ATME College of Engineering

Machine Learning – BCS602

ASHWINI P.,
Assistant Professor
Dept. of Computer Science & Engineering
ATMECE, Mysuru

Course objectives

This course will enable students to,

To introduce the fundamental concepts and techniques of machine learning.

To understanding of various types of machine learning and the challenges faced in realworld applications.

To familiarize the machine learning algorithms such as regression, decision trees, Bayesian models, clustering, and neural networks.

To explore advanced concept like reinforcement learning and provide practical insight into its applications.

To enable students to model and evaluate machine learning solutions for different types of problems.

Overview of syllabus

Module-1
Introduction: Need for Machine Learning, Machine Learning Explained, Machine Learning in Relation to other Fields, Types of Machine Learning, Challenges of Machine Learning, Machine Learning Process, Machine Learning Applications. **Understanding Data – 1**: Introduction, Big Data Analysis Framework, Descriptive Statistics, Univariate Data Analysis and Visualization. Chapter-1, 2 (2.1-2.5)

Course outcomes

At the end of the course, the student will be able to:

- **Describe** the machine learning techniques, their types and data analysis framework.
- **Apply** mathematical concepts for feature engineering and perform dimensionality reduction to enhance model performance.
- **Develop** similarity-based learning models and regression models for solving classification and prediction tasks.
- **Build** probabilistic learning models and design neural network models using perceptrons and multilayer architectures
- **Utilize** clustering algorithms to identify patterns in data and implement reinforcement learning techniques.

MODULE-1

Introduction

NEED FOR MACHINE LEARNING

- BUSINESS ORGANIZATIONS HAVE NUMEROUS DATA
- NEED TO ANALYZE DATA FOR TAKING DECISIONS

Machine learning has become so popular because of three reasons:

1. **High volume of available data to manage:** Big companies such as Facebook, Twitter, and YouTube generate huge amount of data that grows at a phenomenal rate. It is estimated that the data approximately gets doubled every year.
2. **Second reason is that the cost of storage has reduced.** The hardware cost has also dropped. Therefore, it is easier now to capture, process, store, distribute, and transmit the digital information.
3. **Third reason for popularity of machine learning is the availability of complex algorithms now.** Especially with the advent of deep learning, many algorithms are available for machine learning.

NEED FOR MACHINE LEARNING

Before starting the machine learning journey, let us establish these terms - data, information, knowledge, intelligence, and wisdom. A knowledge pyramid is shown in Figure 1.1.

