es easy-solutions

New Syllabus SPPU

As per the New Credit System Syllabus (2019 course) of
Savitribai Phule Pune University w.e.f. academic year 2021-2022

DATA SCIENCE AND BIG DATA ANALYTICS

(Code : 310251)

"QUICK READ SERIES"

Semester VI
Computer Engineering

Chapterwise Solved University Paper Solution

For End Semester Examination

easy-solutions

TechKnowledge™ Publications

easy – solutions

Savitribai Phule Pune University

As per New Credit System Syllabus(Rev. 2019) of Savitribai Phule Pune University with effective from Academic Year 2021-2022

Data Science and Big Data Analytics

(Code : 310251)

"Quick Read Series"

Semester VI - Computer Engineering

TechKnowledge™ Publications

EPE131A Price ₹ 125/-

Data Science and Big Data Analytics (Code : 310251)

(Semester VI - Computer Engineering) (SPPU)

Edition 2022

This edition is for sale in India, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka and designated countries in South-East Asia. Sale and purchase of this book outside of these countries is unauthorized by the publisher.

Published By

TECHKNOWLEDGE PUBLICATIONS

Printed @

37/2, Ashtavinayak Industrial Estate,
Near Pari Company,
Narhe, Pune, Maharashtra State, India.
Pune - 411041

Head Office

B/5, First floor, Maniratna Complex, Taware Colony,
Aranyeshwar Corner, Pune - 411 009.
Maharashtra State, India
Ph : 91-20-24221234, 91-20-24225678.
Email : info@techknowledgebooks.com,
Website : www.techknowledgebooks.com

Subject Code : 310251

Book code : EPE131A

SYLLABUS

In-Sem. Exam

Unit I : Introduction to Data Science and Big Data **07 hrs**

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

Exemplar/Case Studies : Create academic performance dataset of students and perform data pre-processing using techniques of data cleaning and data transformation.

Unit II : Statistical Inference **07 hrs**

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

Exemplar/Case Studies : For an employee dataset, create measure of central tendency and its measure of dispersion for statistical analysis of given data

End-Sem. Exam

Unit III : Big Data Analytics Life Cycle **07 hrs**

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

Exemplar/Case Studies : Case study: Global Innovation Social Network and Analysis (GINA).

Unit IV : Predictive Big Data Analytics with Python **07 hrs**

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

Scikit-learn, Installations, Dataset, mat plotlib, filling missing values, Regression and Classification using Scikit-learn.

Exemplar/Case Studies : Use IRIS dataset from Scikit and apply data preprocessing methods

Unit V : Big Data Analytics and Model Evaluation **07 hrs**

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

Exemplar/Case Studies : Use IRIS dataset from Scikit and apply K-means clustering methods.

Unit VI : Data Visualization and Hadoop **07 hrs**

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

Exemplar/Case Studies : Use IRIS dataset from Scikit and plot 2D views of the dataset

❑❑❑

Table of Contents

❑❑❑

Savitribai Phule Pune University
Semester VI (Computer Engineering)

Data Science and Big Data Analytics

Unit III : Big Data Analytics Life Cycle

Chapter 3 : Big Data Analytics Life Cycle

Q. 1	Explain different phases of data analytics life cycle.	SPPU - Aug. 18, 6 Marks
OR	Explain Data Analytic Life cycle.	SPPU - Dec. 18, 8 Marks
OR	Draw Data Analytics Lifecycle & give brief description about all phases.	SPPU - May 19, 5 Marks
OR	Demonstrate the overview of Data Analytics Life Cycle.	SPPU - Oct. 19, 5 Marks
OR	Why communication is important in data analytics lifecycle projects ?	SPPU - May 19, 8 Marks
OR	Write a short note about the second phase of data analytics life cycle.	(4 Marks)
OR	Draw data analytics life cycle diagram and explain its third phase.	(6 Marks)

Ans. : The data analytics life cycle broadly has six phases. Each of these phases are worked through iteratively with the previous phase before moving to the next phase.

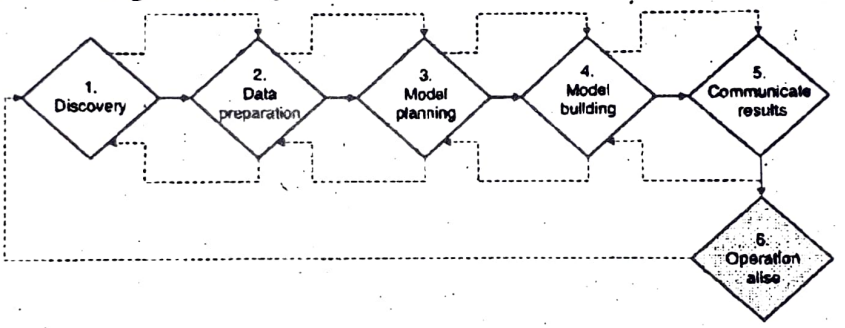


Fig. 3.1 : Data Analytics Life Cycle

1. Phase 1 : Discovery

In the Discovery phase, the data science team

1. Learns about the business problem to solve,
2. Investigates the problem,
3. Develops context and understanding,
4. Examines the available data sources and
5. Formulates the initial hypothesis

The team learns about the business domain in which the problem is to be solved. It assesses the resources available for the project and carries out the feasibility analysis. It spends time in framing the right problem.

Definition : *Framing is the process of stating the analytics problem to be solved.*

As part of the framing activity, the main objectives of the project are ascertained and the success criteria for the project is clearly defined. It also develops the initial hypothesis that can later be substantiated with the data.

2. Phase 2 : Data Preparation

The data preparation phase explores, pre-processes and conditions the data before modelling and analysis could be carried out. In this phase, the following activities are carried out.

1. **Preparing the analytics environment :** In this step, an isolated workspace is created in which the team can explore the data without interfering with the live data. The data from various data sources is collected in the isolated workspace.

2. **Perform ETL process :** ETL stands for Extract, Transform and Load. In this step, the raw data is extracted from the datastore, transformed as deemed right (removing noise, outliers, and biases from data) and then loaded into the datastore again for analysis.
3. **Learn about the data :** Once the ETL process is complete, the team spends time in learning about the data and its attributes. Understanding the data itself is the key to building a good data model in the subsequent phase.
4. **Data conditioning:** In this step, the data is further cleaned and normalized by performing further transformations as required. The data from several sources could be joined or combined as required. The actual data attributes that would be used for analytics are decided.
5. **Data visualisation:** Once the data is in a clean state and ready to be analysed, it is a good idea to visualise it to identify patterns and explore data characteristics. Understanding patterns about the data enables building a perspective about the data model.

Some of the common tools used in this phase are as following. The choice of tools largely depends on the problem at hand, desired outcomes, and the team's skills.

1. **Apache Hadoop :** The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
2. **Apache Kafka :** Apache Kafka is a distributed streaming platform. You can publish and subscribe to streams of records, store streams of records and process streams of records as they occur.
3. **Alpine Miner :** Alpine Miner provides a graphical interface for creating analytics workflows and is optimised for fast experimentation, collaboration, and an ability to work within the database itself.
4. **OpenRefine :** OpenRefine is a powerful tool for working with messy data. It cleans the data and transforms it from one format into another.

3. **Phase 3 : Model Planning**

In this phase, the team explores and evaluates the possible data models that could be applied to the given datasets to get the desired results. The team can try several models before finalising. Some of the major activities carried out in this phase are as following.

1. **Data Exploration :** The team spends time in understanding the available data and the various patterns and relationships amongst its attributes.

   The team could consult subject matter experts, stakeholders, analysts, and others who might have a viewpoint on how the data should be interpreted and examined.
2. **Model Selection :** The goal of this activity to choose an analytical technique based on the given dataset and the desired outcome. Based on the type of data (structured, semi-structured or unstructured) different techniques could be chosen and applied.

   Some of the common tools used in this phase are as following. The choice of tools largely depends on the problem at hand, desired outcomes, and the team's skills.

1. **R :** R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible.
2. **SQL Server Analysis Services :** SQL Server Analysis Services supports tabular models at all compatibility levels, multidimensional models, data mining, and Power Pivot for SharePoint. It provides an analytical data engine used in decision support and business intelligence (BI) solutions, providing the analytical data for business reports and client applications such as Excel, Reporting Services reports, and other third-party BI tools.
3. **SAS/ACCESS :** It provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC and OLE DB. You can access the most popular databases on common platforms without detailed knowledge of the database or SQL.

4. **Phase 4 : Model Building**

In this phase, the team starts to build the data analytics model. The available dataset is divided into

1. Training dataset, 2. Testing dataset and 3. Production dataset

The training dataset is used to train (design) the model. Once the team is confident about the model, it tests the model using the testing dataset. Once the testing is complete, the model is ready to be used in the production (go live). The production dataset or new datasets could be applied to it to get the desired results.

Some of the common tools used in this phase are as following. The choice of tools largely depends on the problem at hand, desired outcomes, and the team's skills.

1. **R :** R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible.
2. **GNU Octave :** It is a scientific programming language with powerful mathematics-oriented syntax with built-in plotting and visualization tools. It is free software that runs on GNU/Linux, mac OS, BSD, and Windows.
3. **WEKA :** It is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
4. **Python :** It is a is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib.
5. **Commercial software :** Apart from the free and open source tools, there are various commercial software available for data analytics such as SAS Enterprise Miner, SPSS Modeler, MATLAB, and Alpine Miner.

5. **Phase 5 : Communicate Result**

- After building, testing, and executing the model, the team compares the outcome with the pre-established success criteria. The results are validated and could be statistically proven.
- The team then articulates the findings and documents the results. The findings are communicated to the project stakeholders.
- The model building exercise could be unsuccessful. The findings are still documented and reported before the team goes on to try and build another model. In data analytics life cycle each phase is an iterative process and works with the previous phase.

6. **Phase 6 : Operationalise**

In the final phase, the model is deployed in the staging environment before it goes live on a wider scale. The staging environment is very similar to the production environment. The idea is to ensure that the model sustains the performance requirements and other execution constraints and any issues are identified before the model is deployed in the production environment. If any changes are required, they are carried out and tested again.

The project outcome is shared with the key stakeholders such as

1. **Business user :** The business user ascertains the benefits and implications of the project findings.
2. **Project sponsor :** The project sponsor asks questions around ROI (return on investment) and any potential risks to maintaining the project.
3. **Project manager :** The project manager determines if the project was timely completed and the goals were met.
4. **Business intelligence analyst :** The business intelligence analyst determines if any of the reports or dashboards needs to be changed to accommodate the new findings.
5. **Database Administrator :** The database administrator needs to plan for backup of datasets and any other code that was written to be run on the database for the analytics project.
6. **Data Engineer :** The data engineer needs to share the code; version control it and maintain it. Any issues or bugs found in the code should be fixed.
7. **Data Scientist :** The data scientist could explain the model to her peers and other stakeholders. She also documents the model and how it was implemented.

Unit IV : Predictive Big Data Analytics with Python

Chapter 4 : Predictive Big Data Analytics with Python

Q. 1 Explain a few data quality issues. **(6 Marks)**

Ans. : Common Data Quality Issues

- Data that is fit for use : For example, if you are working on a cancer project, you would need a dataset that has cancer patients and their health details.
- Data that meets your analytics requirements : For example, if you are trying to relate consumption of meat with probability of having cancer, you would need eating habits of the patients in the dataset.
- Relevance and timeliness : Take an example where you are building an analytics model on modern lifestyle. Could you use a dataset from 18th century? Perhaps not, right? Things and surroundings at that time were quite different than what they are in 21st century. So, the dataset that you pick must be relevant from timeliness or freshness perspective.
- Completeness, correctness, and formatting of data: For example, you would require that the important fields in the data set are fully populated and there are as few rows as possible that have missing values for particular fields. After eliminating such incomplete rows of data, are you left with enough data that you can use for both developing and testing your model?
- There could be other types of errors (or mix ups) as well in the data, such as the following, that require handling or data cleaning before you can use the data for training your machine learning model.
  - **Spelling mistakes :** It is common to find spelling mistakes in names of countries, people, things, etc. There could also be abbreviations such as US, USA, United States of America, America – they all refer to the same country!
  - **Date formatting :** Asian users typically use dd-mm-yyyy date format whereas American users could use mm-dd-yyyy format.
  - **Incorrect labels :** For example, age could be labelled as year born. So, if age column is incorrectly labelled as year born, then age of 56 could be mistakenly assumed to be born in the year 1956.
  - **Scaling and units :** Sometimes the units could be wrongly entered. For example, weight of the person could be entered in Kgs or Pounds. Rows having incorrect units could skew the analysis.
  - **Skewed data (data anomalies) :** For a given field in the dataset, some of the values could be quite high and some of the values could be quite low. Such skewed data could wrongly train the model.

Q. 2 What could you do to fix the poor quality data? **(4 Marks)**

Ans. : Remediating (Fixing) Data Quality Issues

Data cleaning is one of the most time consuming exercise in building a machine learning model. Some of the common measures to clean the data are as following.

1. Delete rows with missing values.
2. Fix any formatting issues.
3. Fix labelling issues.
4. Fix spelling mistakes and abbreviations.
5. Insert new columns based on other columns.
6. Delete rows with skewed values.

Q. 3 List a few data quality metrics used in defining data quality constraints. **(4 Marks)**

Ans. : Some of the common data quality metrics used in defining data quality constraints are as shown in Table 4.1

Table 4.1

Metric	Description	Usage Example
ApproxCountDistinct	Approximate number of distinct value, computed with HyperLogLogPlusPlus sketches.	ApproxCountDistinct("review_id")
ApproxQuantile	Approximate quantile of a distribution.	ApproxQuantile("star_rating", quantile = 0.5)
ApproxQuantiles	Approximate quantiles of a distribution.	ApproxQuantiles("star_rating", quantiles = Seq(0.1, 0.5, 0.9))
Completeness	Fraction of non-null values in a column.	Completeness("review_id")
Compliance	Fraction of rows that comply with the given column constraint.	Compliance("top star_rating", "star_rating >= 4.0")
Correlation	Pearson correlation coefficient measures the linear correlation between two columns. The result is in the range [-1, 1], where 1 means positive linear correlation, -1 means negative linear correlation, and 0 means no correlation.	Correlation("total_votes", "star_rating")
CountDistinct	Number of distinct values.	CountDistinct("review_id")
DataType	Distribution of data types such as Boolean, Fractional, Integral, and String. The resulting histogram allows filtering by relative or absolute fractions.	DataType("year")
Distinctness	Fraction of distinct values of a column over the number of all values of a column. Distinct values occur at least once. Example: [a, a, b] contains two distinct values a and b, so distinctness is 2/3.	Distinctness("review_id")
Entropy	Entropy is a measure of the level of information contained in an event (value in a column) when considering all possible events (values in a column). It is measured in nats (natural units of information). Entropy is estimated using observed value counts as the negative sum of (value_count/total_count) * log(value_count/total_count). Example: [a, b, b, c, c] has three distinct values with counts [1, 2, 2]. Entropy is then (-1/5*log(1/5)-2/5*log(2/5)-2/5*log(2/5)) = 1.055.	Entropy("star_rating")
Maximum	Maximum value.	Maximum("star_rating")
Mean	Mean value; null values are excluded.	Mean("star_rating")
Minimum	Minimum value.	Minimum("star_rating")
MutualInformation	Mutual information describes how much information about one column (one random variable) can be inferred from another column (another random variable). If the two columns are independent, mutual information is zero. If one column is a function of the other column, mutual information is the entropy of the column. Mutual information is symmetric and nonnegative.	MutualInformation(Seq("total_votes", "star_rating"))

- Why do customers buy certain products? (Holidays, Festivals, Seasons)
- What time of the day do they buy it? (Morning, Afternoon, Evening)
- Who are the customers? (Housewives, Seniors, Young adults)

- Based on the analysis, supermarkets can appropriately place the items frequently bought together nearby so as to help the customers to pick these items (even if they might have forgot about it or did not intend to buy it).
- They could run other schemes based on the analysis to make the best use of the known shopping patterns.
- For example, if there is an 80% chance (or confidence) that bread and butter would be bought together, then the store can run discount only on bread (say Rs.5 off) and not put discount on butter.
- One key point to note here is that the association rules do not predict the customer preferences. They rather aim to find relationships between the objects in transactions.

Q. 5 Explain the term itemset taking a suitable example **(4 Marks)**

OR Explain the term frequent itemset taking a suitable example. **(4 Marks)**

OR Explain the term antecedent taking a suitable example. **(4 Marks)**

Ans. :

Definition : *An itemset is a collection of items that have certain relationship.*

- An itemset can contain one or more items. For example, bread, butter, and milk from transaction 1 can form one itemset. It is represented as

  Itemset 1 = {Bread, Butter, Milk}.

  Similarly, Itemset from transaction 3 is

  Itemset 3 = {Bread, Butter}

Definition : *An Itemset that occurs frequently in the dataset is called a frequent itemset.*

For example, {Bread, Butter, Milk} is a frequent itemset. {Bread, Butter} is another frequent itemset.

