.
  3. **Ordinal:** In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have the scores like 0,1,2,3.
- The simplest form of logistic regression is binary or binomial logistic regression in which the target or dependent variable can have only 2 possible types either 1 or 0.

- It allows us to model a relationship between multiple predictor variables and a binary/binomial target variable.
- To sigmoid curve can be represented with the help of following graph (See Fig. 2.28). We can see the values of y-axis lie between 0 and 1 and crosses the axis at 0.5.

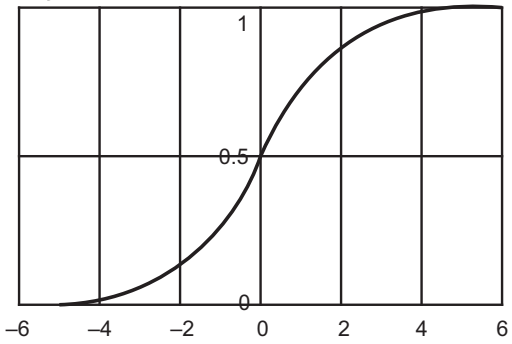


Fig. 2.28

- The classes can be divided into positive or negative. The output comes under the probability of positive class if it lies between 0 and 1.
- For our implementation, we are interpreting the output of hypothesis function as positive if it is ≥ 0.5, otherwise negative.

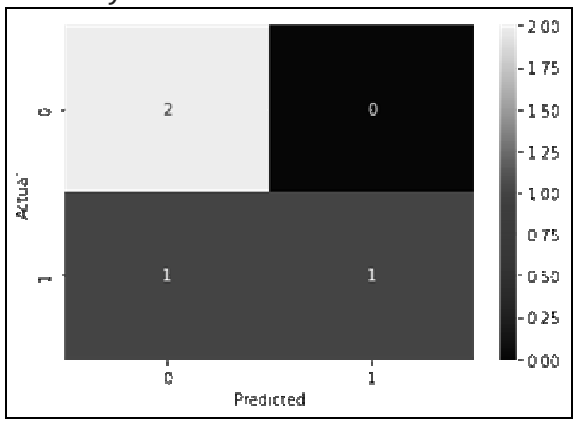
Program for Logistic Regression:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import seaborn as sn
import matplotlib.pyplot as plt
import pandas as pd
students = {'cmat': [680,650,690,700,640,720,680,710,730,640,620,680,700,65
           0,760 ],'gpa': [4,3.9,3.3,3.7,3.9,3.7,2.3,3.3,3.3,1.7,2.7,3.7,
           3.7,3.3,3.3 ], 'work_exp': [3,4,3,5,4,6,1,4,5,1,3,5,6,4,3 ],
           'status': [1,1,0,1,0,1,0,1,1,0,0,1,1,0,1]}
#status is admission status -1 for admitted,0 for not admitted
df = pd.DataFrame(students,columns= ['cmat', 'gpa','work_exp','status'])
print (df)
X = df[['cmat', 'gpa','work_exp']]
y = df['status']
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.25,
                                                 random_state=0)
logistic_regression= LogisticRegression()
logistic_regression.fit(X_train,y_train)
y_pred=logistic_regression.predict(X_test)
```

```
confusion_matrix = pd.crosstab(y_test, y_pred, rownames=['Actual'],
                                                colnames=['Predicted'])
sn.heatmap(confusion_matrix, annot=True)
print('Accuracy is: ',metrics.accuracy_score(y_test, y_pred))
plt.show()
print("Prediction for new students")
new_students = {'cmat': [590,740,680,610,710],
                'gpa': [2,3.7,3.3,2.3,3],
                'work_exp': [3,4,6,1,5]
               }
df2 = pd.DataFrame(new_students,columns= ['cmat', 'gpa','work_exp'])
y_pred=logistic_regression.predict(df2)
print (df2)
print (y_pred)
```

Output:

```
     cmat    gpa  work_exp  status
0    680     4.0     3       1
1    650     3.9     4       1
2    690     3.3     3       0
3    700     3.7     5       1
4    640     3.9     4       0
5    720     3.7     6       1
6    680     2.3     1       0
7    710     3.3     4       1
8    730     3.3     5       1
9    640     1.7     1       0
10   620     2.7     3       0
11   680     3.7     5       1
12   700     3.7     6       1
13   650     3.3     4       0
14   760     3.3     3       1
Accuracy is:  0.75
```