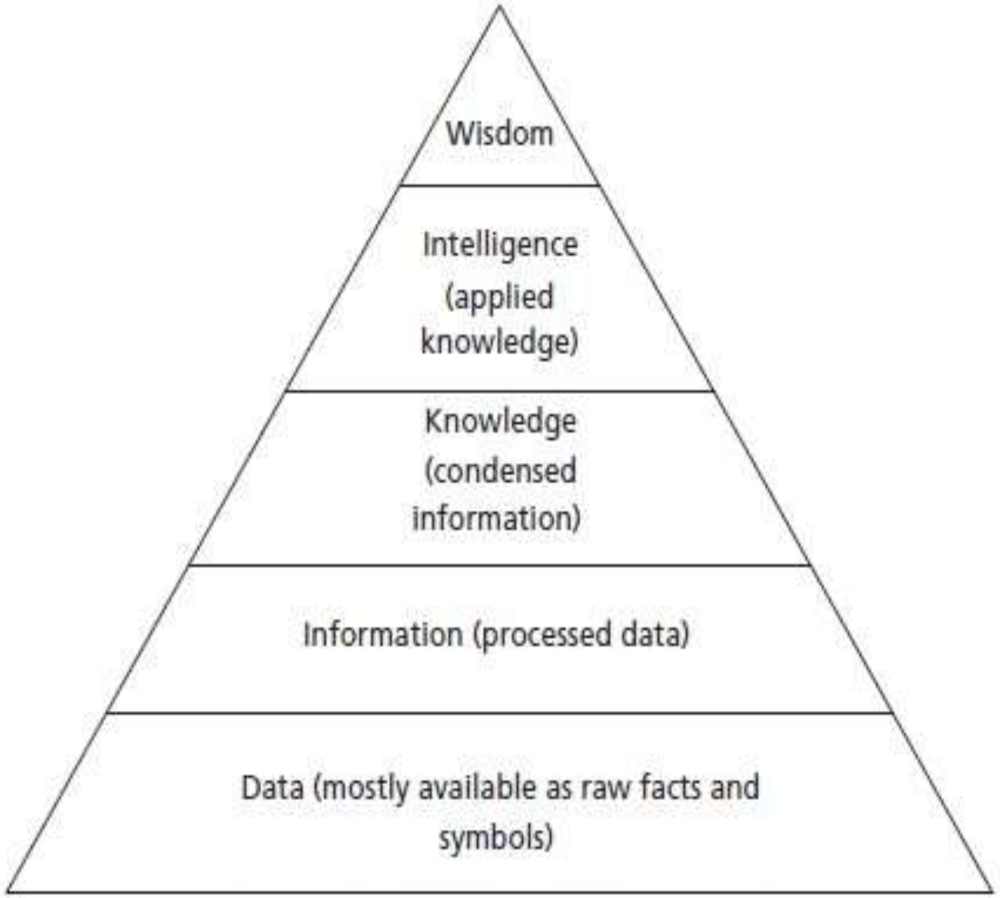

Wisdom
Intelligence (applied knowledge)
Knowledge (condensed information)
Information (processed data)
Data (mostly available as raw facts and symbols)

Figure 1.1: The Knowledge Pyramid

NEED FOR MACHINE LEARNING

What is data?

All facts are data. Data can be numbers or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data with data sources such as flat files, databases, or data warehouses in different storage formats.

Processed data is called information.

This includes patterns, associations, or relationships among data.

For example, sales data can be analyzed to extract information like which is the fast selling product.

NEED FOR MACHINE LEARNING

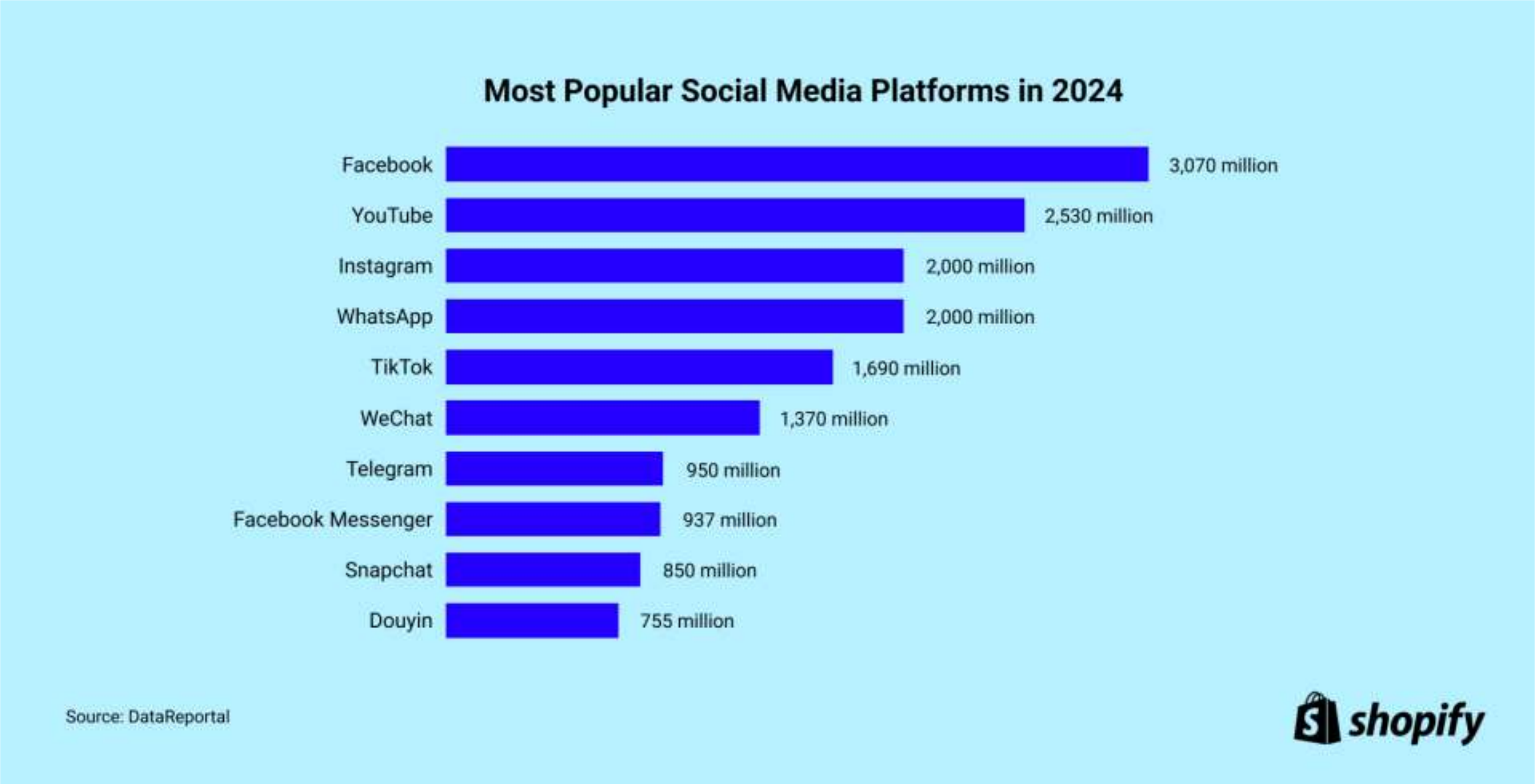
Condensed information is called knowledge.