Any association rule is written like an If...Then statement. For example, If someone buys bread, then she also buys butter. This association rule is denoted as {Bread} → {Butter}

Definition : *The itemset on the left hand side of the association rule is called antecedent.*

Definition : *The Itemset on the right hand side of the association rule is called consequent.*

In the example of association rule {Bread} → {Butter}, {Bread} is antecedent whereas {Butter} is consequent.

Q. 6 Explain the term support with suitable example. **(4 Marks)**

Ans. :

Definition : *Support measures the frequency of a given itemset in the dataset (amongst all transactions).*

- In other words, support is a measure that conveys how popular an itemset is. It is measured as the ratio of the transactions that contain the itemset to the total number of transactions in the dataset.
- For example, {Bread} appears in all the transactions.

Transaction	Item 1	Item 2	Item 3
1	Bread	Butter	Milk
2	Bread	Butter	Milk
3	Bread	Butter	
4	Bread	Butter	Milk
5	Bread	Butter	

Hence, Support {Bread} = $\frac{5}{5}$ = 1 = 100%.

Similarly, Support {Milk} = $\frac{3}{5}$ = 60%

Support {Butter} = $\frac{5}{5}$ = 1 = 100%

Support {Bread, Butter} = $\frac{5}{5}$ = 1 = 100%

Support {Bread, Milk} = $\frac{3}{5}$ = 60%

Support {Butter, Milk} = $\frac{3}{5}$ = 60%

Support {Bread, Butter, Milk} = $\frac{3}{5}$ = 60%

Q. 7 For the given set of transactions, calculate support for all the possible itemsets. **(6 Marks)**

Transaction	Item 1	Item 2	Item 3	Item 4
1	Soap	Oil		
2		Oil	Detergent	
3			Detergent	Biscuits
4	Soap			Biscuits
5		Oil		Biscuits

Ans. : Support measures the frequency of a given itemset in the dataset. Hence, the support for all possible itemsets containing max of two elements is as following. Note here that you could have more itemset with three and four elements in them as well.

{Soap} = $\frac{2}{5}$ = 40% ; {Oil} = $\frac{3}{5}$ = 60%

{Detergent} = $\frac{2}{5}$ = 40% ; {Biscuits} = $\frac{3}{5}$ = 60%

{Soap, Oil} = $\frac{1}{5}$ = 20% ; {Oil, Detergent} = $\frac{1}{5}$ = 60%

{Detergent, Biscuits} = $\frac{1}{5}$ = 20% ; {Soap, Biscuits} = $\frac{1}{5}$ = 20%

{Oil, Biscuits} = $\frac{1}{5}$ = 20%

Q. 8 Explain the term confidence **(2 Marks)**

Ans. :

Definition : *Confidence measures how likely a consequent is true for a given antecedent.*

- In simple words, confidence measures how likely is that an item Y is also purchased when item X is purchased. It is calculated as the ratio of support for itemset {X, Y} to support of itemset {X).

$$\text{Confidence } (\{X\} \rightarrow \{Y\}) = \frac{\text{Support } \{X, Y\}}{\text{Support } \{X\}}$$

Q. 9 Calculate the confidence value for {Bread} → {Milk} from the following transactions. **(6 Marks)**

Transaction	Item 1	Item 2	Item 3
1	Bread	Butter	Milk
2	Bread	Butter	Milk
3	Bread	Butter	
4	Bread	Butter	Milk
5	Bread	Butter	

Ans. : Confidence measures how likely a consequent is true for a given antecedent.

Hence, Confidence

$$(\{Bread\} \rightarrow \{Milk\}) = \frac{\text{Support } \{Bread, Milk\}}{\text{Support } \{Bread\}}$$

$$\text{Support } \{Bread, Milk\} = \frac{3}{5}$$

$$\text{Support } \{Bread\} = \frac{5}{5} = 1$$

$$\text{Confidence } (\{Bread\} \rightarrow \{Milk\}) = \frac{3/5}{1} = \frac{3}{5} = 60\%$$

Hence, you can be 60% confident that someone buying bread will also buy milk.

Q. 10 There are 300 transaction records out of which milk appears on 250 transactions and shoe polish appears on 40 transactions. Out of 40 transactions for shoe polish, 30 contain milk as well. Do you suggest placing shoe polish near milk to improve sales? **(4 Marks)**

Ans. : Since one of the items is frequently selling, calculating lift is more appropriate in this scenario.

$$\text{Support } \{Milk\} = \frac{250}{300}$$

$$\text{Support } \{Shoe\ Polish\} = \frac{40}{300}$$

$$\text{Support } \{Shoe\ Polish, Milk\} = \frac{30}{300}$$

$$\text{Lift } (\{Shoe\ Polish\} \rightarrow \{Milk\}) = \frac{\text{Support } \{Shoe\ polish, Milk\}}{\text{Support } \{Shoe\ polish\} \times \text{support } \{Milk\}} = \frac{\frac{30}{300}}{\left(\frac{40}{300}\right) \times \left(\frac{250}{300}\right)} = 0.9$$

The value of lift < 1.

Hence, there is not a strong association between shoe polish and milk. Hence, you would likely not place shoe polish near milk to improve sales.

Q. 11 Define the term leverage. **(2 Marks)**

Ans. :

Definition : *Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent of each other.*

Q. 12 For the following transactions, leverage for Apple → Orange. What does it indicate? **(6 Marks)**

Transaction	Item 1	Item 2	Item 3	Item 4
1	Apple		Milk	
2	Apple	Orange	Milk	Biscuits
3	Apple	Orange	Milk	Biscuits
4	Apple	Orange		Biscuits
5	Apple		Milk	Biscuits
6		Orange	Milk	Biscuits
7		Orange	Milk	Biscuits
8		Orange		Biscuits
9		Orange		Biscuits
10		Orange		

Ans. :

Let's first calculate the support values.

$$\text{Support } \{\text{Apple}\} = \frac{5}{10}$$

$$\text{Support } \{\text{Orange}\} = \frac{8}{10}$$

$$\text{Support } \{\text{Apple} \rightarrow \text{Orange}\} = \frac{3}{10}$$

$$\text{Leverage } (\{\text{Apple}\} \rightarrow \{\text{Orange}\}) = \text{Support } (\{\text{Apple}\} \rightarrow \{\text{Orange}\}) - \text{Support } \{\text{Apple}\} \times \text{Support } \{\text{Orange}\}$$

$$= \frac{3}{10} - \left(\frac{5}{10} \times \frac{8}{10}\right) = -\frac{1}{10} = -0.1$$

A value of leverage less than 0 indicate that there is not stronger relationship between the itemsets. In this case, Apple → Orange is not a strong relationship.

Q. 13 Define Apriori algorithm. Why is it used? **(4 Marks)**

Ans. :

- Apriori algorithm helps you to carry out the first step of association rule mining that is finding the frequent itemsets from the dataset.

***Definition** : Apriori is an algorithm for finding frequent itemsets.*

- It works on the principle that "if an itemset is a frequent itemset then its subsets must also be frequent".
- This is also called as Apriori principle. It uses this principle to prune (or discard) all the itemsets that do not match the minimum required support threshold value.

Q. 14 How Apriori algorithm works? **(4 Marks)**

Ans. : The Fig. 4.2 gives the high-level view of how Apriori algorithm works.

1. You start with choosing a support threshold below which you would discard the itemsets.
2. You start with one item in the itemset and keep increasing the items in the itemset (after step 3) until complete.
3. Discard the itemsets, generated in step 2, that are below the support threshold. Repeat step 2 until all items are considered.
4. Finally, you would be left with frequent itemsets for which you can generate association rules based on measures such as confidence and lift.

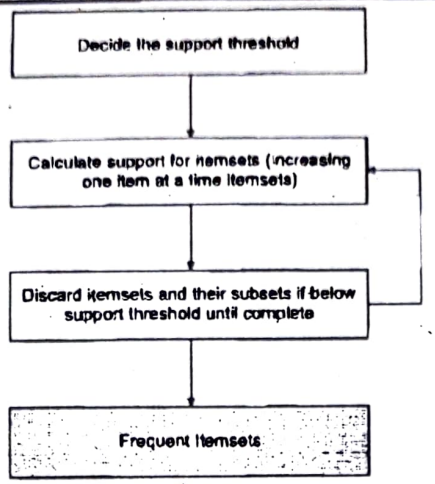


Fig. 4.2 : High-level view of Apriori algorithm works

Q. 15 Transactional Data for an All Electronics Branch is as follows :

TID	List of Item IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Find the frequent item set and generate association rules with confidence values. **(6 Marks)**

Ans. : Assume that support threshold is $\frac{3}{9}$ which means that for an itemset to be considered frequent it must appear in at least three transactions out of the nine transactions given in the dataset.

Iteration 1 (one item in the itemset)

Start with each item in the itemset and calculate their support values and discard the ones that are below the assumed support threshold.

Itemset	Support Value	Action
{I1}	$\frac{6}{9}$	Keep
{I2}	$\frac{7}{9}$	Keep
{I3}	$\frac{6}{9}$	Keep
{I4}	$\frac{2}{9}$	Discard
{I5}	$\frac{2}{9}$	Discard

Support values for {I4} and {I5} are below the chosen support threshold and hence you can discard any itemset containing those items. The items remaining are {I1}, {I2} and {I3}.

Iteration 2 (two items in the itemset)

Take the remaining itemsets and combine them further to contain 2 items each in the itemsets. Re-calculate the support values for the itemsets and discard the ones that are below the support threshold.

Itemset	Support Value	Action
{I1, I2}	$\frac{4}{9}$	Keep
{I1, I3}	$\frac{4}{9}$	Keep
{I2, I3}	$\frac{4}{9}$	Keep

Iteration 3 (three items in the itemset)

Now, take 3 items in the itemset. Re-calculate the support values for the itemsets and discard the ones that are below the support threshold.

Itemset	Support Value	Action
{I1, I2, I3}	$\frac{2}{9}$	Discard

You find that only the itemsets from the iteration 2 can be considered frequent. Next, you can calculate the confidence value for each frequent itemset and generate association rules.

$$\text{Confidence}(\{I1\} \rightarrow \{I2\}) = \frac{\text{Support}\{I1,I2\}}{\text{Support}\{I1\}} = \frac{\frac{4}{9}}{\frac{6}{9}} = \frac{2}{3} = 66.67\%$$

$$\text{Confidence}(\{I2\} \rightarrow \{I1\}) = \frac{\text{Support}\{I2,I1\}}{\text{Support}\{I2\}} = \frac{\frac{4}{9}}{\frac{7}{9}} = \frac{4}{7} = 57.14\%$$

$$\text{Confidence}(\{I1\} \rightarrow \{I3\}) = \frac{\text{Support}\{I1,I3\}}{\text{Support}\{I1\}} = \frac{\frac{4}{9}}{\frac{6}{9}} = \frac{2}{3} = 66.67\%$$

$$\text{Confidence}(\{I3\} \rightarrow \{I1\}) = \frac{\text{Support}\{I3,I1\}}{\text{Support}\{I3\}} = \frac{\frac{4}{9}}{\frac{6}{9}} = \frac{2}{3} = 66.67\%$$

$$\text{Confidence}(\{I2\} \rightarrow \{I3\}) = \frac{\text{Support}\{I2,I3\}}{\text{Support}\{I2\}} = \frac{\frac{4}{9}}{\frac{7}{9}} = \frac{4}{7} = 57.14\%$$

$$\text{Confidence}(\{I3\} \rightarrow \{I2\}) = \frac{\text{Support}\{I3,I2\}}{\text{Support}\{I3\}} = \frac{\frac{4}{9}}{\frac{6}{9}} = \frac{2}{3} = 66.67\%$$

Frequent Itemset	Association Rule	Support Value	Confidence Value
{I1, I2}	{I1 → I2}	$\frac{4}{9}$	66.67%
{I1, I2}	{I2 → I1}	$\frac{4}{7}$	57.14%
{I1, I3}	{I1 → I3}	$\frac{4}{9}$	66.67%
{I1, I3}	{I3 → I1}	$\frac{4}{9}$	66.67%
{I2, I3}	{I2 → I3}	$\frac{4}{7}$	57.14%
{I2, I3}	{I3 → I2}	$\frac{4}{7}$	66.67%

Q. 16 A local retailer has a database that stores 10,000 transactions of last summer. After analysing the data, a data science team has identified the following statistics:

{battery} appears in 6,000 transactions.

{sunscreen} appears in 5,000 transactions.

{sandals} appears in 4,000 transactions.

{bowls} appears in 2,000 transactions.

{battery, sunscreen} appears in 1,500 transactions.

{battery, sandals} appears in 1,000 transactions.

{battery, bowls} appears in 250 transactions.

{battery, sunscreen, sandals} appears in 600 transactions.

Answer the following questions:

(a) What are the support values of the preceding itemsets?

(b) Assuming the minimum support is 0.5, which itemsets are considered frequent?

(c) What are the confidence values of {battery} → {sunscreen} and {battery, sunscreen} → {sandals} ? Which of the two rules is more interesting ? **(6 Marks)**

Ans. :

(a) Support values for the itemsets are as following.

{battery} appears in 6,000 transactions = $\frac{6000}{10000} = \frac{3}{5} = 0.6$

{sunscreen} appears in 5,000 transactions = $\frac{5000}{10000} = \frac{1}{2} = 0.5$

{sandals} appears in 4,000 transactions = $\frac{4000}{10000} = \frac{2}{5} = 0.4$

{bowls} appears in 2,000 transactions = $\frac{2000}{10000} = \frac{1}{5} = 0.2$

{battery, sunscreen} appears in 1,500 transactions = $\frac{1500}{10000} = \frac{3}{20} = 0.15$

{battery, sandals} appears in 1,000 transactions = $\frac{1000}{10000} = \frac{1}{10} = 0.1$

{battery, bowls} appears in 250 transactions = $\frac{250}{10000} = \frac{1}{40} = 0.025$

{battery, sunscreen, sandals} appears in 600 transactions = $\frac{600}{10000} = \frac{3}{50} = 0.06$

(b) The frequent itemsets are the ones that have support value higher or equal to the given support threshold of 0.5. The following itemsets are frequent itemsets.

{battery} appears in 6,000 transactions = $\frac{6000}{10000} = \frac{3}{5} = 0.6$

{sunscreen} appears in 5,000 transactions = $\frac{5000}{10000} = \frac{1}{2} = 0.5$

(c) Confidence values are as following.

$$\text{Confidence (\{battery\} } \rightarrow \text{ \{sunscreen\}) } = \frac{\text{Support } \{\{\text{battery}\} \rightarrow \{\text{sunscreen}\}\}}{\text{Support } \{\text{battery}\}} = \frac{0.15}{0.6} = 25\%$$

$$\text{Confidence (\{battery, sunscreen\} } \rightarrow \text{ \{sandals\}) } = \frac{\text{Support } \{\{\text{battery, sunscreen}\} \rightarrow \{\text{sandals}\}\}}{\text{Support } \{\{\text{battery, sunscreen}\}\}} = \frac{0.06}{0.15} = 40\%$$

The confidence value for {battery, sunscreen} → {sandals} is higher (40 % > 25 %) than the confidence value for {battery} → {sunscreen}. Hence, the association rule {battery, sunscreen} → {sandals} is better based on the confidence value.

Q. 17 Explain the various approaches to improve the efficiency of Apriori algorithm. (6 Marks)

Ans. : Some of the approaches to improve the efficiency of Apriori algorithm are shown in Fig. 4.3(a).

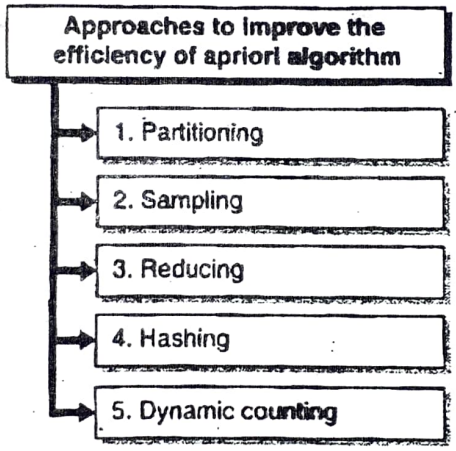


Fig. 4.3(a) : Approaches to improve the efficiency of Apriori algorithm

1. **Partitioning** : Instead of analysing the entire transaction database at once, you can partition the database into multiple smaller databases to increase the speed of processing. You can also run the algorithm on multiple databases in parallel and later join the results of the computation. For example, instead of processing the transaction database by weeks, you could partition the database per day. Within a day, you could further partition the database into first half and second half of the day.
2. **Sampling** : Using sampling, you could look at only those transactions that are good representatives of the overall transaction database. This avoids long processing times and could give similar results. For example, mostly people buy groceries on weekends. So, instead of analysing the transaction database from Monday to Sunday, you could just process the transaction database for Saturday and Sunday.
3. **Reducing** : You can totally eliminate all the transactions that do not contain the frequent itemsets identified in earlier stages in the subsequent stages of computation. Ideally, this is what Apriori algorithm does, but you could do it as well without running through the Apriori algorithm.
4. **Hashing** : Hashing is a technique to get a unique string for a set of strings.