```
Prediction for new students
   cmat  gpa  work_exp
0   590  2.0     3
1   740  3.7     4
2   680  3.3     6
3   610  2.3     1
4   710  3.0     5
[0 1 1 0 1]
```

2.6 CONCEPT OF CLASSIFICATION

- Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels (output variable that is being predicted).
- For example, we can build a classification model to categorize bank loan applications as either safe or risky.
- Classification is the most widely used data science task in business. The objective of a classification model is to predict a target variable that is binary (e.g., a loan decision) or categorical (e.g., a customer type) when a set of input variables are given.
- The model does this by learning the generalized relationship between the predicted target variable with all other input attributes from a known dataset.
- In classification or class prediction, one should try to use the information from the predictors or independent variables to sort the data samples into two or more distinct classes.
- Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.
- The classification algorithm is a supervised learning technique that is used to identify the category of new observations on the basis of training data.
- In classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam and so on. Classes can be called as targets/labels or categories.
- Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data).
- Classification algorithms can be better understood using the Fig. 2.29.
- In the Fig. 2.29 there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.

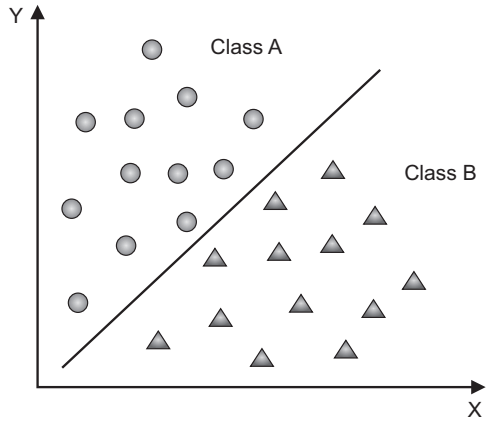


Fig. 2.29

- The algorithm which implements the classification on a dataset is known as a classifier. There are two types of classifiers:
  1. **Binary Classifier:** If the classification problem has only two possible outcomes, then it is called as Binary Classifier.
     Examples: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.
  2. **Multi-class Classifier:** If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.
     Example: Classifications of types of crops, Classification of types of music.

Classification Techniques:

1. **Decision Tree Classification:** Decision trees are also known as classification trees. Decision trees approach the classification problem by partitioning the data into purer subsets based on the values of the input attributes. The attributes that help achieve the cleanest levels of such separation are considered significant in their influence on the target variable and end up at the root and closer-to-root levels of the tree. The output model is a tree framework than can be used for the prediction of new unlabeled data.
2. **k-NN Classification:** The k-NN (k-Nearest Neighbor) classification is based on the principle that any objects in nature that have similarities tend to be in close proximity to each other.
3. **Support Vector Machine (SVM) Classification:** SVM is one of the most widely used classification algorithms both for separating linear or non-linear data. With linear data, the objective is to find that linear hyperplane that separates the instances of two different classes by the maximal distance apart.
4. **Random Forest Classification:** Another standard most-commonly used classification in supervised machine learning is the Random Forest classification. Random Forest can be used both for classification and regression problems. The random forest as the name indicates is a forest, but the forest of what? It is a forest

of decision trees. As seen in Fig. 2.30, decision trees are generated with randomly drawn instances from the training set. The trees are constructed as shown in the algorithm. However, the final class of classification of the input instance is based on the majority voting as shown in Fig. 2.30.

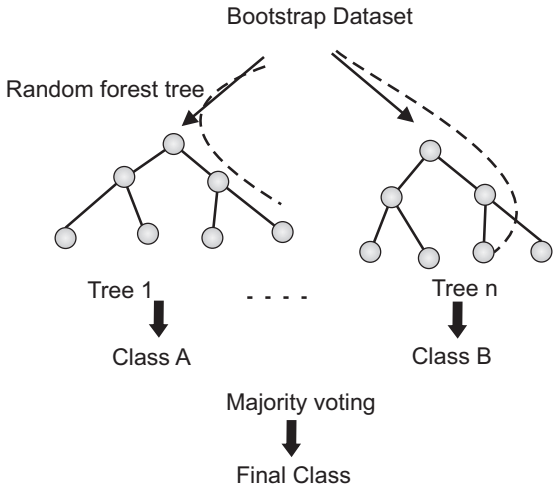


Fig. 2.30: Random Forest

Advantages of Random Forest Algorithm:

(i) Random forests are very flexible and possess very high accuracy.

(ii) It overcomes the problem of overfitting by averaging or combining the results of different decision trees.

(iii) Random forests work well for a large range of data items than a single decision tree does.

(iv) Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling.

Disadvantages of Random Forest Algorithm:

(i) Complexity is the main disadvantage of Random forest algorithms.

(ii) Construction of random forests are much harder and time-consuming than decision trees.

(iii) More computational resources are required to implement Random Forest algorithm.

5. **Naive Bayes Classification:** Naive Bayes classification works on the principle of Bayes theorem. It is a probability based classification approach. However, the probability of interest is not an absolute probability but a conditional probability. In conditional probability, one important thing to note is that the variables should be independent of each other. For dependent variables, conditional probability doesn't hold.

2.7 CONCEPT OF CLUSTERING

- Clustering is the process of identifying the natural groupings in a dataset. Clustering is one of the widely used unsupervised learning techniques in machine learning.
- Clustering is the process of finding meaningful groups in data. In clustering, the objective is not to predict a target class variable, but to simply capture the possible natural groupings in the data.
- Clustering is an attempt to group the data points into distinct “clusters.” The process of dividing the dataset into meaningful groups is clustering.
- Clustering is important in circumstances where identifying learning features for classification from random data distribution could be difficult or would require high expertise.
- In such circumstances, similar data could be easily clustered by applying clustering algorithms.
- Once different clusters are formed the similarity characteristics could be extracted for further analysis.
- Clustering algorithms used to find similarity as well as relationship patterns among data samples and then cluster those samples into groups having similarity based on features.
- The real-world example of clustering is to group the customers by their purchasing behavior. Cluster analysis is based on grouping similar data objects as clusters.
- Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.
- The most common application of clustering is to explore the data and find all the possible meaningful groups in the data.
- Depending on the clustering technique used, the number of groups or clusters is either user-defined or automatically determined by the algorithm from the dataset.
- Since clustering is not about predicting the membership of a customer in a well-defined meaningful group (e.g., frequent high-volume purchaser), the similarities of customers within a group need to be carefully investigated to make sense of the group as a whole.
- Clustering can be defined as "a way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

Types of Clustering:

- Clustering is a way to group a set of data points in a way that similar data points are grouped together. Therefore, clustering algorithms look for similarities or dissimilarities among data points.
- Clustering is an unsupervised learning method so there is no label associated with data points. The algorithm tries to find the underlying structure of the data.
- In Fig. 2.31 all data points in Cluster 2 are closer to other data points in Cluster 2 than to other data points in Cluster 1.

- Before explaining the different ways to implement clustering, the different types of clusters have to be defined.
- Based on a data point's membership to an identified group, a cluster can be:
  1. **Overlapping Clusters:** These are also known as multi-view clusters. The cluster groups are not exclusive and each data object may belong to more than one cluster. For example, a customer of a company can be grouped in a high-profit customer cluster and a high-volume customer cluster at the same time.
  2. **Exclusive or Strict Partitioning Clusters:** Each data object, in this cluster belongs to one exclusive cluster, like the example shown in Fig. 2.32. This is the most common type of cluster.
  3. **Fuzzy or Probabilistic Clusters:** Each data point, in this cluster belongs to all cluster groups with varying degrees of membership from 0 to 1. For example, in a dataset with clusters A, B, C, and D, a data point can be associated with all the clusters with degree A50.5, B50.1, C50.4 and D50. Instead of a definite association of a data point with one cluster, fuzzy clustering associates a probability of membership to all the clusters. Fuzzy C-means algorithm is the example of this type of clustering; it is sometimes also known as the Fuzzy k-means algorithm.
  4. **Hierarchical Clusters:** In this type of cluster, each child cluster can be merged to form a parent cluster. For example, the most profitable customer cluster can be further divided into a long-term customer cluster and a cluster with new customers with high-value purchases.

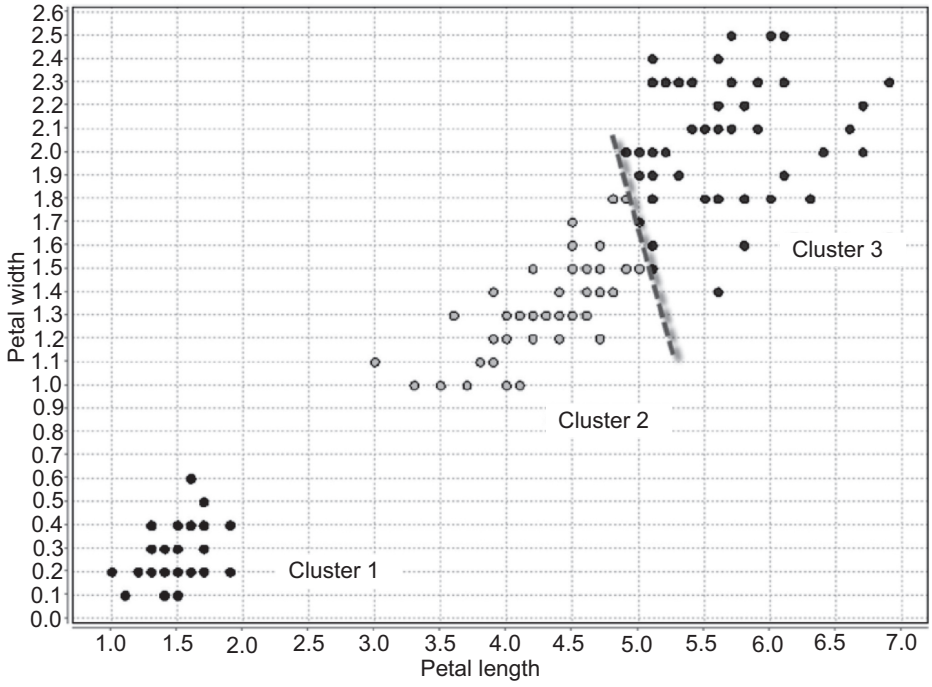


Fig. 2.31

- Clustering techniques in machine learning can also be classified based on the algorithmic approach used to find clusters in the dataset.
- Each of these techniques of clustering algorithms differs based on what relationship they leverage between the data objects. Some of these techniques are explained below:
  1. **Model-based Clustering:** This technique of clustering gets its foundation from statistics and probability distribution models; this technique is also called distribution-based clustering. A cluster can be thought of as a grouping that has the data points belonging to the same probability distribution. Hence, each cluster can be represented by a distribution model such as Gaussian or Poisson, where the parameter of the distribution can be iteratively optimized between the cluster data and the model. With this approach, the entire dataset can be represented by a mixture of distribution models. Mixture of Gaussians is one of the model-based clustering techniques used where a fixed number of distributions are initialized, and parameters are optimized to fit the cluster data.
  2. **Prototype-based Clustering:** In this technique of clustering, each cluster is represented by a central data object, also called a prototype. The prototype of each cluster is usually the center of the cluster, hence, this clustering is also called centroid clustering or center-based clustering. For example, in clustering customer segments, each customer cluster will have a central prototype customer and customers with similar properties are associated with the prototype customer of a cluster.
  3. **Density Clustering:** In Fig. 2.31, it can be observed that clusters occupy the area where there are more data points per unit space and are separated by sparse space. A cluster can also be defined as a dense region where data objects are concentrated surrounded by a low-density area where data objects are sparse. Each dense area can be assigned a cluster and the low-density area can be discarded as noise. In density clustering not all data objects are clustered since noise objects are unassigned to any cluster.
  4. **Hierarchical Clustering:** Hierarchical clustering is a process where a cluster hierarchy is created based on the distance between data points. The output of a hierarchical clustering is a dendrogram (a tree diagram that shows different clusters at any point of precision which is specified by the user). There are two approaches to create a hierarchy of clusters. A bottom-up approach is where each data point is considered a cluster, and the clusters are merged to finally form one massive cluster. The top-down approach is where the dataset is considered one cluster and they are recursively divided into different sub-clusters until individual data objects are defined as separate clusters. Hierarchical clustering is useful when the data size is limited. A level of interactive feedback is required to cut the dendrogram tree to a given level of precision.

k-Means Clustering:

- The k-means clustering technique identifies a cluster based on a central prototype record.
- The k-Means clustering is a prototype-based clustering method where the dataset is divided into k-clusters.

DBSCAN Clustering:

- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering.
- Density-based clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data.
- It is based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.

Self-Organizing Maps:

- A Self-Organizing Map (SOM) is a powerful visual clustering technique that evolved from a combination of neural networks and prototype-based clustering.
- The SOM was introduced by Teuvo Kohonen in the 1980s. This technique is also known as Kohonen networks or Kohonen map. SOM is sometimes also referred to by a more specific name, Self-Organizing Feature Maps (SOFM).
- A SOM is a form of neural network where the output is an organized visual matrix, usually a two-dimensional grid with rows and columns.

2.8 CONCEPT OF REINFORCEMENT LEARNING

- Reinforcement Learning (RL) is a feedback-based Machine Learning (ML) technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions.
- The RL is an agent-based goal seeking technique where an (AI) agent tries to determine the best action to take in a given environment depending on a reward.
- The agent has access to data which correspond to the various states in an environment and a label for each action.
- A deep learning network may be used to take in an observation or state-array and output probabilities for each action (or label).
- The most popular implementation of RL is Google's AlphaGo AI which defeated a top-ranked human Go player.
- Practical applications of RL include route optimization strategies for a self-driving vehicle, for example. Most such applications are experimental as of this publication.
- In RL the agent learns automatically using feedbacks without any labeled data, unlike supervised learning.

- Fig. 2.32 shows the typical framing of a Reinforcement Learning (RL) scenario.
- An agent takes actions in an environment, which is interpreted into a reward and a representation of the state, which are fed back into the agent.

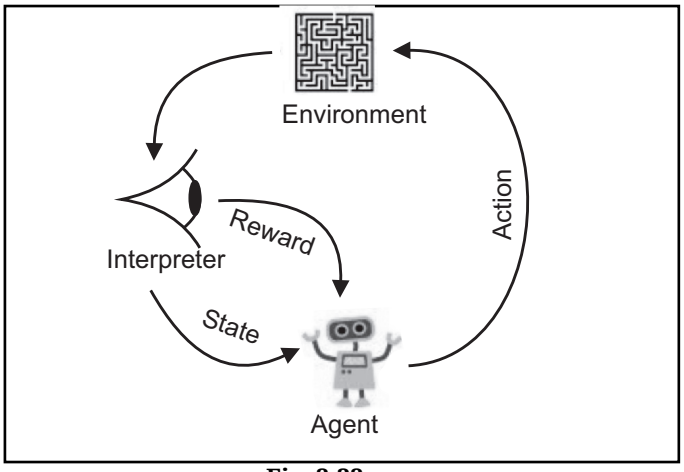


Fig. 2.32

- Fig. 2.33 shows a real-world example of Reinforcement Learning (RL).

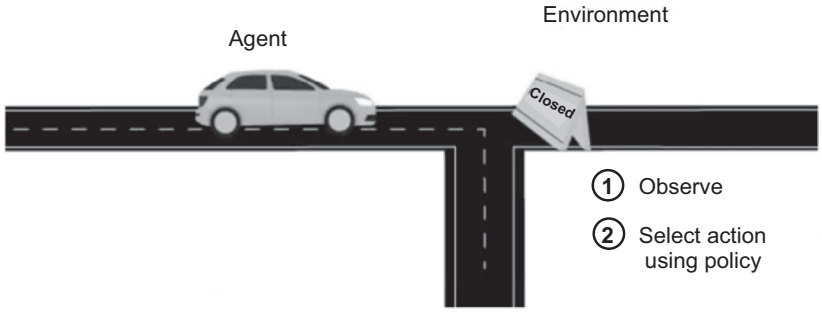


Fig. 2.33: Real-world Example of Reinforcement Learning

Basic Terms used in Reinforcement Learning:

1. **Agent** is an entity that can perceive/explore the environment and act upon it.
2. **Environment** is a situation in which an agent is present or surrounded by. In RL, we assume the stochastic environment, which means it is random in nature.
3. **Actions** are the moves taken by an agent within the environment.
4. **State** is a situation returned by the environment after each action taken by the agent.
5. **Reward** is a feedback returned to the agent from the environment to evaluate the action of the agent. A reward defines the goal of a reinforcement learning problem.
6. **Policy** is a strategy applied by the agent for the next action based on the current state. A policy is a mapping from the states of the environment to actions to the actions the agent takes in the environment.

7. **Value** is expected long-term retuned with the discount factor and opposite to the short-term reward. The value of a state is the total aggregated number of rewards that the agent can expect to get in the future if it starts from that state.

Advantages of Reinforcement Machine Learning:

1. RL is used to solve complex problems that cannot be solved by conventional techniques.
2. The solutions obtained by RL very accurate.
3. RL model will undergo a rigorous training process that can take time. This can help to correct any errors.
4. Due to RL's learning ability, it can be used with neural networks. This can be termed as deep reinforcement learning.
5. When it comes to creating simulators, object detection in automatic cars, robots, etc., reinforcement learning plays a great role in the models.

Disadvantages of Reinforcement Machine Learning:

1. RL needs a lot of data and a lot of computation.
2. Too much reinforcement learning can lead to an overload of states which can diminish the results.
3. RL algorithm is not preferable for solving simple problems. To solving simpler problems won't be correct.
4. RL need lots of data to feed the model consumes time and lots of computational power.
5. RL models require a lot of training data to develop accurate results.
6. When it comes to building RL models on real-world examples, the maintenance cost is very high.

PRACTICE QUESTIONS

Q. I Multiple Choice Questions:

1. Which is the process of converting experience into expertise or knowledge.?
   (a) Learning (b) Writing
   (c) Listening (d) All of the mentioned
2. Machine Learning (ML) techniques are broadly categorized into,
   (a) supervised ML (b) unsupervised ML
   (c) reinforcement ML (d) All of the mentioned
3. Which is the simulation of human intelligence processes by machines, especially computer systems?
   (a) Machine Learning (ML) (b) Artificial Intelligence (AI)
   (c) Deep Learning (d) All of the mentioned

4. Supervised learning is a learning technique that can only be applied on,
   (a) unlabeled data (b) by labeled data
   (c) labeled data (d) All of the mentioned
5. Which algorithm stores all the available data and classifies a new data point based on the similarity?
   (a) k-NN (b) f-NN
   (c) l-NN (d) All of the mentioned
6. Deep learning tools includes,
   (a) Deeplearning4j (b) Cognitive Toolkit
   (c) Tensor Flow (d) All of the mentioned
7. Which is a translator (translates source code to object/target code) A decision tree builds on which models in the form of a tree structure?
   (a) classification (b) regression
   (c) Both (a) and (b) (d) None of the mentioned
8. Which learning a type of machine learning in which models are trained using unlabeled dataset?
   (a) supervised ML (b) unsupervised ML
   (c) reinforcement ML (d) All of the mentioned
9. Which is a data set is a data sample that we select for the model to learn to perform actions from.
   (a) training (b) observing
   (c) writing (d) All of the mentioned
10. The components of Apriori algorithm includes,
    (a) Support (b) Confidence
    (c) Lift (d) All of the mentioned
11. Which is detection is an unsupervised ML method?
    (a) Error (b) Exception
    (c) Anomaly (d) All of the mentioned
12. Which ML is a combination of supervised and unsupervised learning?
    (a) supervised (b) semi-supervised
    (c) unsupervised (d) All of the mentioned
13. Which regression maps an independent variable to a dependent variable by a linear equation?
    (a) Linear (b) Non-linear
    (c) Both (a) and (b) (d) None of the mentioned
14. Logistic regression can be divided into,
    (a) Binomial (b) Multinomial
    (c) Ordinal (d) All of the mentioned

15. Which methods are techniques that create multiple models and then combine them to produce improved/accurate results?
    (a) Machine dependant (b) Ensemble
    (c) Computer machine independent (d) All of the mentioned
16. Which is a form of data analysis that extracts models describing important data classes?
    (a) Regression (b) Binary classifier
    (c) Classification (d) None of the mentioned
17. Boosting takes many forms, including
    (a) AdaBoost (b) XGBoost
    (c) Gradient boosting (d) All of the mentioned
18. Which clustering partitions the data based on variation in the density of records in a dataset?
    (a) HIERARCHICALSCAN (b) TRESCAN
    (c) DBSCAN (d) None of the mentioned

Answers

1. (a)	2. (d)	3. (b)	4. (c)	5. (a)	6. (d)	7. (c)	8. (b)	9. (a)	10. (d)
11. (c)	12. (b)	13. (a)	14. (d)	15. (b)	16. (c)	17. (d)	18. (c)		

Q. II Fill in the Blanks:

1. _____ learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions.
2. Reinforcement learning is a _____ -based Machine Learning (ML) technique.
3. _____ learning is a subset of ML and ML is a subset of AI.
4. The ultimate goal of _____ is to make machines as intelligent as humans.
5. Unsupervised learning is a learning method in which a machine learns _____ any supervision.
6. Classification refers to process of predicting _____ output values for an input.
7. In regression problems the task of machine learning model is to _____ a continuous value.
8. Speech recognition is a process of converting _____ instructions into text and known as computer speech recognition.
9. A _____ consists of constructs of information called features or predictors and a target or response variable.
10. _____ is of the model is extremely important because it determines whether the model works in real-life conditions.
11. The _____ data means some input data is already tagged with the correct output.

12. _____ learning is a process of providing input data as well as correct output data to the machine learning model.
13. The _____ tree is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
14. _____ is the gap between two lines on the closet data points of different classes.
15. _____ Bayes algorithm is a supervised learning algorithm
16. Unsupervised learning is a machine learning technique in which models are not supervised using _____ dataset.
17. _____ cluster analysis is used for exploratory data analysis in order to find the hidden patterns.
18. An _____ rule is an unsupervised learning method which is used for finding the relationships between variables in the large database.
19. Clustering is a method of grouping the objects into clusters such that objects with most _____ remains into a group and has less or no similarities with the objects of another group.
20. The _____ algorithm involves us to telling the algorithms how many possible cluster (or k) there are in the dataset.
21. Basket Data Analysis is to analyze the _____ of purchased items in a single basket or single purchase.
22. Apriori algorithm uses _____ datasets to generate association rules.
23. Anomaly or _____ detection identifies the data points that are significantly different from other data points in a dataset.
24. _____ -supervised learning falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data).
25. Regression analysis is a set of _____ processes for estimating the relationships among variables.
26. Linear regression shows the _____ relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
27. The dataset used in Polynomial regression for training is of _____ -linear nature.
28. _____ regression is a supervised learning classification algorithm used to predict the probability of a target variable.
29. The _____ function maps any real value into another value within a range of 0 and 1.
30. Random _____ can be built using bagging in tandem with random attribute selection.
31. The algorithm which implements the classification on a dataset is known as a _____.

32. The output of a hierarchical clustering is a _____ (a tree diagram).
33. _____ is the process of finding meaningful groups in data.
34. The DBSCAN algorithm starts with calculation of a _____ for all data points in a dataset, with a given fixed radius ε (epsilon).
35. A SOM is a form of neural network where the output is an organized visual _____, usually a two-dimensional grid with rows and columns.
36. _____ is a feedback returned to the agent from the environment to evaluate the action of the agent.

Answers

1. Machine	2. feedback	3. Deep	4. AI
5. without	6. discrete	7. predict	8. voice
9. model	10. Validation	11. labeled	12. Supervised
13. decision	14. Margin	15. Naïve	16. training
17. hierarchical	18. association	19. similarities	20. k-means
21. association	22. frequent	23. outlier	24. Semi
25. statistical	26. linear	27. non	28. Logistic
29. sigmoid	30. forests	31. classifier	32. dendrogram
33. Clustering	34. density	35. matrix	36. Reward

Q. III State True or False:

1. Supervised machine learning is a field of learning where the machine learns with the help of a supervisor and instructor.
2. Deep learning models are trained by using large sets of unlabeled data.
3. In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree.
4. Naïve Bayes is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
5. The k-means clustering is an unsupervised learning algorithm, which groups the unlabeled dataset into different clusters.
6. Linear regression is the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables.
7. Association rule mining discovers strong association or correlation relationships among data.
8. Lift is an indication of how frequently the itemset appears in the dataset.
9. The semi-supervised learning hybrids technique which combines supervised and unsupervised learning.
10. Logistic regression predicts the output of a categorical dependent variable.

11. The sigmoid function is a mathematical function used to map the predicted values to probabilities.
12. Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.
13. If the classification problem has only two possible outcomes, then it is called as Binary Classifier.
14. If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.
15. Clustering is supervised learning techniques in machine learning.
16. Decision trees are also known as classification trees.
17. Support Vector Machine (SVM) classification is used for separating linear or non-linear data.
18. The k-means clustering is a hierarchical-based clustering method where the dataset is divided into k-clusters.
19. Hierarchical clustering is a process where a cluster hierarchy is created based on the distance between data points.

Answers

1. (T)	2. (F)	3. (T)	4. (T)	5. (T)	6. (T)	7. (T)	8. (F)	9. (T)	10. (T)
11. (T)	12. (T)	13. (T)	14. (T)	15. (F)	16. (T)	17. (T)	18. (F)	19. (T)	

Q. IV Answer the following Questions:

(A) Short Answer Questions:

1. Define machine learning.
2. Define deep learning?
3. List types of machine learning.
4. Enlist three parameters for machine learning.
5. Define classification and regression.
6. Define reinforcement machine learning.
7. State any two uses of machine learning.
8. Define Neural Networks (NNs).
9. Define Artificial intelligence (AI).
10. List AI applications. Any two.
11. Define model.
12. Define supervised machine learning.
13. Give purpose of k-NN algorithm.
14. Define decision tree.
15. What is the purpose of SVM?
16. Give use of Naïve Bayes.

17. Define unsupervised machine learning.
18. Define clustering.
19. Define association rule mining.
20. What is the purpose of Apriori algorithm?
21. Define anomaly detection.
22. Differentiate between supervised and unsupervised machine learning.
23. Define semi-supervised machine learning.
24. Define regression analysis.
25. Define regression model.
26. What is logistic regression?
27. Define linear regression.
28. Define polynomial regression.
29. List ensemble techniques.
30. Define classification.
31. Define cluster and clustering.
32. Enlist types of clustering.
33. Give long form of DBSCAN.
34. Define SOM.

(B) Long Answer Questions:

1. What is machine learning? States its advantages and disadvantages. Also list its various applications.
2. What is deep learning? How it works? Explain diagrammatically.
3. What is AI? What is its purpose? State its advantages and disadvantages.
4. With the help of diagram describe relationship between AI, ML and DL.
5. Write a short note on: Learning models for algorithms.
6. List application of machine learning in data science.
7. With the help of suitable diagram describe machine learning model.
8. What is model? What is its use? How to select it? How to engineer features of model?
9. How to train and validate a model? Describe in detail.
10. What are the types of machine learning? Compare them.
11. What is supervised learning? How it works? State its advantages and disadvantages.
12. What is k-NN? How it works? Explain diagrammatically. Also state its advantages and disadvantages.
13. What is decision tree? How it works? State its advantages and disadvantages.
14. Explain Support Vector Machine (SVM) with the help of diagram.

15. Write a short note on: Naïve Bayes.
16. Describe unsupervised learning with diagram and advantages and disadvantages.
17. With the help of example explain k-means clustering algorithm.
18. What is association rule mining? Describe with the example.
19. Explain polynomial regression diagrammatically.
20. What is semi-supervised machine learning? With the help of diagram explain its basic idea. Also state its advantages and disadvantages.
21. What is regression model? Explain linear regression with diagram.
22. Define logistic regression with assumptions.
23. Write a short note on: Ensemble techniques.
24. With the help of example explain concept of classification. Also list various classification techniques.
25. What is random forest? Describe diagrammatically.
26. What is clustering? How it works? Explain with example.
27. Describe various clustering techniques. Describe two of them in short.
28. What is DBSCAN clustering? Explain with example.
29. Define SOM with example.
30. What is reinforcement learning? Explain diagrammatically. Also state its advantages and disadvantages.
31. Differentiate between ssupervised, unsupervised, semi-supervised and reinforcement machine learning.
32. What is meant by predicting new observations? Explain in detail.

■■■

CHAPTER 3

Mining Frequent Patterns, Associations and Correlations

Objectives...

- To understand Concept of Data Mining
- To learn Mining Frequent Patterns
- To study Concepts of Associations and Correlations

3.0 INTRODUCTION

- Today's information age a huge amount of data available and it increasing day by day. This data is of no use until it is converted into useful information.
- It is necessary to analyze this huge amount of data and extract useful information from it.
- Data mining is a technique that extracting information from huge sets of data. Data mining is the procedure of mining knowledge from data.

Overview of Data Mining:

- We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. Data mining can meet this need by providing tools to discover knowledge from data.
- Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.
- Data mining is defined as, "to extracting or mining knowledge from massive amount of datasets."
- Some people view data mining as an essential step in the process of knowledge discovery.
- Steps involved in Fig. 3.1 are explained below:

  Step 1: Data Cleaning: In this step, the noise and inconsistent data is removed and/or cleaned.

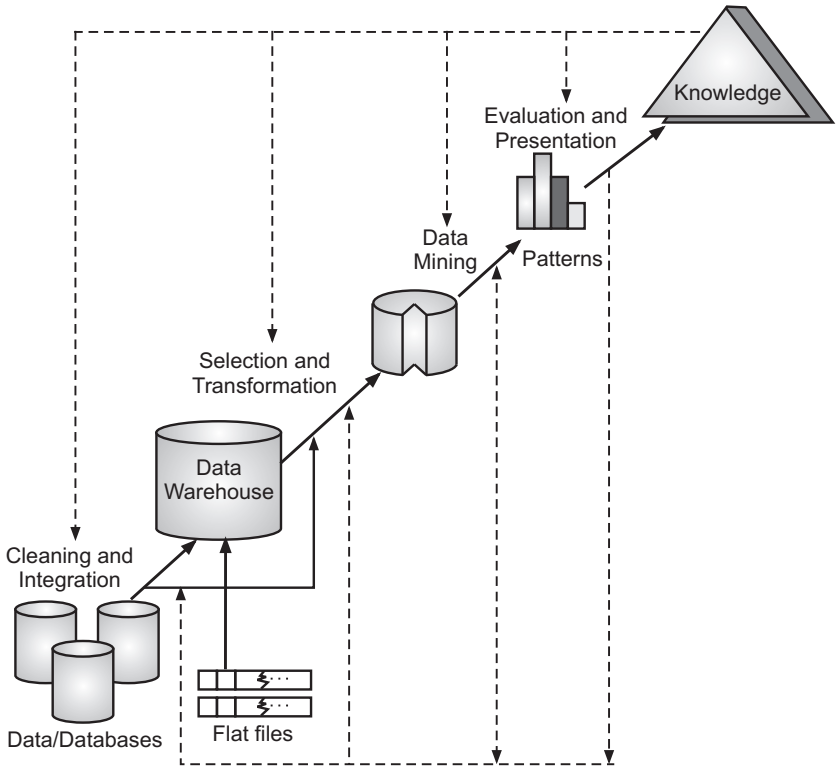


Fig. 3.1: Data Mining as a Step in the Process of Knowledge Discovery

Step 2: Data Integration: In this step, multiple data sources are combined.

Step 3: Data Selection: In this step, data relevant to the analysis task are retrieved from the dataset.

Step 4: Data Transformation: In this step, data is transformed or consolidated into forms appropriate for mining by performing aggregation or summary operations.

Step 5: Data Mining: In this step, intelligent methods are applied in order to extract data patterns.

Step 6: Pattern Evaluation: In this step, data patterns are evaluated.

Step 7: Knowledge Presentation: In this step, knowledge is represented.

- The most basic forms of data for mining applications are database data, data warehouse data and transactional data.
- Data mining can also be applied to other forms of data such as data streams data, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW data.

Advantages of Data Mining:

1. Data mining is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

2. The data mining technique enables organizations to obtain knowledge-based data.
3. Compared with other statistical data applications, data mining is efficient and cost-efficient.
4. Data mining helps the decision-making process of an organization.
5. Data mining facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
6. Data mining can be helpful while predicting future trends.

Disadvantages of Data Mining

1. Data mining system violates the privacy of its user and that is why it lacks in the matters of safety and security of its users.
2. Identity theft is a big issue when using data mining. If adequate security is not provided, it could pose vulnerabilities in the security. The personal information of the users is collected in the data mining. With such huge amount of data, hackers could easily access them and steal critical information.
3. Data mining technique is not a 100 percent accurate and may cause serious consequences in certain conditions.
4. Data mining involves lots of technology in use for the data collection process. Every data generated needs its own storage space as well as maintenance. This can greatly increase the implementation cost.

- Data mining is widely used in diverse areas such as Financial Banking, Telecommunication Industry, Healthcare, Education, Fraud detection, Bioinformatics, Retail Industry, Intrusion Detection and so on.

3.1 WHAT KIND OF PATTERNS CAN BE MINED?

- Frequent patterns, as the name suggests, are patterns that occur frequently in data. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data.
- Let us now examine the kinds of patterns that can be mined. There are a number of data mining functionalities such as characterization and discrimination, the mining of frequent patterns, associations and correlations, classification and regression, clustering analysis and outlier analysis.
- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories namely, descriptive and predictive.
  1. **Descriptive mining** tasks characterize properties of the data in a target data set.
  2. **Predictive mining** tasks perform induction on the current data in order to make predictions.

3.1.1 Class/Concept Description

- Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders.
- The descriptions of a class or a concept are called class/concept descriptions. The two ways from the descriptions are derived are characterization and discrimination.

3.1.1.1 Characterization and Discrimination

- These descriptions can be derived by the following two ways:
  1. **Data Characterization** refers to summarizing data of class under study. This class under study is called as target class. Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query. For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database. There are several methods for effective data summarization and characterization like statistical measures and plots, data cube-based OLAP, attribute-oriented induction and so on.
  2. **Data Discrimination** refers to the mapping or classification of a class with some predefined group or class. Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries. For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period. The methods used for data discrimination are similar to those used for data characterization.

3.1.2 Mining Frequent Patterns, Associations and Correlations

- Frequent patterns are patterns that occur frequently in data. In this section we will study mining of Frequent patterns, Associations and Correlations.

Mining of Frequent Patterns:

- There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns) and frequent substructures.
- A frequent itemset a set of items that often appear together in a transactional dataset. For example, butter and bread, which are frequently bought together in grocery stores by number of customers.
- A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card is a (frequent) sequential pattern.

- A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.
- If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Mining of Association:

- Associations are used in retail sales to identify patterns that are frequently purchased together.
- This process refers to the process of uncovering the relationship among data and determining association rules.
- For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

Mining of Correlations:

- It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

3.1.3 Classification and Regression for Predictive Analysis

- Classification is the process of finding a model that describes the data classes or concepts.
- For example, we can build a classification model to categorize bank loan applications as either safe or risky.
- Such analysis can help provide us with a better understanding of the data at large and to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data.
- The derived model can be presented in the forms of Classification (IF-THEN) Rules, Decision Trees, Mathematical Formulae and Neural Networks.
- A decision tree is a tree-like structure (also known as classification trees), where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distributions.
- Decision trees can easily be converted to classification rules. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.
- Classification predicts categorical (discrete, unordered) labels, regression models continuous-valued functions i.e., regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.
- The term prediction refers to both numeric prediction and class label prediction. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.

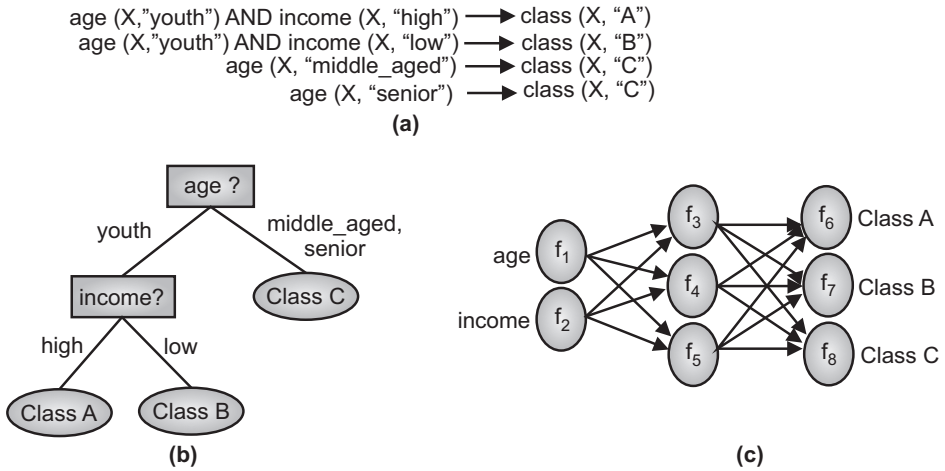


Fig. 3.2: A Classification Model can be represented in various Forms
(a) IF-THEN Rules (b) A Decision Tree (c) Neutral Network

- Regression also encompasses the identification of distribution trends based on the available data.
- Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification and regression process.
- Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration.
- Classification predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts.
- The derived model is based on the analysis set of training data i.e. the data object whose class label is well known.
- Prediction is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction.
- Prediction can also be used for identification of distribution trends based on available data.

3.1.4 Cluster Analysis

- Cluster refers to a group of similar kind of objects. Clustering is the process of identifying the natural groupings in a dataset.
- Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.
- Cluster analysis has been widely used in many applications such as Business Intelligence (BI), Image pattern recognition, Web search, Security and so on.
    - In business intelligence, clustering can be used to organize a large number of customers into groups, where customers within a group share strong similar characteristics. This facilitates the development of business strategies for enhanced customer relationship management.

- Moreover, consider a consultant company/firm with a large number of projects. To improve project management, clustering can be applied to partition projects into categories based on similarity so that project auditing and diagnosis (to improve project delivery and outcomes) can be conducted effectively.

- The set of clusters resulting from a cluster analysis can be referred to as a clustering. Clustering analyzes data objects without consulting class labels.
- Clustering can be used to generate class labels for a group of data. For example, cluster analysis can be performed on All Electronics customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.
- Fig. 3.3 shows a 2D (two-Dimensional) plot of customers with respect to customer locations in a city. Three clusters of data points are evident.

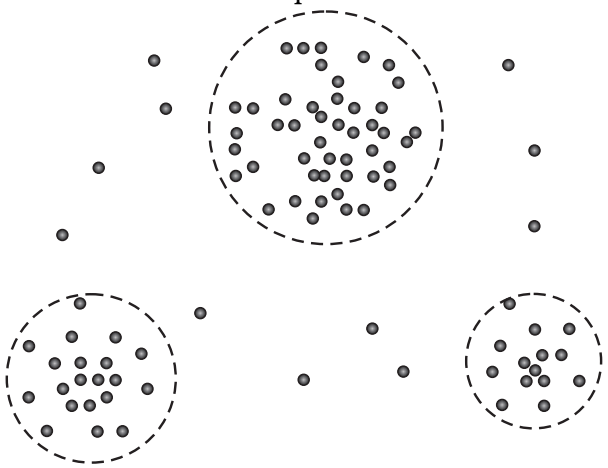


Fig. 3.3: A 2D Plot of Customer Data with respect to Customer Locations in a City, showing Three Data Clusters

3.1.5 Outlier Analysis

- Outlier detection also known as anomaly detection. Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies.
- Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance and intrusion detection.
- Outlier analysis is the process of identifying outliers or abnormal observations in a dataset.
- Outlier analysis is a process that involves identifying the anomalous observation in the dataset.
- Outliers may be defined as, the data objects that do not comply with the general behavior or model of the data available.

- However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier analysis or anomaly mining.
- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers.
- Rather than using statistical or distance measures, density-based methods may identify outliers in a local region, although they look normal from a global statistical distribution view.
- **For example,** outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the locations and types of purchase or the purchase frequency.

3.2 MINING FREQUENT PATTERNS

- Frequent pattern mining searches for recurring relationships in a given data set. It is a data mining technique with the objective of extracting frequent itemsets from a data set.
- Due to increase used of digital and Internet activities, large amounts of data continuously being collected and stored.
- Industries/Organizations are interested to mined frequent itemsets from this data which leads to the detection of associations and correlations among items in large transactional data sets.
- The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes such as catalog design, cross-marketing, and customer shopping behavior analysis.
- A typical example of frequent itemset mining is market basket analysis, market basket analysis process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”.
- The discovery of these associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.
- Association analysis is one of the functionality of data mining which discovers the association relationships among huge amounts of data.
- Association analysis measures the strength of co-occurrence between one item and another.
- Association rules suggest a strong relationship that exists between two items, frequent items are a collection of items that frequently occur together.

- Association rule mining finds interesting associations and/or correlation relationships among large sets of such data items. Association rules show attribute value conditions that occur frequently together in a given dataset.

  $A \rightarrow B$, where A and B are data sets.
- It means that the transaction which tends to contain A also tends to contain B.

3.2.1 Market Basket Analysis

- Frequent pattern mining is also called as association rule mining.
- Association rule mining is an analytical process that finds frequent patterns, associations or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases and other data repositories.
- Association rule mining searches for interesting relationships among items in a given dataset. The strengths of association rule analysis are:
  1. It produces clear and understandable results.
  2. It supports undirected data mining.
  3. It works on variable-length data.
  4. The computations it uses are simple to understandable.
- Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.
- Market basket analysis involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.
- Market basket analysis is a process that looks for relationships among entities and objects that frequently appear together, such as the collection of items in a shopper's cart.
- Suppose, a marketing manager of Electronic shop, would like to determine which items are frequently purchased together within the same transactions.
- An example of such a rule,

  buys (X; "computer") $\rightarrow$ buys(X; "Antivirus software") [support = 1%; confidence = 50%]

  where, X is a variable representing a customer. Here, support and confidence are two measures of rule interestingness.
- A Confidence, or Certainty, of 50% means that if a customer buys a computer, there is a 50% chance that customer will buy antivirus software as well.
- A 1% Support means that 1% of all of the transactions under analysis showed that computer and antivirus software were purchased together. This association rule involves a single attribute or predicate (i.e., buys) that repeats.
- Association rules that contain a single predicate are referred to as single-dimensional association rules. Dropping the predicate notation, the above rule can be written simply as

  "Computer $\rightarrow$ antivirus software [1%, 50%]".

- Suppose, that we are given the Electronic store relational database relating to purchases. A data mining system may find association rules like,

  age (X, " ")^income(X, "20K:::29K") →buys(X, "Laptop")

  [support = 2%, confidence = 60%]
- The rule indicates that of the customers under study, 2% are 20 to29 years of age with an income of 20,000 to 29,000 and have purchased a laptop at Electronic store.
- There is a 60% probability that a customer in this age and income group will purchase a laptop.
- Note that there is an association between more than one attribute or predicate (i.e., age, income, and buys).
- Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a multidimensional association rule.
- Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.
- Additional analysis can be performed to uncover interesting statistical correlations between associated attribute-value pairs.

Association Rule Metrics/Measures:

- Various metrics are in place to help us understand the strength of association between these two itemsets.

1. **Support:** The support of a rule $x \rightarrow y$ (where x and y are each items/events etc.) is defined as the proportion of transactions in the data set which contain the item set x as well as y. So,

$$\text{Support } (x \rightarrow y) = \frac{\text{Number of transactions which contain the item set x and y}}{\text{Total number of transactions}}$$

Whereas,

$$\text{Support } (x) = \frac{\text{Number of transactions which contain the item set x}}{\text{Total number of transactions}}$$

2. **Confidence:** The confidence of a rule $x \rightarrow y$ is defined as:

$$\text{Confidence } (x \rightarrow y) = \frac{\text{Support } (x \rightarrow y)}{\text{Support } (x)}$$

- So, it is the ratio of the number of transactions that include all items in the consequent (y in this case), as well as the antecedent (x in this case) to the number of transactions that include all items in the antecedent (x in this case).
- In the table below, Support (milk → bread) = 0.4 means milk and bread are purchased together occur in 40% of all transactions. Confidence (milk → bread) = 0.5 means that if there are 100 transactions containing milk then there will be 50 that will also contain bread.

TID	Milk	Bread	Butter	Jam
1	1	0	1	1
2	1	1	1	0
3	0	1	1	0
4	1	0	0	1
5	1	1	1	1

Example: The following database has four transactions:

T1 {K, A, D, B}
T2 {D, A C, E, B}
T3 {C, A, B, E}
T4 {B, A, D}

Find the support and confidence for following rules:

(i) A -> D
(ii) D -> A
(iii) B -> D
(iv) D -> B
(v) AB -> D
(vi) D -> AB
(vii) AD -> B
(viii) B - > AD
(ix) BD -> A
(x) A -> BD

Solution:

(i) $A \rightarrow D$

Support = $\frac{3}{4}$ = 75%

Confidence = $\frac{3}{4}$ = 75%

(ii) $D \rightarrow A$

Support = $\frac{3}{4}$ = 75%

Confidence = $\frac{3/4}{3/4}$ = 100%

(iii) $B \rightarrow D$

Support = $\frac{3}{4}$ = 75%

Confidence = $\frac{3/4}{4/4}$ = 75%

(iv) $D \rightarrow B$

Support = $\frac{3}{4}$ = 75%

Confidence = $\frac{3/4}{3/4}$ = 100%

(v) $AB \rightarrow D$

$$\text{Support} = \frac{3}{4} = 75\%$$

$$\text{Confidence} = \frac{3/4}{4/4} = 75\%$$

(vi) $D \rightarrow AB$

$$\text{Support} = \frac{3}{4} = 75\%$$

$$\text{Confidence} = \frac{3/4}{3/4} = 100\%$$

(vii) $AD \rightarrow B$

$$\text{Support} = \frac{3}{4} = 75\%$$

$$\text{Confidence} = \frac{3/4}{3/4} = 100\%$$

(viii) $B \rightarrow AD$

$$\text{Support} = \frac{3}{4} = 75\%$$

$$\text{Confidence} = \frac{3/4}{4/4} = 75\%$$

(ix) $BD \rightarrow A$

$$\text{Support} = \frac{3}{4} = 75\%$$

$$\text{Confidence} = \frac{3/4}{3/4} = 100\%$$

(x) $A \rightarrow BD$

$$\text{Support} = \frac{3}{4} = 75\%$$

$$\text{Confidence} = \frac{3/4}{4/4} = 100\%$$

Lift:

- Lift measures dependency between X and Y.

$$\text{Lift}(x \rightarrow y) = \frac{\text{Support}(x \rightarrow y)}{\text{Support}(x) \times \text{Support}(y)} = \frac{\text{Confidence}(x \rightarrow y)}{\text{Support}(y)}$$

- The numerator for lift is the proportion of transactions where x and y occur jointly. The denominator is an estimate of the expected joint occurrence of x and y, assuming that they occur independently.
- A lift value of 1 indicates that x and y jointly occur in transactions with the frequency that would be expected by chance alone.
- The preferred values are much larger or smaller than 1, whereby values less than 1 mean x and y are negatively correlated and greater values indicate positive correlation.
- Lift is vulnerable to noise in small datasets, because infrequent itemset can have very high lift values.

3.3 FREQUENT ITEMSET, CLOSED ITEMSET AND ASSOCIATION RULE

- A frequent itemset typically refers to a set of items that frequently appear together in a transactional data set.
- Maximal Frequent Itemset (MFI) is a frequent item set for which none of its immediate supersets are frequent.
- Closed Frequent Itemset (CFI) is a frequent item set for which none of its immediate supersets has the same support as that of the itemset.
- It is important to point out the relationship between frequent itemsets, closed frequent itemsets and maximal frequent itemsets.
- As mentioned earlier closed and maximal frequent itemsets are subsets of frequent itemsets but maximal frequent itemsets are a more compact representation because it is a subset of closed frequent itemsets.
- The Fig. 3.4 shows the relationship between these three types of itemsets.

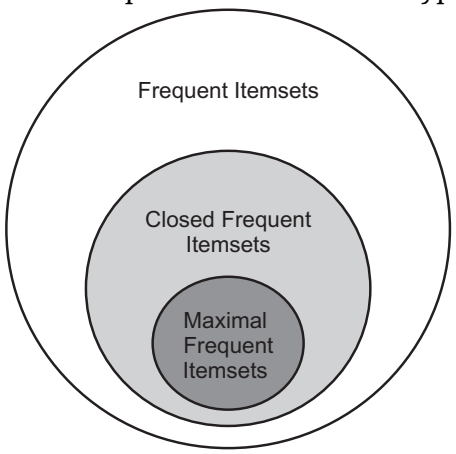


Fig. 3.4

- Closed frequent itemsets are more widely used than maximal frequent itemset because when efficiency is more important that space, they provide us with the support of the subsets so no additional pass is needed to find this information.
- The analytical process that finds frequent itemset and associations from data sets is called frequent pattern mining or association rule mining.
- Let us see an example to understand the concept of frequent itemset, closed itemset and maximal itemset. Suppose we have four documents in corpus CS= {D1, D2, D3, D4} also documents D1, D2, D3 and D4 are represented by set of terms as depicts in Table 3.1. The Table 3.2 shows all itemsets generated from corpus with their support.
- In given example if 50% is the threshold value (minimum support) for the item set to be frequent. The Table 3.3 shows list of frequent itemsets for corpus C.

Table 3.1: Documents Dataset

Did	Terms in Documents
D1	{APPLE,BEETROOT,COCONUT,DOCTOR}
D2	{APPLE,DOCTOR}
D3	{APPLE,FRUITS}
D4	{COCONUT,FRUITS}

Table 3.2: Itemsets with Support

{APPLE}-3	{COCONUT,DOCTOR}-1
{BEETROOT}-1	{COCONUT,FRUITS}-1
{COCONUT}-2	{DOCTOR,FRUITS}-0
{DOCTOR}-2	{APPLE,BEETROOT,COCONUT}-1
{FRUITS}-2	{APPLE,BEETROOT,DOCTOR}-1
{APPLE,BEETROOT}-1	{APPLE,BEETROOT,FRUITS}-0
{APPLE,COCONUT}-1	{BEETROOT,COCONUT,DOCTOR}-1
{APPLE,DOCTOR}-2	{BEETROOT,COCONUT,FRUITS}-0
{APPLE,FRUITS}-1	{COCONUT,DOCTOR,FRUITS}-0
{BEETROOT,COCONUT}-1	{APPLE,BEETROOT,COCONUT,DOCTOR}-1
{BEETROOT,DOCTOR}-0	{APPLE,BEETROOT,COCONUT,FRUITS}-0
{BEETROOT,FRUITS}-0	{BEETROOT,COCONUT,DOCTOR,FRUITS}-0
{APPLE,BEETROOT,COCONUT,DOCTOR,FRUITS}-0	

Table 3.3: Frequent Itemsets

Itemset	Support
{APPLE}-3	75%
{COCONUT}-2	50%
{DOCTOR}-2	50%
{FRUITS}-2	50%
{APPLE,DOCTOR}-2	50%

- In this example {APPLE}, {COCONUT}, {DOCTOR},{FRUITS},{APPLE,DOCTOR} are frequent itemsets. In above case {APPLE, DOCTOR}, {COCONUT}, {FRUITS} are maximal frequent itemsets.
- The {APPLE, DOCTOR} is a maximal frequent item set, because it is the immediate superset of frequent item set {APPLE} and is frequent.
- As {APPLE} and {DOCTOR} are frequent but they have superset (APPLE, DOCTOR}, which is frequent. So, {APPLE} and {DOCTOR} are not considered as MFI.

Example 1: If total no. of items is 7 then how many possible itemsets can be generated?

Solution: Total no. of itemsets which can be generated using 7 items are $2^7 = 128$.

Example 2: Consider the following Database:

Customer	Items
1	Orange Juice, Soda
2	Milk, Orange Juice, Window Cleaner
3	Orange Juice, Detergent
4	Orange Juice, Detergent, Soda
5	Window Cleaner, Soda

If min_support count is 2 then identify the patterns which are frequent.

Solution: As minimum support is 2, so patterns which are frequent are:

(i) Orange Juice, Soda

(ii) Orange Juice, Detergent.

Example 3: Consider the given dataset "Shopping" with 6 transactions, each transaction consist of items which were purchase together.

Transaction ID	Items
1	A,B,C,E
2	A,C,D,E
3	B,C,E
4	A,C,D,E
5	C,D,E
6	A,D,E

- Let us say min_support=0.5. This is fulfilled if min_support_count>= 3.
  - Frequent item set $X \in FX \in F$ is maximal if it does not have any frequent supersets.
  - Frequent item set $X \in F$ is closed if it has no superset with the same frequency.

Find Frequent Item Set:

- A frequent itemset is simply a set of items occurring a certain percentage of the time. In a dataset, a itemset is considered frequent if its frequency/occurrence is equal or more than on min_support % or min_support_count.

- Following Table 3.4 is showing the all frequent item set on shopping dataset.

Table 3.4: Frequent Itemsets

Itemsets	Frequency/Occurrence	Decision
1-itemset	{A} = 4	Frequent.
	{B} = 2	Not frequent => ignore.
	{C} = 5	Frequent.
	{D} = 4	Frequent.
	{E} = 6	Frequent.
2-itemset	{A,B} = 1	Not frequent => ignore.
	{A,C} = 3	Frequent.
	{A,D} = 3	Frequent.
	{A,E} = 4	Frequent.
	{B,C} = 2	Not frequent => ignore.
	{B,D} = 0	Not frequent => ignore.
	{B,E} = 2	Not frequent => ignore.
	{C,D} = 3	Frequent.
	{C,E} = 5	Frequent.
	{D,E} = 4	Frequent.
Itemsets	**Frequency/Occurrence**	**Decision**
3-itemset	{A,B,C} = 1	Not frequent => ignore.
	{A,B,D} = 0	Not frequent => ignore.
	{A,B,E} = 1	Not frequent => ignore.
	{A,C,D} = 2	Not frequent => ignore.
	{A,C,E} = 3	Frequent
	{A,D,E} = 3	Frequent
	{B,C,D} = 0	Not frequent => ignore
	{B,C,E} = 2	Not frequent => ignore
	{C,D,E} = 3	Frequent
4-itemset	{A,B,C,D} = 0	Not frequent => ignore
	{A,B,C,E} = 1	Not frequent => ignore
	{B,C,D,E} = 0	Not frequent => ignore

Finding Closed Itemset:

- From Table 3.4 consider all the item sets which are frequent and then check for the property of closed itemset i.e., Frequent item set $X \in F$ is closed if it has no superset with the same frequency.
- Following Table 3.5 is showing the all closed frequent item set on shopping dataset.

Table 3.5: Closed Itemsets

Item sets	Frequency/Occurrence	Closed Decision
1-itemset	{A} = 4	Not closed due to {A,E}as having same frequency.
	{C} = 5	Not closed due to {C,E} as having same frequency.
	{D} = 4	Not closed due to {D,E} as having same frequency.
	{E} = 6	Closed.
2-itemset	{A,C} = 3	Not closed due to {A,C,E} as having same frequency.
	{A,D} = 3	Not closed due to {A,D,E} as having same frequency.
	{A,E} = 4	Closed.
	{C,D} = 3	Not closed due to {C,D,E} as having same frequency.
	{C,E} = 5	Closed.
	{D,E} = 4	Closed.
3-itemset	{A,C,E} = 3	Not closed as no superset with the same frequency.
	{A,D,E} = 3	Not closed as no superset with the same frequency.
	{C,D,E} = 3	Not closed as no superset with the same frequency.

Finding Maximal Itemset

- From Table 3.4 consider all the item sets which are frequent and then check for the property of maximal itemset i.e., frequent item set $X \in FX \in F$ is maximal if it does not have any frequent supersets.

- Following Table 3.6 is showing the all maximal frequent item set on shopping Dataset.

Table 3.6: Maximal Itemsets

Item Sets	Frequency/ Occurrence	Decision
1-itemset	{A} = 4	Not Maximal due to its frequent superset.
	{C} = 5	Not Maximal frequent due to its frequent superset.
	{D} = 4	Not Maximal frequent due to its frequent superset.
	{E} = 6	Not Maximal frequent due to its frequent superset.
2-itemset	{A,C} = 3	Not Maximal frequent due to its frequent superset.
	{A,D} = 3	Not Maximal frequent due to its frequent superset.
	{A,E} = 4	Not Maximal frequent due to its frequent superset.
	{C,D} = 3	Not Maximal frequent due to its frequent superset.
	{C,E} = 5	Not Maximal frequent due to its frequent superset.
	{D,E} = 4	Not Maximal frequent due to its frequent superset.
3-itemset	{A,C,E} = 3	Maximal frequent.
	{A,D,E} = 3	Maximal Frequent.
	{C,D,E} = 3	Maximal Frequent.

- Association rules are the statements which are used to find the relationships between unrelated data in a database. These rules are useful for analyzing and predicting customer behavior.
- Support and Confidence are two methods for generating the association rules. The frequent items are calculated by support threshold and confidence threshold. The values for these thresholds are predefined by the users.
- Mining association rules consist of following two-step approach:
  1. **Frequent Itemset Generation:** Generate all itemsets whose support ≥ minsup.
  2. **Rule Generation:** Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset.

3.4 FREQUENT ITEMSET MINING METHODS

- Several algorithms for generating rules have been used in the research work carried out by various authors.
- Here, we will study the Apriori Algorithm and FP Growth algorithm for generating the frequent itemsets. Apriori algorithm finds interesting association along with a huge set of data items.
- The association rule mining problem was firstly given by Apriori algorithm by Rakesh Agrawal and Ramakrishnan Srikant in 1994 in the following manner:

- Let I = {I_1,I_2.....I_m}, m ∈ I be a set of m different attributes, T be the transaction that covers a set of items and D be a database. A rule is a relationship in the form of X=>Y, where X and Y are called itemsets.
- For example, it could be functional for the video store manager to recognize what movies are frequently borrowed together or if there is a correlation between renting a certain type of movies or buying popcorn.

3.4.1 Apriori Algorithm

- Apriori algorithm is widely used algorithm for generate frequent itemsets. Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994.
- The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. It is a classic algorithm for learning association rules.
- Apriori algorithm is easy to execute and very simple is used to mine all frequent itemsets in database.
- The algorithm makes many searches in database to find frequent itemsets where k itemsets are used to generate k + 1 itemsets. Each k itemset must be greater than or equal to minimum support threshold to be frequency.
- Otherwise, it is called candidate itemsets. In the first, the algorithm scan database to find frequency of 1 itemsets that contains only one item by counting each item in database.
- The frequency of 1 itemsets is used to find the itemsets in 2 itemsets which in turn is used to find 3 itemsets and so on until there are not any more k itemsets.
- If an itemset is not frequent, any large subset from it is also non-frequent; this condition prune from search space in database.
- Apriori algorithm uses Apriori property which specifies that all subsets of a frequent itemset must also be frequent.
- Apriori algorithm is based on the antimonotonicity property, which means If any set is not satisfied any test conditions then all of its superset will also not satisfied the same test conditions.

Apriori Property:

- If an itemset is infrequent, then all its supersets must also be infrequent. So, according to Apriori property if {A} is infrequent item set then all its superset like {A,B} ,{A,C} or {A,B,C} etc., will also be infrequent.
- It is called antimonotonicity because the property is monotonic in the context of failing a test.
- Apriori algorithm used step wise search approach, it means k-itemsets are used to discover (k+1) itemsets.

- Initially Transactional Dataset (also called as 'Corpus') is scanned to uncover the set of 1-itemsets (singleton sets).1-itemsets which do not satisfied the minimum support condition are removed.
- The consequential set is represented by S_1. Now S_1 is utilized to uncover S_2, the set of frequent 2-itemsets, which is utilized to uncover S_3, and so on.
- The process will be continued until no new k-itemsets can be determined.

Algorithm for Apriori:

Input: D, database.

min_sup: Minimum Support Threshold.

Output: L, Frequent itemsets in database.

1. Scan the database to determine the support of each one-itemset, compare it with min_sup and get the frequent one-itemset, (L_1).
2. Use L_{K-1} property, Join L_{K-1} to find the candidate k itemset.
3. Scan the database to find the support of each candidate k-itemset, compare it with min_sup and get the frequent K-itemset.
4. Repeat the Steps 2 to 3 until candidate itemset is null. If null, generate all subsets for each frequent itemset.

Example 1: Consider, total number of transactions is 15 and Min Support = 20%.

TID	List of Items
1	A1, A5, A6, A8
2	A2, A4, A8
3	A4, A5, A7
4	A2, A3
5	A5, A6, A7
6	A2, A3, A4
7	A2, A6, A7, A9
8	A5
9	A8
10	A3, A5, A7
11	A3, A5, A7
12	A5, A6, A8
13	A2, A4, A6, A7
14	A1, A3, A5, A7
15	A2, A3, A9

Calculate Minimum Support:

- First should calculate the minimum SupportCount/threshold. Question says minimum support should be 20%. It calculate as follows:
  Minimum support count(20/100 × 15) = 3.
  Scan 1: Calculate SupportCount for frequency itmset (C_1)

Itemset	SupportCount
A1	2
A2	6
A3	6
A4	4
A5	8
A6	5
A7	7
A8	4
A9	2

Prune 1: Prune C_1 by comparing with min, SupportCount i.e. 3 and keep those items in L_1 which > 3 support count.

Itemset	SupportCount
A2	6
A3	6
A4	4
A5	8
A6	5
A7	7
A8	4

Scan 2: Calculate SupportCount for frequent two itemset (C_2) using L_1.

Itemset	SupportCount	Itemset	Support Count
{A2, A3}	3	{A4, A5}	1
{A2, A4}	3	{A4, A6}	1
{A2, A5}	0	{A4, A7}	2
{A2, A6}	2	{A4, A8}	1
{A2, A7}	2	{A5, A6}	3
{A2, A8}	1	{A5, A7}	5
{A3, A4}	1	{A5, A8}	2
{A3, A5}	3	{A6, A7}	3
{A3, A6}	0	{A6, A8}	2
{A3, A7}	3	{A7, A8}	0
{A3, A8}	0		

Prune 2: Prune C_2 by comparing with min, SupportCount i.e., 3. and keep those items in L_2 which > 3 SupportCount.

Itemset	SupportCount
{A2, A3}	3
{A2, A4}	3
{A3, A5}	3
{A3, A7}	3
{A5, A6}	3
{A5, A7}	5
{A6, A7}	3

Scan 3: Calculate SupportCount for frequent three itmeset (C_3) using L_2.

Itemset	SupportCount
{A2, A3, A4}	1
{A3, A5, A7}	3
{A5, A6, A7}	1

Prune 3: Prune C_3 by comparing with min, SupportCount i.e. three itemset (C_3) in L_3 which > 3 SupportCount.

Itemset	SupportCount
{A3, A5, A7}	3

At the end {A3, A5, A7} represent a frequent itemset in transaction.

Example 2: Trace the results of Apriori algorithm for following transaction with minimum support threshold 3.

TID	Items Brought
1	M, T, B
2	E, T, C
3	M, E, T, C
4	E, C
5	J

From this calculate the frequency of all items,

Items Brought	Support
M	2
E	3
T	3
C	3
J	1
B	1

Discard the items with minimum support less than 3.

Items Brought	Support
E	3
T	3
C	3

Combine two items.

Items Brought
E, T
E, C
T, C

Calculate frequency of all items.

Items Brought	Support
E, T	2
E, C	3
T, C	2

Discard the items with minimum support less than 3.

Items Brought	Support
E, C	3

Result: Only one items {E, C} have minimum support 3.

Program for Apriori algorithm using Python programming language.

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
#creating dataset
   dataset = [
               ["Milk", "Eggs", "Bread"],
               ["Milk", "Eggs"],
               ["Milk", "Bread"],
               ["Eggs", "Apple"],
            ]
#: Convert list to dataframe with boolean values
te = TransactionEncoder()
te_array = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_array, columns=te.columns_)
```

```
#result after preprocessing
print("Result after preprocessing:")
print(df)
#Find frequently occurring itemsets using Apriori Algorithm
frequent_itemsets_ap = apriori(df, min_support=0.01, use_colnames=True)
#result after  applying apriori algorithm
print("\n Result  after  apriori algorithm ")
print(frequent_itemsets_ap)
#Mine the Association Rules
rules_ap = association_rules(frequent_itemsets_ap, metric="confidence",
                                                          min_threshold=0.8)
frequent_itemsets_ap['length'] =
                frequent_itemsets_ap['itemsets'].apply(lambda x: len(x))
print("\n Frequent 2  Item sets")
print(frequent_itemsets_ap[frequent_itemsets_ap['length'] >= 2])
print("\n  Frequent 3  Item sets")
print(frequent_itemsets_ap[frequent_itemsets_ap['length'] >= 3])
```