For example, the historical patterns and future trends obtained in the above sales data can be called knowledge. Unless knowledge is extracted, data is of no use. Similarly, knowledge is not useful unless it is put into action.

Intelligence is the applied knowledge for actions.

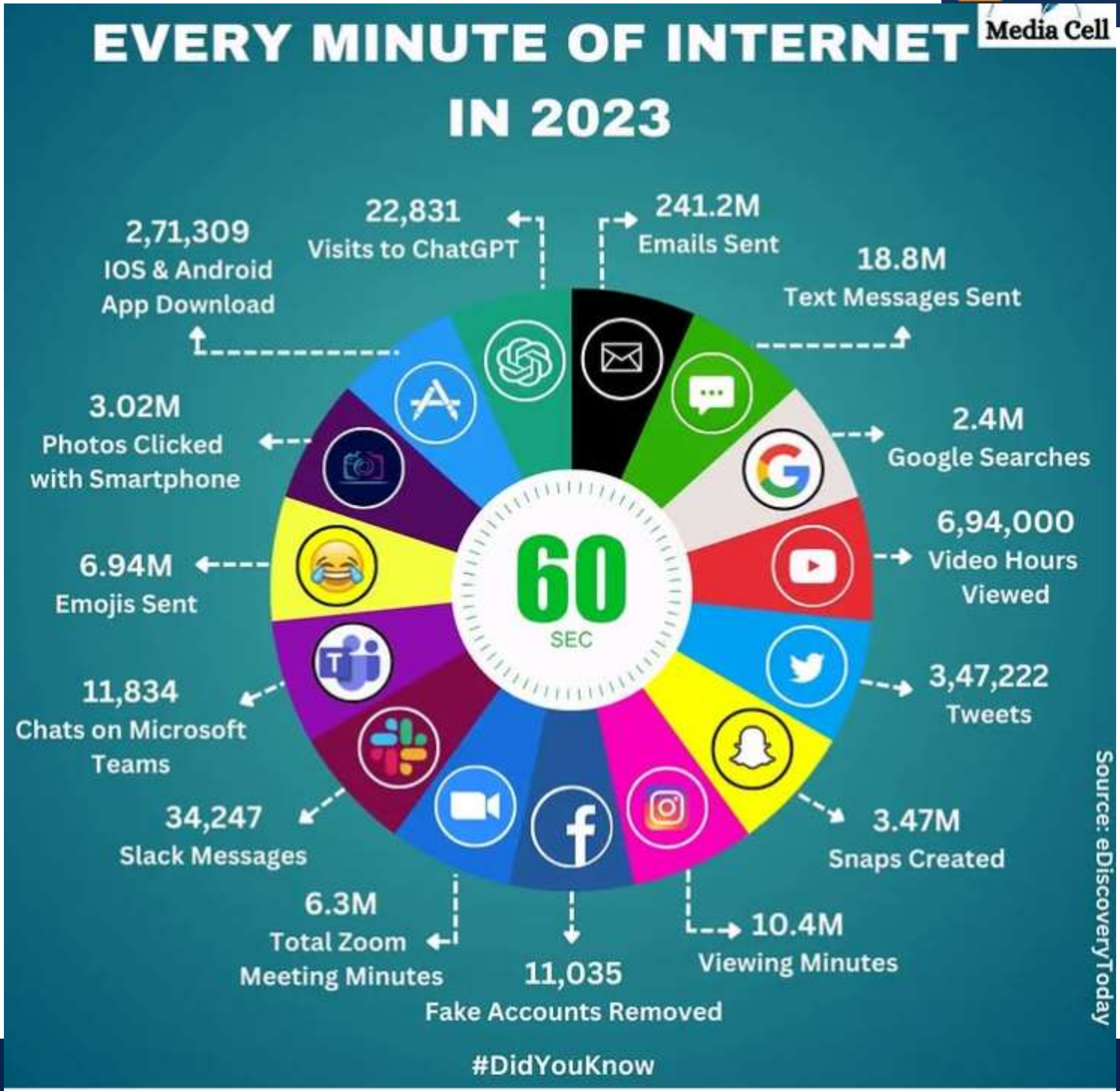
An actionable form of knowledge is called intelligence. Computer systems have been successful till this stage.