   So, instead of using multiple items in an analysis, you can just calculate its hash value and use that for analysis. This drastically reduces the number of items in the analysis and also saves you from multiple iterations.

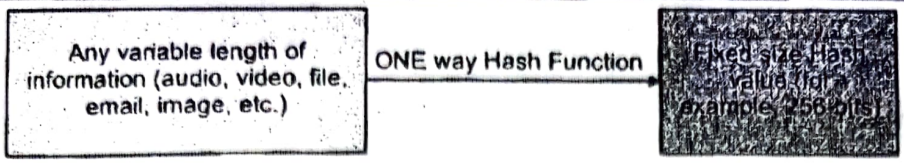


Fig. 4.3(b)

5. **Dynamic Counting** : In this technique, you only add new candidate itemsets (to test) when all of their subsets are estimated to be frequent. You could use other estimation techniques to populate the list of items that you wish to analyse for being frequent itemsets.

Q. 18 Write a short note on FP-growth. **(4 Marks)**

Ans. : The Apriori candidate generate-and-test method significantly reduces the size of candidate sets, leading to good performance gain. However, it can suffer from two nontrivial costs as following.

1. It may still need to generate a huge number of candidate sets. For example, if there are 104 frequent 1-itemsets, the Apriori algorithm will need to generate more than 107 candidate 2-itemsets.
2. It may need to repeatedly scan the whole database and check a large set of candidates by pattern matching. It is costly to go over each transaction in the database to determine the support of the candidate itemsets.

Definition : Frequent pattern growth or simply FP-growth is a method that mines the complete set of frequent itemsets without such costly candidate generation process as in Apriori algorithm.

It adopts a divide-and-conquer strategy as follows.

1. First, it compresses the database representing frequent items into a frequent pattern tree, or FP-tree, which retains the itemset association information.
2. 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 high-level steps to carry out FP-growth based mining are as following.

1. Scan the entire database to find the possible occurrences of the itemsets in the database. This step is the similar to the first step of Apriori algorithm.
2. Construct the FP tree. Create the root of the tree where the root is represented by null.
3. Scan the database once again and observe the transactions. Examine the first transaction and find the itemsets in the database. The itemset with the maximum count is taken at the top and the itemsets with lower count are taken at bottom and so on. It means that the branch of the tree is constructed with transaction itemsets in descending order of count.
4. Examine the transaction in the database. The itemsets are sorted in descending order of count. If any itemset of this transaction is already present in any other branch, then this transaction branch may share a common prefix to the root of the FP Growth algorithm. This means that the common itemset is connected to the new node of another itemset in this transaction.
5. The count of the itemset is increased as it occurs in the transactions. Both the common node and new node count is incremented by 1 as they are created and linked according to transactions.
6. Mine the created FP Tree. The lowest node is examined first along with the connections of the lowest nodes. The lowest node represents the frequency pattern of length 1. Traverse the path in the FP Tree. These paths are called as a conditional pattern base. Conditional pattern base is a sub-database consisting of prefix paths in the FP tree with the lowest node as suffix.

7. Construct a Conditional FP Tree, which is formed by the count of itemsets in the path. The itemsets which satisfy the threshold support are considered in the Conditional FP Tree.
8. Finally, generate Frequent Patterns from the conditional FP Tree to get frequent itemsets.

Q. 19 Define regression analysis with example. **(4 Marks)**

Ans. : Regression Analysis

The statistical definition of regression analysis is

***Definition** : A functional relationship between two or more correlated variables that is often empirically determined from data and is used specially to predict values of one variable when given values of the others.*

- So, for example,
  - Petrol required (X) = 5 litres
  - Kilometres driven (Y) = 50
  - Relationship : Y = 10X
  - Prediction : for 10 litres of petrol, you could go around 10× 10 = 100 Kms.

Q. 20 Define dependent and independent variables with a suitable example. **(4 Marks)**

Ans. :

1. Dependent variables

***Definition** : The variables that have influence on the outcome are called input, independent, explanatory or predictor variable.*

2. Independent variables

***Definition** : The variable whose outcome (or value) depends on other variables is called response, outcome, or dependent variable.*

Q. 21 What is linear regression? **(4 Marks)**

Ans. : Linear Regression

- In simple words,

***Definition** : The process of finding a straight line that best approximates a set of points on a graph is called linear regression.*

- The word linear signifies that the type of relationship that you try to establish between the variables tends to be a straight line. Linear regression is one of the most simple and popular techniques of regression analysis.
- There are two types of linear regression.

1. **Simple Linear Regression (SLR) :** This has only one independent (or input) variable. For example, number of litres of petrol and kilometres driven. It is also called as univariate regression.
2. **Multiple Linear Regression (MLR) :** This has more than one independent (or input) variables. For example, number of litres of petrol, age of the vehicle, speed and kilometres driven:

   A variation of Multiple Linear Regression is Multivariate Regression. In Multivariate regression, there are multiple inputs and multiple possible outputs.

   The general formula (or model) for linear regression analysis is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_n X_n + \varepsilon$

   Where,

   - Y is the outcome (dependent) variable

- X_i are the values of independent variables
- β_0 is the value of Y when each X_i is equal to 0. It is also called as y-intercept
- β_i is the change in Y based on the unit change in X_i. It is also called as regression coefficient or slope of the regression line.
- ε is the random error or noise that represents the difference between the predicted value (based on the regression model) and actual value.

- For Simple Linear Regression (just one input or independent variable), the formula is exactly what you learnt in your school (remember y = mx + c)? It is

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

- Most of the times, ε is omitted from the calculation to give the simple formula as

$$Y = \beta_0 + \beta_1 X_1$$

- The goal is to find the regression line that best approximates (or fits) the relation between the input variable and output variable. You are required to calculate the values of β_0 and β_1(the regression coefficients). To do so, you could use the least square method in which you calculate the square of distance between each observed point and the probable regression line. The objective is to find a linear model where the sum of squares of the distances is minimal.
- To calculate β_0 and β_1, you can use the following formulae.

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

where $\bar{x}$ is the mean of x and $\bar{y}$ is the mean of y. $\beta_0 = \bar{y} - \beta_1 \bar{x}$

Q. 22 Given the following data for the sales of car of an automobile company for six consecutive years. Predict the sales for next two consecutive years. **(6 Marks)**

Year	Sales
2013	110
2014	100
2015	250
2016	275
2017	230
2018	300

Ans. : Calculate the mean of Year and Sales and then calculate the values as required.

Year	Sales	$X_i - X_{mean}$	$Y_i - Y_{mean}$	$(X_i - X_{mean}) \times (Y_i - Y_{mean})$	$(X_i - X_{mean})^2$
2013	110	- 2.5	- 100.8333333	252.0833333	6.25
2014	100	- 1.5	- 110.8333333	166.25	2.25
2015	250	- 0.5	39.16666667	- 19.58333333	0.25
2016	275	0.5	64.16666667	32.08333333	0.25
2017	230	1.5	19.16666667	28.75	2.25
2018	300	2.5	89.16666667	222.9166667	6.25
$\bar{X}$(Mean of X) = 2015.5	$\bar{Y}$(Mean of Y) = 210.83			682.5	17.5

Hence, $\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$\beta_1 = \frac{682.5}{17.5} = 39$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$

$\beta_0 = 210.83 - 39 \times 2015.5 = -78,393.67$

Hence, the regression line is

$Y = \beta_0 + \beta_1 X_1]$

$Y = -78,393.67 + 39X$

For the year 2019

Sales $Y = -78,393.67 + 39 \times 2019 = 347$ units

For the year 2020

Sales $Y = -78,393.67 + 39 \times 2020 = 386$ units

Q. 23 Describe some of the use cases/applications for Linear Regression. **(4 Marks)**

Ans. : Use Cases (or Applications of) for Linear Regression

Some of the common use cases (or applications) of linear regression are as following.

1. **Healthcare :** As you understand by now, healthcare industry is evolving and there are several researches that are going on. Linear regression could be used to establish the relationship between treatment and its effects or to understand complex operations of the human body to derive certain relationships. For example, you could study the effect of a particular drug chemical substance to reduce the level of infection in blood. 1 mg of substance could reduce the infection by 20% and 3 mg could reduce by 50% and so on.
2. **Demand forecasting :** Businesses are always looking to maximise sales and reduce inventory. Sales might depend on several factors and it could be really helpful to determine the relationship of sales with those factors. Businesses can then try to modify those factors (through various promotional schemes) and appropriately forecast sales.
3. **Other predictions :** There are several other areas where predictions can be made using the established linear relationship between the variables. It could be sports outcomes, crop output, machinery performance, fitness, and other similar areas.

Q. 24 What are the applications of logistic regression? **(4 Marks)**

Ans. : Use Cases (or Applications of) for Logistic Regression

Some of the common applications are as following.

1. **Finance:** Based on various input parameters such as credit score, potential income, age, etc. the probability of a loan application getting approved or rejected could be estimated. This probability prediction can be used to make the right judgement after a review.
2. **Sports:** Based on various variables such as score, weather condition, pitch condition, player's past track record, etc. you could predict the chances of the team winning or losing a match.
3. **Maintenance:** You can build up proactive maintenance schedules based on various factors such as machine age, hours of operations, working conditions, etc. The machinery breakdown in plants can cause downtimes that affects business. Having a predictive maintenance schedules (say when the probability of breakdown goes higher than 50% based on the input parameters) could ensure that the breakdown is minimal.

4. **Classification or categorisation:** Logistic regression could also be used to classify the entities and objects based on input variables. For example, based on the size of mangoes, their colour and weight, you could classify them into export quality or not export quality.

Q. 25 Write a short note on decision trees. **(4 Marks)**

OR Explain why decision trees are used. Also, draw a decision tree and explain its parts. **(6 Marks)**

Ans. : Decision Trees

***Definition** : Decision tree uses the concept of trees to structure the given information in the sequence of decisions and consequences.*

- Decision trees can work for both
  1. Categorical variables (Classification trees)
  2. Continuous variables (Regression trees)
- The basic idea here is that the given information (or datapoints or input variables) is structured in a tree fashion having branches and leaves. The information is split into branches based on the most significant input variables until further branching is not possible and you just end up with leaves.
- For example, the Fig. 4.4 shows a simple animal classification tree.

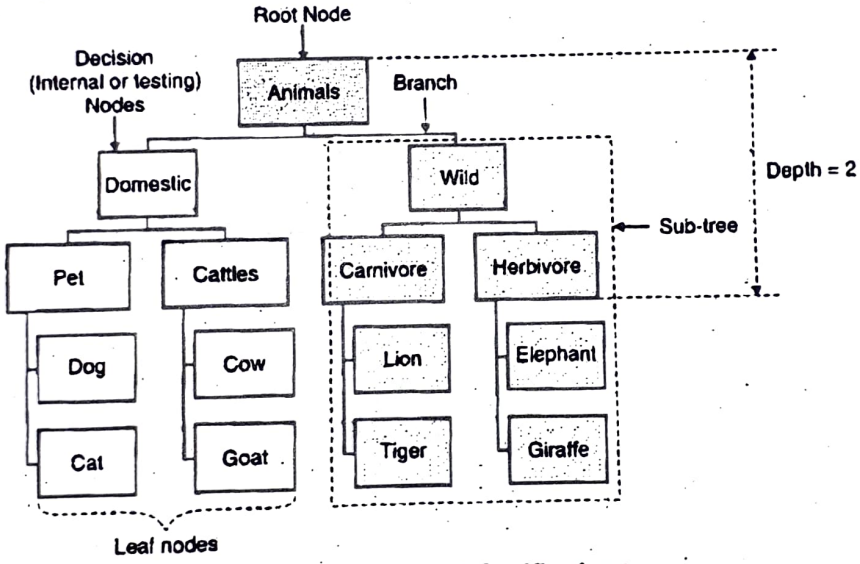


Fig. 4.4 : Simple animal classification tree

- The top of the tree is called the root node. Yes, look at decision trees as inverted trees where root is on the top!
- Each split point is called a branch. A branch holds similar items.
- A branch could have a name, label or value assigned to it as well. Each branch can be considered as a sub-tree which could be further branched out.
- A branch connects two nodes. Note here that there could be several branches coming out of a node. It need not be always two as shown in the example here.
- For example, you could add a third branch to the animal root node as *Aquatic* and add water-based animals such as fish, crocodile, and frog. You could also add a fourth branch called *Insects* and add honeybee, mosquito, and cockroach to it.
- The node attached to the branch is called as internal, testing or decision node. A decision node is used to traverse (navigate) the tree downwards and make decisions, classification, or predictions.
- Finally, nodes that cannot be further branched out are called leaf nodes. Leaf nodes provide the ultimate prediction.

- The depth of a node is the minimum distance required from the root node to reach it. For example, the node *Herbivore* is at a depth of 2 from *Animals* root node.

Q. 26 Write a short note on entropy. **(4 Marks)**

OR With an example, explain entropy. **(6 Marks)**

Ans. : Entropy

The dictionary meaning of entropy is "the degree of disorder or uncertainty in a system". In decision tree context,

***Definition** : Entropy measures randomness or impurity of an attribute (or datapoint) in a decision tree.*

Entropy generally lies between 0 and 1. A lower value of entropy is desired.

A lower value of entropy signifies a homogeneous dataset with less randomness (hence better predictions). A high entropy indicates high disorder. Entropy could also be more than 1. It then signifies that the dataset is very random and is not good for creating a prediction or classification model.

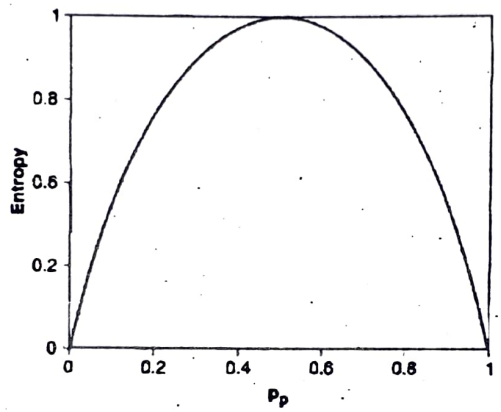


Fig. 4.5 : Entropy

- A probability of 0.5 makes entropy equal to 1 which means equally divided samples (not good for predicting)
- A probability of 1 makes entropy 0 which means prediction with 100% accuracy
- The goal is to have low entropy to ensure that you can predict the event either happening or not happening
- Entropy is calculated using the following formula :

$$\text{Entropy } H_x = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Where x is the random variable with n outcomes. The log in the formula is taken as log to the base 2 ($\log_2$). $p(x_i)$ is the probability or distribution of the input variable.

Q. 27 Calculate the entropy of the following distribution. **(4 Marks)**

Gender	Count
Male	9
Female	5

Ans. :

$$p(\text{Male}) = \frac{9}{14}$$

$$p(\text{Female}) = \frac{5}{14}$$

So, for calculating the entropy of the distribution you would use the following formula.

$$H_x = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$$\text{Entropy} = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right)$$

$$\text{Entropy} = -(-0.53 - 0.40)$$

$$\text{Entropy} = 0.93$$

Q. 28 Write a short note on Information Gain. **(4 Marks)**

OR With an example, explain Information Gain. **(8 Marks)**

Ans. :

Information Gain

To construct a good decision tree, you need to make optimal choices for splitting the data into branches. But, how do you decide which attributes to choose for splitting the data? information gain is a measure that helps you to make the splitting decision.

Definition : Information gain is a measure of purity produced by an attribute.

It is a representation of how well the given attribute separates out the given dataset based on the desired target classification.

- Information gain is high if the chosen attribute produces near pure subset of datapoints (subset containing majority of one class of datapoints) after split.
- Information gain is low if the chosen attribute does not produce near pure subset of datapoints (subset containing majority of one class of datapoints) after split.
- To calculate how impure a split subset is, you calculate its entropy.

Information gain is calculated as

IG = Entropy (before split) – Entropy (after split for all subsets)

At each step in constructing the decision tree, you choose to split the data on the attribute with the highest value in information gain as this leads to the purest subsets. Information gain compares the degree of purity of the parent node before a split with the degree of purity of the child node after a split. At each split, an attribute with the greatest information gain is considered the most informative attribute.

Q. 29 Write a short note on Gini Index. **(4 Marks)**

Ans. :

- Gini index is another purity measurement method used for splitting the datapoints in a decision tree. It only works with categorical variables ("Success" or "Failure").
- Gini index is calculated by squaring the probabilities of "success" and "failure" of finding the datapoints in the subset after a split and multiplying them with the weighted size of the datapoints in the subset after the split to the size of the datapoints in the parent node before split.
- Higher the value of the Gini index, higher is the homogeneity of the sub-node.