Output:

```
Result after preprocessing:
       Apple  Bread   Eggs   Milk
   0  False   True   True   True
   1  False  False   True   True
   2  False   True  False   True
   3   True  False   True  False
Result  after  apriori algorithm
       support             itemsets
   0     0.25              (Apple)
   1     0.50              (Bread)
   2     0.75               (Eggs)
   3     0.75               (Milk)
   4     0.25        (Eggs, Apple)
   5     0.25        (Eggs, Bread)
   6     0.50        (Milk, Bread)
   7     0.50         (Eggs, Milk)
   8     0.25  (Eggs, Milk, Bread)
```

```
Frequent 2  Item sets
      support           itemsets       length
   4     0.25        (Eggs, Apple)        2
   5     0.25        (Eggs, Bread)        2
   6     0.50        (Milk, Bread)        2
   7     0.50         (Eggs, Milk)        2
   8     0.25  (Eggs, Milk, Bread)        3
Frequent 3  Item sets
      support             itemsets     length
   8     0.25   (Eggs, Milk, Bread)        3
```

3.5 GENERATING ASSOCIATION RULE FROM FREQUENT ITEMSET

- Once, the frequent itemsets are generated, identifying rules out of them is comparatively very easy. Rules are formed by binary partition of each itemset.
- Since, only relevant rules are of interest, the question arises how to measure this relevance. The most common way is by using confidence.
- The confidence of a rule, $X \Rightarrow Y$, is known as the conditional probability of a transaction containing Y given that it contains X. This can be calculated according to this formula: Confidence $(X \Rightarrow Y) = Sup(X \cup Y)/Sup(X)$.
- Given two disjoint itemsets X and Y. An implication expression of the form $X \Rightarrow Y$ is referred to as an association rule with a minimum support level minsup and a minimum confidence level minconf, if both following criteria are satisfied:
  1. $Sup(X \cup Y) \geq$ minsup
  2. $Conf(X \Rightarrow Y) \geq$ minconf
- Since, only rules of certain relevance are of interest, similar to minsup in frequent itemset mining, a specific minimum confidence threshold is needed to generate only the most relevant rules.
- For generation of any rule, first thing is to determine the frequent itemset. Given a frequent itemset L, find all non-empty subsets $f \in L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
- If {A,B,C,D} is a frequent itemset, then possible candidate rules are:

  $ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$, $A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$

  $AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$, $BD \rightarrow AC$, $CD \rightarrow AB$.
- If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \phi$ and $\phi \rightarrow L$). In above case $|L|=4$ then there are $2^4 - 2 = 14$ subsets.

- In general, confidence does not have an anti-monotone property (as already explained in Apriori algorithm) c(ABC →D) can be larger or smaller than c(AB →D).
- But confidence of rules generated from the same itemset have an anti-monotone property for example,

  L = {A, B, C, D} : c(ABC → D) ≥ c(AB → CD) ≥ c(A → BCD)
- If we will follow the antimonotonic property then we can prune most of the low confidence rules.
- We start with a frequent itemset {A,B,C,D} and start forming rules with just one consequent. Remove the rules failing to satisfy the minconf condition.
- Now, start forming rules using a combination of consequents from the remaining ones. Keep repeating until only one item is left on antecedent. This process has to be done for all frequent itemsets.
- In Apriori algorithm, at the end of example, we get a frequent itemset X = {A3, A5, A7} but the task of generation of rule is not still achieved.
- Association rules can be generated from X. The non-empty subsets of X are {A3, A5, A7}, {A3, A5}, {A5, A7}, {A3, A7}, {A3}, {A5} and {A7}. The resulting association rules are as shown below, each listed with its confidence.

Association Rules	**Confidence** = $\frac{\textbf{Sub (X} \cup \textbf{Y)}}{\textbf{Sup (X)}}$
{A3, A5} → A7	3/3 = 100%
{A3, A7} → A5	3/3 = 100%
{A5, A7} → A3	3/5 = 60%
A3 → {A5, A7}	3/6 = 50%
A5 → {A3, A7}	3/8 = 37.5%
A7 → {A3, A5}	3/7 = 42.8%

- Now as per confidence value we can select the rule from frequent time set. For example, confidence threshold is 75 then two rules will be generated from frequent itemset {A3, A5, A7}.

  {A3, A5} → A7

  {A3, A7} → A5

3.6 IMPROVING THE EFFICIENCY OF APRIORI

- Since, the amount of the processed data in mining frequent itemsets tends to be huge or massive, it is important to devise efficient algorithms to mine such data.
- The basic Apriori algorithm scans the database several times, depending on the size of the largest frequent itemset.

- Since, Apriori algorithm was first introduced and as experience was accumulated, there have been many attempts to devise more efficient algorithms of frequent itemset mining including approaches such as hash-based technique, partitioning, sampling, and using vertical data format.
- Several refinements have been proposed that focus on reducing the number of database scans, the number of candidate itemsets counted in each scan or both.
- "How can we further improve the efficiency of Apriori-based mining?" Many variations of the Apriori algorithm have been proposed that focus on improving the efficiency of the original algorithm.
- The Apriori algorithm is the most classic association rule mining algorithm. Association rule mining leads to the discovery of associations and correlations among items in large transactional or relational data sets and the discovery of interesting and valuable correlation relationship among huge amounts of transaction records.
- The objective of association rule mining is to find the rule that the occurrence of one event can lead to another incident happened.
- Many variations of Apriori algorithm have been proposed that focus on improving the efficiency of the original algorithm.
- Some improvements techniques in Apriori algorithm are given below:
  1. **Hash-based technique (hashing itemsets into corresponding buckets):** A hash-based technique can be used to reduce the size of the candidate k-itemsets, C_k, for $k > 1$.
  2. **Transaction reduction (reducing the number of transactions scanned in future iterations):** The goal of transaction reduction technique is to reduce the number of transactions in future that do not have any frequent itemset. A transaction

     that does not contain any frequent k-itemsets cannot contain any frequent $(k + C_1)$-itemsets. Therefore, such a transaction can be marked or removed from further consideration because subsequent database scans for j-itemsets, where $j > k$, will not need to consider such a transaction.
  3. **Partitioning (partitioning the data to find candidate itemsets):** A partitioning technique can be used that requires just two database scans to mine the frequent itemsets. In data set partitioning technique, data or a set of transactions is partitioned into smaller segments for the purpose of finding candidate itemsets.
  4. **Sampling (mining on a subset of the given data):** Sampling is a technique is important when efficiency is most important than accuracy. It is based on the mining on a subset of the given data. 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. In this way, we trade off some degree of accuracy against

efficiency. The S sample size is such that the search for frequent itemsets in S can be done in main memory, and so only one scan of the transactions in S is required overall. Because we are searching for frequent itemsets in S rather than in D, it is possible that we will miss some of the global frequent itemsets.

5. **Dynamic itemset counting (adding candidate itemsets at different points during a scan):** The dynamic itemset counting technique during scanning, candidate itemset would be added at different start point, if all their subsets are estimated. A dynamic itemset counting technique was proposed in which the database is partitioned into blocks marked by start points. In this variation, new candidate itemsets can be added at any start point, unlike in Apriori, which determines new candidate itemsets only immediately before each complete database scan. The technique uses the count-so-far as the lower bound of the actual count. If the count-so-far passes the minimum support, the itemset is added into the frequent itemset collection and can be used to generate longer candidates. This leads to fewer database scans than with Apriori for finding all the frequent itemsets.

- Although above improvements were used to improve the efficiency of Apriori algorithm, reduce the size of candidate itemsets and lead to good performance gain, still they have following two limitations:
  1. Difficult to handle a large number of itemsets.
  2. It is tedious to repeatedly scan the datasets and check a large dataset of candidates by pattern matching.

3.7 FREQUENT PATTERN GROWTH (FP-GROWTH) ALGORITHM

- Number of improvements techniques such as Partitioning, Hash-based technique, Transaction reduction, Sampling, Dynamic itemset counting etc. were used to improve the efficiency of Apriori algorithm, which lead to good performance gain and reduce the size of candidate item sets, but still Apriori algorithm have following two major limitations:
  1. Using Apriori needs a generation of candidate itemsets. These itemsets may be large in number if the itemset in the database is huge.
  2. Apriori needs multiple scans of the database to check the support of each itemset generated and this leads to high costs.
- The above shortcomings of Apriori algorithm can be overcome using the FP (Frequent Pattern)-growth algorithm.
- FP-growth algorithm adopts a divide-and-conquer strategy as follows. First, it compresses the database representing frequent items into a frequent pattern tree, or FP-tree, which retains the itemset association information.