The ultimate objective of knowledge pyramid is **wisdom** that represents the maturity of mind that is, so far, exhibited only by humans.



Most Popular Social Media Platforms in 2024

Source: DataReportal



EVERY MINUTE OF INTERNET IN 2023

60 SEC

22,831
Visits to ChatGPT

241.2M
Emails Sent

18.8M
Text Messages Sent

2.4M
Google Searches

6,94,000
Video Hours Viewed

3,47,222
Tweets

3.47M
Snaps Created

10.4M
Viewing Minutes

11,035
Fake Accounts Removed

6.3M
Total Zoom Meeting Minutes

34,247
Slack Messages

11,834
Chats on Microsoft Teams

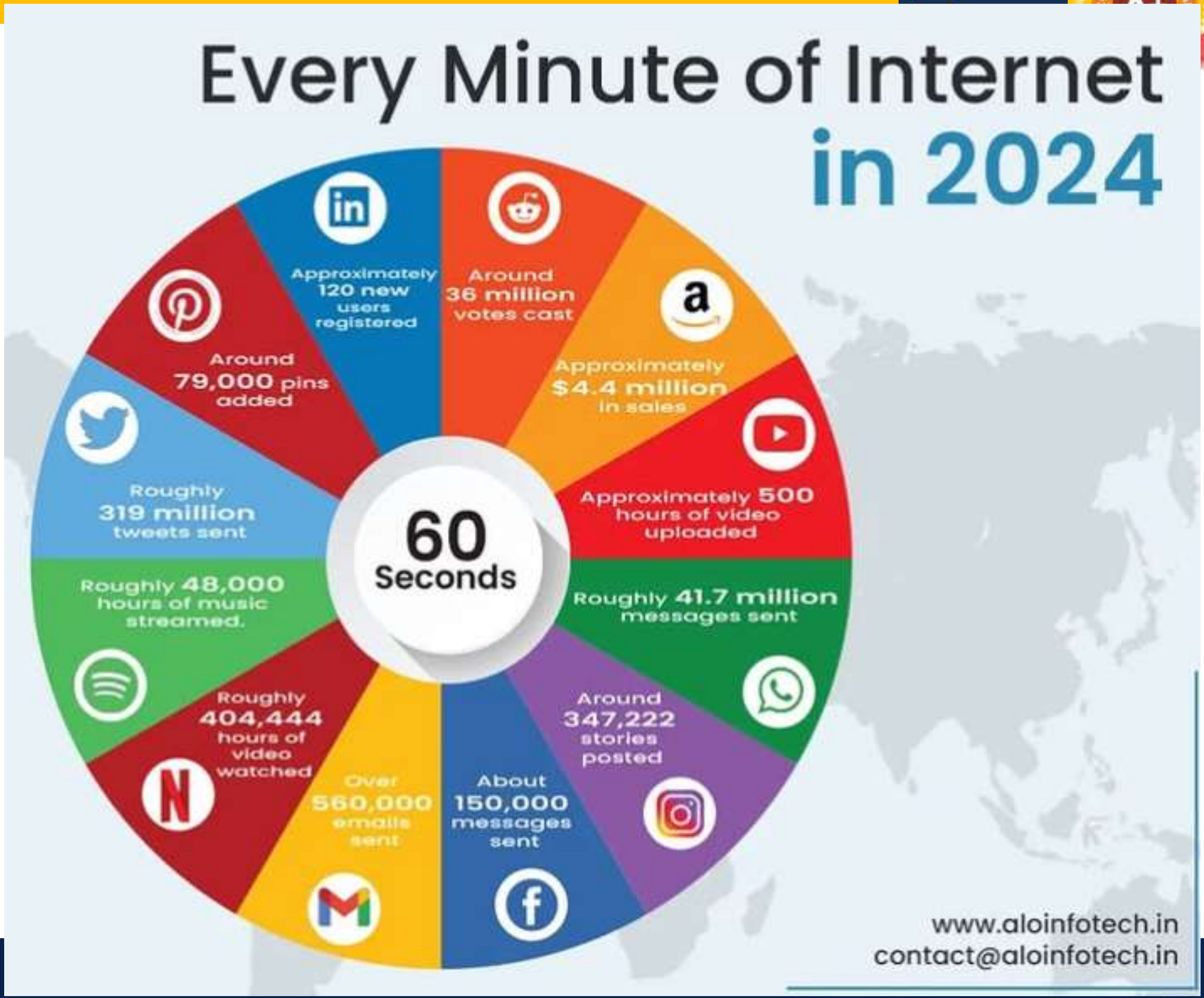
6.94M
Emojis Sent

3.02M
Photos Clicked with Smartphone

2,71,309
IOS & Android App Download

Source: eDiscoveryToday

#DidYouKnow



Every Minute of Internet in 2024

60 Seconds

Approximately 120 new users registered

Around 36 million votes cast

Approximately $4.4 million in sales

Approximately 500 hours of video uploaded

Roughly 41.7 million messages sent

Around 347,222 stories posted

About 150,000 messages sent

Over 560,000 emails sent

Roughly 404,444 hours of video watched

Roughly 48,000 hours of music streamed.