Gini for sub – nodes $= (p_i)^2$

$$\text{Gini Index} = \sum_{i=1}^{k} w_i (p_i)^2$$

Where P_i is the probability of the datapoints in a set and w_i is the weight of the sub-node after the split. k is the number of sub-nodes formed.

- There is also a related term called Gini Impurity which is given as

  Gini Impurity = 1 – Gini Index

Q. 30 Explain and write the general algorithm for drawing decision trees. **(6 Marks)**

Ans. : The General Algorithm

This is not an algorithm in itself but just for your conceptual understanding on how the actual decision tree algorithms work at a high level.

Assume that

- You need to form a tree, T
- From dataset, S
- So as to get leaf nodes (classification), C
- You take help from purity information attributes, A, to decide the split

The general algorithm is

1. Start with root node to form a tree T containing all datapoints in the dataset S
2. If the node(s) purity is below the purity threshold or not all the records of S belong to class C, then use the purity information attribute A to split the node. This creates sub-tress.
3. Repeat step 2 until
   (a) All the leaf nodes satisfy minimum purity threshold
   (b) The tree cannot be further split
   (c) Any other stopping criteria is reached (such as maximum tree depth desired)

- Following these steps you get the desired decision tree that you can use for predictive classification or regression.

Q. 31 Explain and write the ID3 algorithm for drawing decision trees. **(6 Marks)**

Ans. :

- ID3 is one of the earliest and popular decision tree algorithms. It is an acronym for Iterative Dichotomiser 3. The dictionary meaning of the word dichotomise is to "divide into two parts, classes, or groups" – that is specifically what ID3 algorithm (or for that matter any other decision tree algorithm) does.
- ID3 algorithm uses information gain as the purity attribute to decide the split.
- The algorithm works as following.
  1. Create a root node for the tree with all the datapoints in the dataset. If all the datapoints have the same classification, then no further branching is required. Stop processing. The tree would just have one node with the given classification in the dataset. (So, in this scenario, no matter what input variables you have, you have the same output classification. So, the decision tree is not required. This is a corner and fictious case. Don't think too much about it, move on.)
  2. Iteratively, compute information gain for all the available input attributes with respect to the parent node. Pick the attribute with the highest information gain and split the tree based on it. Keep working for the same branch until it has reached the leaf node before moving to the next branch. If you do not have any further input data attributes and just the classification field in the dataset then pick the most frequently occurring classification in the dataset and create the node for it.
  3. Once step 2 is complete (no further branching is possible), you get the desired decision tree.

Q. 32 Explain and write the C4.5 algorithm for drawing decision trees. **(6 Marks)**

Ans. :

C4.5 is an extension to ID3 algorithm. It uses gain ratio to decide on the split. C4.5 algorithm provides several improvements over ID3 algorithm. Those improvements are as following.

1. It can handle missing data attributes from the dataset
2. It can handle missing values for data attributes
3. It uses pruning to simplify the decision tree
4. It can handle continuous variables such as temperature and cost
5. It provides better error handling

C4.5 algorithm works as following.

1. Create a root node for the tree with all the datapoints in the dataset. If all the datapoints have the same classification, then no further branching is required. Stop processing. The tree would just have one node with the given classification in the dataset.
2. Iteratively, compute gain ratio for all the available input attributes with respect to the parent node.

   Pick the attribute with the highest gain ratio and split the tree based on it. Keep working for the same branch until it has reached the leaf node before moving to the next branch. If you do not have any further input data attributes and just the classification field in the dataset then pick the most frequently occurring classification in the dataset and create the node for it.
3. Once step 2 is complete (no further branching is possible), you get the desired decision tree.

Q. 33 Explain and write the CART algorithm for drawing decision trees. **(6 Marks)**

Ans. :

CART is an acronym for Classification and Regression Trees. CART can handle categorical as well as continuous variables. Unlike ID3 and C4.5 that use entropy-based criteria for split, CART uses Gini Index to decide on the split.

CART algorithm works as following

1. Create a root node for the tree with all the datapoints in the dataset. If all the datapoints have the same classification, then no further branching is required. Stop processing. The tree would just have one node with the given classification in the dataset.
2. Iteratively, compute Gini Index for all the available input attributes with respect to the parent node. Pick the attribute with the highest Gini Index and split the tree based on it. Keep working for the same branch until it has reached the leaf node before moving to the next branch. If you do not have any further input data attributes and just the classification field in the dataset then pick the most frequently occurring classification in the dataset and create the node for it.
3. Once step 2 is complete (no further branching is possible), you get the desired decision tree.

Q. 34 Write a short note on Naïve Bayes. List its applications. **(4 Marks)**

Ans. : Naïve Bayes provides another approach for classifying the data based on input variables.

Definition : Naïve Bayes is a family of probabilistic classifiers based on Bayes' theorem (or Bayes' law) that uses the relationship between the probabilities of events for classification.

- Given an observation of an input, a probabilistic classifier predicts a probability distribution over a set of classes. It does not just provide the most likely class that the datapoint should belong to but rather the probability distribution of the observation falling under multiple classes.

- Naïve Bayes takes a naive approach towards classification. The dictionary meaning of the word naive is "unaffected simplicity" or simply put "foolish".
- The naive assumption is that the occurrence of a certain feature (or data attribute) is independent of other features (or data attributes).
- It uses Bayes' algorithm and takes a naive approach towards classification and hence the name Naïve Bayes. Just remember that for now.
- Naïve Bayes has similar usage and application as other classifiers have such as
  1. Email spam filtering
  2. Fraud detection
  3. Diagnosing healthcare problems
  4. Text classification
  5. General classification predictions

Q. 35 You have designed a spam email detector based on Naïve Bayes classifier that marks emails as spam if the email body has both the words "gift" and "won". It is observed that for 100 emails

(i) 20 out of 25 spam emails have the word "gift" in the email body

(ii) 5 out of 75 non-spam emails have the word "gift" in the email body

(iii) 15 out of 25 spam emails have the word "won" in the email body

(iv) 10 out of 75 non-spam emails have the word "won" in the email body

What would be the accuracy of your spam detector? **(6 Marks)**

Ans. :

$$P(\text{Spam} \mid \text{Gift}) = \frac{P(\text{Gift} \mid \text{Spam}) \times P(\text{Spam})}{P(\text{Gift})} = \frac{\frac{20}{25} \times \frac{25}{100}}{\frac{25}{100}} = \frac{4}{5}$$

$$P(\text{Spam} \mid \text{Won}) = \frac{P(\text{Won} \mid \text{Spam}) \times P(\text{Spam})}{P(\text{Won})} = \frac{\frac{15}{25} \times \frac{25}{100}}{\frac{25}{100}} = \frac{3}{5}$$

$$P(\text{Not Spam} \mid \text{Gift}) = \frac{P(\text{Gift} \mid \text{Not Spam}) \times P(\text{Not Spam})}{P(\text{Gift})} = \frac{\frac{5}{75} \times \frac{75}{100}}{\frac{25}{100}} = \frac{1}{5}$$

$$P(\text{Not Spam} \mid \text{Won}) = \frac{P(\text{Won} \mid \text{Not Spam}) \times P(\text{Not Spam})}{P(\text{Won})} = \frac{\frac{10}{75} \times \frac{75}{100}}{\frac{25}{100}} = \frac{2}{5}$$

According to Naïve Bayes classifier,

$$P(\text{Spam} \mid \text{Gift, Won}) = P(\text{Spam} \mid \text{Gift}) \times P(\text{Spam} \mid \text{Won})$$

$$P(\text{Spam} \mid \text{Gift, Won}) = \frac{4}{5} \times \frac{3}{5} = \frac{12}{25}$$

Hence, out of 25 spam emails, $\frac{12}{25} \times 25 = 12$ emails have gift and won

$$P(\text{Not Spam} \mid \text{Gift, Won}) = P(\text{Not Spam} \mid \text{Gift}) \times P(\text{Not Spam} \mid \text{Won})$$

$$P(\text{Not Spam} \mid \text{Gift, Won}) = \frac{1}{5} \times \frac{2}{5} = \frac{2}{25}$$

Hence, out of 75 not spam emails,

$\frac{2}{25} \times 75 = 6$ emails have gift and won

Hence, accuracy of the spam detector is

$$\text{Accuracy} = \frac{\text{Spam}}{\text{Spam + Not Spam}} = \frac{12}{12+6} = \frac{12}{18} = \frac{2}{3} = 66.67\ \%$$

Q. 36 Write a short note on Smoothing. **(4 Marks)**

Ans. :

- To calculate the probability of the overall event, Naïve Bayes classifier multiplies the probabilities of respective data attribute values. A problem with this approach is that if the probability of any data attribute value is 0 (or very close to 0), then the entire probability of the event becomes 0 (or too low) even if there are high probabilities for other data attribute values.

***Definition** : Smoothing is a technique that adjusts the probabilities of rare or non-occurring attribute values to avoid overall probability of the event becoming 0 (or too low).*

- There are several smoothing techniques. A common smoothing technique is Laplace smoothing. It adds 1 to all the probabilities before multiplying them for Naïve Bayes classifier. It is represented as

$$P^*(x) = \frac{\text{Count}(x) + 1}{\sum_x (\text{Count}(x) + 1)}$$

Or simply, $$P(x)_{smooth} = \frac{x+1}{N+V}$$

Where,

- N is the total number of observations in the particular classification and
- V is the total number of distinct classifications

Q. 37 List the advantages and disadvantages of Naïve Bayes classifier. **(6 Marks)**

Ans. : Advantages of Naïve Bayes Classifier

Naïve Bayes classifier has the following advantages.

1. It can handle missing data attribute values.
2. It can handle irrelevant input variables.
3. It is simple to implement without requiring special software libraries.
4. It is computationally fast.
5. It performs better with categorical variables when compared with decision trees.

Disadvantages of Naïve Bayes Classifier

Naïve Bayes classifier has the following disadvantages.

1. It assumes that the data attributes are conditionally independent.
2. In general, it is not very reliable for probability estimation.
3. Usually, it can be used only with the categorical variables. Continuous variables need to be converted into categorical variables for working with them.
4. If there are a lot of data attributes, probability calculations tend to be very low or near zero. You require additional steps such as taking logarithms or apply smoothing.

Q. 38 Explain confusion matrix with an example. **(6 Marks)**

Ans. :

***Definition** : A confusion matrix lays out the comparison between the predicted classification and actual classification to provide various performance measures.*

The number of correct and incorrect predictions are written for each class. Confusion matrix provides calculations for various types of errors in prediction and helps to improve your classification model.

Following is how a confusion matrix looks like for two possible classifications.

Confusion Matrix for Class 1 and Class 2		Actual Class	
		Class 1	Class 2
Predicted Class	Class 1	True Positives (TP)	False Positives (FP)
	Class 2	False Negative (FN)	True Negative (TN)

- **True Positive** = Correct True Prediction for Class 1. This is desired.
  - Predicted Class = Actual Class : Correct classification
- **True Negative** = Correct False Prediction for Class 1 (True Prediction for Class 2). This is desired.
  - Predicted Class = Actual Class : Correct classification
- **False Positive** = Incorrect True Prediction for Class 2. This is Type I error.
  - Predicted Class ≠ Actual Class : Incorrect classification
- **False Negative** = Incorrect False Prediction for Class 2. This is Type II error.
  - Predicted Class ≠ Actual Class : Incorrect classification.

□□□

Unit V : Big Data Analytics and Model Evaluation

Chapter 5 : Big Data Analytics and Model Evaluation

Q. 1 Write a short note on clustering. **(4 Marks)**

Ans. :

- The dictionary meaning of cluster is "a number of similar things that occur together". For example, cluster of stars in the galaxy.
- With respect to machine learning,

***Definition** : Clustering is a technique in which the data points are arranged in similar groups dynamically without any pre-assignment of groups.*

- It is the task in which the data points are grouped in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters).
- It is a data exploration technique commonly used to understand how the data should be interpreted and grouped to be best analysed. There is no prior learning of groupings required. Instead, groups are implicitly arranged (or created) based on the data attributes. How the data is grouped depends on the type of algorithm used. The same dataset can be used to form various clusters depending upon the data attributes and then you can decide which clusters make sense to be used for further data analytics.
- For example, here is a simple plot of data points. As you can see, some set of points are closer to each other when compared with others. These sets could possibly form a group (or a cluster) for further data analytics.

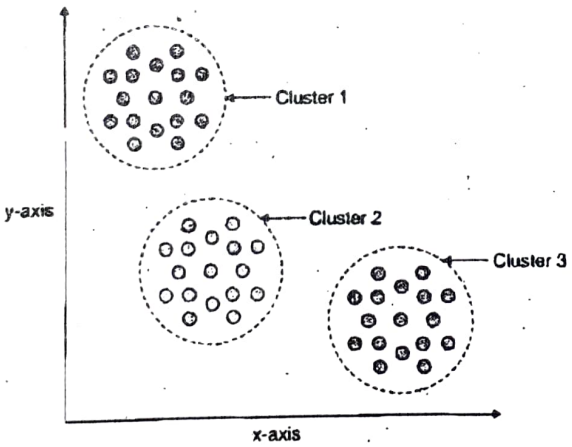


Fig. 5.1

Q. 2 Explain the properties of a cluster. **(4 Marks)**

Ans. : Properties of a Cluster

Typically clusters have the following properties.

1. **All the data points in a cluster should be similar to each other :** By definition, a cluster is a grouping of similar objects. So, a good cluster must have data points that indeed have some convincing similarities on the basis of which they are grouped. For example, phone preference cluster mentioned in the example in the previous section could be a convincing enough cluster.
2. **The data points from different clusters should be as different as possible :** To make clusters distinct enough, the data points from different clusters should be far apart or, in other words, must be clearly distinguishable. For example, a cluster containing all customers whose preference is iPhone Model X is clearly distinguishable from another cluster containing customers preferring Motorola X.

Q. 3 Explain types of clustering. **(4 Marks)**

Ans. : Types of Clustering

At a high-level, there are two types of clustering.

1. **Hard Clustering :** In this, each data point belongs to only one cluster at a time. For example, customer A (in the previously given example) could only be in iPhone X cluster if you clustered based on phone model preference.
2. **Soft Clustering :** In this, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, customer A would be in both the clusters -iPhone X and Motorola X with respective probabilities of 1 and 0.

Q. 4 Explain 3 Use Cases of Clustering. **(6 Marks)**

OR With example, describe customer segmentation using clustering. **(4 Marks)**

OR With example, describe image processing using clustering. **(4 Marks)**

OR With example, describe the use of clustering in the healthcare industry. **(4 Marks)**

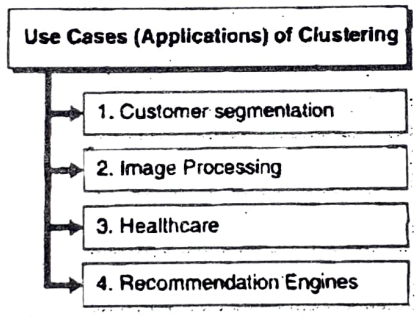
OR With example, describe the use of clustering in recommendation engines. **(4 Marks)**

Ans. : Use Cases (Applications) of Clustering

There are several use cases or applications of clustering technique. These applications are not specific to K-means or any other clustering algorithm.

Depending on the dataset to be clustered and the desired clustering outcome any algorithm could be chosen.

Some of the common applications of clustering are as following.



1. Customer Segmentation

- One of the most common applications of clustering is customer segmentation. Various companies build customer segment profiles to understand the customer's shopping behaviour and accordingly build strategies for predicting demand and improving sales.
- **Example :** A woman buying clothes for her 3-month old infant. Based on the size chosen, the company could send her promotional offers for next size of clothes she is likely to purchase. For Example, if she bought cloth size fitting 3 months old infant, after 9 months, she could be sent promotional offers on cloth size for 1 year old baby. The chances of her buying clothes for 1 year old baby are quite high. The company could identify such clusters of mothers who would require 1 year old baby clothes and accordingly send them promotional offers.

2. Image Processing

- Using clustering technique, you could identify objects in an image. You can cluster similar pixels in the image together and then match that with known objects to identify the objects in the image.
- Additionally, you can use this technique to capture image frames from a video and then determine objects in the video. Based on the objects you can categorise the video (for example, a video having animals), probable location of the video (if you could identify a famous location such as the Gateway of India) or identify people in the video. Image and video based analytics have reached very advanced level these days where you could carry out various useful tasks.