- It then divides the compressed database into a set of conditional databases (a special kind of projected database), each associated with one frequent item or "pattern fragment" and mines each database separately.
- For each "pattern fragment," only its associated data sets need to be examined. Therefore, this approach may substantially reduce the size of the data sets to be searched, along with the "growth" of patterns being examined.
- The FP-growth algorithm is used to mine the complete set of frequent itemsets. It creates FP tree to compress a large dataset.
- In FP tree nodes, frequent items are arranged in such a manner that more frequently occurring nodes have better chances of sharing nodes than the less frequently occurring ones.
- The FP algorithm uses divide and conquer method to decompose mining task after that avoid candidate generation by consider sub-database only.
- The FP-growth method performance shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm.
- FP-growth algorithm preserves whole information for frequent pattern mining. FP-growth algorithm is as follows:

1. **Construct Condition Pattern Base for Each Node in the FP-Tree:**
   (i) Starting at the frequent header table in the FP-tree.
   (ii) Traverse the FP-tree by following the link of each frequency item.
   (iii) Accumulate all of transformed prefix paths of that item to form a conditional pattern base.
2. **Construct Conditional FP-Tree from each Conditional Pattern-Base:**
   (i) Accumulate the count for each item in the base.
   (ii) Construct the FP-tree for the frequency items of the pattern base.
3. **Recursively Mine Conditional FP-trees and Grow Frequency Patterns obtained so Far:**
   (i) If the conditional FP-tree contains a single path, simply enumerate all the patterns.

Example 1: Consider following table,

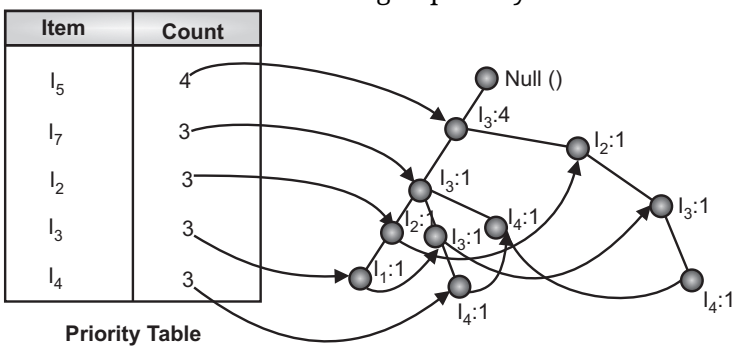
Database D

Transaction ID	Item
T_1	I_1, I_2, I_3, I_5
T_2	I_2, I_3, I_4, I_5
T_3	I_1, I_4, I_5
T_4	I_1, I_2, I_3, I_4, I_5

Scan Database →

Item	Count
I_1	3
I_2	3
I_3	3
I_4	3
I_5	5

After that, all the itemset arrange in a sequential order or give the priority and construct FP-tree. And new consider the items as a suffix for the database D and arrange the items or Transaction according to priority.



After that FP-tree of priority table, insert this priority table except item I_5 because it contains higher priority and mining the above FP-tree as summarized in Table 3.7.

Table 3.7: FP-Tree of Priority Table

Item	Conditional Pattern Base FP-Tree	Conditional Generated	Frequent Pattern
I_4	(I_1, I_2, I_3 : 1), (I_2, I_3 : 1), (I_1 : 1)	(I_2, I_3 : 2), (I_1 : 2)	(I_2, I_3, I_4 : 1) (I_1, I_4 : 2)
I_3	(I_1, I_2 : 2) (I_2 : 1)	(I_1, I_2 : 2), (I_2 : 3)	(I_1, I_2, I_3 : 3) (I_2, I_3) : 3
I_2	(I_1 : 2)	(I_1 : 2)	(I_1, I_2 : 2)
I_1	(I_5 : 3)	(I_5 : 3)	(I_5, I_1 : 3)

Example 2: Generate the FP tree for following transaction dataset with minimum support 3.

TID	Item Brought
100	F, A, C, D, G, L, M, P
200	A, B, C, F, L, M, O
300	B, F, H, I, O
400	B, C, K, S, P
500	A, F, C, E, L, P, M, N

Solution: First calculate the frequency of each item.

Item	Frequency
A	3
B	3
C	4
D	1
E	1
F	4
G	1
H	1
I	1

Item	Frequency
J	1
K	1
L	2
M	3
N	1
O	2
P	3
S	1

Now frequent pattern set (L) is built which will contain all elements whose frequency is greater than or equal or 3.

L = {(F : 4), (C : 4), (A : 3), (M : 3), (P : 3)}

Next we will build ordered frequent items.

Item	Frequency	Item
100	F, A, C, D, G, L, M, P	F, C, A, M, P
200	A, B, C, F, L, M, O	F, C, A, B, M
300	B, F, H, J, O	F, B
400	B, C, K, S, P	C, B, P
500	A, F, C, E, L, P, M, N	F, C, A, M, P

Now we will build data structure for first transaction.

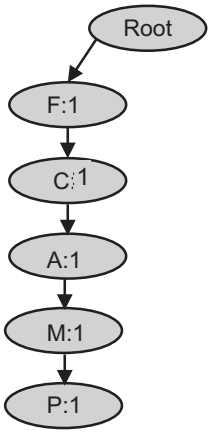


Fig. 3.5

For next transaction we will increase the frequency of present items and add the present items in data structure.

So for TID - 200.

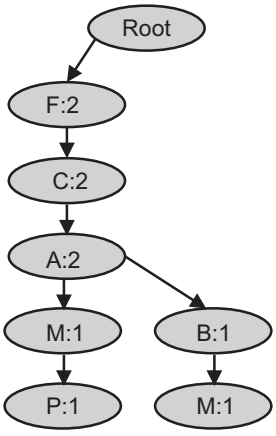


Fig. 3.6

Like this we will increase the frequency count of each item for next transactions. After all transactions final tree will be shown in Fig. 3.7.

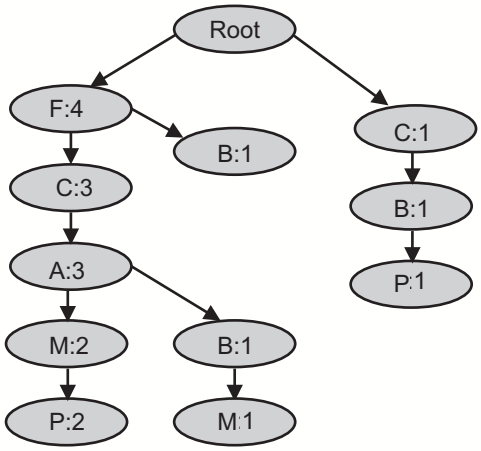


Fig. 3.7

From this we will build conditional pattern base starting from bottom to top.

Item	Conditional Pattern Base
P	{{F, C, A, M : 2}, {C, B : 1}}
M	{{F, C, A : 2}, {F, C, A, B : 1}}
B	{{F, C, A : 1}, {F : 1}, {C : 1}}
A	{{F, C : 3}}
C	{{F : 3}}
F	ϕ

Now built conditional frequency pattern tree,

Item	Conditional Pattern Base	Conditional FP-Tree
P	{{F, C, A, M : 2}, {C, B : 1}}	{C : 3}
M	{{F, C, A : 2}, {F, C, A, B : 1}}	{F, C, A : 3}
B	{{F, C, A : 1}, {F : 1}, {C : 1}}	ϕ
A	{{F, C : 3}}	{F, C : 3}
C	{{F : 3}}	{F : 3}
F	ϕ	ϕ

From these we build frequency pattern rule.

Item	Conditional Pattern Base	Conditional FP-Tree	Frequent Pattern Generated Rule
P	{{F, C, A, M : 2}, {C, B : 1}}	{C : 3}	{<C, P : 3>}
M	{{F, C, A : 2}, {F, C, A, B : 1}}	{F, C, A : 3}	{<F, M : 3>, <C, M : 3> <A, M : 3>, <F, C, M : 3> <F, A, M : 3>, <C, A, M : 3>}
B	{{F, C, A : 1}, {F : 1}, {C : 1}}	ϕ	{ }
A	{{F, C : 3}}	{F, C : 3}	{<F, A : 3>, <C, A : 3>, <F, C, A : 3>
C	{{F : 3}}	{F : 3}	{<F, C : 3>}
F	ϕ	ϕ	{ }

After this select those rules who satisfy minimum support.

Program for FP growth algorithm.

```
!pip install pyfpgrowth
# Sample code to do FP-Growth in Python
import pyfpgrowth
# Creating Sample Transactions
transactions = [
       ['Milk', 'Bournvita', 'Saffron'],
       ['Milk', 'Saffron'],
       ['Bournvita', 'Saffron','Wafer'],
       ['Bournvita','Wafer'],
    ]
#Finding the frequent patterns with min support threshold=0.5
FrequentPatterns=pyfpgrowth.find_frequent_patterns(transactions=
                                    transactions,support_threshold=0.5)
print("\n Frequent Pattern")
print(FrequentPatterns)
# Generating rules with min confidence threshold=0.5
Rules=pyfpgrowth.generate_association_rules(patterns=FrequentPatterns,
                                                    confidence_threshold=0.5)
print("\n Rules ")
Rules
```

Output:

```
Requirement already satisfied: pyfpgrowth in /usr/local/lib/python3.7/dist-packages (1.0).
Frequent Pattern
   {('Milk',): 2, ('Milk', 'Saffron'): 2, ('Bournvita', 'Milk'): 1,
   ('Bournvita', 'Milk', 'Saffron'): 1, ('Wafer',): 2, ('Saffron',
   'Wafer'): 1, ('Bournvita', 'Wafer'): 2, ('Bournvita', 'Saffron',
   'Wafer'): 1, ('Bournvita',): 3, ('Saffron',): 3, ('Bournvita',
   'Saffron'): 2}
Rules
   {('Bournvita',): (('Saffron',), 0.6666666666666666),
    ('Bournvita', 'Milk'): (('Saffron',), 1.0),
    ('Bournvita', 'Saffron'): (('Wafer',), 0.5),
    ('Bournvita', 'Wafer'): (('Saffron',), 0.5),
    ('Milk',): (('Bournvita', 'Saffron'), 0.5),
    ('Milk', 'Saffron'): (('Bournvita',), 0.5),
    ('Saffron',): (('Bournvita',), 0.6666666666666666),
    ('Saffron', 'Wafer'): (('Bournvita',), 1.0),
    ('Wafer',): (('Bournvita', 'Saffron'), 0.5)}
```

Advantages of FP-growth Algorithm:

1. The FP-growth algorithm scans the database only twice which helps in decreasing computation cost.
2. The FP-growth algorithm uses divide and conquer method so the size of subsequent conditional FP-tree is reduced.
3. The FP-growth method transforms the problem of finding long frequent patterns into searching for shorter ones in much smaller conditional databases recursively.

Disadvantages of FP-growth Algorithm:

1. The FP-growth algorithm is difficult to be used in an interactive mining process as users may change the support threshold according to the rules which may lead to repetition of the whole mining process.
2. The FP-growth algorithm is not suitable for incremental mining.
3. When the dataset is large, it is sometimes unrealistic to construct a main memory based FP-tree.

PRACTICE QUESTIONS

Q. I Multiple Choice Questions:

1. Which is the process as extracting information from huge sets of data?
   (a) Data Mining (b) Big Data Mining
   (c) Data Processing (d) None of the mentioned

2. Finding frequent patterns plays an essential role in, , and so on.
   (a) mining associations (b) mining correlations
   (c) mining frequent itemset (d) All of the mentioned
3. Which is the process of finding a model that describes the data classes or concepts?
   (a) Regression (b) Classification
   (c) Mining (d) Discovery
4. Which patterns are patterns that occur frequently in data?
   (a) Mining (b) Discovery
   (c) Frequent (d) None of the mentioned
5. Improvements techniques in Apriori algorithm include,
   (a) Sampling (b) Partitioning
   (c) Transaction reduction (d) All of the mentioned
6. Which algorithm is used to mine the complete set of frequent itemsets?
   (a) FP-growth (b) PF-growth
   (c) FG-growth (d) None of the mentioned
7. If a supermarket has L items, the number of possible itemsets is,
   (a) 2L-1 (b) 2^L-1
   (c) L/2 (d) L-1
8. Consider an association rule of the form A→ B, where A and B are itemsets. Support of the rule is defined as,
   (a) Fraction of transactions that contain both A and B
   (b) Fraction of transactions that contain A
   (c) Fraction of transactions that contain b
   (d) None of the mentioned
9. A store sells 7 items. MaXimum possible number of candidate 3-itemsets is
   (a) 15 (b) 25
   (c) 35 (d) 45
10. An itemset satisfying the support criterion is known as,
    (a) Frequent Itemset (b) Confident Itemset
    (c) Accurate itemset (d) Reliable itemset
11. If X, Y are two sets of items, and X ⊆ Y. Which of the following statement is always true?
    (a) support(X) ≤ support(Y) (b) support(X) ≥ support(Y)
    (c) support(X) = support(Y) (d) support(X) ≠ support(Y)

12. Consider the itemset {p,q,r,s}. Which of the following statements is always true?
    (a) confidence(pqr → s) ≥ confidence (pq → rs)
    (b) confidence(pqr → s) ≥ confidence (pq → s)
    (c) confidence(pqr → s) ≤ confidence (pq → rs)
    (d) confidence(pqr →s) ≤ confidence(pq→s)
13. Consider three itemsets V_1 = {oil, soap, toothpaste}, V_2 = {oil,soap}, V_3 = {oil}. Which of the following statements are correct?
    (a) support(V_1) > support(V_2)
    (b) support(V_3) > support(V_2)
    (c) support(V_1) > support(V_3)
    (d) support(V_2) > support(V_3)
14. In the following data table, if the support threshold is (greater than or equal to) 0.2 the frequent 4-itemsets are:

Transaction ID	Itemsets
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, c}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, c}

    (a) {a, b, c, d}
    (b) {a, b, c, e}
    (c) {a, c, d, e}
    (d) {a, b, d, e}
15. What does Apriori algorithm do?
    (a) mines all frequent patterns through pruning rules with lesser support
    (b) mines all frequent patterns through pruning rules with higher support
    (c) Both (a) and (b)
    (d) None of the mentioned
16. What is meant by support(A)?
    (a) Total number of transactions containing A
    (b) Total Number of transactions not containing A
    (c) Number of transactions containing A / Total number of transactions
    (d) Number of transactions not containing A / Total number of transactions

17. Which algorithm requires fewer scans of data?
    (a) FP growth (b) Apriori
    (c) Both (a) and (b) (d) None of the mentioned
18. What is the principle on which Apriori algorithm work?
    (a) If a rule is infrequent, its generalized rules are also infrequent
    (b) If a rule is infrequent, its specialized rules are also infrequent
    (c) Both (a) and (b) (d) None of the mentioned
19. What are closed frequent itemsets?
    (a) A closed itemset (b) A frequent itemset
    (c) Itemset with closed and frequent (d) None of the mentioned
20. 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) None of the mentioned
21. The Apriori algorithm works in which fashion.
    (a) top-down and depth-first (b) top-down and breath-first
    (c) bottom-up and depth-first (d) bottom-up and breath-first
22. Which Association Rule would we prefer?
    (a) High support and medium confidence
    (b) High support and low confidence
    (c) Low support and high confidence
    (d) Low support and low confidence
23. If an item set 'XYZ' is a frequent item set, then all subsets of that frequent item set are,
    (a) undefined (b) frequent
    (c) not frequent (d) None of the mentioned
24. If {A,B,C,D} is a frequent itemset, candidate rules which is not possible is,
    (a) C –> A (b) A –> BC
    (c) D –> ABCD (d) B –> ADC
25. What is not true about FP growth algorithms?
    (a) It mines frequent itemsets without candidate generation
    (b) There are chances that FP trees may not fit in the memory
    (c) FP trees are very expensive to build
    (d) It expands the original database to build FP trees

Answers

1. (a)	2. (d)	3. (b)	4. (c)	5. (d)	6. (a)	7. (b)	8. (a)	9. (c)	10. (a)
11. (b)	12. (a)	13. (b)	14. (d)	15. (a)	16. (c)	17. (a)	18. (b)	19. (c)	20. (b)
21. (d)	22. (c)	23. (b)	24. (c)	25. (d)					

Q. II Fill in the Blanks:

1. Data mining is the process of collecting massive amounts of raw data and transforming that data into useful _____.
2. Frequent pattern growth is a method of mining frequent itemsets _____ candidate generation.
3. _____, adopts a divide-and-conquer strategy for finding the complete set of frequent itemsets.
4. All nonempty subsets of a frequent itemset must also be _____.
5. To improve the efficiency of the level-wise generation of frequent itemsets, the Apriori _____ is used to reduce the search space..
6. Association _____ are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.
7. Frequent _____ are patterns that appear frequently in a data set.
8. The descriptions of a class or a concept are called _____ descriptions.
9. Data _____ refers to summarizing data of class under study.
10. _____ analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster).
11. Frequent pattern mining is also called as _____ rule mining.

Answers

1. information	2. without	3. FP-growth	4. frequent
5. property	6. rules	7. patterns	8. class/concept
9. Characterization	10. Cluster	11. association	

Q. III State True or False:

1. The discovery of frequent patterns, associations and correlation relationships among enormous amounts of data is helpful in decision analysis and business management.
2. The Apriori algorithm is not use for generate frequent itemsets.
3. FP algorithm is more efficient algorithm as compare to apriori as it required less database scan to generate frequent itemsets.

4. Variations in Apriori algorithms like involving hashing and transaction reduction can be used to make the procedure more efficient.
5. FP growth algorithm avoids costly candidate generation, resulting in greater efficiency.
6. Data mining also known as Knowledge Discovery in Data (KDD).
7. Data discrimination refers to the mapping of a class with some predefined group or class.
8. Frequent substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.
9. Outlier analysis is a process that involves identifying the anomalous observation in the dataset.
10. FP growth algorithm creates FP tree to compress a very small dataset.

Answers

1. (T)	2. (F)	3. (T)	4. (T)	5. (T)	6. (T)	7. (T)	8. (T)	9. (T)	10. (F)

Q.IV Answer the following Questions:

(A) Short Answer Questions:

1. Define data mining.
2. Define frequent pattern.
3. What is itemsets?
4. Define frequent itemsets?
5. Define support and confidence.
6. What is the purpose of FP growth algorithm?
7. What is the purpose of Apriori property?
8. Define outliner analysis?
9. What is closed and maximal itemsets?
10. What is data characterization?
11. Define data discrimination?
12. List any two applications of data mining.
13. Define association analysis?
14. What is meant by minimum threshold and minimum confidence threshold?
15. List applications of outlier analysis.
16. What is relevance analysis?
17. List frequent itemset mining methods.
18. What is the purpose of Apriori algorithm?

(B) Long Answer Questions:

1. What is data mining? Explain with diagram? Also state its advantages and disadvantages.
2. Explain usage of Market Basket Analysis with example?
3. Explain Apriori algorithm in detail.
4. What are frequent itemsets, closed itemsets, and association rules? Describe in detail.
5. What is outliner analysis? Describe in detail.
6. What kind of patterns can be mined? Explain I deatail.
7. How to mine following:
   (i) Frequent Patterns.
   (ii) Associations.
   (iii) Correlations.
8. What are different types of data? Explain in detail with appropriate examples.
9. What are different sources of data in data science? Describe in detail.
10. Explain different data formats in brief.
11. What is meant by followings and Explain in reference of mining frequent patterns:
    (i) Predictive analysis.
    (ii) Cluster analysis.
    (iii) Outlier analysis.
12. What is structured and unstructured data? Distinguish between them.
13. How to generating association rules from frequent itemsets? Explain in detail.
14. With the help of example describe FP growth algorithm. Also state its advantages and disadvantages.
15. Which techniques are used for improving efficiency of Apriori algorithm? Describe two of them in detail.
16. Write a short note on: Class/concept description.
17. A database has five transactions. Let min sup =60%andmin conf =80%.

TID	Items Bought
T100	{M,O,N,K,E,Y}
T200	{D,O,N,K,E,Y}
T300	{C,A,K,E}
T400	{D,U,C,K,Y}
T500	{C,O,O,K,I,E}

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

18. A database has six transactions. Let min-sup = 50% and min-conf = 75%. Find all frequent item sets using Apriori algorithm. List all the strong association rules.

TID	List of Items
001	Pencil, Sharpener, Eraser, Chart papers, Sketch pen
002	Chart papers, Charts, Glue sticks
003	Pencil, Glue stick, Eraser, Pen, Sketch pen
004	Oil pastels, Poster colors, Correction tape
005	Whitener, Pen, Pencil, Charts, Glue stick
006	Color pencils, Crayons, Eraser, Pen

19. Consider the following transaction database:
    transaction1 = (product1, product2, product7)
    transaction2 = (product4, product5, product7)
    transaction3 = (product6, product7, product8, product9)
    transaction4 = (product1, product3, product4, product6, product7
    Find the frequent product set using FP Algorithm with minimum support 50%.
20. Consider following database and find out the frequent item sets using Apriori algorithm with min-sup = 50%.

Transaction ID	Items Purchased
T1	{Apple, Mango, Banana}
T2	{Mango, Banana, Cabbage, Carrots}
T3	{Banana, Carrots, Mango}
T4	{Carrots, Mango}

21. Consider the following transaction database with a min sup of 40 % (2 transactions) Using FP algorithm find the frequent itemset.

Transaction ID	Items Purchased
t1	$\{I_1, I_3, I_4\}$
t2	$\{I_2, I_3, I_5\}$
t3	$\{I_1, I_2, I_3, I_5\}$
t4	$\{I_2, I_5\}$
t5	$\{I_1, I_2, I_3, I_5\}$

22. Consider following database and find out the frequent item sets using Apriori algorithm with min-sup = 50%.

Transaction ID	Items Purchased
T1	{Green Peas, Broccoli, Baby Corn}
T2	{Broccoli, Baby Corn, Red Cabbage, Tomatoes}
T3	{ Baby Corn, Tomatoes, Broccoli}
T4	{ Tomatoes, Broccoli}

■■■

CHAPTER 4

Social Media and Text Analytics

Objectives...

- To understand Concept of Social Media
- To learn Text Analytics

4.0 INTRODUCTION

- In the information age communication is the act of exchanging information by speaking, writing or using some other medium has exploded.
- The most basic communication theory states that communication consists of a sender, a message, a channel where the message travels, noise or interference and a receiver.
- In recent years, social media has gained significant popularity and become an essential medium of communication.
- Merriam-Webster (America's most trusted online dictionary for English word definitions, meanings and pronunciation) defines social media as, "forms of electronic communication through which users create online communities to share information, ideas, personal messages and other content". **OR**
- As per guidelines given by Government of India, 'Department of Electronics and Information Technology': "Social Media in recent times has become synonymous with Social Networking sites such as Facebook or MicroBlogging sites such as Twitter. However, very broadly social media can be defined as, any web or mobile based platform that enables an individual or agency to communicate interactively and enables exchange of user generated content."
- Critical characteristics of social media are:
  1. **Connectedness:** This characteristic of social media basically evaluate any social media's ability through that how much it is able to connect and re-connect people interested in same topics and domains. Connectedness property of social media also ensured with facility of all time availability for the users using variety of media and access devices including PCs, Laptops, mobile phones etc.
  2. **Collaboration:** The connections achieved on this media, enable people to collaborate and create knowledge. Such collaborations can be either open or

closed. Wikipedia is an example of open collaboration which enabled creation of an open web based encyclopedia through contribution from hundreds of thousands of people.

3. **Community:** Connectedness and collaboration helps create and maintain communities. These communities can create awareness about various issues and can be used for seeking inputs into policy making, building goodwill or even seeking feedback into delivery of public services.

- All three characteristics have mutual dependency and are interconnected with each other.
- Kaplan and Haenlein in 2010 classified social media into six different types namely, collaborative projects, blogs and microblogs, content communities, social networking sites, virtual game worlds and virtual social worlds.
- There are several web based social network services are available as given below:
  1. **Twitter:** Twitter allows the user to send and reply messages in form of tweets. These tweets are the small messages, generally include 140+ characters.
  2. **Whatsapp:** It is a mobile based messaging app. It allows to send text, video, and audio messages
  3. **Facebook:** Allows to share text, photos, video etc. It also offers interesting online games.
  4. **Flickr:** Flickr offers image hosting and video hosting.
  5. **LinkedIn:** Linkedin is a business and professional networking site.
  6. **Google+:** It is pronounced as Google Plus. It is owned and operated by Google.
  7. **Hike:** It is also mobile based messager allows to send messages and exciting emoticons.
- Text analytics is the automatic discovery of new, previously unknown, information from unstructured textual data.
- The text analytics is process involves following three tasks:
  1. Information retrieval (gathering the relevant documents).
  2. Information extraction (unearthing information of interest from these documents).
  3. Data mining, (discovering new associations among the extracted pieces of information).

4.1 OVERVIEW OF SOCIAL MEDIA ANALYTICS

- A social network is a type of complex network and can be described as a social structure composed of a set of social actors or users and the inter-relations and social interactions between them.

- These social networks are useful to study the relationships between individuals, groups, social units or societies.
- Social media analytics is the process of collecting, tracking and analyzing data from social networks.
- Fig. 4.1 illustrates the five year growth figures for the number of social media users that has been conveyed by Digital 2019 reports.
- The statistical figures reveal drastic increase in the growth of social media users (almost double) throughout the world considering from year 2014 to year 2019.

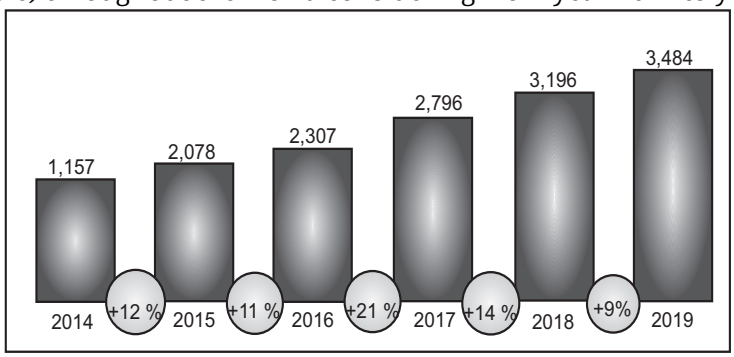


Fig. 4.1: Number of Social Media Users (in millions) from Year 2014 to Year 2019

- Web 2.0 and social media facilitate the creation of vast amounts of digital content that represents a valuable data source for researchers and companies alike.
- Social media analytics relies on new and established statistical and machine learning techniques to derive meaning from large amounts of textual and numeric data.
- The analysis of social networks is centered on the fundamental theory that the social network is made up of the relations and interaction between users and within units rather than by the properties of the user itself.
- Social Network Analysis (SNA) is the general reference to the process of investigating social networks or structures within social media through the use of networks, knowledge graphs and graph theory.
- In this section we will social media analytics methods and to demonstrate how social media analytics can be applied in a variety of contexts to deliver useful insight.