Roughly 319 million tweets sent

Around 79,000 pins added

www.aloinfotech.in
contact@aloinfotech.in

NEED FOR MACHINE LEARNING

The objective of machine learning is to process these archival data for organizations to take better decisions to design new products, improve the business processes, and to develop effective decision support systems.

MACHINE LEARNING EXPLAINED

- Machine learning is an important **sub-branch** of Artificial Intelligence (AI).
- A frequently quoted definition of machine learning was by **Arthur Samuel**, one of the pioneers of Artificial Intelligence.
- He stated that “**Machine learning is the field of study that gives the computers ability to learn without being explicitly programmed**.”
- The key to this definition is that the systems should learn by

MACHINE LEARNING EXPLAINED

- ➢ In **conventional programming**, after understanding the problem, a detailed design of the program such as a flowchart or an algorithm needs to be created and converted into programs using a suitable programming language.
- ➢ This approach could be difficult for many real-world problems such as puzzles, games, and complex image recognition applications.
- ➢ Initially, artificial intelligence aims to understand these problems and develop general purpose rules manually.
- ➢ Then, these **rules are formulated into logic** and implemented in a program to create intelligent systems.

MACHINE LEARNING EXPLAINED

- This idea of developing intelligent systems by using **logic and reasoning** by converting an expert’s knowledge into a set of rules and programs is called an expert system.
- The focus of AI is to develop intelligent systems by using **data-driven approach,** where data is used as an input to develop intelligent models.
- The models can then be used to predict new inputs.
- Thus, the aim of **machine learning** is to **learn** a model or set of **rules from the given dataset automatically** so that it can predict the unknown data correctly.

MACHINE LEARNING EXPLAINED

- As **humans** take **decisions** based on an **experience**, computers **make models based on extracted patterns** in the input data and then use these data-filled models for prediction and to take decisions.
- For computers, the learnt model is equivalent to human experience. This is shown in Figure 1.2.

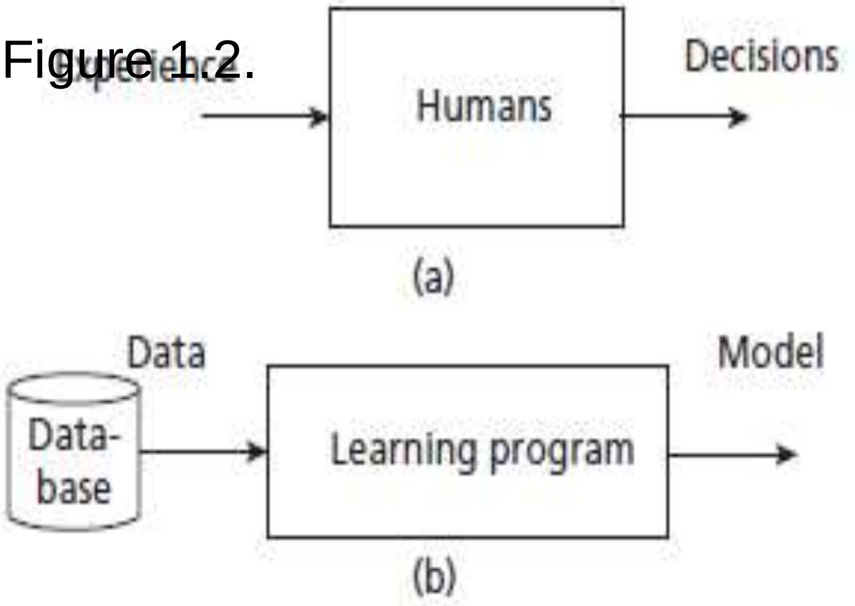


Figure 1.2: (a) A Learning System for Humans (b) A Learning System for Machine Learning

MACHINE LEARNING EXPLAINED

- Often, the quality of data determines the quality of experience and, therefore, the quality of the learning system.
- In **statistical learning**, the relationship between the input x and output y is modeled as a function in the **form y = f(x).**
- Here, **f is the learning function that maps the input x to output y.**
- Learning of function f is the crucial aspect of forming a model in statistical learning.
- In **machine learning**, this is simply called **mapping of input to output.**

MACHINE LEARNING EXPLAINED

- The learning program summarizes the raw data in a model.
- Formally stated, a model is an explicit description of patterns within the data in the form of:

1. Mathematical equation
2. Relational diagrams like trees/graphs
3. Logical if/else rules, or
4. Groupings called clusters

MACHINE LEARNING EXPLAINED

Another pioneer of AI, Tom Mitchell's definition of machine learning states that, "**A computer program is said to learn from experience E, with respect to task T and some performance measure P, if its performance on T measured by P improves with experience E.**"

The important components of this definition are

experience E,

task T, and

performance measure P.