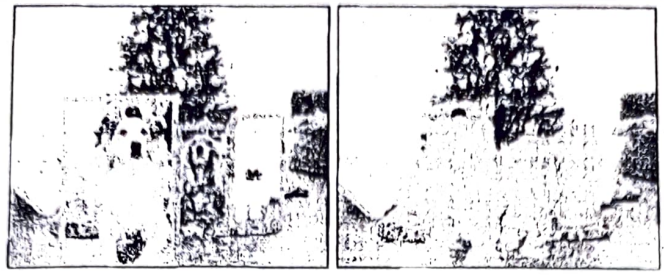


Fig. 5.2(a)

3. **Healthcare**

- Patient data, containing attributes such as blood pressure, height, weight, cholesterol level and glucose level, could be used to create clusters and detect any early signs of health problems based on historical records of other patients who had similar characteristics and then they were diagnosed with a particular health problem. For example, if someone has high cholesterol level and weight, it might be possible to detect if she is closer to getting a heart attack. There could be several other uses of clustering in the healthcare industry.
- For example, detecting cancer, its type and severity and the possible treatment plan at early stage could have a huge impact on number of years the patient is expected to survive. The shape of the cancerous cells plays a vital role in determining the severity of the cancer. Using clustering, you can determine the shape of the cancerous cells and accordingly create a treatment plan.

Fig. 5.2(b)

4. **Recommendation Engines** : Some of the examples are as shown in Fig. 5..2(c)

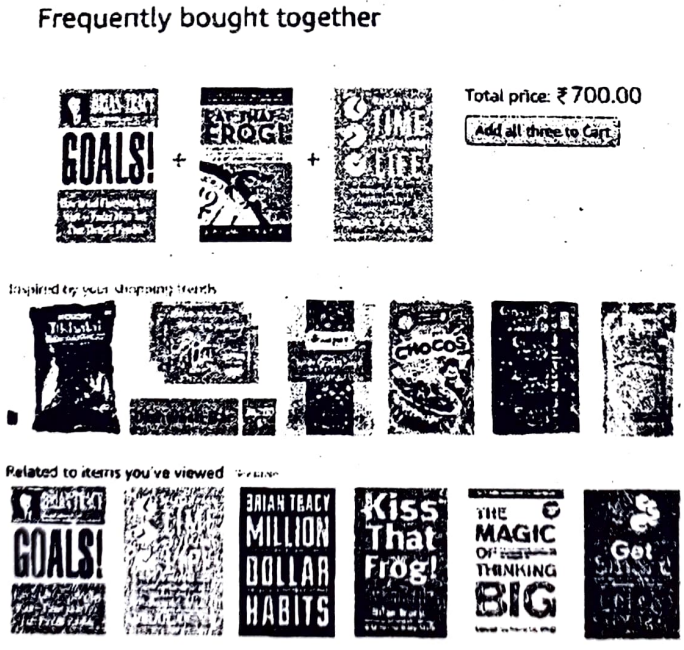


Fig. 5.2(c)

- The organisations have the data for you (such as what you bought, what items you looked at, where you clicked, etc.) and several of their other site or app visitors. Based on how similar you are when compared with others (creating clusters), recommendations are made to you.
- The collected data is clustered and analysed to answer questions such as "how likely it is that someone who has bought the product X will also buy product Y?". If the probability is quite high, then the recommendation is made. For example, if customer *A* has bought Book 1 and also Book 2 historically and then customer *B* has just bought Book 1 now, she is recommended that Book 2 might be of interest to her.

Q. 5 Explain the K-means algorithm. **(6 Marks)**

OR Explain the steps involved in the K-means algorithm. **(6 Marks)**

Ans. :

K-means

- K-means is one of the popular clustering techniques (or algorithm). It helps to form clusters from the given dataset for further data analytics. Each data point belongs to only one cluster.
- You pre-decide the number of clusters, *k*, that you want to group your datapoints into before executing the algorithm steps. Let's learn about how it works.

Overview of the Method

- K-means algorithm is designed to minimise the sum of distances between the data points and their respective cluster centroid. Each cluster is associated with a centroid.

In coordinate geometry,

Definition: *Centroid is a point whose coordinates are the averages of the corresponding coordinates of a given set of points.*

At a high-level, the following steps are taken for clustering the data using K-means clustering algorithm.

1. Decide the number of clusters, *k*, that you desire to group your data points into.
2. Select *k* random data points as centroids.
3. Compute the distance from each data point to each centroid. Assign all the data points to the closest cluster centroid.

   The distance *d* between any two points, (x_1, y_1) and (x_2, y_2) is calculated as

   $$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

4. Recompute the centroids of newly formed clusters. The centroid (X_c, Y_c) of the *m* data points in a cluster is calculated as

   $$(X_c, Y_c) = \left(\frac{\sum_{i=1}^{m} X_i}{m}, \frac{\sum_{i=1}^{m} Y_i}{m} \right)$$

   It is a simple arithmetic mean of all X coordinates and Y coordinates of the *m* data points in the cluster.
5. Repeat steps 3 and 4 until any of the following criteria is met
   a. Centroids of newly formed clusters do not change.
   b. Points remain in the same cluster.
   c. Maximum number of iterations are reached as desired.

Q. 6 A bank has received the following loan applications. Which of the applications could be risky to approve? **(6 Marks)**

Credit Score (out of 1000)	Amount in Lakhs
500	10
726	25
430	5
678	15
780	30
380	10
645	15
890	50
900	65
450	10

Ans. : Let us use K-means clustering for grouping the loan applications. Let's choose the desired number of clusters, k = 2 and the number of iterations for centroid calculation and cluster assignment to be 2 as well. Let's put credit score on X-axis and loan amount on Y-axis. The first plot of data points is as shown in Fig. 5.3(a)

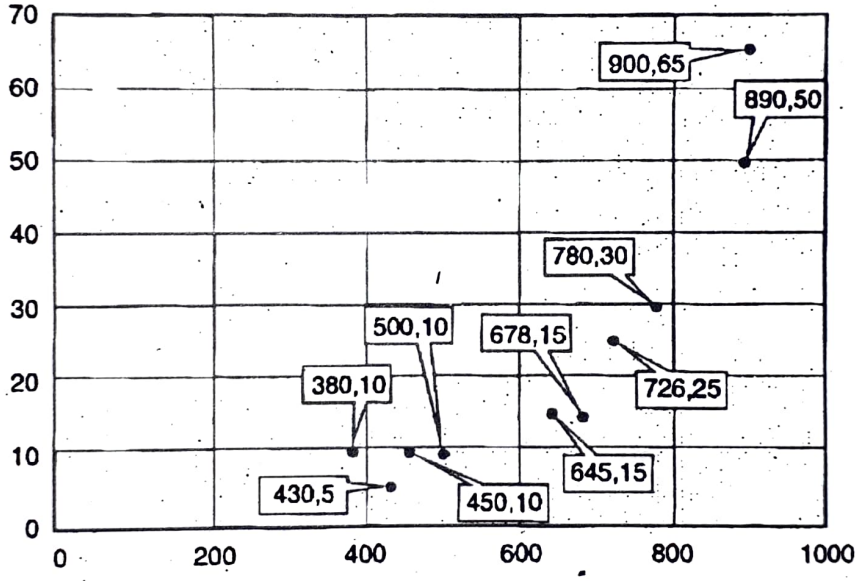


Fig. 5.3(a)

Iteration 1

Let's randomly choose two centroids.

Centroid 1 = (430, 5) and Centroid 2 = (726, 25).

Now, let's calculate the distance of the data points from the chosen centroids and complete the data points table. The data point is assigned to the cluster based on the closest centroid.

A sample distance calculation is as following for the first data point.

Distance from Centroid 1 (430, 5) for (500, 10) =

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$d = \sqrt{(430 - 500)^2 + (5 - 10)^2}$$

$$d = 70.18$$

Credit Score	Amount	Distance from Centroid 1	Distance from Centroid 2	Assigned Cluster
500	10	70.18	226.50	1
726	25	296.67	0.00	2
430	5	0.00	296.67	1
678	15	248.20	49.03	2
780	30	350.89	54.23	2
380	10	50.25	346.32	1
645	15	215.23	81.61	2
890	50	462.20	165.89	2
900	65	473.81	178.54	2
450	10	20.62	276.41	1

Now re-calculate the centroids for the next iteration.

For Centroid 1 calculation, you have four data points that fall in cluster 1.

Hence, Centroid 1 is the mean of the data points which is

$$\left(\frac{500 + 430 + 380 + 450}{4}, \frac{10 + 5 + 10 + 10}{4}\right) = (440, 8.75).$$

For Centroid 2 calculation, take the remaining six data points that fall in cluster 2.

Hence, Centroid 2 is the mean of these six data points which is

$$\left(\frac{726 + 678 + 780 + 645 + 890 + 900}{6}, \frac{25 + 15 + 30 + 15 + 50 + 65}{6}\right) = (769.83, 33.33)$$

Now, you have both the centroids ready for the next and final iteration.

Iteration 2

Centroid 1 = (440, 8.75) and Centroid 2 = (769.83, 33.33).

Now, let's calculate the distance of the data points from the centroids and complete the data points table. The data point is assigned to the cluster based on the closest centroid.

Credit Score	Amount	Distance from Centroid 1	Distance from Centroid 2	Assigned Cluster
500	10	60.01	270.84	1
726	25	286.46	44.61	2
430	5	10.68	341.01	1
678	15	238.08	93.64	2
780	30	340.66	10.70	2
380	10	60.01	390.53	1
645	15	205.10	126.17	2
890	50	451.89	121.32	2
900	65	463.43	133.97	2
450	10	10.08	320.68	1

You stop here because the number of iterations that you decided are completed and also you see that the clusters assigned in the iteration 1 for the data points did not change.

Hence, you got the two clusters as shown in Fig. 5.3(b).

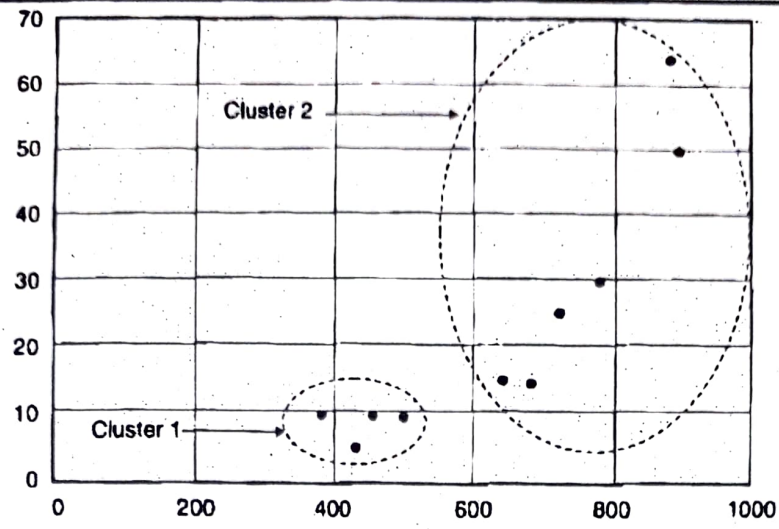


Fig. 5.3(b)

From the clusters formed, it can be concluded that people in cluster 2 seem to be less risky for granting loan.

Q. 7 A cluster has the following data points. Calculate its inertia. Assume 2nd data point to be the centroid for the cluster. **(6 Marks)**

Age	Income in Thousand
33	12
33	15
35	13
34	14
32	16

Ans. :

- To calculate inertia of a cluster, you need to find the distance of all data points from the centroid of the cluster and add them.
- Let's assume age on X-axis and income on Y-axis. Now, let's calculate the distance of the data points from the cluster centroid and complete the data points table.

A sample distance calculation is as following for the first data point. Distance from the Cluster Centroid (33, 15) for (33, 12)

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$d = \sqrt{(33 - 33)^2 + (15 - 12)^2}$$

$$d = 3$$

Age	Income	Distance from Cluster Centroid
33	12	3.00
33	15	0.00
35	13	2.83
34	14	1.41
32	16	1.41
Total		8.66

Hence, the cluster inertia is 8.66.

Q. 8 What are some of the drawbacks of K-means algorithm. **(6 Marks)**

OR. What are some of the challenges with K-means algorithm. **(6 Marks)**

Ans. :

Reasons to Choose and Cautions (Drawbacks / Challenges)

A few things to consider for K-means clustering are as following.

1. **Object attributes :** The data points could have several attributes such as age, weight, height, income, etc.

   Once your clustering is complete, you need to ensure that the attributes would be available for new data points as and when added for future analysis. For example, if your existing clustering is based on customer ratings on a particular product, such rating may not be immediately available for a customer who has just completed the purchase. It might require a week or a month before the customer is comfortable rating the product judiciously.

2. **Units of measurement and scaling :** You need to be careful in choosing the units for your data point attributes. While it may not impact K-means clustering a lot, it could actually shift a few data points here and there. For example, suppose you have two data points age and income. Age is expressed in double digits. But, if you choose income to be in thousands and someone earning 10,000 is represented as 10, then both age and income are in double digits and equally contribute towards distance calculation. But, suppose the income was not in thousands, then the income data attribute would dominate the distance calculation as it has many more digits than age. This might shift the data points more with respect to income than age and hence skew (bias or distort) the cluster assignment.

3. **Initial Centroid position :** K-means is sensitive to initial centroid position. Hence, you should run the K-means analysis several times to ensure that clusters created are optimum.

4. **Number of clusters :** As discussed earlier, you must be careful while deciding on the number of clusters required for optimum placement of data points. Cluster Inertia is a key parameter to consider for ensuring that you have the right number of clusters.

Q. 9 Write a short note on k-Nearest Neighbours (kNN) classification algorithm. **(4 Marks)**

Ans. :

k-Nearest Neighbours (kNN) Classification Algorithm

- One of the popular variations of K-means clustering algorithm is k-Nearest Neighbours (kNN) classification algorithm.
- It works on the same principle as K-means clustering algorithm that a data point is likely to resemble its neighbours and would possibly have the same classification. kNN is a supervised learning method.
- The way it works is simple and straightforward.
  1. Assume that you have already assigned labels to the existing data points in a given data set.
  2. Then you are given a new data point to classify.
  3. You calculate the distance of the new data point with respect to its *k* neighbours. The number *k* is chosen based on your data set and requirements but in general higher the better to reduce the noise and avoid incorrect labelling.
  4. The new data point is classified based on the classification of the majority of the surrounding neighbour's classification.

Q. 10 Following is a dataset for weight of teens and whether they like Pizza.

Weight (Kgs)	Like Pizza?
78	Yes
54	No
69	Yes
73	Yes
59	No
48	No
82	No
65	Yes

Using k-Nearest Neighbours (kNN) Classification Algorithm determine if a teen weighing 63 Kgs likely to like Pizza? Use k = 3. **(6 Marks)**

Ans. :

Calculate the distance of the new data point (63 Kgs) with respect to other data points in the data set.

A sample distance calculation is as following for the first data point.

$$d = \sqrt{(x_1 - x_2)^2}$$

$$d = \sqrt{(78 - 63)^2} = 15$$

$$d = 15$$

Weight (Kgs)	Distance for 63 Kgs	Like Pizza?
78	15	Yes
54	9	No
69	6	Yes
73	10	Yes
59	4	No
48	15	No
82	19	No
65	2	Yes

Sort the table based on the distance.

Weight (Kgs)	Distance for 63 Kgs	Like Pizza?
65	2	Yes
59	4	No
69	6	Yes
54	9	No
73	10	Yes
78	15	Yes
48	15	No
82	19	No

Now, given that k = 3.

So pick top 3 rows of the sorted table.

Weight (Kgs)	Distance for 63 Kgs	Like Pizza?
65	2	Yes
59	4	No
69	6	Yes

Now, you have got 2 Yes, and 1 No. The majority classification is "Yes". Hence, a teen weighing 63 Kgs is likely to like Pizza.

Q. 11 Write a short note on hierarchical clustering. **(6 Marks)**

OR Write a short note on dendrogram. **(4 Marks)**

Ans. : Hierarchical Clustering

As shown in the Fig. 5.4(a), for some data, the "other" category is further split into sub-groups. Putting data together in this form helps to quickly understand the relationship within the data and also appropriate group and categorise them.

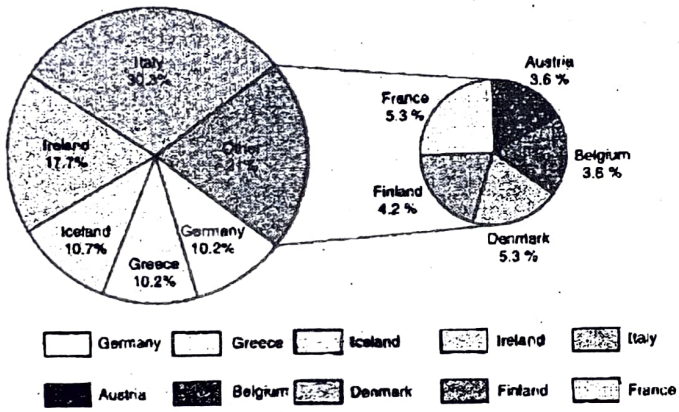


Fig. 5.4(a)

Hierarchical clustering is a similar concept.

Definition: *Hierarchical clustering, also called as hierarchical cluster analysis or HCA, is a method of cluster analysis in which the data points are arranged in a hierarchy of clusters.*

So, you not only cluster the datapoints, but you also have a hierarchy applied on the clusters that can be visually looked at to understand their relationships. The results of hierarchical clustering are usually presented in a dendrogram.