Benefits of Social Media Analytics:

1. The continuous monitoring, capturing and analyzing of social media data can become the valuable information for decision-making.
2. The social media is that it gives us the ability to track and analyze the growth of the community on social media sites and the activities and behavior of the people using the sites.
3. Governments from around the world are starting to realize the potential of data analytics in making timely and effective decisions.

4.1.1 Social Media Analytics Process

- The social media analytics process comprises of three stages namely, data capturing, data understanding and data presentation.
- These stages in social media analytics process are shown in Fig. 4.2.
- Fig. 4.2 shows the work done in each stage of the process. Each and every of stage of social media analytics process has its own importance and relevance.

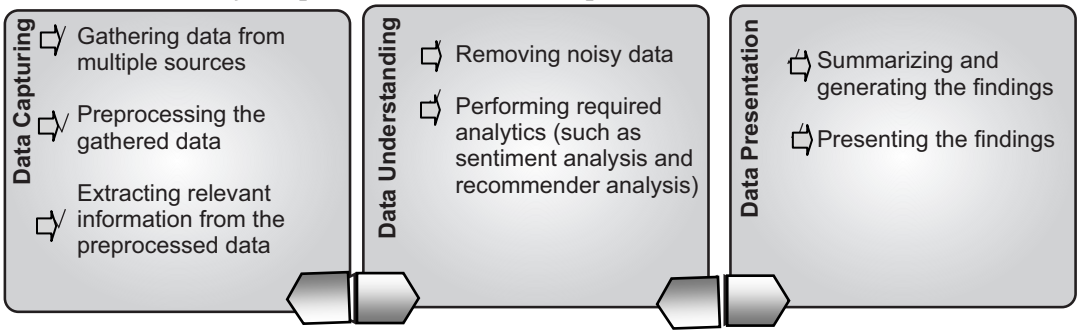


Fig. 4.2: Social Media Analytics Process

- The stages in social media analytics process are explained below:

1. Data Capturing:

- Data capture means valid data identification. Data identification is the process of identifying the subsets of available data to focus on for analysis.
- Raw data is useful once it is interpreted. Any data that conveys a meaningful message becomes information.
- In this stage (data capturing) a data analyst procures relevant social media data by monitoring and assessing the social media content to understand the users' reactions, likes and dislikes, interests, reviews, and comments about a topic, agenda, brand, item or product.
- Capturing of data may comprise of collecting data from various platforms/applications like Facebook, Instagram, LinkedIn, Twitter, Wikis, Microblogs and many such popular sites.
- Generally data collected by theses social networking sites are unstructured, which required preprocessing of data. This data is very crucial for proper analysis of data.
- Data preprocessing may involve data cleaning, data integration, data transformation, data reduction and data discretization.
- Once the preprocessing is end, only relevant information is retained and the unwanted data are discarded so that proper analysis can be performed in the next stage of only on relevant or appropriate data.

2. Data Understanding:

- The next stage after data capturing is Data understanding stage, which utilizes the data capturing stage and analyzing the captured data for gaining meaningful insights.

- Before analyzing the data, the process of noise removal from data may be required for better accuracy in data analysis.
- Preprocessing of data covered treatment of noisy, missing or corrupted data, which involve several fields and techniques such as statistical analysis, machine learning, deep learning, computer vision, natural language processing, and/or data mining.
- The data understanding stage, in social media analytics process lies in the middle of the social media analytics process and forms the core and most important stage in the entire process.
- Data analysis is the set of activities/tasks that assist in transforming raw data into insight, which in turn leads to a new base of knowledge and business value.
- In other words, data analysis is the phase that takes filtered data as input and transforms that into information of value to the analysts.
- Once the data analysis is performed, the analyzed data is further presented to the next stage in social media analytics process, which is the data presentation stage.
- The evaluation of results or outcomes in the last stage mostly depends on the findings in the data understanding stage.
- If proper analysis techniques are not used in the data understanding stage in social media analytics process, the findings may lead to fully incorrect output generation.
- Hence, immense care needs to be taken in the data understanding stage, in social media analytics process so that the right tools and techniques are used while analyzing the relevant data.

3. Data Presentation:

- This stage is the final stage in the social media analytics process. In data presentation stage, the results are summarized and evaluated to gain significant insights.
- These final results or outcomes are then presented mostly using proper data visualization tools to present the output or result in an easy and simple interpretable form.
- Note that data presentation is what the users get as output/result at the end and hence, no matter how big the data is in volume, the data visualization graphic(s) should make the output easily understandable for the data analysts.
- Interactive data visualization has led us to a new era of revolution where graphics have led us to easy data analysis and decision making.
- For example, data visualization can help in identifying outliers in data, improving response time of analysts to quickly identify issues, displaying data in a concise format, providing easier visualization of patterns etc.
- The most challenging part in data representation is to learn how data visualization works and which visualization tool serves the best purpose for analyzing precise information in a given case.

- It is most important to understand in data representation stage that the three stages of the social media analytics process most of the time work iteratively rather than linearly.
- If the models generated in this stage (data understanding), in social media analytics process fails to uncover useful results during analysis, the process turns back to the data capturing stage to further capture additional data that may increase the predictive and analysis power.
- Similarly, if in this stage (data understanding), the results that are generated is not convincing or have low predictive power, then there is a necessity to turn back to the data capturing or data understanding stage to tune the data and/or the parameters used in the analytics model.
- Thus, this entire process of social media analytics may go through several iterations before the final results are generated and presented.

4.1.2 Seven Layers of Social Media Analytics

- Social media analytics refers to the approach of collecting data from social media sites and blogs and evaluating that data to make business decisions.
- Social media analytics help in business/organization analysis to a large extent which traditional business analytics would not have managed single handedly.
- Conventional (or traditional) business analytics focuses only on structured data, while today we are content rich with semi-structured or unstructured data found in social media.
- These semi-structured or unstructured data that are used in social media analytics are real-time, content-rich and dynamic which contributes to valuable and informative decision making in businesses/organizations.
- For this, the seven layers of social media analytics are needed to be thoroughly understood and analyzed to carry out social media analytics for insightful decision making.
- Social media consist of seven layers of data that contain useful information that is often garnered for Business Intelligence (BI).
- These seven layers of data may be either visible (say, textual data) or invisible (say, hyperlink network).
- All these seven layers of the social media play a vital role in contributing to social media input for gaining useful insights in businesses.
- Fig. 4.3 shows the seven layers of data found in social media.

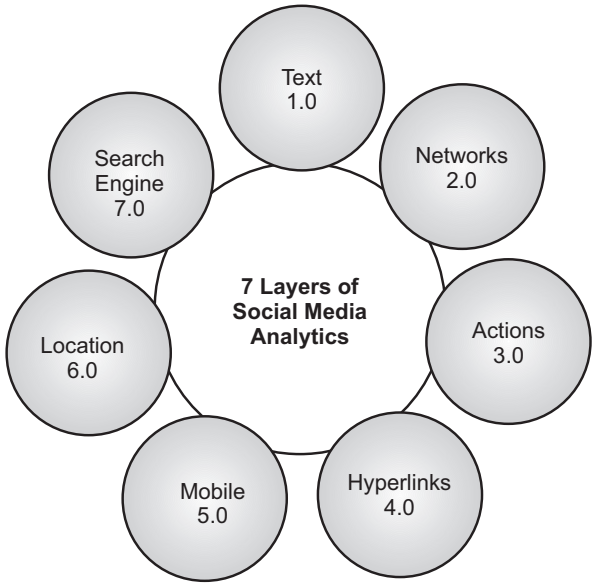


Fig. 4.3: Seven Layers of Social Media Analytics

- Let us see the seven layers of social media analytics in detail:

1. Layer 1 (Text):

- The textual messages of social media include textual posts, tweets, comments, status updates, blog posts and so on.
- These texts of social media are often in business analytics to identify user opinion or sentiment regarding particular product, topic or individual.
- Social media text analytics consist of extracting, analyzing and interpreting the textual elements of social media contents such as comments.

2. Layer 2 (Networks):

- Social media network analytics extracting, analyzing and interpret personal and professional social network such as Facebook friendship network, Twitter follower network and so on.
- Social media network analytics focuses on the networking structure of the social media data.
- The social media data indicates the connection between users based on the concept of friends and followers. Such connections/interactions are often found in various networking sites such as Twitter, Facebook, LinkedIn and so on.
- The network analysis is mainly done using graph theory which consist of nodes (considered as the users) and the edges (considered as the links or connection among users).
- Network analysis are often done to identify or finding influential users, predict new links and for various other analysis.

- The Facebook friendship network and Twitter Follower network Network analytics seeks to identify influential nodes (e.g., people and organization) and their position in the network.

3. Layer 3 (Actions):

- Actions in social media mainly include the actions performed by users while using social media such as clicking on like or dislike button, sharing posts, creating new events or groups, accepting a friend request and so on.
- Data analysts often carry out actions analytics using social media data for measuring various factors such as popularity of a product or person, recent trends followed by users and popularity of user groups.
- Social media actions analytics deals with extracting, analyzing and interpreting the action performed by social media users, including, likes, dislikes, shares, mentions and endorsement.
- Action analytics in social media are mostly used to measure popularity and influence over social media.

4. Layer 4 (Mobile):

- Mobile analytics is comparatively a recent trend in social media analytics. Mobile analytics focuses on analysis of user engagement with mobile applications.
- Mobile analytics are usually carried out for marketing analysis to attract those users who are highly engaged with a mobile application.
- The in-app analysis is another common analysis carried out in mobile analytics. The in-app analysis concentrates on the kind of activities and interaction of users with an application.

5. Layer 5 (Hyperlinks):

- Hyperlinks (or links) are commonly found in almost all Web pages that allow navigation of one Web page to another.
- Hyperlinks analytics, extracting, analyzing and interpreting social media hyperlinks (e.g. in-links and out-links).
    - The hyperlink into a Web page is called as in-link. The number of in-links to a Web page is referred to as in-degree.
    - The hyperlink out of a Web page is called as out-link. The number of out-links from a web page is referred to as out-degree.
- Hyperlink analysis can reveal, for example, Internet traffic patterns and sources of the incoming or outgoing traffic to and from a source.
- In simple words, mobile analytics is all about analyzing and interpreting social media hyperlinks.

6. Layer 6: (Location):

- Location analytics is concerned with mining and mapping the locations of social media users, contents and data.

- Location analytics is also known as geospatial analysis or simply spatial analytics. This analytics is carried out to gain insight from the geographic content of social media data.
- Real-time location analytics is often carried out by data analysts for Business Intelligence (BI). For example, the courier services used by social media sites need to keep track of the locations of delivery in real-time.
- In location analytics, historical geographic data is also often used to bring an increase in sales and profit in businesses.

7. **Layer 7 (Search Engines):**

- The search engines analytics focuses on analyzing historical search data for gaining a valuable insight into a range of areas, including trends analysis, keyword monitoring, search result and advertisement history and advertisement spending statistics.
- Search engine analytics pays attention to analyzing historical search data to generate informative search engine statistics and these statistical results can then be used for Search Engine Optimization (SEO) and Search Engine Marketing (SEM).

4.1.3 Social Media Analytics Life Cycle

- The life cycle of social media analytics is shown in Fig. 4.4.
- The cycle of social media analytics consists of six steps namely, identification, extraction, cleaning, analyzing, visualization, and interpretation.
- Each step in the life cycle of social media analytics is followed one after another to mine the required business insights based on the social media data chosen.
- The life cycle of social media analytics consists of the core desired business objectives that notify each step of the social media analytics journal.

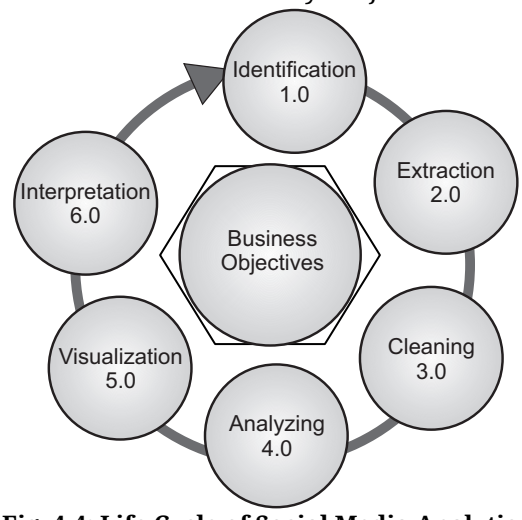


Fig. 4.4: Life Cycle of Social Media Analytics

- Fig. 4.4 shows six general steps common for all social media analytics processes. In the beginning, the business objectives need to be clearly defined and then all the six stages of the social media analytics cycle are carried out one after another until the business objectives are all fully satisfied.
- Let us now understand the contribution of each of the steps in the analytics life cycle.

Step 1 - Identification:

- The identification step is mainly concerned in identifying the correct source of data for carrying out analysis.
- The identification step is a crucial step in the social media analytics life cycle that decides which data to consider among the vast, varied and diverse data that is collected from various social media platforms.
- The decision on which data is to be considered for mining is mainly governed by the business/organization objectives that are needed to be achieved.

Step 2 - Extraction:

- Once the accurate and/or appropriate social media data source is identified for a specific analysis, the next step is to use a suitable API for data extraction.
- Through these APIs (Application Programming Interfaces) that almost every social media service companies have, it is possible to access a required small portion of social media data hosted in the database.
- Number of other specialized tools can help in data extraction from social media sites. However, the extraction of data from social media sites is to be done following all privacy and ethical issues concerned with mining data from social media platforms.

Step 3 - Cleaning:

- Data cleaning is done as a data preprocessing step used to remove the unwanted and/or unnecessary data so as to reduce the quantity of data to be analyzed. Cleaning data is usually done to handle irrelevant or missing data.
- In this step, the data is cleaned by filling in the missing values, smoothing any noisy data, identifying and removing outliers, and resolving any inconsistencies. Filling up the missing values in data is known as the imputation of missing data.
- The data cleaning method to be adapted for filling up the missing values depends on the pattern of data used and the nature of analysis to be performed with the data.
- The technique, smoothing is also used during data preprocessing for data analysis. Smoothing is intended to identify trends in the presence of noisy data when the pattern of the trend is unknown.
- For data cleaning, outliers should also be excluded from the data set as much as possible as these outliers may mislead the analysis process resulting in incorrect results.
- For this reason, an important preprocessing step is to correct the data by following some data cleaning techniques.

Step 4 - Analyzing:

- In analyzing step, the clean data is used by data analysts for analyzing data for business insights.
- The main objective in analyzing step is to generate meaningful insights from the social media data fed as input for analysis. Choosing the right tools and techniques is an important factor for accurate data analysis.
- The analyzing step often involves a detailed step-by-step procedure to arrive at a meaningful insight for correct and fruitful decision-making.

Step 5 - Visualization:

- Once the data is analyzed, it is preferred to be presented in a graphical and/or pictorial format as it is said that the human brain processes visual content better than plain textual information. This is where the role of data visualization in data analytics comes into play.
- Interactive data visualization has led to a new era of revolution where graphics lead to easy and simple analysis and decision making.
- The right data visualization tool to be used depends on which layer of social media data analytics is to be dealt with.
- For example, if it is required to display network data visualization, once can display it through a network chart, the 'word cloud' visualization can be used for representing textual data.

Step 6 - Interpretation:

- The final stage/step in the social media analytics life cycle is the interpretation of results. Data interpretation step involves translating outcomes in meaningful business solutions.
- At data interpretation step, the various data visualizations generated from the previous stage are studied to understand the results and finally give a meaningful summary of the entire analytics that is carried out on the data.

4.1.4 Accessing Social Media Data

- Social media data is the collected information from social media networks that show how users share, view with the content or profiles.
- Social media data is the information that is collected from different social media networks such as Facebook, Instagram, Twitter, LinkedIn, YouTube and so on.
- Social media applications or apps and Websites not only provide much information but also allow users to interact in various ways to access as well as contribute to the information.
- Social media analytics is the process of dealing with social media data by identifying meaningful structural properties and media content and analyzing these data to gain meaningful insights.

- Social media data refers to all of the raw insights and information collected from individuals social media networks.
- To work with social media data, it is required to get access to data that is generated from various social media sites.
- Social media scraping is the task of extracting the unstructured data generated from these social media sites.
- There are various social media scrapers developed which act as an excellent tool for extracting social media data.
- Few such prominent scrapers often used by data analysts and data scientists include Octoparse, OutWit Hub, Scrapinghub, Parsehub, and Dexi.io.
- Most of the social media network sites have their APIs (Application Programming Interfaces) that can be used by data analysts or data scientists to access the social media data found in these sites and also integrate the various other features of the APIs into applications.
- The data is often collected with tools which communicate with the respective API of the social media platform, if one exists and crawl the data.
- The APIs can be differentiated among each other based on the features provided by the APIs, the popularity of the social media site to which the API is connected, the cost of using the APIs and the ease in which each API can be used for data analysis.
- Some of the prominent social media APIs used for accessing social media data are explained below:

1. Facebook API:

- Facebook provides a platform, where people come to socialize, talk and share their views with each other.
- Facebook social networking sites commonly used by a large number of the people to interact with their families and friends and also making business appearances or meeting online with other users.
- Facebook content can be accessed through Facebook APIs free of cost. Facebook's one of the commonly used APIs is the Facebook Graph API.
- Facebook Graph API is commonly used by social media researchers to access Facebook data. Facebook APIs also help in posting images, creating new stories, accessing Facebook pages and so on.
- In Facebook there is also a provision of using the Facebook Marketing API that helps create applications for services and product marketing.

2. YouTube API:

- The YouTube platform is owned by Google. The YouTube's basic functionality is video and music sharing.

- The API of YouTube provides options to YouTube users to play or upload videos, manage playlists, search music and videos and several other functionalities.
- The YouTube API also has a provision for analysis of the videos, subscribe to the videos, and schedule live streaming broadcasts.

3. Instagram API:

- Instagram is a photo and video sharing social networking platform owned by Facebook.
- The Instagram social media platform has a provision of photo and video sharing among users. The sharing of data can be done either publically or only among followers of Instagram users.
- The concept of hashtags became popular on Instagram API that allows users to provide highlights of the topics portrayed in their feeds.
- Instagram also has number of APIs built for specific purposes. One most popular API is the Instagram Graph API that is used by analysts or data scientists to access the data of Instagram accounts used for business operations.
- Mostly, Instagram content can be accessed through Instagram APIs free of cost. Such Instagram content is often used by researchers from the social media analytics community to generate meaningful insights about some particular topic, product or individual.

4. Twitter API:

- Twitter is a social networking platform which enables registered users to read and post short messages called tweets. The tweet contains rich social media content.
- The length of the tweet messages is limited to 140 characters and users are also able to upload photos or short videos.
- It is the place, where people are more inclined to listen than to speak. Twitter is one of the popular social media services provided online.
- The bulk amount of Twitter API users post messages called tweets that contain rich social media content. Twitter content can be accessed through Twitter APIs for free of cost.
- However, there are paid versions of Twitter APIs which provide more accessibility to data as well as reliability. The APIs provided by Twitter are categorized based on the type of service it provides.
- For example, the Search API allows access of tweets to retrieve historical content, the advertisement API help creating advertisement campaigns, and the Account Activity API is used to access Twitter account activities.

5. LinkedIn API:

- LinkedIn is the world's largest professional social networking site. It offers a platform, where users or company's do B2B marketing.

- LinkedIn allows the profile owners to share employment and personal information with others. This site focuses on user's professional identities.
- Every popular social media network sites, as well as online discussion forums and new sites, develop their APIs to provide access to content that can be used for data analysis to generate interesting results that can help in product promotion, political campaign, influence maximization, information diffusion, business profit uplifts, and many more.
- Social media crawling has become a buzzword even in the student community as students carry out projects on social media analysis by using real-time social media datasets.

4.2 KEY SOCIAL MEDIA ANALYTICS METHODS

- Social media analytics is the ability to gather and find meaning in data gathered from social networks to support business decisions making.
- There are several challenges involved in dealing with social media data. These challenges are tackled using multi-disciplinary techniques like data mining, statistics, graph-based mining, computational linguistics and so on.
- In the context of social media analytics, there are mainly three methods of analysis that have a variety of applications or platforms. These three methods are social network analysis, text analysis/mining and trend analytics.
- Let us quickly study each of these methods in detail to understand the pattern of work carried out in social media analytics.

4.3 SOCIAL NETWORK ANALYSIS

- Social Network Analysis (SNA) is one of the methods of social media analytics. The SNA mainly involves studying the relationships between media users, organizations, user communities, users from a particular demographic group and so on.
- SNA is the process of investigating social structures through the use of social media networks.
- In social media analytics, SNA emphasizes analyzing the users in a network (often referred to as nodes) and their connections among each other (often termed as edges).
- By concentrating on the structure of connections or interactions among users, SNA can help identify opinion leaders, influential users or user communities in social media.
- It is also essential and important to note that online social network data is huge, dynamic, noisy and scattered.
- Social network analysis and mining allow analyzing such complex, expensive and dynamic data to generate required outcomes. In doing so, one major decision to be taken is whether to apply static structure mining or dynamic structure mining.

- **Static Structure Mining:** It works with snapshots of data of a social network that is stored within a specified time period. In this case, the analysis is thus carried out on a static social network and focus is given on the structural regularities of the static network graph.
- **Dynamic Structure Mining:** It uses dynamic data that constantly keeps changing with time. In this case, the analysis is thus carried out on a dynamic social network and focus is given on unveiling the changes in the pattern of data with the change in time.

- Following program shows the source code to display a simple social network graph that consists of six nodes and nine edges. To display the graph, the networkx Python library is imported and used.
- The edges between nodes are created one by one and then the graph is displayed for visualization of the network.
- The various other network information such as the number of nodes and edges, network density, etc. is also provided in the output.
- The program also displays the degree value, the clustering coefficient value and the eccentricity value of each node in the graph.
- The clustering coefficient value of each node in the program help in assessing the degree to which the nodes in the social graph tend to cluster together. The eccentricity value determines the maximum graph distance to be traveled between a node and any other nodes in the graph.

```
** Program for Social Network Graph
import networkx as nx
from operator import itemgetter
G = nx.Graph()
G.add_edge("A", "B")
G.add_edge("A","B")
G.add_edge("A","C")
G.add_edge("A","D")
G.add_edge("A","E")
G.add_edge("B","C")
G.add_edge("B","D")
G.add_edge("B","E")
G.add_edge("F","E")
G.add_edge("F","D")
nx.draw_networkx(G)
#Displaying Graph Information
```

```
print(nx.info(G))
density = nx.density(G)
print('Network density:', density)
# Displaying Degree of each node
print(nx.degree(G))
#Displaying Top 3 Nodes Based on Highest Degree
degree_dict = dict(G.degree(G.nodes()))
nx.set_node_attributes(G,degree_dict,G.degree)
#nx.set_node_attributes(G, hist_sig_dict, 'historical_significance')
sorted_degree = sorted(degree_dict.items(), key=itemgetter(1),
                                                        reverse=True)
print('Top 3 nodes by degree:')
for d in sorted_degree[:3]:
   print(d)
# Displaying Clustering Coefficients of each node
print('Clustering Coefficients of iven graph  : ', nx.clustering(G))
# Displaying eccentricity of each node
print('Eccentricity of given graph:', nx.eccentricity(G))
```

Output:

```
Graph with 6 nodes and 9 edges
Network density: 0.6
[('A', 4), ('B', 4), ('C', 2), ('D', 3), ('E', 3), ('F', 2)]
Top 3 nodes by degree:
('A', 4)
('B', 4)
('D', 3)
Clustering Coefficients :  {'A': 0.5, 'B': 0.5, 'C': 1.0, 'D':
                    0.3333333333333333, 'E': 0.3333333333333333, 'F': 0}
Eccentricity : {'A': 2, 'B': 2, 'C': 3, 'D': 2, 'E': 2, 'F':
```

- Fig. 4.5 shows the network graph that is displayed using the draw_networkx() function of the networkx library.
- The output of the above program consists of basic social network graph information such as the number of nodes, the number of edges, average degree of all nodes considered together, network density and the top three nodes based on the highest degree.

- Also, the above program output shows the degree value, the clustering coefficient value, and the eccentricity value of each node.

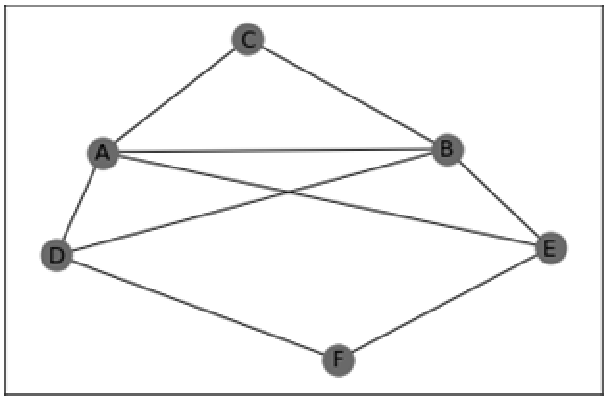


Fig. 4.5: Seven Layers of Social Media Analytics

- Few of the mining issues (or problems) dealt in social network analysis and mining include link prediction, community detection, influence maximization, expert finding and prediction of trust and distrust among individuals.
- Each of these issues in social network analysis and mining is studied extensively for research to find optimum algorithms and techniques for the same.