MACHINE LEARNING EXPLAINED

- For example, the **task T** could be **detecting an object in an image**.
- The machine can gain the knowledge of object using **training dataset of thousands of images**. This is called **experience E**.
- So, the focus is to use this experience E for this task of object detection T.
- The ability of the system to detect the object is measured by **performance measures like precision and recall**.
- Based on the performance measures, course correction can be done to improve the performance of the system.
- Models of computer systems are equivalent to human experience.
- Experience is based on data.
- Humans gain experience by various means.

MACHINE LEARNING EXPLAINED

- Once the knowledge is gained, when a new problem is encountered, humans search for similar past situations and then formulate the heuristics and use that for prediction.
- In systems, **experience is gathered** by these steps:

  1. **Collection of data**
  2. Once data is gathered, **abstract concepts are formed** out of that data. Abstraction is used to generate concepts. This is equivalent to humans' idea of objects, for example, we have some idea about how an elephant looks like.

MACHINE LEARNING EXPLAINED

3. **Generalization converts the abstraction into an actionable form of intelligence**.

- ordering of all possible concepts.
- involves ranking of concepts,
- inferencing from them and
- formation of heuristics, an actionable aspect of intelligence.
- Heuristics are educated guesses for all tasks.

For example, if one runs or encounters a danger, it is the resultant of human experience or his heuristics formation. In machines, it happens the same way.

4. **Heuristics** - The course correction is done by taking evaluation measures. Evaluation checks the thoroughness of the models and to-do course correction, if necessary, to generate better formulations

MACHINE LEARNING IN RELATION TO OTHER FIELDS

- Machine learning uses the concepts of **Artificial Intelligence, Data Science, and Statistics** primarily.
- It is the resultant of **combined ideas of diverse fields.**

MACHINE LEARNING IN RELATION TO OTHER FIELDS

Machine Learning and Artificial Intelligence

- Machine learning is an important branch of AI, which is a much broader subject.
- AI aims to develop intelligent agents.
- An agent can be a robot, humans, or any autonomous systems
- The resurgence in AI happened due to development of data driven systems.
- The **aim is to find relations and regularities present in the data**.
- Machine learning is the subbranch of AI, whose aim is **to extract the patterns for prediction.**
- It is a broad field that includes **learning from examples** and other areas like reinforcement learning.
- The relationship of AI and machine learning is shown in Figure 1.3.

MACHINE LEARNING IN RELATION TO OTHER FIELDS

- **Deep learning** is a subbranch of machine learning.
- In deep learning, the models are constructed using neural network technology.
- Neural networks are based on the human neuron models.
- Many neurons form a network connected with the activation functions that trigger further neurons to perform tasks.