Dendrogram

Definition: *A dendrogram is a diagram representing a tree or hierarchy.*

- It is a branching diagram representing a hierarchy of categories based on degree of similarity or number of shared characteristics. For example, the Fig. 5.4(b) illustrates a simple dendrogram for six observations.

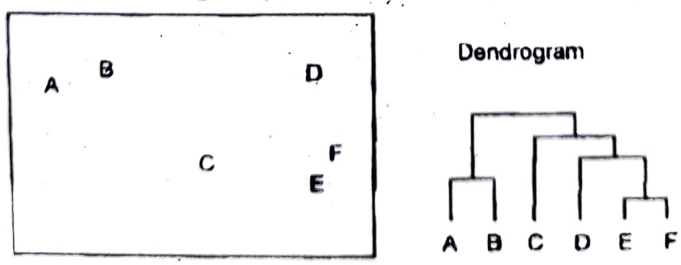


Fig. 5.4(b)

- The key to interpreting a dendrogram is to focus on the height at which any two objects are joined together. In the given example, you can see that E and F are most similar, as the height of the link that joins them together is the smallest. The next two most similar objects are A and B.
- The height of the dendrogram indicates the order in which the clusters were joined. A more informative dendrogram can be created where the heights reflect the distance between the clusters as shown in the Fig. 5.4(c).

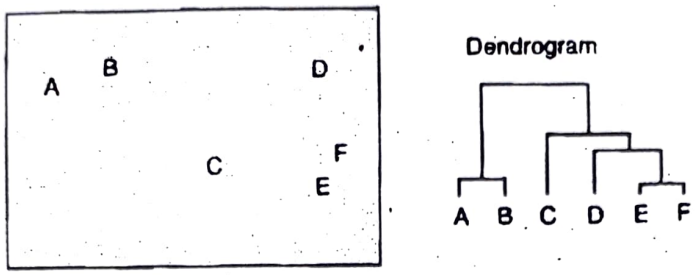


Fig. 5.4(c)

- In this case, the dendrogram shows that the big difference between clusters is between the cluster of A and B versus that of C, D, E, and F.
- For example, the dendrogram suggests that C and D are much closer to each other than is C to B, but the original data (shown in the scatter plot), shows that this is not true. The consequence of the information loss is that the dendrograms are most accurate at the bottom, showing which items are very similar.

Q. 12 List a few Agglomeration (Linkage) Methods. **(4 Marks)**

Ans. : The most common methods are as following.

1. **Maximum or complete linkage clustering :** This method computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the largest value of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters.
2. **Minimum or single linkage clustering :** This method computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the smallest of these dissimilarities as a linkage criterion. It tends to produce long, "loose" clusters.
3. **Mean or average linkage clustering :** This method computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2 and considers the average of these dissimilarities as the distance between the two clusters. The compactness of the clusters it creates can vary.
4. **Centroid linkage clustering :** This method computes the dissimilarity between the centroid for cluster 1 and the centroid for cluster 2.
5. **Ward's minimum variance method :** This method minimises the total within cluster variance. At each step, the pair of clusters with the smallest between-cluster distance are merged. It tends to produce more compact clusters.

Q. 13 Describe the applications of time series analysis. **(6 Marks)**

Ans. : Applications of Time Series Analysis

1. **Retail sales :** Time series analysis is useful in predicting future sales. These forecasts need to account for the seasonal aspects of the customer's purchasing decisions. For example, during the winter season, sweater sales are typically more, and swimsuit sales are the highest during the summer season. An appropriate time series model accounts for fluctuating demand over the calendar year.

2. **Spare parts planning :** Time series analysis is useful in forecasting future spare part demands to ensure an adequate supply of parts to repair customer products. Often the spares inventory consists of thousands of distinct part numbers. To forecast future demand, complex models for each part number can be built using input variables such as expected part failure rates, service diagnostic effectiveness, forecasted new product shipments, and forecasted trade-ins/decommissions. However, time series analysis can provide accurate short-term forecasts based simply on prior spare part demand history.

3. **Stock trading :** Stock traders typically utilise a technique called pairs trading. In pairs trading, an identified strong positive correlation between the prices of two stocks is used to detect a market opportunity. Suppose the stock prices of Company A and Company B consistently move together. Time series analysis can be applied to the difference of these companies' stock prices over time. A statistically larger than expected price difference indicates that it is a good time to buy the stock of Company A and sell the stock of Company B, or vice versa. Of course, this trading approach depends on the ability to execute the trade quickly and be able to detect when the correlation in the stock prices is broken. Pairs trading is one of many techniques that falls into a trading strategy called statistical arbitrage.

Q. 14 Describe various characteristics (components) of time series analysis. **(4 Marks)**

Ans. : Characteristics (Components) of Time Series Analysis

A time series can consist of the following components.

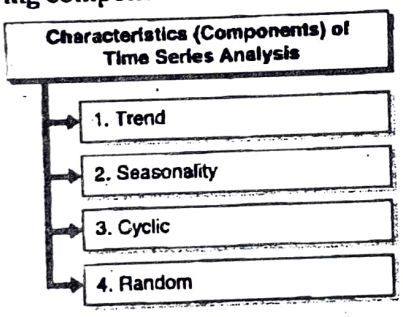


Fig. 5.5

1. **Trend :** The trend refers to the long-term movement in a time series. It indicates whether the observation values are increasing or decreasing over time. Examples of trends are a steady increase in sales month over month or an annual decline of fatalities due to car accidents.

2. **Seasonality :** The seasonality component describes the fixed, periodic fluctuation in the observations over time. As the name suggests, the seasonality component is often related to the calendar.

3. **Cyclic :** A cyclic component also refers to a periodic fluctuation, but one that is not as fixed as in the case of a seasonality component.

4. **Random :** After accounting for the other three components, the random component is what remains. Although noise is certainly part of this random component, there is often some underlying structure to this random component that needs to be modelled to forecast future values of a given time series. Randomness does not follow a specific pattern.

Q. 15 Explain the high-level steps involved in text analysis. **(4 Marks)**

Ans. : Steps in Text Analysis

The high-level steps involved in text analysis are as following.

1. **Parsing :** It is the process that takes unstructured text and imposes a structure for further analysis. The unstructured text could be a plain text file, a weblog, an Extensible Markup Language (XML) file, a Hypertext Markup Language (HTML) file, or a Word document. Parsing deconstructs the provided text and renders it in a more structured way for the subsequent steps.

2. **Search and retrieval :** It is the identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or entities like people or organizations. These search items are generally called key terms. Search and retrieval originated from the field of library science and is now used extensively by web search engines.

3. **Text mining :** It uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest. With the proper representation of the text, many of the techniques you learnt earlier, such as clustering and classification, can be adapted to text mining. For example, the k-means clustering can be modified to cluster text documents into groups, where each group represents a collection of documents with a similar topic. The distance of a document to a centroid represents how closely the document talks about that topic. Classification tasks, such as sentiment analysis and spam filtering, are prominent use cases for the naïve Bayes classifier. Text mining may utilise methods and techniques from various fields of study, such as statistical analysis, information retrieval, data mining, and natural language processing.

Q. 16 Describe a few text pre-processing techniques. **(6 Marks)**

Ans. : Text Pre-Processing Techniques

Some of the common text pre-processing techniques are as shown in Fig. 5.6.

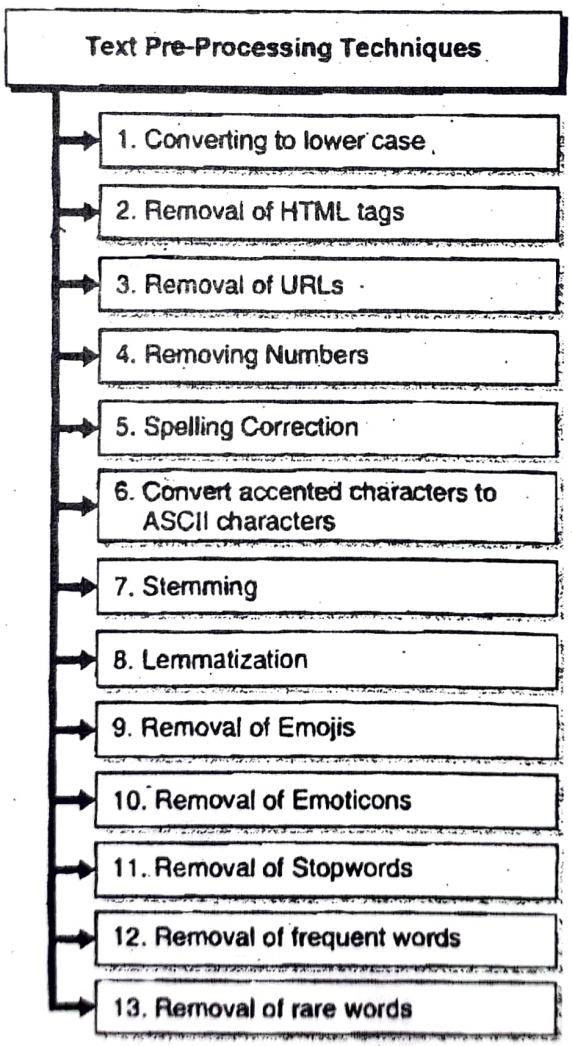


Fig. 5.6

1. **Converting to Lower case (Lowercasing) :** Converting all the text into the lower case is a simple and effective approach for text analysis. If you are not applying lower case conversion on words like DOG, doG, dog, dOG, then these all words would be treated as different.

Before lowercasing	After lowercasing
DOG	dog
dOG	dog
dOg	dog

2. **Removal of HTML tags :** The chances to get HTML tags in your text data is quite common specially when you are extracting or scraping data from different websites. You don't get any valuable information from these HTML tags. So, it is better to remove them from textual data.

Before HTML tags removal	After HTML tags removal
Data science project	Data science project

3. **Removal of URLs :** URL is the short-form of Uniform Resource Locator. The URLs within the text refer to the location of another website or anything else. These URLs are of no use to you in textual analysis. You can remove them.

Before URL removal	After URL removal
For more information go to https://www.google.com	For more information go to

4. **Removing Numbers :** If your analysis does not require numbers, you can remove them.

Before number removal	After number removal
The weight of panda is 650 Kgs	The weight of panda is Kgs

5. **Spelling Correction :** Similar to lowercasing, spelling correction is another important pre-processing technique that avoids treating wrongly spelled words different from correctly spelled words.

Before spelling correction	After spelling correction
Team India wno today	Team India won today

6. **Convert accented characters to ASCII characters :** You might have seen special characters at the top of the common letter or characters. These are accented characters. For example, e in the word résumé has accents. If you don't remove these, then the text analysis model will consider resume and résumé as different words, even if both are the same.

With accented characters	Without accented characters
I submitted my résumé on the portal	I submitted my resume on the portal

7. **Stemming :** Stemming is reducing words to their base or root form by removing a few suffix characters from words. Stemming is a text normalisation technique. There are various stemming algorithms but the most widely used one is porter stemming.

Before stemming	After stemming
Learning	Learn
Books	Book
Caring	Car
Obesity	Obes
Causes	Caus

But stemming doesn't always provide the correct form of words because it blindly follows the rules like removing suffix characters to get base words irrespective of ensuring correctness. Sometimes, stemmed words don't relate to original ones and sometimes they may also give non-dictionary or improper words.

8. **Lemmatization :** Lemmatization is similar to the stemming technique that aims to get the base words. But, unlike stemming that might produce improper words, the lemmatization process does not only trim the suffix characters but also uses lexical knowledge bases to get original words in their right forms.

   Hence, the result of lemmatization is better than stemming.

Before lemmatization	After lemmatization
Learning	Learn
Books	Book
Caring	Care
Obesity	Obesity
Causes	Cause

9. **Removal of Emojis :** In today's online communication, emojis play a very crucial role. Emojis are small images using which users may express their feelings. Until and unless you are making sense out of these emojis, you can remove them for text analysis.

With emojis	Without emojis
I am super happy ☺	I am super happy

10. **Removal of Emoticons :** Unlike emojis which are tiny images, an emoticon portrays a human facial expression using just keyboard characters, such as letters, numbers, and punctuation marks without using any images. Until and unless you are making sense out of these emoticons, you can remove them for text analysis.

With emoticons	Without emoticons
I am super happy ;)	I am super happy

11. **Removal of Stopwords :** Stopwords are common words that are mostly irrelevant for text analysis. For example, "a", "an", "the", "is", "for", etc.

With stopwords	Without stopwords
I ate an apple	I ate apple

12. **Removal of frequent words :** Stopwords are language specific. If you are working on text analysis for a particular domain, it may involve frequent words that may not give a lot of useful information. For example, the word "experiment" may appear several times if you are performing text analysis on a scientific research or thesis.

    You can remove such frequent words from text analysis.

With frequent word	Without frequent word
In the experiment it was found that	In the it was found that

13. **Removal of rare words :** You can also remove rare words from text analysis as it is unlikely to be found multiple times or provide any useful information.

With rare word	Without rare word
Petrichor was all around	Was all around

Q. 17 Write a short note on Bag-of-Words. **(4 Marks)**

Ans. : Bag-of-Words

- In bag-of-words (BoW), a text document is converted into a vector of counts. The vector contains an entry for every possible word in the vocabulary.
- If the word say, "aardvark" appears three times in the document, then the feature vector has a count of 3 in the position corresponding to that word. If a word in the vocabulary doesn't appear in the document, then it gets a count of 0.

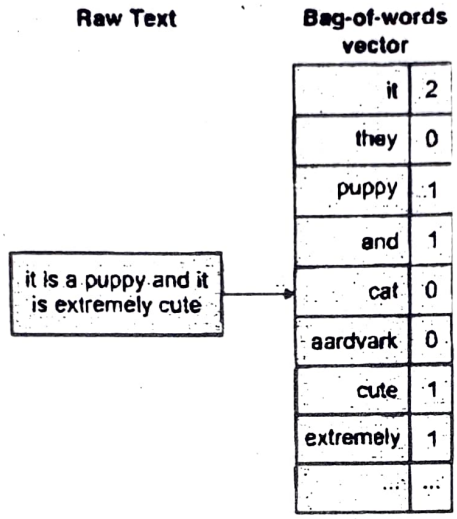


Fig. 5.7

- Bag-of-words converts a text document into a flat vector. It is "flat" because it does not contain any of the original textual structures. The original text is a sequence of words. But a bag-of-words has no sequence. It just remembers how many times each word appears in the text.
- The ordering of words in the vector is not important, as long as it is consistent for all documents in the dataset. Neither does bag-of-words represent any concept of word hierarchy. For example, the concept of "animal" includes "dog," "cat," "raven," etc. But in a bag-of-words representation, these words are all equal elements of the vector.

Q. 18 Write a short note on Bag-of-n-Grams. **(4 Marks)**

Ans. : Bag-of-n-Grams

- Bag-of-n-Grams is a natural extension of bag-of-words. An n-gram is a sequence of n tokens. A word is essentially a 1-gram, also known as a unigram. After tokenisation, the counting mechanism can group individual tokens into word counts or count overlapping sequences as n-grams. For example, the sentence "Rahul knocked on the door" generates the n-grams "Rahul knocked," "knocked on," "on the," and "the door" for n = 2.

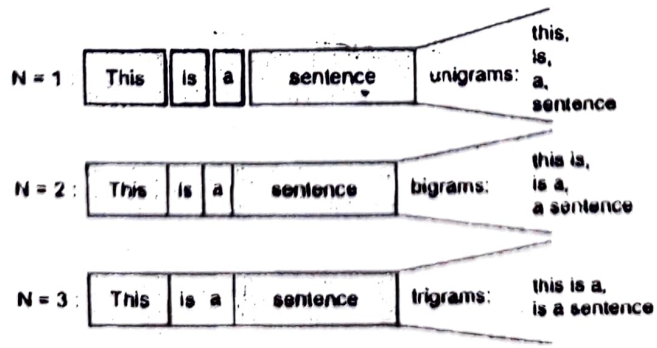


Fig. 5.8

- n-grams retain more of the original sequence structure of the text, and therefore the bag-of-n-grams representation can be more informative. However, this comes at a cost. Theoretically, with k unique words, there could be k^2 unique 2-grams (also called bigrams). In practice, there are not nearly so many, because not every word can follow every other word. Nevertheless, there are usually a lot more distinct n-grams (n > 1) than words.
- This means that bag-of-n-grams is a much bigger and sparser feature space. It also means that n-grams are more expensive to compute, store, and model. The larger n is, the richer the information, and the greater the cost.

Q. 19 Explain Term Frequency and Inverse Document Frequency. **(4 Marks)**

Ans. :

1. Term Frequency (TF)

Term Frequency (TF) measures the frequency of a word in a given document. It highly depends on the length of the document and the generality of word. For example, a very common word such as "the" can appear multiple times in a document. But if you take two documents, the one which has 100 words and the other which has 10,000 words, then there is a high probability that the count of "the" would be much higher in the 10,000 worded document. But it would be incorrect to assume that the longer document is more important than the shorter document just on the basis of term frequency. Hence, you normalise the term frequency value. You divide the frequency with the total number of words in the document to get the normalised term frequency value.