4.3.1 Link Prediction

- Link prediction is the problem of predicting the existence of a link between two entities in a social network.
- The link prediction problem is one common research issue in social network analysis and mining.
- The link prediction issue studies a static snapshot of the nodes and edges of a social network at a given time T1 and based on the study, predicts the future links of the social network for a future time T2.
- The link prediction problem is a common feature found in many social networking sites for possible friends' suggestions as found on Facebook or Twitter.
- This feature, in turn, allows a user to increase the personal or professional friends circle to broaden the social links and connections.
- This will increase the social networking activities as each user will be then connected to more users on the social network.
- The objective of link prediction is to identify pairs of nodes that will either form a link or not in the future.
- The Fig. 4.6 shows a standard link prediction framework that feeds a static social network as input and then applies either a similarity-based approach or a learning-based approach for prediction of future links in the social media network.

- **Similarity-based Approach:** This approach of link prediction calculates the similarities of non-connected pair of nodes in a social network and a score is accordingly assigned for each non-connected pair. Based on the descending order of similarity score, a list is prepared to choose the top-N ranked links from the list for link prediction.
- **Learning-based Approach:** This approach of link prediction is a classifier that uses some standard machine learning models to assign a label that is binary positive or negative.
  - A **positive value** indicates that there is a chance of better connectivity between the non-connected pair of nodes.
  - A **negative value** indicates that there is very little chance of connectivity between the non-connected pair nodes.

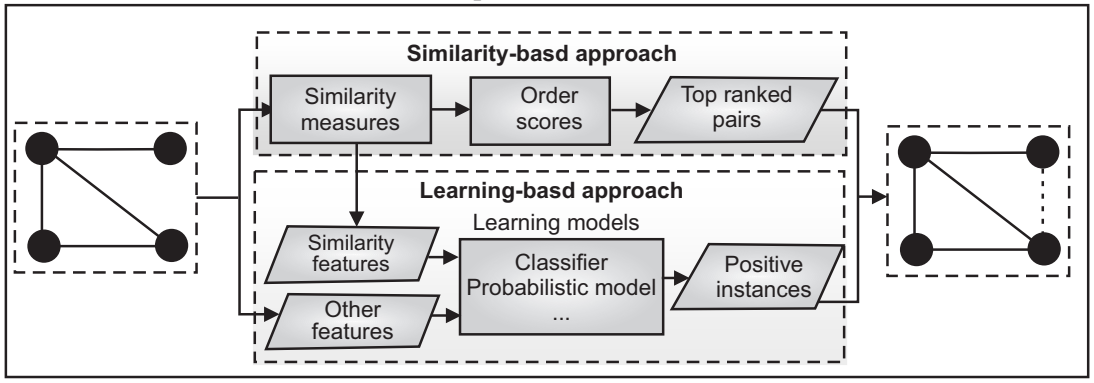


Fig. 4.6: The Generic Link Prediction Framework

4.3.2 Community Detection

- Community detection plays a major role in social media networks. Community detection techniques are useful for social media algorithms to discover users with common interests and keep them tightly connected.
- Community detection in social media network can be used in machine learning to detect groups with similar properties and extract groups for various reasons.
- For example, this technique can be used to discover manipulative groups inside a social network or a stock market.
- For community detection in social media networks, a study is carried out to find the correlation between nodes in the network to assess the strength of the connection or interactions between nodes.
- With community detection, intra-communities are formed in which a group of users (or nodes) with close correlation belong to the same community and in inter-communities, users (or nodes) belong to different communities.

- A user belonging to the same community is expected to share similar tastes, likes and dislikes which helps in the prediction of what products a user is likely to buy, which movie a user is likely to watch, what services a user may be interested in, and so on.
- Fig. 4.7 shows the various categories of approaches followed for community detection in social media networks.
- The community detection categories are broadly divided into, the traditional clustering community detection methods, the link-based community detection methods, the topic-based community detection methods and the topic-link based community detection methods.

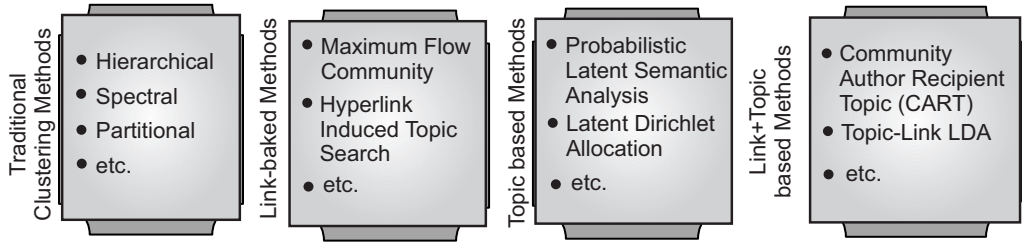


Fig. 4.7: Various Categories of Community Detection Methods

- Various community detection categories are explained below:

1. **Traditional Clustering Methods:**
   - These methods of community detection are mainly divided into hierarchical, spectral and partitional methods.
     - **Hierarchical Clustering:** This method either gradually merges or splits the groups to create nested clusters. Two such standard hierarchical clustering methods are the Clustering Using Representatives (CURE) and the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH).
     - **Spectral Clustering:** This method creates groups of communities by using the spectral properties of the similarity matrix.
     - **Partitioning Clustering:** This method divides all nodes into n clusters, where the value of n is provided as a parameter well in advance.
2. **Link-based Clustering Methods:**
   - These methods emphasize the study of edges of the social network in order to form communities.
   - In these community detection methods, what is mainly explored is the strength of connections between nodes and not the basic semantics such as the common topic of interests or likings among nodes.
   - The Two standard link-based community detection methods are Hyperlink Induced Topic Search (HITS) and Maximum Flow Community (MFC).

3. **Topic-based Methods:**
   - The topic-based community detection methods emphasize the generation of communities based on the common topic of interests.
   - In topic-based community detection methods, what is mainly explored is finding communities that are topically similar and do not consider any emphasis on the strength of connections between nodes.
   - Two standard topic-based community detection methods are Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA).
4. **Topic-link Based Methods:**
   - The topic-link based methods community detection methods are the most common approaches used nowadays for community detection in social networks.
   - These community detection methods are hybrid as these approaches consider both the strength of connections between nodes as well as finding communities that are topically similar.
   - Thus, the topic-link based methods consider the disadvantages of using only one single method i.e., link-based or topic-based for community detection, and combine both the methods to give more accurate results.
   - Two standard topic-link based community detection methods are Community-Author-Recipient-Topic (CART) and Topic-Link LDA.

- Community detection in social media analytics plays a major role not only in social networks but also in various other fields such as information networks, sociology, biology, politics and economics.
- The main challenges of community detection lie when the network of nodes to consider is vast and dynamic which needs special methods to deal with the complexity of the network.
- An important application that community detection in social media analytics has found in network science is the prediction of missing links and the identification of false links in the network.

4.3.3 Influence Maximization

- Influence maximization is the problem (or issue) of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence.
- Influence propagation is the task of choosing a set of proficient users who can prove to be very efficient for viral marketing.
- This set of efficient users in a social media network is called a seed set which is considered as valuable nodes to target for promotion or publicity as these online users have the highest reach of spreading information.
- Indirectly, the seed set of users can help other users to decide in choosing as to which movie to watch, which political party to follow, which product to buy, which community to join and so on.

- Nowadays the viral marketing is considered an effective tool being adapted by all companies and organizations for the promotion of brands or companies and publicity of organizations.
- Viral marketing is a strategy that uses existing social networks to promote a product mainly on various social media platforms.
- What is done in case of viral marketing is to initially use an influence maximization technique to,
  - find a set of few influential users of a social network, and
  - influence those users about the goodness and usefulness of a product so that it can create a cascade of influence of buying the same product by the users' friends.
- The user's friends will again, in turn, recommend or publicize the same product to their friends and this helps in easy and simple product promotion.
- This product promotion strategy is also adopted in other domains such as political campaigns, movie recommendations, company publicity and so on.
- The main challenge in influence maximization is to generate the best influential users (the seed set) in the social media network.
- Fig. 4.8 shows the generic influence maximization model in which an unweighted social graph is fed as input to the model.
- The social graph contains past action propagation traces which are then used by an influence diffusion technique to learn the weights of each edge. Now the unweighted social graph is converted to a weighted graph.
- A weighted graph is again provided to a standard influence maximization algorithm to generate the seed set. This seed set is considered as output for the entire influence maximization model.

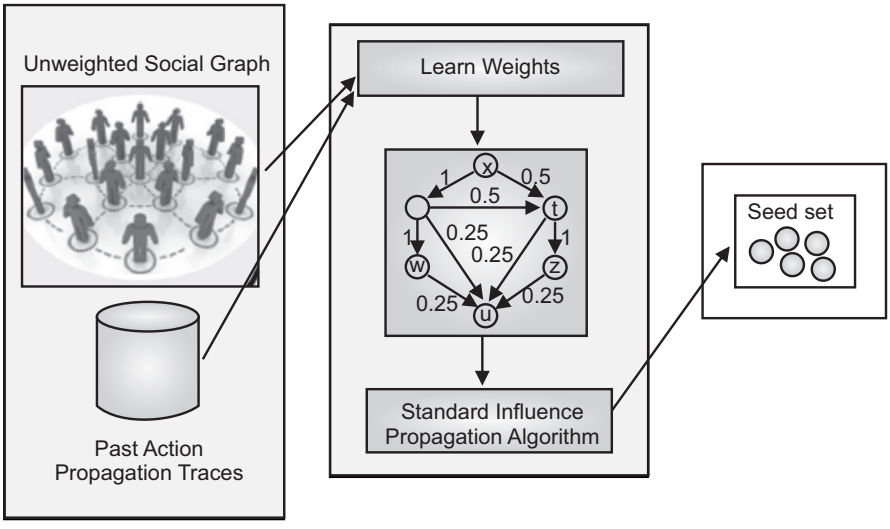


Fig. 4.8: The Generic Influence Maximization Framework

- There are number of standard information diffusion models that can be used to assign weights to the un-weighted social graph.
- Some of information diffusion models include the Independent Cascade (IC) model, the Linear Threshold (LT) model, and the Weighted Cascade (WC) model.
- In general, an information diffusion model considers the entire social media network as a graph G consisting of vertices (or nodes) and edges (or connection between nodes).
- This graph can be represented as G = (V, E) where, V denotes the vertices of the social network and E denotes the edges between the nodes.
- The influence diffusion model to be used for influence maximization is chosen based on the nature of the complexity of the social networking site.
- In real cases, it is not easy to decide which diffusion model will work the best for a particular social media network and a standard model is usually chosen to provide an optimum result.

4.3.4 Expert Finding

- The task of expert finding, as one of the most important research issues in social networks. The expert finding is aimed at identifying persons with relevant expertise or experience for a given topic.
- Some social media networking sites provide a pool of experts for certain topics and discussions. The expert finding is the process of generating and grouping experts of a social network based on his/her expertise on certain topics.
- For doing so, the main task adapted in expert findings in social media analytics is to retrieve a ranked list of top-N experts who are well conversant on a given specified topic.
- The basic idea behind finding such experts is to take their help in questions answering and problem solving.
- Such experts are a blessing in disguise for the information retrieval community for sharing the expertise on various topics that need discussion and learning.
- Expert finding is concerned about finding persons who are knowledgeable on a given topic.
- Fig. 4.9 shows an example of an expert finding model that considers a set of users and a specific task as input.
- The task usually consists of a set of skills, which in the given Fig. 4.9 is the set semantic Web, Machine Learning (ML) and Natural Language Processing (NPL).
- A standard expert finding model in social media analytics is used for generating the results and the skills are highlighted (in box of Fig. 4.9) and the corresponding expert name is also reflected based on his/her proficiency in at least one out of the three mentioned skills.

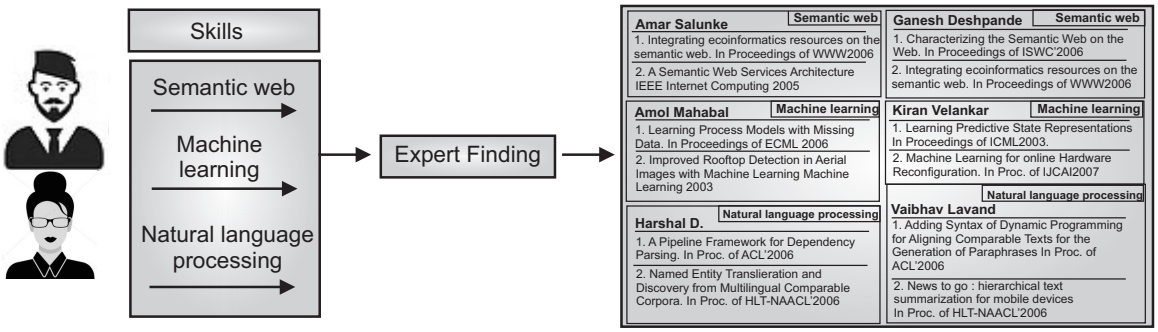


Fig. 4.9: An Example of Expert Finding Process

- An expert finding model in social media analytics considers as input a group of user's μ and a task T which consists of a set of skills S.
- The main task in expert finding in social media analytics is to find individual nodes (or users) U' μ U that has at least one essential skill S' S, where S μ T. In this way, a team of experts is formed for each skill of a task.
- The expert finding in social media analytics has gained a lot of importance in the research community as it helps in forming communities or collaborations in social networks based on the expert domain.

4.3.5 Prediction of Trust and Distrust among Individuals

- With the development of the Internet, more and more individuals or organizations tend to communicate and interact on the social network platform.
- Through social network platforms, users can not only share their own feelings about different products but also express their views on others, which greatly enriches user's social activities.
- In some online social network services, users are allowed to label users with similar interests as "trust" to get the information they want and use "distrust" to label users with opposite interests to avoid browsing content they do not want to see.
- Social media networks are growing at an exponential rate, and in each year more and more users are joining the network and contributing some activities and content into the network.
- This remarkable growth in the number of users for social media network usage has indirectly raised a question of trust and distrust among the connected users.
- Trusted users in social media networks spread the right information and positive effects on a social network.
- However, the distrusted users in social media networks pose a threat to the social network as there is a likelihood that such users may cause a disturbance or threat in the near future.
- As every social media networking site wants to build a reputed network that can be fully trusted by users, it has become essential to trace such individual users that have a likelihood of conducting harmful or mischievous online activities in the near future.

- For maintaining the reputation of a social media network, it is important to partition users into two groups - trusted users and distrusted users of a social network.
- This will help in preventing any kind of malicious activity occurring via the social media network and in turn, will bring an increase in the number of users joining the particular networking site based on the level of trust.
- Fig. 4.10 shows a simple example of generating trusted users of a social media network.
- The social network in Fig. 4.10 consists of four nodes (or users) and the solid line connectivity shows the trust between nodes (or users).
- For example, in Fig. 4.10 node A trusts node Y and node D while node B trusts node Y. This indicates that there is a strong likelihood that node B will also trust node X (shown with dotted line).
- The detection of trust is usually carried out in a backward-forward process as can be seen how node B trusts node Y and then going backward direction to node A and then in the forward direction to node X.
- Finding distrusted users is a challenging and complex problem compared to finding trusted users in a network.
- This is because the trust factor is considered transitive, i.e., if user A trusts user B and user B trusts user Y, then it can be concluded that there is a maximum likelihood that user A also trusts user Y.
- However, the distrust factor cannot be considered transitive. This is so because if user A distrusts user B, and user B distrusts user Y, this does not guarantee that user A will distrust user Y.

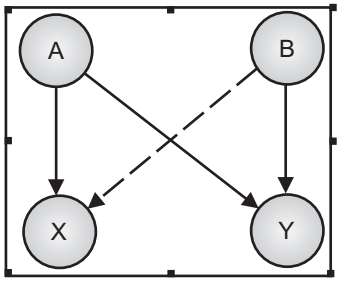


Fig. 4.10: An example of Direct Trust Propagation

4.4 INTRODUCTION TO NATURAL LANGUAGE PROCESSING

- Language is the basic and primary means of communication used by humans. Language is the tool we use to express our ideas, thoughts and emotions.
- Natural Language Processing (NLP) is concerned with the development of computational models of aspects of human language processing.
- Natural Language Processing (NLP) is the convergence between linguistics, computer science and Artificial Intelligence (AI).

- NLP mainly aims for the interconnection between natural languages and computers that means how to analyze and model a high volume of natural language data.
- NLP will ever remain a standard requirement in the field of data science. NLP focuses on bridging the gap between human communication and computer understanding.
- NLP is a collection of techniques for working with human language. Examples would include flagging e-mails as spam, using Twitter to assess public sentiment and finding which text documents are about similar topics.
- Natural Language Understanding helps machines “read” text (or another input such as speech) by simulating the human ability to understand a natural language such as English.
- NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.
- The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.
- The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.
- In recent years, online social networking has revolutionized interpersonal communication.
- The newer research on language analysis in social media has been increasingly focusing on the latter's impact on our daily lives, both on a personal and a professional level. NLP is one of the most promising avenues for social media data processing.
- NLP is a scientific challenge to develop powerful methods and algorithms which extract relevant information from a large volume of data coming from multiple sources and languages in various formats.
- The field of NLP is found to be highly beneficial for resolving ambiguity in the various languages spoken worldwide and is a key area of study for text analytics as well as speech recognition.
- NLP is the technology that is used by machines to understand, analyze, manipulate, and interpret human's languages.
- NLP, as an important branch of data science, plays a vital role in extracting insights from the input text. Industry experts have predicted that the demand for NLP in data science will grow immensely in the years to come.
- One of the key areas where NLP is playing a pivotal role in data science is while dealing with multi-channel data like mobile data or any social media data.
- Through the use of NLP, these multichannel data are being assessed and evaluated to understand customer sentiments, moods, and priorities.
- Language is a method of communication with the help of which we can speak, read and write.

- NLP is a subfield of Computer Science that deals with Artificial Intelligence (AI), which enables computers to understand and process human language.
- Text processing has a direct application to Natural Language Processing, also known as NLP.
- NLP is aimed at processing the languages spoken or written by humans when they communicate with one another.
- Fig. 4.11 shows the phases or logical steps in natural language processing.

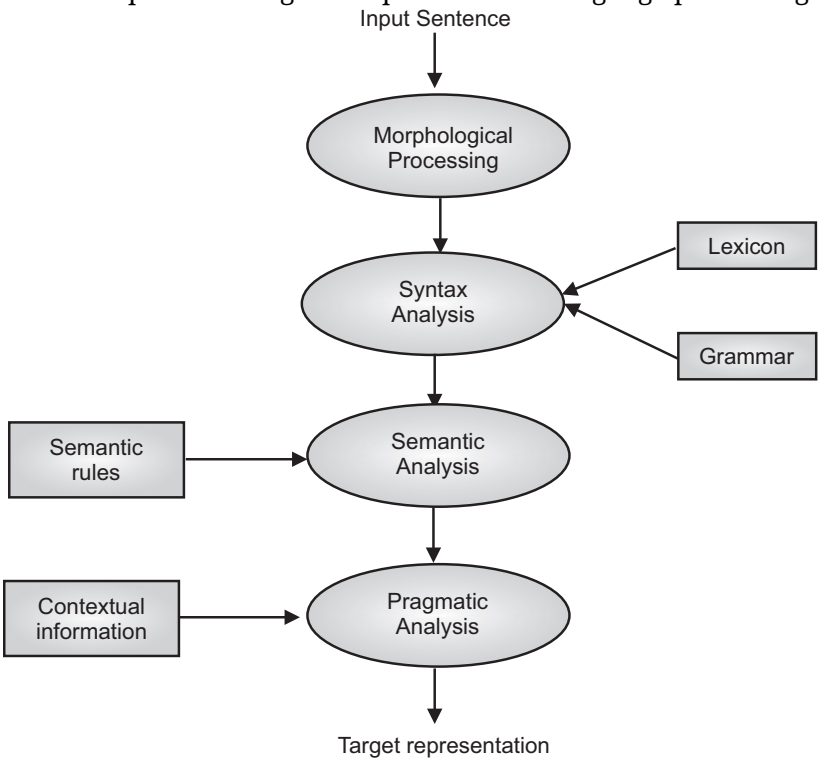


Fig. 4.11

- Let us discuss the phases or logical steps in natural language processing:
  1. Morphological Processing is the first phase of NLP. The purpose of this phase is to break chunks of language input into sets of tokens corresponding to paragraphs, sentences and words. For example, a word like “uneasy” can be broken into two sub-word tokens as “un-easy”.
  2. Syntax Analysis is the second phase of NLP. The purpose of this phase is two folds: to check that a sentence is well formed or not and to break it up into a structure that shows the syntactic relationships between the different words. For example, the sentence like “The school goes to the boy” would be rejected by syntax analyzer or parser.

3. Semantic Analysis is the third phase of NLP. The purpose of this phase is to draw exact meaning, or you can say dictionary meaning from the text. The text is checked for meaningfulness. For example, semantic analyzer would reject a sentence like "Hot ice-cream".
4. Pragmatic Analysis is the fourth phase of NLP. Pragmatic analysis simply fits the actual objects/events, which exist in a given context with object references obtained during the last phase (semantic analysis). For example, the sentence "Put the banana in the basket on the shelf" can have two semantic interpretations and pragmatic analyzer will choose between these two possibilities.

- There are two very different schools of thought in NLP, namely, statistical NLP and linguistic NLP.
  - The statistical school of NLP solves this problem by using massive corpuses of training data to find statistical patterns in language.
  - The linguistic school focuses on understanding language as language, with techniques such as identifying which words are verbs or parsing the structure of a sentence.

Examples of NLP Applications:

- Today, NLP an emerging technology, derives various forms of AI we used to see these days.
- For today's and tomorrow's increasingly cognitive applications, the use of NLP in creating a seamless and interactive interface between humans and machines will continue to be a top priority.
- Some common applications of NLP are explained below:
  1. Automatic text summarization is a technique which creates a short, accurate summary of longer text documents.
  2. Machine Translation (MT) is basically a process of translating one source language or text into another language
  3. Spelling correction and grammar correction is a very useful feature of word processor software like Microsoft Word. NLP is widely used for this purpose.
  4. Question-answering, another main application of natural language processing (NLP), focuses on building systems which automatically answer the question posted by user in their natural language.
  5. Sentiment analysis is among one other important applications of NLP. Online E-commerce companies like Amazon, ebay, etc., are using sentiment analysis to identify the opinion and sentiment of their customers online.
  6. Speech engines like Siri, Google Voice, Alexa are built on NLP so that we can communicate with them in our natural language.