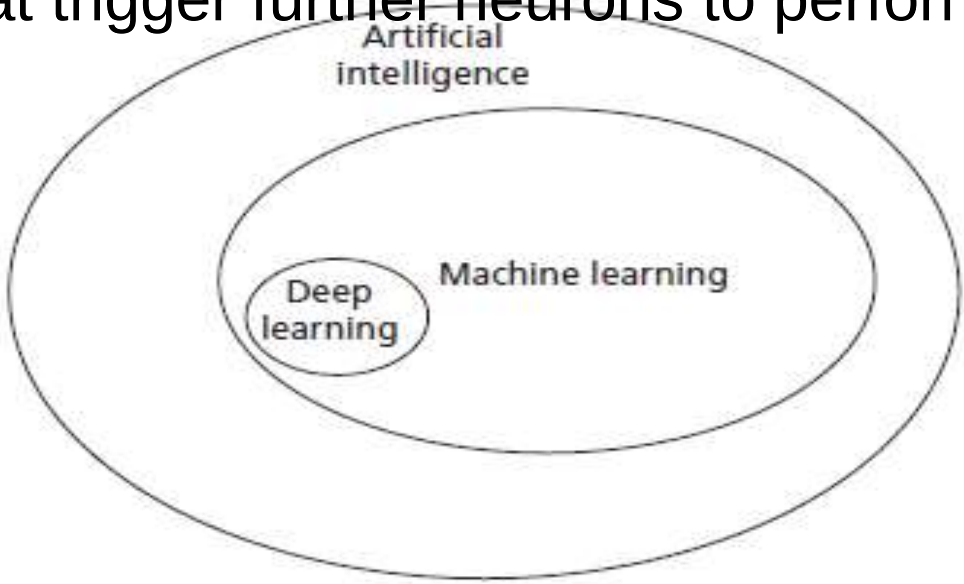


Figure 1.3: Relationship of AI with Machine Learning

MACHINE LEARNING IN RELATION TO OTHER FIELDS

Machine Learning, Data Science, Data Mining, and Data Analytics

- Data science is an ‘Umbrella’ term that encompasses many fields.
- Machine learning starts with **data**.
- Therefore, **data science and machine learning are interlinked**.
- Machine learning is a branch of data science.
- Data science deals with **gathering of data for analysis**.

MACHINE LEARNING IN RELATION TO OTHER FIELDS

Big Data : Data science concerns about **collection of data.**

Big data is a field of data science that deals with data’s following characteristics:

1. **Volume:** Huge amount of data is generated by big companies like Facebook, Twitter, YouTube.
2. **Variety:** Data is available in variety of forms like images, videos, and in different formats.
3. **Velocity:** It refers to the speed at which the data is generated and processed.

MACHINE LEARNING IN RELATION TO OTHER FIELDS

- Big data is used by many machine learning algorithms for applications such as **language translation and image recognition.**
- Big data **influences** the growth of subjects like **Deep learning**. Deep learning is a branch of machine learning that deals with constructing models using neural networks.

MACHINE LEARNING IN RELATION TO OTHER FIELDS

Data Mining

- Data mining’s original genesis is in the **business.**
- unearthing of the data produces **hidden information** that otherwise would have eluded the attention of the management.
- Nowadays, many consider that data mining and machine learning are same. There is no difference between these fields except that data mining aims to extract the hidden

MACHINE LEARNING IN RELATION TO OTHER FIELDS

Data Analytics Another **branch of data science** is data analytics.

- It aims to **extract useful knowledge** from crude data. There are different types of analytics.
- Predictive data analytics is used for **making predictions**. Machine learning is closely related to this branch of analytics and shares almost all algorithms**.**

Pattern Recognition It is an engineering field.

- It uses machine learning algorithms to extract the features for pattern analysis and pattern classification.
- One can view pattern recognition as a specific application of

MACHINE LEARNING IN RELATION TO OTHER FIELDS

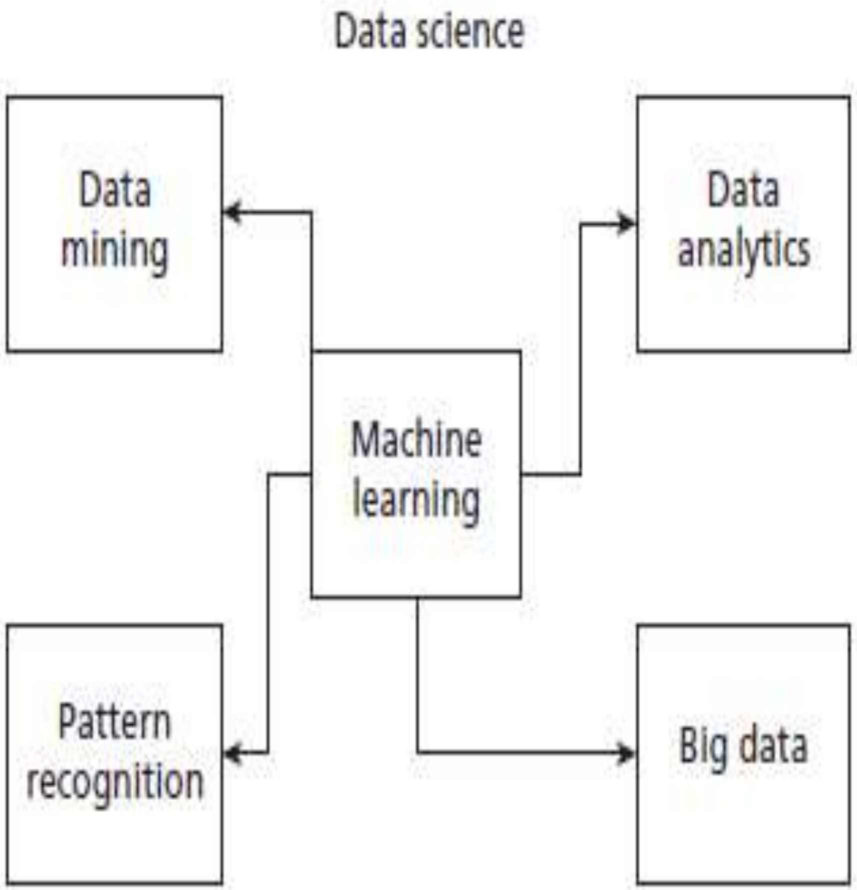


Figure 1.4: Relationship of Machine Learning with Other Major Fields

MACHINE LEARNING IN RELATION TO OTHER FIELDS

Machine Learning and Statistics

- Statistics is a branch of mathematics that has a solid theoretical foundation regarding statistical learning.
- Like machine learning (ML), it can learn from data.
- But the difference between statistics and ML is that statistical methods look for regularity in data called patterns.
- Initially, statistics sets a hypothesis and performs experiments to verify and validate the hypothesis in order to find relationships among data.