You can calculate term frequency as following.

$$\text{Term Frequency (TF)} = \frac{\text{count of the term in the document (t)}}{\text{total number of terms in the document (d)}}$$

One thing to keep in mind is that you need to finally vectorise the document. When you are planning to vectorise the documents, you cannot just consider the words that are present in that particular document. If you do that, then the vector length will be different between the documents, and it will not be feasible to compute the similarity. So, you vectorise the documents on the vocab. Vocab is the entire list of all possible words in the corpus.

When you are vectorising the documents, you check for each words count. In worst case if the term doesn't exist in the document, then that particular TF value will be 0 and in other extreme case, if all the words in the document are same, then it will be 1. The final value of the normalised TF value will be in the range of [0 to 1].

2. Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) is the measure of the importance of a word. Term frequency (TF) does not consider the importance of words. Some words such as "of", "and", etc. can be most frequently present but are of little significance. IDF provides weightage to each word based on its frequency in the corpus D.

IDF can be calculated as following.

$$\text{IDF} = \log\left(\frac{\text{Total number of documents N in corpus D}}{\text{number of documents containing the term t}}\right)$$

When you calculate IDF, it will be very low for the most occurring words such as stopwords ("the", "is", "was", etc.). IDF is constant per corpus whereas TF is document specific.

Q. 20 Write a short note on Social Network Analysis (SNA). **(4 Marks)**

Ans. :

Social Network Analysis (SNA)

The dictionary meaning of graph is "the collection of all points whose coordinates satisfy a given relation". Social networks use graph theory to model pairwise relations between objects.

Definition: *Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory.*

Each graph has nodes (or vertex) and edges.

Definition: *Vertex or node represents an object in the graph.*

Definition: *The connection between the nodes is called an edge or a link.*

The nodes in the graph represent the users or objects and the edges represent the relationship between the nodes.

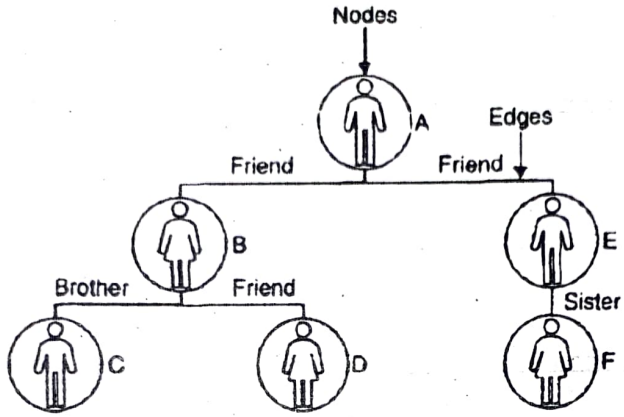


Fig. 5.9

Some of the basic properties of a graph that are used for Social Network Analysis.

1. **Degree of a Node :** Degree of a node is the number of edges it has. For example, degree of node A is 2 whereas degree of node F is 1.
2. **Path Length :** Path length is the distance between two nodes. For example, the path length between node A and node D is 2 (node A→ node B → node D) whereas the path length between node D and node F is 4 (node D → node B → node A → node E → node F).
3. **Centrality :** Centrality refers to the "importance" or "influence" of a particular node within a network. It is often a key node that joins several networks. For example, node A can be treated to have high centrality because it joins the two groups formed by node B and node E.
4. **Density :** Density refers to the proportion of actual direct connections versus total number of direct connections possible in the network. For example, node A is only directly connected to node B and node E. Had it been also directly connected to other nodes (node C, node D and node F), then the network would be considered denser.
5. **Closeness :** Closeness of a node is the average length of the shortest path between the node and all other nodes in the graph.
6. **Betweenness :** Betweenness refers to the number of times a node acts as a bridge along the shortest path between two other nodes. For example, node A is a bridge between node B and node E and also node C and node F.

Q. 21 What are some of the major applications of social network analysis? **(4 Marks)**

Ans. : Need (Applications) of Social Network Analysis

Some of the major applications of SNA are as following.

1. **Identifying the Influencers :** In social networks, influencers are people who have the ability to influence potential buyers of a product or service by promoting or recommending the items on social media. Influencer marketing (also known as influence marketing) is a form of social media marketing involving endorsements and product placement from influencers, people and organisations who have a purported expert level of knowledge or social influence in their field.

Influencers are someone with the power to affect the buying habits or quantifiable actions of others by uploading some form of original-often sponsored-content to social media platforms like Instagram, YouTube, Snapchat or other online channels. Influencer marketing is when a brand enrols influencers who have an established credibility and audience on social media platforms to discuss or mention the brand in a social media post. Influencer content may be framed as testimonial advertising.

Social network analysis helps in identifying such influencers based on several criteria such as geography, demographics, topics, etc.

2. **Human Resource Management (HRM) :** HRM often strives to identify critical resources and understand their contribution to the organisation flow, collaboration, participation, and information flow. Using SNA, an organisation can optimise the talent connections, productivity, and utilisation. It also helps to identify the reach of an individual, identify accelerators of growth and poorly connected resources, and decide whom to give more opportunity.
3. **Contact tracing :** SNA could also be used for contract tracing for infectious diseases (such as Covid-19). SNA could help to identify and isolate individuals and groups with high betweenness and out-degree centrality (transmitters of disease) and implement sound contact tracing activities to reduce the impact and spread.
4. **Identify themes and connections :** SNA can also identify dominant themes and relations between keywords and identify the sentiments. For example, the Figs. 5.10, from Journal of Medical Internet Search, shows the connection between the top 10 words for COVID-19 themes.

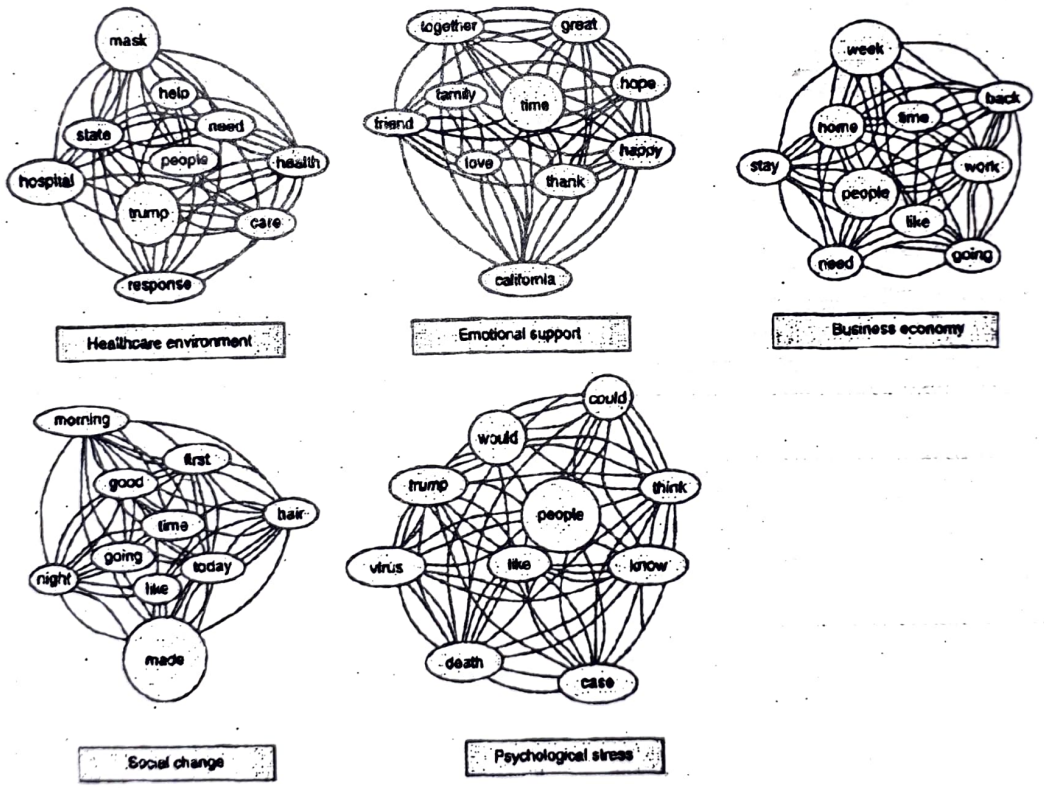


Fig. 5.10

Q. 22 Write a short note on business analysis. **(4 Marks)**

Ans. : The International Institute of Business Analysis (IIBA) defines business analysis as following.

***Definition:** Business Analysis is the practice of enabling change in an organisational context, by defining needs and recommending solutions that deliver value to stakeholders.*

- The Business Analyst is an agent of change. Business Analysis is a disciplined approach for introducing and managing change to organisations, whether they are for-profit businesses, governments, or non-profits.

- Job titles for business analysis practitioners include not only business analyst, but also business systems analyst, systems analyst, requirements engineer, process analyst, product manager, product owner, enterprise analyst, business architect, management consultant, business intelligence analyst, data scientist, and more. Many other jobs, such as management, project management, product management, software development, quality assurance and interaction design rely heavily on business analysis skills for success.
- Business analysis is used to identify and articulate the need for change in how organisations work, and to facilitate that change. As a business analyst, you identify and define the solutions that will maximise the value delivered by an organisation to its stakeholders. Business analysts work across all levels of an organisation and may be involved in everything from defining strategy, to creating the enterprise architecture, to taking a leadership role by defining the goals and requirements for programs and projects or supporting continuous improvement in its technology and processes.
- Business analysts have the specialised knowledge to act as a guide and lead the business through unknown or unmapped territory, to get it to its desired destination. The value of business analysis is in realisation of benefits, avoidance of cost, identification of new opportunities, understanding of required capabilities and modelling the organisation. Through the effective use of business analysis, you can ensure an organisation realises these benefits, ultimately improving the way they do business. Business analysis helps businesses do business better.

Q. 23 Explain Holdout Method of cross-validation. **(4 Marks)**

Ans. : Holdout Method

Holdout method is a non-exhaustive approach for cross-validation. In this approach, you divide the entire dataset into two parts - training dataset and testing dataset.

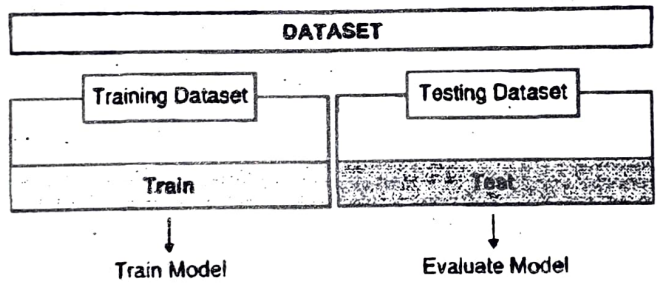


Fig. 5.11

- You train the model on the training dataset and then evaluate the model on the testing dataset. Usually, the size of training data is set more than twice that of testing data. A ratio of 70:30 or 80:20 is quite common.
- In this approach, the data is first shuffled randomly before splitting. Hence, it should be used with caution because you may achieve highly misleading results. This method is often termed as "the simplest kind of cross-validation" but in reality it is just a validation step like you perform for any machine learning model and not necessarily a great cross-validation technique.

Q. 24 Write a short note on Grid Search hyperparameter tuning technique. **(4 Marks)**

Ans. : Some of the common techniques and algorithms to tune hyperparameters are as shown in Fig. 5.12(a).

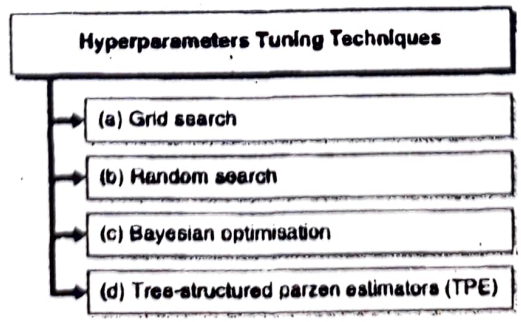


Fig. 5.12(a)

1. Grid Search

In the grid search method, you create a grid of all the possible values for hyperparameters. Each iteration tries a combination of hyperparameters in a specific order. It fits the model on each and every combination of hyperparameter possible and records the model performance. Finally, it returns the best model with the best hyperparameters.

For example, if the hyperparameter is the number of leaves in a decision tree, then the grid could be 10, 20, 30, ..., 100. Some educated guesswork is necessary to specify the minimum and maximum values for hyperparameters.

Grid search is very simple to set up and trivial to parallelise for fast computing. It is the most expensive method in terms of total computation time. However, if run in parallel, it is fast in terms of wall clock time.

2. Random Search

In the random search method, you create a grid of possible values for hyperparameters. Each iteration tries a random combination of hyperparameters from this grid, records the performance, and lastly returns the combination of hyperparameters which provided the best performance.

Random search is a slight variation on grid search. Instead of searching over the entire grid, random search only evaluates a random sample of points on the grid. This makes random search a lot cheaper than grid search and still finding good hyperparameter values. The Fig. 5.12(b) visualises the differences between grid search and random search.

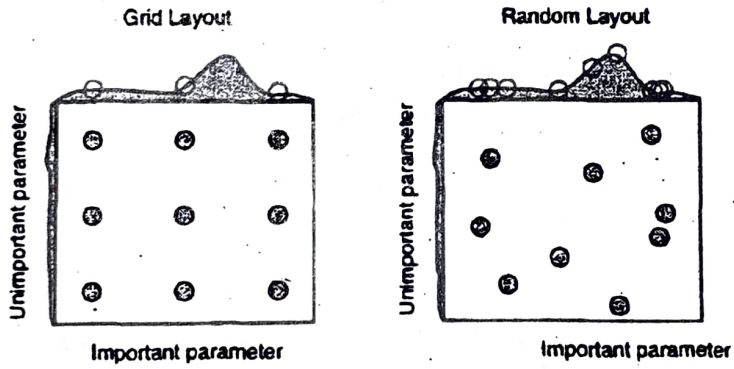


Fig. 5.12(b)

3. Bayesian Optimisation

Tuning and finding the right hyperparameters for your model is also an optimisation problem. You want to minimise the loss function of your model by changing model parameters. Bayesian optimisation helps to find the minimal point in the minimum number of steps. Bayesian optimisation also uses an acquisition function that directs sampling to areas where an improvement over the current best observation is likely.

4. Tree-structured Parzen estimators (TPE)

The idea of Tree-based Parzen optimisation is similar to Bayesian optimisation.

Instead of finding the values of p (y | x) y where is the function to be minimised (e.g., validation loss) and x is the value of hyperparameter, the TPE models P (x | y) and P (y). One of the drawbacks of tree-structured Parzen estimators is that they do not model interactions between the hyperparameters. However, TPE works extremely well in practice and is commonly used.

□□□

Unit VI : Data Visualization and Hadoop

Chapter 6 : Data Visualization and Hadoop

Q. 1 What is data visualisation ? **SPPU - Dec. 18, 9 Marks, May 19, 8 Marks**

OR What is mean by Data conditioning and data visualisation ? **SPPU - Dec. 19, 5 Marks**

Ans. :

- It is one thing to process the massive amount of data and another to make it human friendly to be able to comprehend it, even from a distance, without getting into nitty-gritty of the complex calculations.

Table 6.1

Sr. No.	Height	Weight
1.	185	72
2.	170	56
3.	168	60
4.	179	68
5.	182	72
6.	188	77
7.	180	71
8.	180	70
9.	183	84
10.	180	88
11.	180	67
12.	177	76

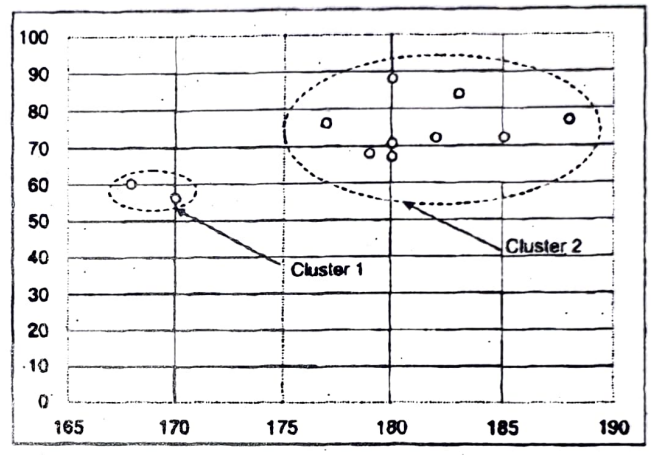


Fig. 6.1

***Definition** : Data visualisation is a graphical or pictorial representation of data that makes it easy to communicate the information to humans.*

- Data visualisation uses various forms of representations to match the data and the relationship amongst its data attributes so as to communicate the desired information effectively.