4.5 TEXT ANALYTICS

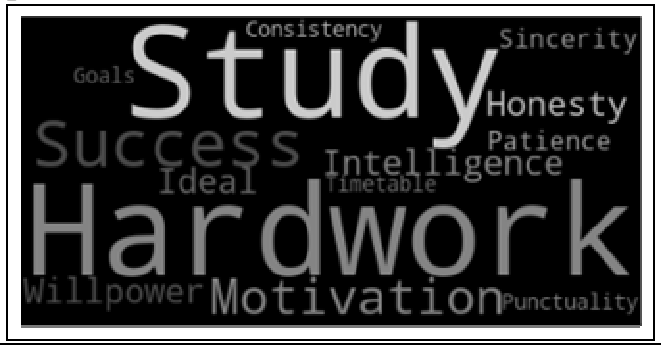
- Text analytics combines a set of machine learning, statistical and linguistic techniques to process large volumes of unstructured text or text that does not have a predefined format, to derive insights and patterns.
- Text mining (also referred to as text analytics) is an artificial intelligence (AI) technology that uses NLP to processes the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning algorithms.
- Following program shows the interesting use of text analytics, it takes sam.txt file as input and on the basis of text frequency , display word cloud:

Sam.txt file:

Hardwork
Sincerity
Success
Intelligence
Study
Success
Honesty
Patience
HardWork
Success
Study
Study
Hardwork
Intelligence
Hardwork
Consistency
Motivation
Ideal
Honesty
Punctuality
Hardwork
Willpower
Motivation
Goals
Study
Timetable
Willpower
Ideal
Study
Motivation

```
# importing the necessary modules:
import numpy as np
import pandas as pd
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import matplotlib.pyplot as plt
from PIL import Image
with open("/content/sam.txt") as f:
   text=f.read()
#text = open("sam.text").read()
wordcloud = WordCloud().generate(text)
# Display the generated image:
plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```

Output:



- Pre-processing of text documents play vital role for any Text analytics based applications.
- The pre-processing steps have enormous consequence on the quality knowledge output during text analytics.
- As the applicability of phrase “Garbage-in, garbage-Out” on each text analytics process, Input for text data supposed to be improved to enhance quality of text analytics.
- Although the applying of good text analytics algorithm, but availability of redundant, irrelevant and noisy input data hamper the quality output of text analytics.
- So preparation and filtration of raw input data makes the sense before inputting it to the text analytics process.
- In general text pre-processing steps include Tokenization, Bag of words, Word weighting: TF-IDF, n-Grams, stop words, Stemming and lemmatization, synonyms and parts of speech tagging.

4.5.1 Tokenization

- The first part of any NLP process is simply breaking a piece of text into its constituent parts (usually words) this process is called tokenization.
- This extracts unwanted element from a text document, converts all the letters to lower case, and removes punctuation marks.
- The output of tokenization is a representation of the text document as a stream of terms.
    - **Convert Lower Case:** Each term of the document is covert into the lower case so that words like "hello" and "Hello" are reflected the same term for analysis.
    - **Remove White Space, Numbers and Punctuations:** White space, numbers and punctuation marks like ",", "?" etc. are remove from text documents as they have no significant contribution for clustering.
- In tokenization the text is cut into pieces called "tokens" or "terms." These tokens are the most basic unit of information we will use for yhe model.

4.5.2 Bag of Words

- Once the tokenization is done, based on the process we may end up with a "bag of words."
- Bag of words is essentially a vector for the document/sentence telling us to whether a specific word is there or not and how many times.
- The most basic concept in NLP (aside from some very high-level applications of it) is that of a "bag-of-words," also called a frequency distribution.
- It is a way to turn a piece of free text (a tweet, a Word document, or whatever else) into a numerical vector that we can plug into a machine learning algorithm.
- The idea is quite simple – there is a dimension in the vector for every word in the language and a document's score in the nth dimension is the number of times the nth word occurs in the document.
- The piece of text then becomes a vector in a very high-dimensional space. Tokenization setting splits the documents into words/terms, constructing a word vector known as Bag-of-Words (BoW).
- The BoW one of the simplest models in NLP, is used to extract the features from piece of text or document so that it can be used in modeling such that in ML algorithms.
- The BoW basically, constructs a word presence feature set from all the words of an instance.
- The majority of this section will be about extensions and refinements of the basic idea of bag-of-words – we will discuss some of the more intricate (and error-prone) sentence parsing toward the end.

- Right off the cuff, we might want to consider the following extensions to the word vector:
  1. There are a staggering number of words in English and an infinite number of potential strings that could appear in text. We need some way to cap them off.
  2. Some words are much more informative than others – we want to weight them by importance.
  3. Some words don't usually matter at all. Things such as "I" and "is" are often called "stop words," and we may want to just throw them out at the beginning.
  4. The same word can come in many forms. We may want to turn every word into a standardized version of itself, so that "ran," "runs," and "running" all become the same thing. This is called "lemmatization."
  5. Sometimes, several words have the same or similar meanings. In this case, we don't want a vector of words so much as a vector of meanings. A "synset" is a group of words that are synonyms of each other, so we can use synsets rather than just words.
  6. Sometimes, we care more about phrases than individual words. A set of *n* words in order is called an "*n*-gram," and we can use *n*-grams in place of words.
- NLP is a deep, highly specialized area. If we want to work on a cutting-edge NLP project that tries to develop real understanding of human language, the knowledge we will need goes well outside the scope.
- However, simple NLP is a standard tool in the data science toolkit and unless we end up specializing in NLP, this chapter should give us what we need. When it comes to running code that uses bag-of-words, there is an important thing to note.
- Mathematically, we can think of word vectors as normal vectors (an ordered list of numbers, with different indices corresponding to different words in the language.
- There is often no need to actually specify which words correspond to which vector indices and doing so would add a human-indecipherable layer in the data processing, which is usually a bad idea.
- In fact, for many applications, it is not even necessary to explicitly enumerate the set of all words being captured.
- This is not just about human readability: the vectors being stored are often quite sparse, so it is more computationally efficient to only store the nonzero entries.
- We might wonder why we would bother to think of a map from strings to floats as a mathematical vector. The reason is that vector operations, such as dot products, end up playing a central role in NLP.
- NLP pipelines will often include stages where they convert from map representations of word vectors to more conventional sparse vectors and matrices, for more complicated linear algebra operations such as matrix decompositions.

4.5.3 Word Weighting: TF-IDF

- To weight the terms, term frequency–inverse document frequency measure (TF-IDF) is used. The TF-IDF stands for the "Term Frequency-Inverse Document Frequency".
- According to the term frequency concepts, more occurrences of terms in the document specified more significance of the term for the document.
- Simultaneously, occurrence of any term in more documents indicates the less worth of its for a single document.
- The TF- IDF method helps to normalize the term frequency to control the relative effect of terms to specific document vs. its presence in entire corpus.
- The idea is, the more a term occurs in a document, the more contributing it is but may not have discriminative nature.
- The TF-IDF supports to condensed text dimensionality. The tf-idf weight of a term is the product of its tf weight and its idf weight.
- So in easy terms the TF-IDF is a numerical statistic that reflects how important a word is to a document in a corpus.
- The first correction to bag-of-words is the idea that some words are more important than others.
- In some cases, we know a priori which words to pay attention to, but more often we are faced with a corpus of documents and have to determine from it which words are worth counting in the word vector and what weight we should assign to them.
- So weight for any term *t* in a document *d* can be represented by $W_{t,d}$ as:

$$W_{t,d} = \log(1 + tf_{t,d}) \times \log_{10}\left(\frac{N}{df_t}\right) \quad \ldots (4.1)$$

where,

$tf_{t,d}$ is the term frequency of term t in document d.

N is the number of documents that contain t.

df_t is the document frequency of t.

4.5.4 n-Grams

- An n-gram means a sequence of n words. An n-gram is a piece of text containing M words can be broken into a collection of M − n + 1 n-grams, as shown in Fig. 4.12 (contains 2-grams).
- We can create a bag-of-words out of n-grams, run TF-IDF on them, or model them with a Markov chain, just as if they were normal words. The problem with n-grams is that there are so many potential ones out there.
- The problem with n-grams is that there are so many potential ones out there. Most n-grams that appear in a piece of text will occur only once, with the frequency decreasing, the larger n is.

- The general approach here is to only look at n-grams that occur more than a certain number of times in the corpus.

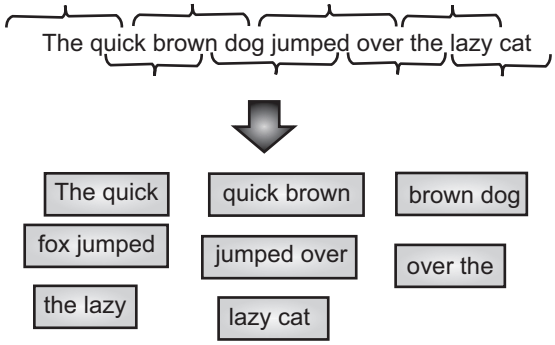


Fig. 4.12

- In NLP for short, n-grams are used for a variety of things. Some examples include auto completion of sentences (such as the one we see in Gmail these days), auto spell check (yes, we can do that as well), and to a certain extent, we can check for grammar in a given sentence.

Basic Concept of n-gram with Example:

- An n-gram is a contiguous sequence of n items from a given sequence of text or speech.
- The items can be phonemes, syllables, letters, words or base pairs according to the application.
- The n-grams typically are collected from a text or speech corpus. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram"; size 3 is a "trigram".
- In this project we will use 2-gram(bigram), 3-gram(trigram), and 4-gram (quadgram) in the prediction.

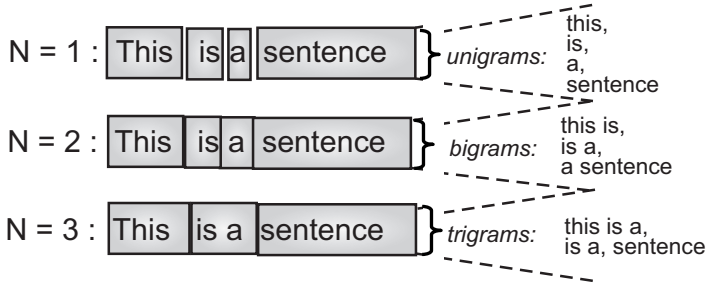


Fig. 4.13

4.5.5 Stop Words

- Basically the words which are excessively frequent in the corpus can be eliminate because they are typically common words such as "a", "an", "the", or "of".
- These terms are not discriminative from a Text analytics perception. Such words are also referred to as stop words.

- So stop words is a list of words that doesn't have the potential to contribute characterize the content in the text.

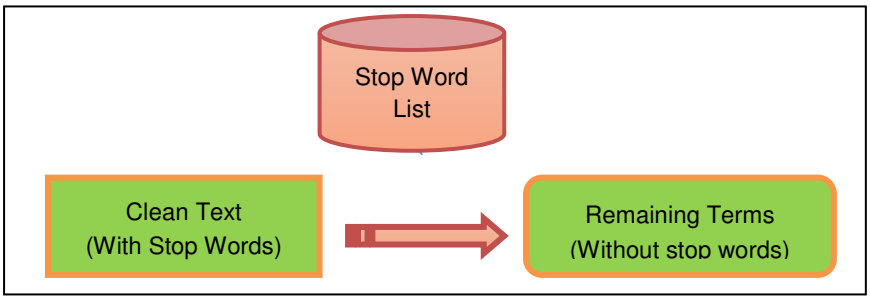


Fig. 4.14: Stop Word Removing Process

- Stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc.
- Stop words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information.
- Stop words are any word in a stop list (or stop-list or negative dictionary) which is filtered out (i.e. stopped) before or after processing of natural language data (text).[
- The Bag-of-words, TF-IDF, and n-grams are fairly general processing techniques, which can be applied to many other areas.
- The simplest version is to remove what are called “stop words.” These are words such as “the,” “it,” and “and” that aren’t really informative in and of themselves.
- They are critically important if you’re trying to parse the structure of a sentence, but for bag-of-words, they are just noise.
- There is no absolute definition of “stop words.” Frequently, they are found by taking the most common words in a corpus, then going through them by hand to determine which ones aren’t really meaningful.
- In other cases (such as the aforementioned example script), there is a list of them prebuilt into an NLP library.
- Stop words become problematic when you are using n-grams. For example, the very informative phrase “to be or not to be” is likely to get stopped out entirely!

4.5.6 Stemming and Lemmatization

- Stemming and Lemmatization have been developed in the 1960s. These are the text normalizing and text mining procedures in the field of NLP that are applied to adjust text, words, documents for more processing.

  1. **Stemming:**
     - It is the process of reduces the terms to their linguistic root, to obtain the index terms. According to English grammar, words found in text documents can have different forms of a word, such as organize, organization, and organizing.

- However, if searching query has word organize then all the documents which contain any form of the words are expected to be match.
- A word can be view as combination of Stem and affixes. So such query required that affixes should be removing and convert words to their stem form.
- Stemming generally refers to an experiential process which works for removal of derivational affixes. So, Stemming is the simplified form of morphological analysis, which just find the stem of the word.
- Text miners are working with many stemming algorithm like Porter stemmer, Snowball stemmer etc.
- Stemming is a rather crude method for cataloging related words; it essentially chops off letters from the end until the stem is reached.
- This works fairly well in most cases, but unfortunately English has many exceptions where a more sophisticated process is required.

2. Lemmatization:

- Lemmatization technique is like stemming. The output we will get after lemmatization is called 'lemma', which is a root word rather than root stem, the output of stemming.
- After lemmatization, we will be getting a valid word that means the same thing. A method that switches any kind of a word to its base root mode is called Lemmatization.
- In contrast to stemming, lemmatization looks beyond word reduction and considers a language's full vocabulary to apply a morphological analysis to words. The lemma of 'was' is 'be' and the lemma of 'mice' is 'mouse'.
- Lemmatization is typically seen as much more informative than simple stemming. Lemmatization looks at surrounding text to determine a given word's part of speech, it does not categorize phrases.

Difference between Stemming and Lemmatization:

Sr. No.	Stemming	Lemmatization
1.	Stemming is faster because it chops words without knowing the context of the word in given sentences.	Lemmatization is slower as compared to stemming but it knows the context of the word before proceeding.
2.	It is a rule-based approach.	It is a dictionary-based approach.
3.	Accuracy is less.	Accuracy is more as compared to Stemming.

contd. …

4.	When we convert any word into root-form then stemming may create the non-existence meaning of a word.	Lemmatization always gives the dictionary meaning word while converting into root-form.
5.	Stemming is preferred when the meaning of the word is not important for analysis. Example: Spam Detection	Lemmatization would be recommended when the meaning of the word is important for analysis. Example: Question Answer
6.	For example: "Studies" => "Studi"	For example: "Studies" => "Study"

- Stemming and Lemmatization are broadly utilized in Text mining where Text Mining is the method of text analysis written in natural language and extricate high-quality information from text.

4.5.7 Synonyms

- Generally, all Natural language processing (NLP) application deals with text frequency analysis and text indexing.
- In such text analysis activities it is always useful to compress the vocabulary without losing meaning because it saves lots of memory. To accomplish this, we must have to define mapping of a word to its synonyms.
- Intuitively, when we work with text analysis, the words themselves are less important than the "meaning."
- This suggests that when we might want to collapse related terms such as "big" and "large" into a single identifier. These identifiers are often called "synsets" for sets of synonyms.
- Most of the NLP packages utilized synsets as a major component of their ability to understand a piece of text and making semantic relations among the text.
- The simplest use of synsets is to take a piece of text and replace every word with its corresponding synset.
- Ultimately, we think of synsets as constituting the vocabulary of a "clean" language, one in which there is a one-to-one matching between meanings and words.
- Replacing word with synset usually deal with some major problems because in general, a word can belong to several synsets, because a word can have several distinct meanings. So doing a translation from the original language to "synset-ese" is not always feasible.

4.5.8 Parts of Speech Tagging

- As we move from purely computational techniques toward ones based more closely on language, the next stage is Part-Of-Speech tagging (POS tagging).

- Part-of-Speech (POS) tagging, it may be defined as the process of converting a sentence in the form of a list of words, into a list of tuples.
- In POS tagging the tuples are in the form of (word, tag). We can also call POS tagging a process of assigning one of the parts of speech to the given word.
- So POS tagging is the process of assigning a part-of-speech such as noun, verb, pronoun, prepositions, adverbs and adjective to each word in a sentence.
- The tagging algorithm takes sequence of words of a natural language sentence and specified tag sets (a finite list of part-of-speech tags).
- This can be done in NLTK (Natural Language Toolkit (is a Python package/module that we can use for NLP)) as follows:

```
>>> nltk.pos_tag(["I", "eat", "mango"])
[('I', 'PRP'), ('eat', 'VBP'), ('mango', 'NN')]
```

- In this case, the PRP tag tells us that “I” is a prepositional phrase, “drink” is a verb phrase, and “milk” is a common noun in singular form. A complete list of the POS tags nltk uses can be seen by calling,

```
>>> nltk.help.upenn_tagset()
```

- Following example shows the frequency distribution plot generated for all words in the input text.
- The output displays the input text, the tokenized words of the text, the total frequency distribution of all words, and the frequency distribution of five words having the highest frequency.
- Also, a list of stop words is displayed in the output which is then compared with the input text to form two lists – one having stop words and another having no stop words.
- Lastly, the POS tagging is carried out on the input text and the grammatical group of each word is accordingly displayed in the output.

```
#Program: Python Code for Performing Text Analytics
#Importing necessary library
import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('averaged_perceptron_tagger')
import matplotlib.pyplot as plt
# sample text for performing tokenization
sample_text = ' It was cute. I looked cute, because it was my Birthday.
Birthdays are for being happy and making happy to others also.'
print('The Text Is : ',sample_text)
# importing word_tokenize from nltk
```

```
from nltk.tokenize import word_tokenize
# Passing the text into word tokenize for breaking the sentences
tkn = word_tokenize(sample_text)
print('The Individual Words in the Text Is : ', tkn)
# Importing FreqDist library from nltk and passing token into FreqDist
from nltk.probability import FreqDist
# finding the frequency distribution of the tokenized text
frqdist = FreqDist(tkn)
print(' Total Frequency Distribution : ', frqdist)
# To find the frequency of top 5 words in the text
frqdist1 = frqdist.most_common(5)
print('Frequency Distribution of Top 5 Words : ', frqdist1)
# Frequency Distribution Plot
frqdist.plot(20,cumulative=False,title="Frequency Distribution Plot")
plt.show()
# Generating List of Stopwords
from nltk.corpus import stopwords
stop_words=set(stopwords.words('english'))
print(' List of Stopwords : \n', stop_words)
# Segregating Words from the text which are Stopwords and
                                                        not Stopwords
s1=[]
s2=[]
for w1 in token:
if w1 not in stop_words:
s1.append(w1)
print(' Filtered Sentence:', s1)
```

Output:

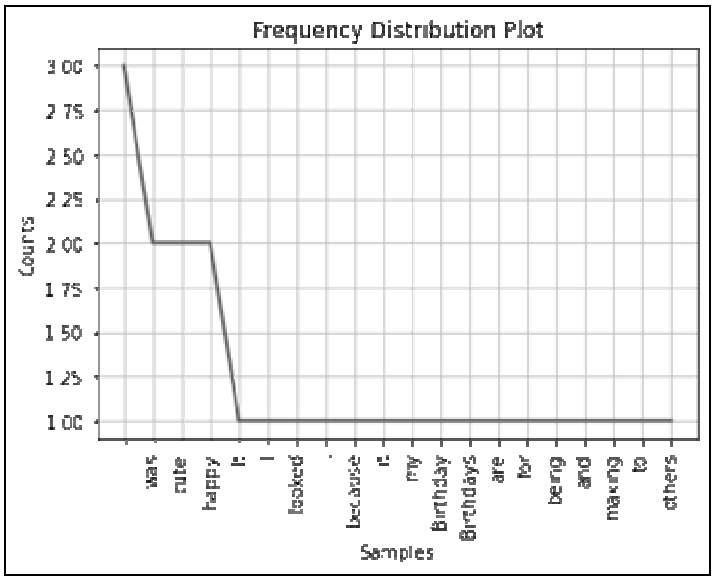
```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

The Text Is : It was cute. I looked cute, because it was my Birthday. Birthdays are for being happy and making happy to others also.

The Individual Words in the Text Is : ['It', 'was', 'cute', '.', 'I', 'looked', 'cute', ',', 'because', 'it', 'was', 'my', 'Birthday', '.', 'Birthdays', 'are', 'for', 'being', 'happy', 'and', 'making', 'happy', 'to', 'others', 'also', '.']

Total Frequency Distribution : <FreqDist with 21 samples and 26 outcomes>

Frequency Distribution of Top 5 Words : [('.', 3), ('was', 2), ('cute', 2), ('happy', 2), ('It', 1)]



List of Stopwords :

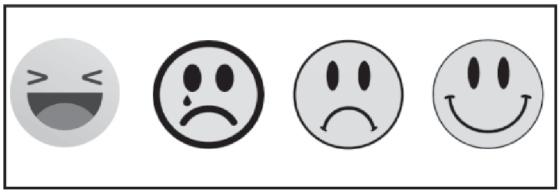
{"didn't", "wouldn't", "she's", 'too', 'their', 'is', 'off', 'them', 'doing', 'themselves', 'again', 'any', 'out', 'mustn', 'a', 'her', 'how', "you've", 'mightn', 'itself', 'not', 'such', 's', 'from', 'which', 'those', 'shan', 'yours', 'i', 'him', 'there', 'your', 'below', "shan't", 't', 'myself', 'his', 'of', 'hasn', 'having', 'in', 'had', 'so', 'have', 'other', "won't", 'aren', "that'll", 'm', "it's", "you're", "wasn't", 'should', 'weren', "couldn't", 'did', 'down', 'are', 'has', 'y', 'was', 'over', "should've", 'wouldn', "hadn't", "mustn't", 'no', 'yourselves', 'what', "hasn't", 'the', 'between', "don't", 'nor', 'do', 'ours', 'through', 'each', 'you', 'can', 'me', 'it', 'ain', 'once', 'to', 'when', 'd', 'because', 'she', 'or', 'here', 'couldn', 'been', 'won', "you'll", 'with', 'at', 'all', 'further', 'herself', 'then', 'but', 'hadn', 'most', 'both', 'until',

```
"isn't", 'these', 'ourselves', 'by', 'just', 'this', "you'd",
'himself', 'whom', 'hers', 'don', 'my', 'only', 'than', 'our',
'shouldn', 'who', 'were', "doesn't", 'its', 'as', 'and', 'we',
'against', 'wasn', 'into', 'll', 'for', 'about', 'very', 'yourself',
'ma', "aren't", 'didn', 'up', 'own', 'while', 'more', "weren't",
'isn', 'o', 'am', 'now', 'needn', 'they', 'be', 'same', 'doesn', 'on',
'he', "mightn't", 'does', 'during', 'where', 'an', 'under', 'theirs',
'that', 'being', 're', 'if', "shouldn't", 'some', 'above', 'why',
"haven't", 've', 'few', 'will', 'haven', 'after', 'before', "needn't"}
 Filtered Sentence: ['It', 'cute', '.', 'I', 'looked', 'cute', ',',
'Birthday', '.', 'Birthdays', 'happy', 'making', 'happy', 'others',
'also', '.']
 Tokenized Sentence: ['was', 'because', 'it', 'was', 'my', 'are',
'for', 'being', 'and', 'to']
Part-Of-Speech Tagging :
 [('It', 'PRP'), ('was', 'VBD'), ('cute', 'JJ'), ('.', '.'), ('I',
'PRP'), ('looked', 'VBD'), ('cute', 'NN'), (',', ','), ('because',
'IN'), ('it', 'PRP'), ('was', 'VBD'), ('my', 'PRP$'), ('Birthday',
'NNP'), ('.', '.'), ('Birthdays', 'NNS'), ('are', 'VBP'), ('for',
'IN'), ('being', 'VBG'), ('happy', 'JJ'), ('and', 'CC'), ('making',
'VBG'), ('happy', 'JJ'), ('to', 'TO'), ('others', 'NNS'), ('also',
'RB'), ('.', '.')]
```

4.5.9 Sentiment Analysis

- Sentiment analysis is used to identify the emotions conveyed by the unstructured text. Sentiment analysis basically means to analyze and find the emotion or intent behind a piece of text or speech or any mode of communication.
- Sentiment analysis is one of the NLP techniques, which can be used to determine the sensibility behind the texts, i.e. tweets, movie reviews, product reviews, customer interactions, YouTube comments, forum discussion, blogs or any incoming message, etc.
- Generally, such reactions are taken from social media and clubbed into a file to be analyzed through NLP.
- A prominent social media site often used for sentiment analysis is Twitter where users can tweet messages to express their opinion on any topic or person.
- Sentiment analysis is used on the web to analyze the attitude, behavior and emotional state of the sender.
- The application is implemented through a combination of NLP and statistics by assigning the values to the text (positive, negative or natural),

- There are more complicated versions of sentiment analysis that can, for example, determine complicated emotional content such as anger, fear, and elation.
- But the most common examples focus on "polarity," where on the positive–negative continuum a sentiment falls.



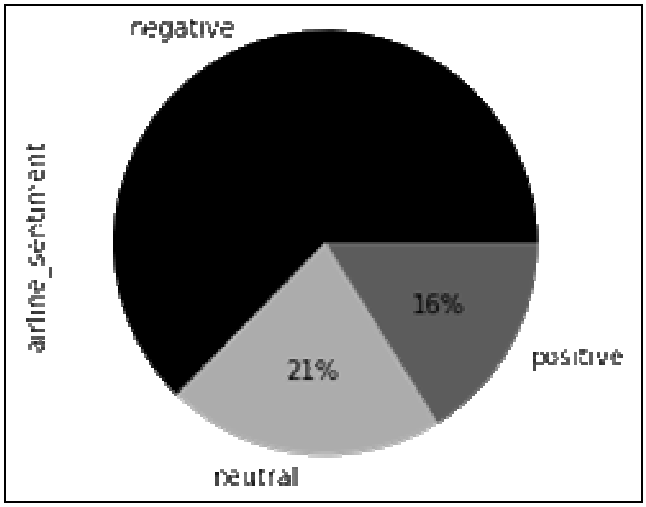
- The following program of sentiment analysis, takes Airline tweets records as input. Program categorized tweets into positive, negative and neutral category and displays it with the help of pie and Bar chart.

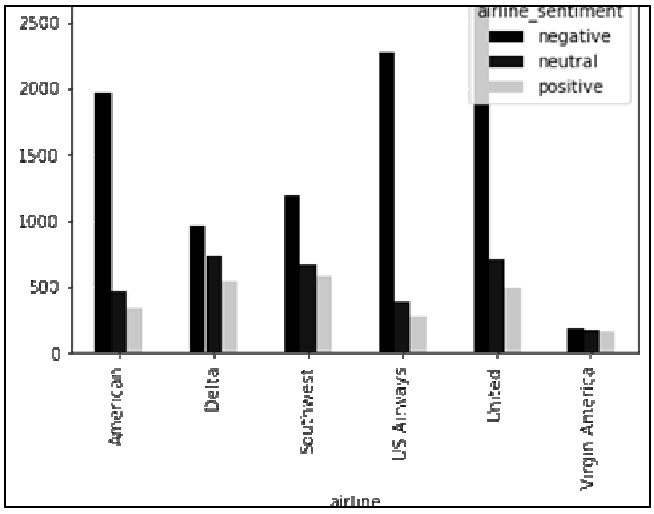
```
import numpy as np
import pandas as pd
#import nltk
import matplotlib.pyplot as plt
data_source_url = "https://raw.githubusercontent.com/kolaveridi/kaggle-
                  Twitter-US-Airline-Sentiment-/master/Tweets.csv"
airtweets = pd.read_csv(data_source_url)
plsize = plt.rcParams["figure.figsize"]
plt.rcParams["figure.figsize"] = plsize
print("Data in CSV file")
print(airtweets.head())
print("Distribution of sentiments across all tweets ")
airtweets.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.
                          0f%%', colors=["Black", "Orange", "green"])
airline_sentiment = airtweets.groupby(['airline', 'airline_sentiment']).
                                  airline_sentiment.count().unstack()
airline_sentiment.plot(kind='bar',color=['black', 'blue', 'cyan'])
```

Output:

```
Data in CSV file
tweet_id  ...               user_timezone
0  570306133677760513  ...  Eastern Time (US & Canada)
1  570301130888122368  ...  Pacific Time (US & Canada)
2  570301083672813571  ...  Central Time (US & Canada)
3  570301031407624196  ...  Pacific Time (US & Canada)
4  570300817074462722  ...  Pacific Time (US & Canada)

[5 rows x 15 columns]
Distribution of sentiments across all tweets
<matplotlib.axes._subplots.AxesSubplot at 0x7f93e058d950>
```





- Simple sentiment analysis is often done with manual lists of keywords. If words such as “bad” and “terrible” occur a lot in a piece of text, it strongly suggests that the overall tone is negative. That is what above stock market articles analysis program did.
- Slightly more sophisticated versions are based on plugging bag-of-words into machine learning pipelines.
- Commonly, we will classify the polarity of some pieces of text by hand and then use them as training data to train a sentiment classifier.
- This has the massive advantage that it will implicitly identify key words you might not have thought of and will figure out how much each word should be weighted.
- The extensions of the bag-of-words model, similar to n-grams, can also utilized to identify phrases such as “nose dive,” which deserve a very large weight in sentiment analysis but whose constituent words don’t mean much.

- There are mainly following three classification levels in sentiment analysis:
  1. **Document-level Sentiment Analysis**: In this level, the entire document is chosen as one basic information unit that aims to classify each document based on positive or negative opinions.
  2. **Sentence-level Sentiment Analysis**: In this level, the entire document is partitioned into sentence level. Each sentence of the document is then classified based on the positive or negative opinions it expresses.
  3. **Aspect-level Sentiment Analysis**: In this level, the entire document is classified based on specific aspects of entities. The most challenging task in this level is to identify the entities and the different aspects of each entity. For example, if a review of a product is given as, The mobile is heavy, but the navigation facility of the mobile is good, the entity mobile is being described both on the positive and the negative aspects.

4.5.10 Document or Text Summarization

- A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form.
- Text summarization involves generating a summary from a large body of text which somewhat describes the context of the large body of text.
- Text summarization in NLP is the process of summarizing the information in large texts for quicker consumption.
- Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.
- Document or text summarization is the technique of finding a precise summary of the bulk amount of data fed in as input to be able to interpret the key idea of the content of the text based on the generated summary.
- In short, text summarization technique allows for creating a much-shortened version of lengthy documents.
- If done manually, the entire process of text summarization will be a very complex and tedious task and sometimes maybe almost impossible within a short time frame.
- In such a case, one easy solution is to use machine learning algorithms that can be trained to identify the important sections of a document and accordingly produce a summary of the document.
- The method of extracting these summaries from the original huge/massive text without losing vital information is called as Text Summarization.
- There are mainly two types of text summarization approaches followed in text analytics, namely the extraction-based text summarization and the abstraction-based text summarization.

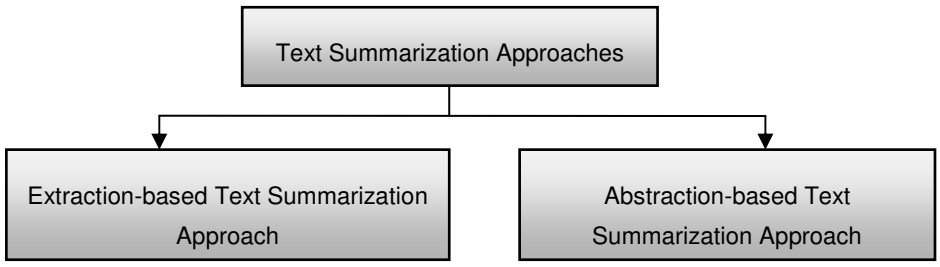


Fig. 4.15: Text Summarization Approaches

- The text summarization approaches in Fig. 4.15 are explained below:
  1. **Extraction-based text summarization** uses a machine learning approach to weigh the vital sections of sentences in the text and use the results to generate text summaries. Each section of sentences is analyzed to accordingly assign weights and then each section is ranked based on the importance and relevance of the text. The extracted sections are combined to ultimately form a text summary. The objective of extraction-based text summarization objective is to identify the significant sentences of the text and add them to the summary. An extraction summarization consists of selecting important sentences, paragraphs etc., from the original document and concatenating them into shorter form.
  2. **Abstraction-based text summarization** uses a deep learning approach to paraphrase a document to generate the text summary. The advantage of the abstraction-based text summarization approach compared to the extraction-based text summarization approach is that there are fewer grammatical inaccuracies found in the generated summarized textual output. The abstraction-based text summarization ensures that the core information is conveyed through shortest text possible.
- Text summarization has been found to be an immensely helpful technique in social media analytics as social media generates a bulk amount of resourceful data that needs to be often interpreted by a data analyst or a data scientist.
- Text summarization has recently gained much attention in the research world of social media analytics.
- It is a prominent field of study in machine learning that is often used for quickly assessing the text content.
- With time, more and more document summarization tool is expected to be built that can lead to more accurate and crisp results.
- In the following program the first step followed for text summarization is splitting paragraphs into sentences. Next step is from each sentence, all the special characters, extra spaces, stopwords and numbers are removed.
- Then, the sentences are tokenized for storing all the words in the sentences. The weighted frequency of each of the words is found by dividing the frequency of each word by the frequency of the most occurring word.

- Lastly in the following program, the weighted frequency is plugged in place of the corresponding words found in the original sentences. The sum of the weighted frequencies is then found and then the sentences are sorted and arranged in descending order of sum. The summarized text is finally displayed as output.

```
** Program for Text Summarization
import bs4 as bs
import urllib.request
import re
import nltk
import heapq
!pip install beautifulsoup4
!pip install lxml
nltk.download('punkt')
nltk.download('stopwords')
url_data=urllib.request.urlopen("https://en.wikipedia.org/wiki/
                                  Indian_Space_Research_Organisation")
article = url_data.read()
parsedart  = bs.BeautifulSoup(article,'lxml')
para = parsedart.find_all('p')
arttext = ""
for p in para  :
  arttext += p.text
# Removing Square Brackets and Extra Spaces
arttext  = re.sub(r'\[[0-9]*\]', ' ',arttext)
arttext = re.sub(r'\s+', ' ', arttext)
# Removing special characters and digits
formart_text = re.sub('[^a-zA-Z]', ' ', arttext )
formart_text = re.sub(r'\s+', ' ',formart_text)
# Tokenize sentences
slist = nltk.sent_tokenize(arttext)
stopwords = nltk.corpus.stopwords.words('english')
# Find the weighted frequency of occurrences of all words
word_frequencies = {}
for word in nltk.word_tokenize(formart_text):
  if word not in stopwords:
    if word not in word_frequencies.keys():
      word_frequencies[word] = 1
```

```
        else:
          word_frequencies[word] += 1
    # Calculating weighted frequency of each word
    max_frequncy = max(word_frequencies.values())
    for word in word_frequencies.keys():
      word_frequencies[word] = (word_frequencies[word]/max_frequncy)
    # Calculating Sentence Scores
    sentence_scores = {}
    for sent in slist:
        for word in nltk.word_tokenize(sent.lower()):
            if word in word_frequencies.keys():
                if len(sent.split(' ')) < 30:
                    if sent not in sentence_scores.keys():
                        sentence_scores[sent] = word_frequencies[word]
                    else:
                        sentence_scores[sent] += word_frequencies[word]
    summary_sentences = heapq.nlargest(7, sentence_scores,
                                                    key=sentence_scores.get)
    summary = ' '.join(summary_sentences)
    # Displaying the summarized text
    print('\n\n Summarized Text : \n', summary)
```

Output:

```
Summarized Text :
```

It is one of six government space agencies in the world which possess full launch capabilities, deploy cryogenic engines, launch extra-terrestrial missions and operate large fleets of artificial satellites. Polar Satellite Launch Vehicle or PSLV is the first medium-lift launch vehicle from India which enabled India to launch all its remote-sensing satellites into Sun-synchronous orbit. Parallelly, another solid fueled rocket Augmented Satellite Launch Vehicle based upon SLV-3 was being developed technologies to launch satellites into geostationary orbit. It undertakes the design and development of space rockets, satellites, explores upper atmosphere and deep space exploration missions. Augmented or Advanced Satellite Launch Vehicle (ASLV) was another small launch vehicle realised in 1980s to develop technologies required to place satellites into geostationary orbit. ISRO is the primary agency in India to perform

tasks related to space based applications, space exploration and development of related technologies. Alongside, technologies for Indian National Satellite System for communication satellites and Indian Remote Sensing Programme for earth observation satellites were developed and launches from overseas initiated.

4.5.11 Trend Analytics

- With trend analysis underlying idea is that machine learning which observing data of a given time period and this data can be used to predict the trend of a future.
- Trend analytics mainly involves determining the possible drifts or trends over a period of time.
- Usually, historical trends are analyzed to determine future trends for a given phenomenon or feature. So, trend analytics is used to predict future events.
- The main task in trend analytics is comparing data stored during a period of time to analyze and visualize the change of trend in this data with time.
- For example, predicting the stock market trends to allow users to understand and analyze in which company or organization it will be suitable to invest money.
- Market trend analysis is also an important concept adapted in businesses and by companies to analyze the possibility of profit in the business.
- In project management, trend analysis is a mathematical technique that uses past results to forecast future outcomes i.e., achieved by tracking various factors related to the project to control the quality of the project.
- Various tools are used for trend analysis ranging from simple linear based tools such as linear regression to many complex non-linear tools like the Mann-Kendall test.
- Trend analysis has proved to be very helpful in many applications, be it a study on climatic change or stock market trends.
- However, it should be noted that the results generated by trend analysis are not foolproof.
- This is so because the sample data considered for a fixed period of time is liable to sampling error as the entire census population is not considered for the study of the trends.
- Also, there might be an issue of measurement error or observation error which usually occurs while carrying out an experiment and gives a difference in result between the measured value and the actual value.
- There are three types of methods in trend analysis namely, temporal trend analysis, geographic trend analysis and intuitive trend analysis.

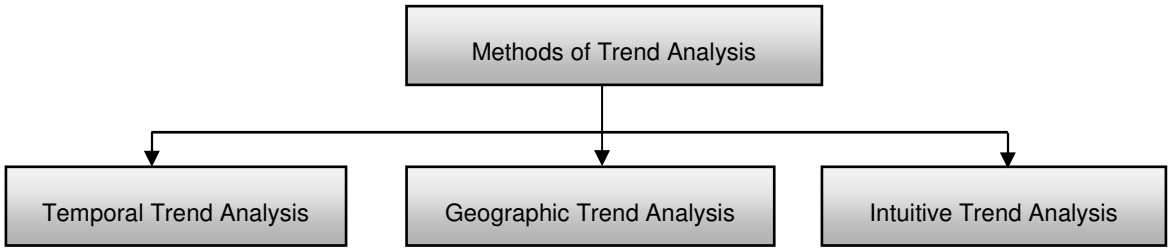


Fig. 4.16: Methods of Trend Analysis

- Each of these trend analysis methods is studied extensively for research to find optimum algorithms and techniques for the same.
- Let us now have a basic understanding of each of the above three trend analysis methods:

1. **Geographic Trend Analysis**:
   - This trend analysis is mainly involved in analyzing the trend of products, users or other elements within or across geographic locations.
   - This Geographic trend analysis helps in finding the pattern of trends within a specific geographic location.
   - Usually, this kind of trend analysis is limited to geographic boundaries and the trend is usually dependent on various geographic related factors such as culture, climate, food habits, etc.
   - While temporal trend analysis focuses more on predicting future events based on historic data, geographic trend analysis is limited to geography (a region, a city, a state, a country, etc.) and is comparatively easy to analyze and interpret.
2. **Intuitive Trend Analysis:**
   - This trend analysis is more often used when there is a lack of large statistical data required to carry out trend analysis.
   - In intuitive trend analysis, the data analyst or data scientist need to behave like a futurist and based on logical explanations and study of behavioral patterns, a prediction of future trends is made.
   - However, this method is prone to biases of the analyst and is comparatively difficult to analyze and interpret compared to temporal and geographic trend analysis. Trend analysis has paved its importance in the business, financial, and health sectors.
   - Trend analysis methods help data analysts to compare between firms and companies to analyze and understand the differences in the profit-making features of each workplace.
   - Trend analysis also helps data analysts to analyze and assess the short-term liquidity position and the long-term solvency position of an organization or company for a fixed period of time.

- Every business sectors nowadays follow trend analysis for prediction of its profitability in the upcoming years.
- This has made trend analysis a vital research topic to focus on generating more accurate predictions of the future.

3. **Temporal Trend Analysis:**
   - This trend analysis allows one to examine and model the change in the value of a feature or variable in a dataset over time.
   - The best example of temporal trend analysis is the time-series analysis which has been elaborately.
   - Time series analysis deals with statistical data that are placed in chronological order, that is, in accordance with time.
   - It deals with two variables, one being time and the other being a particular phenomenon.

- Time series can be constituted by three components namely short-term movement (periodic changes), long-term movement (secular trend), and random movement (irregular trend).
- Time series modeling and forecasting have vital significance in various practical analysis domains. A good amount of real research work is being done on a daily basis in this research area for several years.
- The collected data for time series analysis is considered as the past observations based on which forecasting is done for analyzing future trends.
- Various fitting models have been developed for time series forecasting that suits well for a particular area such as business, finance, engineering, and economics. One of the important characteristics of the time series is its stationarity.
- A time series is considered stationary if its behavior does not change over time. This indicates that the observed values vary about the same level, i.e., the variability is constant over time.
- In turn, the statistical properties of the stationary time series such as the mean and variance also remain constant over time.
- However, most of the time-series problems that are encountered are non-stationary. Non-stationary time series do not have a constant mean or variance and follows a trend by either drifting upward or backward.

4.6 CHALLENGES TO SOCIAL MEDIA ANALYTICS

- Social media has evolved over the last decade to become an important driver for acquiring and spreading information in different domains, such as business/ organizations, entertainment, science and so on.
- The enormous growth of social media usage has led to an increasing accumulation of data. The growth of social media usage opens up new opportunities for analyzing several aspects of, and patterns in communication.

- For example, social media data can be analyzed to gain insights into issues, trends, influential actors and other kinds of information.
- Social media analytics tools are used for gathering data from social platforms to help guide marketing strategies. Examples of some of the Social Media analytical tools include Keyhole, AgoraPulse, Google Analytics and many more.
- Social media data are governed by the properties of high volume, high velocity, as well as high variety.
- This makes social media complex to deal with though it has a plus point as it carries valuable insights that can be analyzed for fruitful decision-making.
- Some of the major challenges faced in dealing with such complex data in social media analytics are explained below:
  1. **Unstructured Data**: Unstructured Data (or unstructured information) refers to information that either does not have a predefined data model. Unstructured data is not an organized data. Social media data can be of various forms – textual (e.g., comments, tweets, etc.), graphical (e.g., images, videos, etc.), action-oriented (clicking like button, accepting friend-request, etc.), and relation-based (e.g., friends, followers, etc.). This makes social media data to be highly unstructured and poses a challenge to the data analyst or data scientist for intelligent decision-making. The unstructured and uncertain nature of this kind of big data presents a new kind of challenge: how to evaluate the quality of data and manage the value of data within data analytics. Shared files, images, videos, audio, comments, and messages are all unstructured data types.
  2. **High Volume and Velocity of Data:** The volume refers to the size of Data. Velocity refers to the speed at which the data is getting accumulated. Social media data gets generated every flicker of a second and capturing and analyzing data that is high in volume and velocity is a real challenge. Imagine the number of likes given in Facebook posts by users per second and the number of tweets posted by users on Twitter. Capturing and analyzing such bulk amounts of data requires special sophisticated tools that are often used by data analysts or data scientists to generate required results.
  3. **Diversity of Data**: Both social media users and the data these users generate are diverse. The users belong to various cultures, regions, and countries and the data generated are of various data types and multilingual. Not every data is crucial to be studied and analyzed. Finding and capturing only important content from such noisy diverse data is again a challenging and time-consuming task.
  4. **Organizational Level Issues;** Nowadays, many organizations are spending huge amount of money in developing their resources for collecting, managing and analyzing data sets from the social media. However, they do not clearly

understand how to ethically use social media analytics. Most organizations lack ethical data control practices like well-defined standards and procedures for sourcing, analyzing and sharing big data.

- To meet up the above-mentioned challenges related to social media data, a large number of tools have been developed and are still undergoing better advancements.
- Each tool may prove to be advantageous in dealing with one of the layers of social media analytics.
- For example, the Netminer tool is often used for dealing with social network data, the Google Analytics tool is powerful in dealing with action-oriented social media data, and the Lexalytics tool is good for handling textual social media data.

PRACTICE QUESTIONS

Q. I Multiple Choice Questions:

1. Which refers to the means of interactions among people in which they create, share, and/or exchange information and ideas in virtual communities?
   (a) Social media (b) Social communication network
   (c) Social data transfer (d) All of the mentioned
2. Web based social network services include,
   (a) Twitter (b) Facebook
   (c) LinkedIn (d) All of the mentioned
3. Which is a type of complex network and can be described as a social structure composed of a set of social users with social interactions between them?
   (a) social media (b) social network
   (c) social relation (d) None of the mentioned
4. Which is the process of tracking, collecting and analyzing data from social networks?
   (a) Social media analysis (b) Social network analytics
   (c) Social media analytics (d) None of the mentioned
5. Which is the automatic discovery of new, previously unknown, information from unstructured textual data?
   (a) text analytics (b) graphics analytics
   (c) video analytics (d) data analytics
6. Social media APIs includes,
   (a) Facebook API (b) YouTube API
   (c) Twitter API (d) All of the mentioned

7. How many layers play a vital role in contributing to social media input for gaining useful insights?
   (a) 5 (b) 6
   (c) 7 (d) 9
8. Which is a collection of techniques for working with human language?
   (a) NDP (b) NLP
   (c) NML (d) NDLP
9. Methods in trend analysis includes,
   (a) geographic trend analysis (b) intuitive trend analysis
   (c) temporal trend analysis (d) All of the mentioned
10. Which is the process of identifying the most important meaningful information in a document and compressing them into a shorter version preserving its overall meanings?
    (a) Text reporting (b) Text observations
    (c) Text summarization (d) Text planning
11. Which analytics mainly involves determining the possible drifts or trends over a period of time?
    (a) Trend (b) Social media
    (c) Social network (d) None of the mentioned
12. To handle challenges related to social media data following which tools are used.
    (a) Google Analytics tool (dealing with action-oriented social media data)
    (b) Lexalytics tool (handling textual social media data)
    (c) Netminer tool (dealing with social network data)
    (d) All of the mentioned
13. Which is the problem of predicting the existence of a link between two entities in a social network?
    (a) Link observation (b) Link prediction
    (c) Link analysis (d) Link analytics
14. Which can be used in machine learning to detect groups with similar properties and extract groups for various reasons?
    (a) Community transfer (b) Community processing
    (c) Community detection (d) None of the mentioned
15. Which analytics is carried out to gain insight from the geographic content of social media data?
    (a) Location (b) Community
    (c) Text (d) Trend

16. Which analysis means to identify the view or emotion behind a situation?
    (a) Trend (b) Location
    (c) Sentiment (d) None of the mentioned
17. Which uses data (unstructured or semistructured) from a variety of sources such as Media, Web and so on?
    (a) Location mining (b) Trend mining
    (c) Sentiment mining (d) Text mining

Answers

1. (a)	2. (d)	3. (b)	4. (c)	5. (a)	6. (d)	7. (c)	8. (b)	9. (d)	10. (c)
11. (a)	12. (d)	13. (b)	14. (c)	15. (a)	16. (c)	17. (d)			

Q. II Fill in the Blanks:

1. _____ media are interactive technologies that allow the creation or sharing/exchange of information, ideas, interests, and other forms of expression via virtual communities and networks.
2. _____ analytics applications require clear, interpretable results and actionable outcomes to achieve the desired result.
3. Data _____ means valid data identification.
4. In POS we identify whether words in a sentence are nouns, verbs, adjectives etc., and can be done in _____ for NLP.
5. Social media has gained significant popularity and become an essential medium of _____.
6. The text _____ also referred to as text data mining or text mining is the process of obtaining high-quality information from text.
7. Social media analytics is the process of gathering and analyzing data from social _____ such as Facebook, Instagram, LinkedIn and Twitter.
8. _____ analysis is the set of activities that assist in transforming raw data into meaning insight.
9. Interactive data visualization has led us to a new era of revolution where _____ have led us to easy analysis and decision making.
10. The textual _____ of social media mainly include tweets, textual posts, comments, status updates, and blog posts.
11. SNA is the process of investigating social _____ through the use of networks and graph theory.
12. _____ structure mining uses dynamic data that constantly keeps changing with time.
13. _____ detection techniques are useful for social media algorithms to discover people with common interests and keep them tightly connected.

14. _____ -based community detection methods emphasize the generation of communities based on the common topic of interests.
15. Link-based _____ community detection methods emphasize the study of edges of the social network in order to form communities.
16. The _____ finding is the task of generating and grouping experts of a social network based on his/her expertise on certain topics.
17. _____ is the technology that is used by machines to understand, analyze, manipulate, and interpret human's languages.
18. Tokenization is the process of dividing the text into individual pieces usually words or _____.
19. Bag of words is essentially a _____ for the document/sentence telling us to whether a specific word is there or not and how many times.
20. The _____ weighting is a standard approach to feature vector construction.
21. An _____ means a sequence of n words.
22. _____ words are commonly used in Text Mining and Natural Language Processing (NLP) to eliminate words that are so commonly used that they carry very little useful information.
23. Lemmatization is a method responsible for grouping different inflected forms of words into the root form, having the _____ meaning.
24. Part-of-Speech (POS) tagging is the process of converting a sentence in the form of a list of words, into a list of _____.
25. Text analytics incorporates tools and techniques that are used to derive _____ from unstructured data.
26. Text _____ also known as text tagging or text categorization is the process of categorizing text into organized groups.
27. A _____ is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form.
28. Social Media analytical _____ include Keyhole, AgoraPulse, Google Analytics and many more.
29. Social media data are governed by the properties of high _____, high velocity, as well as high variety.
30. Classification levels in _____ analysis includes Document-level, Sentence-level and Aspect-level.
31. The tasks that involve _____ engine analytics include advertisement spending statistics, keyword monitoring, and trends analysis.

Answers

1. Social	2. Text	3. capture	4. NLTK
5. communication	6. analytics	7. networks	8. Data
9. graphics	10. messages	11. structures	12. Dynamic
13. Community	14. Topic	15. clustering	16. expert
17. NLP	18. tokens	19. vector	20. TF-IDF
21. n-gram	22. Stop	23. same	24. tuples
25. insights	26. classification	27. summary	28. tools
29. volume	30. sentiment	31. search	

Q. III State True or False:

1. Social media refers to the means of interactions among people in which they create, share, and/or exchange information and ideas in virtual communities such as public posts, video chat etc. and networks.
2. Social media analytics is the ability to gather and find meaning in data gathered from social networks.
3. Data identification is the process of identifying the subsets of available data to focus on for analysis.
4. The social media analysis process which comprises of three stages namely data capturing, data understanding and data presentation.
5. Geographic trend analysis is mainly involved in analyzing the trend of products, users or other elements within or across geographic locations.
6. Social media network analytics focuses on the networking structure of the social media data which indicates the connection between users based on the concept of friends and followers.
7. Most of the social media Websites such as Facebook, Twitter, LinkedIn, YouTube etc., have their APIs (Application Programming Interfaces) that can be used by data analysts to access the social media data found in these sites.
8. Search engine analytics pays attention to analyzing historical search data to generate informative search engine statistics.
9. Tokenization setting splits the documents into words constructing a word vector known as Bag-of-Words (BoW).
10. TF-IDF stands for the Term Frequency-Inverse Document Frequency.
11. Stemming is a technique used to extract the base form of the words by removing affixes from them.
12. Location analysis is used to identify the emotions conveyed by the unstructured text.

13. Text analytics helps analysts extract meanings, patterns, and structure hidden in unstructured textual data.
14. Rule based text classification categorizes text into organized clusters by using a set of linguistic rules.
15. Historical trends are analyzed to determine future trends for a given phenomenon or feature.
16. Structured data refers to information that either does not have a predefined data format or model.
17. The volume refers to the size of Data. Velocity refers to the speed at which the data is getting accumulated.
18. Sentiment analysis is very commonly used in social media analytics by considering as input the comments, tweets, reviews, discussions, emails or feedbacks provided in social media by several online users.
19. The cycle of social media analytics consists of six steps namely, identification, extraction, cleaning, analyzing, visualization, and interpretation.
20. Influence maximization is the problem of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence.
21. NLP enables computers to understand and process human language.
22. Expert finding is concerned about finding persons who are knowledgeable on a given topic.
23. Social media data is the information that is collected from a organization's profiles across different social media networks.

Answers

1. (T)	2. (T)	3. (T)	4. (F)	5. (T)	6. (T)	7. (T)	8. (T)	9. (T)	10. (T)
11. (T)	12. (F)	13. (T)	14. (T)	15. (T)	16. (F)	17. (T)	18. (F)	19. (T)	20. (T)
21. (T)	22. (T)	23. (T)							

Q. IV Answer the following Questions:

(A) Short Answer Questions:

1. Define social media.
2. What is the purpose of social media?
3. Define text analytics.
4. Define social media analytics.
5. Define social network.
6. List social media sites.
7. Define location text analytics.
8. Define tokenization.
9. What is the purpose of n-grams?

10. List challenges for social media.
11. Define NLP.
12. Give seven layers of social media analytics.
13. List example of stop words.
14. Define the terms stemming and lemmatization.
15. Define community detection.
16. Define text categorization.
17. What is the purpose of social network analysis?
18. What is meant by expert finding?
19. List applications of NLP.
20. Define link prediction.
21. Define text summarization.
22. Define trend analytics.

(B) Long Answer Questions:

1. What is social media? State its advantages and disadvantages.
2. What is text analytics? What is its purpose? Also states its tasks.
3. What is social media analytics? What is its purpose? List its benefits.
4. Explain process of social media analytics diagrammatically.
5. Describe layers of social media analytics with the help of diagram.
6. What is social network? List any four examples of it. Explain two of the m in short.
7. What is social media data? List its types. Also state how to accessing social media data in detail.
8. What is social network analysis? Define it? Describe in detail.
9. With the help of suitable diagram describe lifecycle of social media analytics.
10. What is link prediction? Explain with example.
11. What is community detection? What are its different methods? Explain four of the in short.
12. What is influence maximization? Explain its framework diagrammatically.
13. What is expert finding? How to find an expert? Decribe with example.
14. Write a short note on: Prediction of trust.
15. Explain the term distrust among individuals in detail.
16. What is NLP? What is its purpose? Describe its phases with the help of diagram.
17. What is text analytics? Explain in detail.
18. What is tokenization: How it used in text analytics?
19. What is bag of words? How to use it NLP? Explain in detail.
20. What is Word Weighting (TF-IDF)? Describe in detail.

21. Explain n-gram with example.
22. What is stemming and lemmatization? How they are differ from each other?
23. Describe the term synonyms with respect to NLP.
24. Write a short note on: Parts of speech tagging.
25. What is sentiment analysis? Explain with its classification?
26. What is text analytics? Explain its steps diagrammatically. Also states its advantages, disadvantages and applications.
27. What is text categorization? Describe diagrammatically. Also list approaches.
28. What is text summarization? Explain its two types in detail.
29. What is trend analytics? Describe its methods in detail.
30. Write a short note on: Challenges to social media analytics.

■■■