Q. 2 Describe the goals of data visualisation. **(6 Marks)**

Ans. : Goals (Objectives) of Data Visualisation

Fig. 6.2(a) shows some of the major goals or objectives of data visualisation.

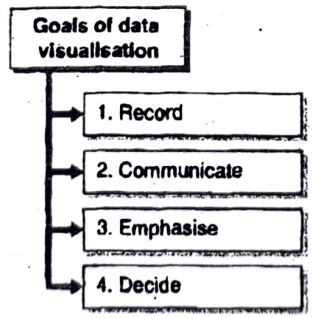


Fig. 6.2(a) : Goals (Objectives) of Data Visualisation

1. **Record** : Data visualisation helps you to record the information that might be lying in various forms such as tables, logs, emails, conversations, audio, video, or any other form of information sharing. A user may not have to dig through several of these information sources to get to the data she desires. For example, a music player could provide a visual representation of the musical notes being played on your phone and you could according tune various parameters such as bass and treble as you like it.

Unit VI : Data Visualization and Hadoop

Chapter 6 : Data Visualization and Hadoop

Q. 1 What is data visualisation ? SPPU - Dec. 18, 9 Marks, May 19, 8 Marks

OR What is mean by Data conditioning and data visualisation ? SPPU - Dec. 19, 5 Marks

Ans. :

- It is one thing to process the massive amount of data and another to make it human friendly to be able to comprehend it, even from a distance, without getting into nitty-gritty of the complex calculations.

Table 6.1

Sr. No.	Height	Weight
1.	185	72
2.	170	56
3.	168	60
4.	179	68
5.	182	72
6.	188	77
7.	180	71
8.	180	70
9.	183	84
10.	180	88
11.	180	67
12.	177	76

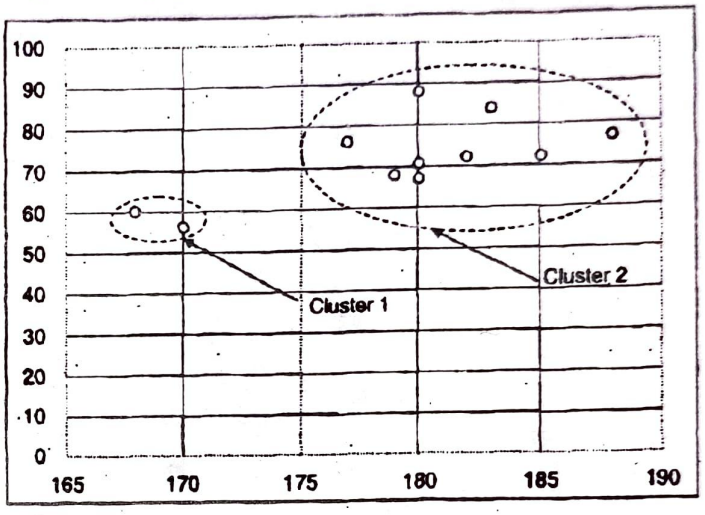


Fig. 6.1

***Definition** : Data visualisation is a graphical or pictorial representation of data that makes it easy to communicate the information to humans.*

- Data visualisation uses various forms of representations to match the data and the relationship amongst its data attributes so as to communicate the desired information effectively.

Q. 2 Describe the goals of data visualisation. **(6 Marks)**

Ans. : Goals (Objectives) of Data Visualisation

Fig. 6.2(a) shows some of the major goals or objectives of data visualisation.

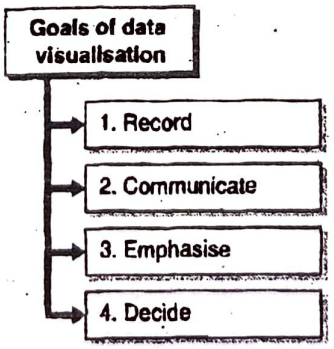


Fig. 6.2(a) : Goals (Objectives) of Data Visualisation

1. **Record :** Data visualisation helps you to record the information that might be lying in various forms such as tables, logs, emails, conversations, audio, video, or any other form of information sharing. A user may not have to dig through several of these information sources to get to the data she desires. For example, a music player could provide a visual representation of the musical notes being played on your phone and you could according tune various parameters such as bass and treble as you like it.

1. **Grid Search**

In the grid search method, you create a grid of all the possible values for hyperparameters. Each iteration tries a combination of hyperparameters in a specific order. It fits the model on each and every combination of hyperparameter possible and records the model performance. Finally, it returns the best model with the best hyperparameters.

For example, if the hyperparameter is the number of leaves in a decision tree, then the grid could be 10, 20, 30, ..., 100. Some educated guesswork is necessary to specify the minimum and maximum values for hyperparameters.

Grid search is very simple to set up and trivial to parallelise for fast computing. It is the most expensive method in terms of total computation time. However, if run in parallel, it is fast in terms of wall clock time.

2. **Random Search**

In the random search method, you create a grid of possible values for hyperparameters. Each iteration tries a random combination of hyperparameters from this grid, records the performance, and lastly returns the combination of hyperparameters which provided the best performance.

Random search is a slight variation on grid search. Instead of searching over the entire grid, random search only evaluates a random sample of points on the grid. This makes random search a lot cheaper than grid search and still finding good hyperparameter values. The Fig. 5.12(b) visualises the differences between grid search and random search.

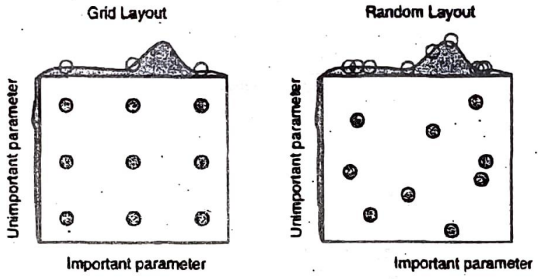


Fig. 5.12(b)

3. **Bayesian Optimisation**

Tuning and finding the right hyperparameters for your model is also an optimisation problem. You want to minimise the loss function of your model by changing model parameters. Bayesian optimisation helps to find the minimal point in the minimum number of steps. Bayesian optimisation also uses an acquisition function that directs sampling to areas where an improvement over the current best observation is likely.

4. **Tree-structured Parzen estimators (TPE)**

The idea of Tree-based Parzen optimisation is similar to Bayesian optimisation.

Instead of finding the values of p (y | x) y where is the function to be minimised (e.g., validation loss) and x is the value of hyperparameter, the TPE models P (x | y) and P (y). One of the drawbacks of tree-structured Parzen estimators is that they do not model interactions between the hyperparameters. However, TPE works extremely well in practice and is commonly used.

□□□

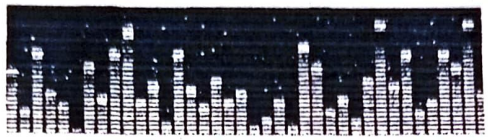


Fig. 6.2(b) : Examples of Record

2. **Communicate :** A primary goal of data visualisation is to communicate the information in the most effective way for the given data. There are several types of charts, maps and graphs that could be effectively used to visualise different forms of data as well as relationship amongst its data attributes as suitable. For example, Microsoft® Excel provides various types of charts for plotting your data.

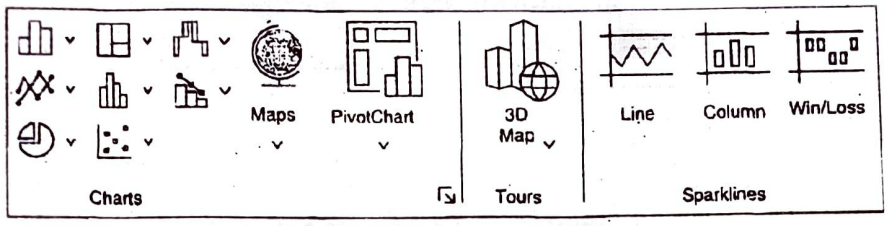


Fig. 6.2(c) : Various types of charts provided by Microsoft Excel

3. **Emphasise :** Using data visualisation, you can emphasise or highlight a portion of data, find patterns, show trends, or depict relationships between various data attributes. For example, the Fig. 6.2(d) gives a quick view of rainfall in various states of India in a particular year.

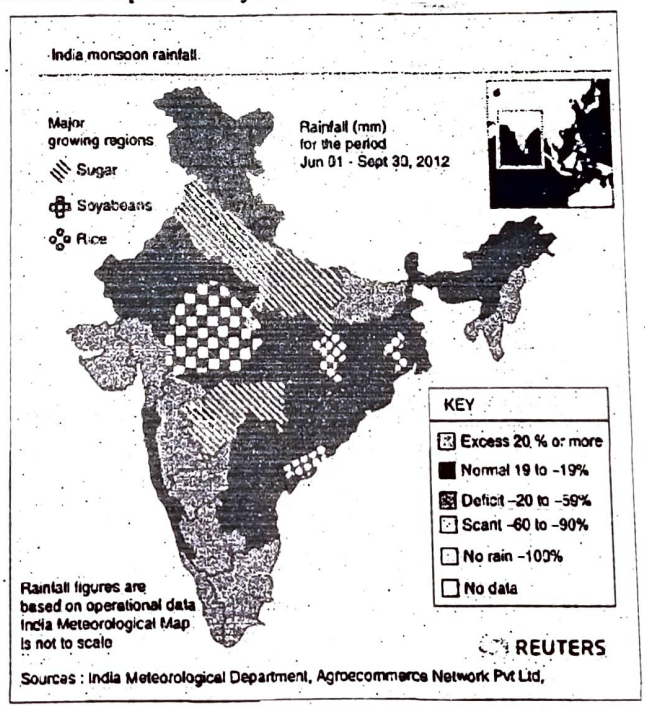


Fig. 6.2(d) : Example of a chart emphasising the data points

As shown in Fig. 6.2(d) if the country has received adequate rainfall or shortage of rainfall. Which states were most affected, and which were least affected. This representation is much easier to understand and explain than to go through 25+ entries for rainfall measurement for each state collected for each day of the rainy season.

4. **Decide :** Data visualisation makes it easier to quickly make decisions and take actions. You do not have to go through length reports and complex data to understand what needs to be done next. For example, from surveying 1,000 customers about the customer service, you might have detailed responses.

But, representing the cause visually makes you to quickly decide the action plan that training is required for the customer service representative to improve customer satisfaction.

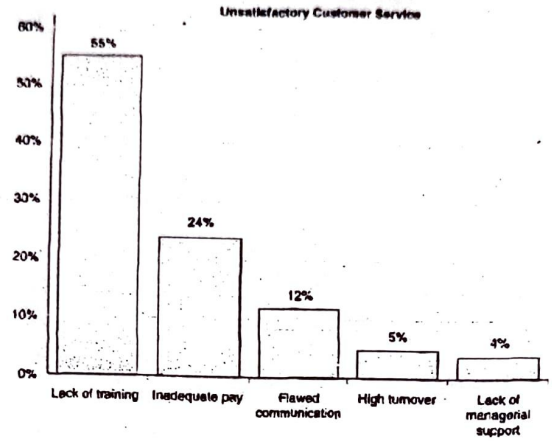


Fig. 6.2(e) : Example of a chart helping in decision making

Q. 3 What are the challenges in Big data visualization ? SPPU - Dec. 18, 8 Marks

OR Why it is difficult to visualize Big Data ? SPPU - May 19, 9 Marks , Dec. 19, 8 Marks

Ans. : Some of the major challenges or difficulties with visualising Big Data are as shown in Fig. 6.3.

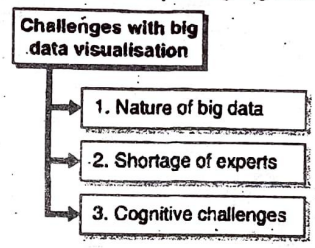


Fig. 6.3 : Challenges with Big Data Visualisation

1. **Nature of Big Data :** The Five Vs of Big Data – Volume, Variety, Velocity, Veracity and Value. The nature of Big Data itself is one of the biggest challenges of visualising it. The massive volume of data requires special software and hardware for handing the visualisation. The heterogeneity (variety) of data attributes makes it further hard to conceptualise the right forms of visualisation to show relationship between the data attributes. The velocity of data is so fast that you require to update your visualisation very frequently to keep it accurate. You may want to interact with it in the real time as well which makes it further complex to visualise it. The veracity and value characteristics require that your visualisation meets the data quality and usefulness as well. If the visualisation is not useful or was formed with poor quality data, it may not fulfill the desired objectives of the visualisation.
2. **Shortage of Experts :** Data analytics is an emerging field and there are not many experts around the world. You require experts who can
   (a) Understand the wide variety of data
   (b) Model the data correctly so as to meet the desired objectives
   (c) Build and manage software and hardware tools and techniques required for Big Data processing
   (d) Design appropriate visual interfaces and
   (e) Also communicate the findings effectively

   Building a team of experts that have all the required capabilities is challenging.

3. **Cognitive Challenges :** Finally, irrespective of what you have got, visualisation is something that a human needs to understand and make sense about. Overloaded charts with full of various colours and gauges make it difficult to get the real sense of data. Also, plots like regression line, ROC curve etc. are difficult to understand and make sense about if you do not know how to read and interpret them.

Q. 4 Explain any four data visualization techniques. **SPPU - Dec 18, 9 Mark, May 19, 8 Marks**

OR Write a short note on Bubble chart. **(4 Marks)**

OR With a suitable example, draw a Histogram and explain its usage. **(8 Marks)**

OR With a suitable example, draw a Word Cloud and explain its usage. **(8 Marks)**

Ans. : At a high-level, various data visualisations could be grouped as shown in Fig. 6.4(a).

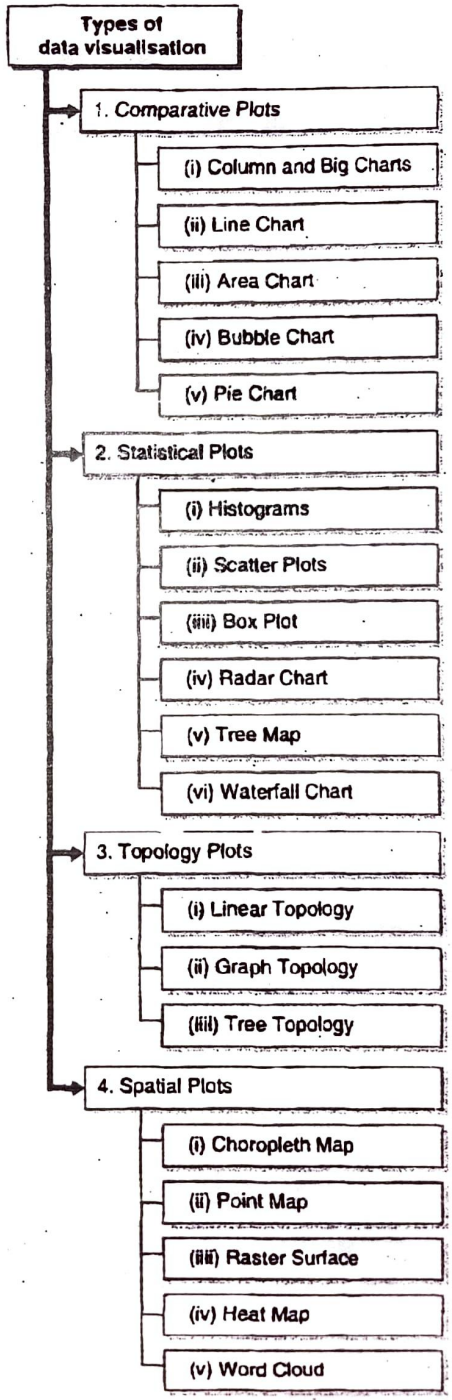


Fig. 6.4(a) : Types of Data Visualisation

(A) Comparative Plots

Comparative plots are used for comparing the datapoints. Some of the commonly used comparative plots.

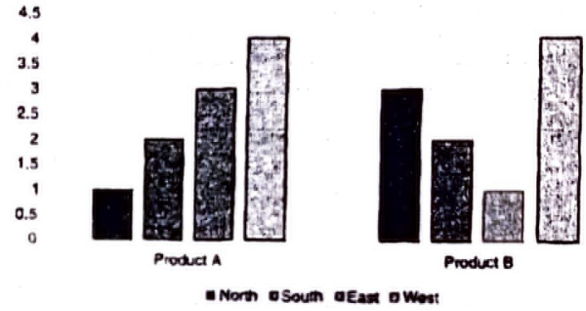
1. Column and Bar Charts

- Column and Bar charts are perhaps the most common, most simple, and most popular chart that you might have ever seen. It is used for comparing two or more values in the same category.
- There could be several variations of the chart such as column chart, bar chart, stack chart and their 2D and 3D plots. Some of the column and bar charts are as following for the following data shown in Table 6.2(a).

Table 6.2(a)

Sales Region	Product A	Product B
North	1	3
South	2	2
East	3	1
West	4	4

(b) Example of a column chart



(c) Example of a column chart