TYPES OF MACHINE LEARNING

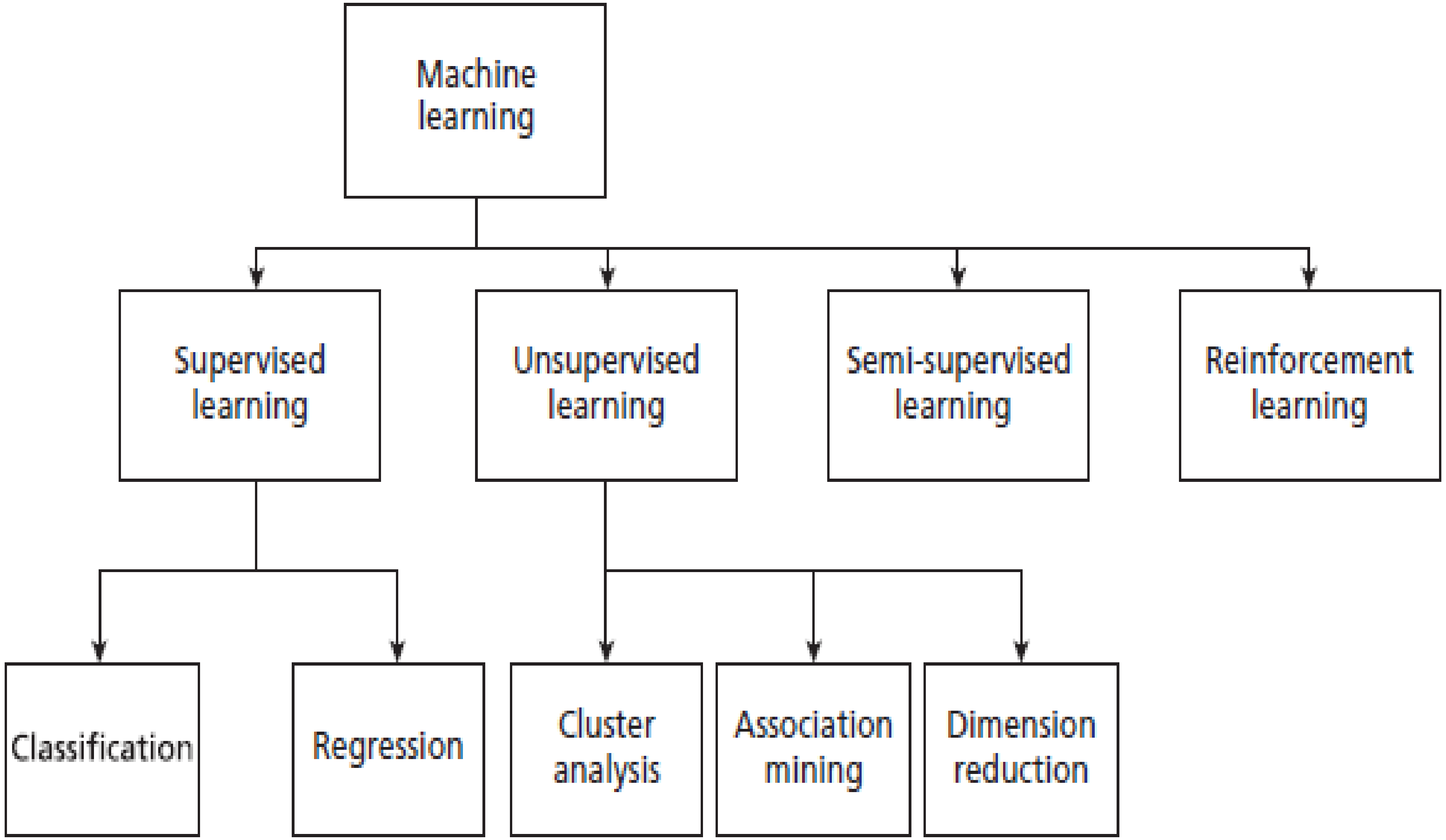


Figure 1.5: Types of Machine Learning

TYPES OF MACHINE LEARNING

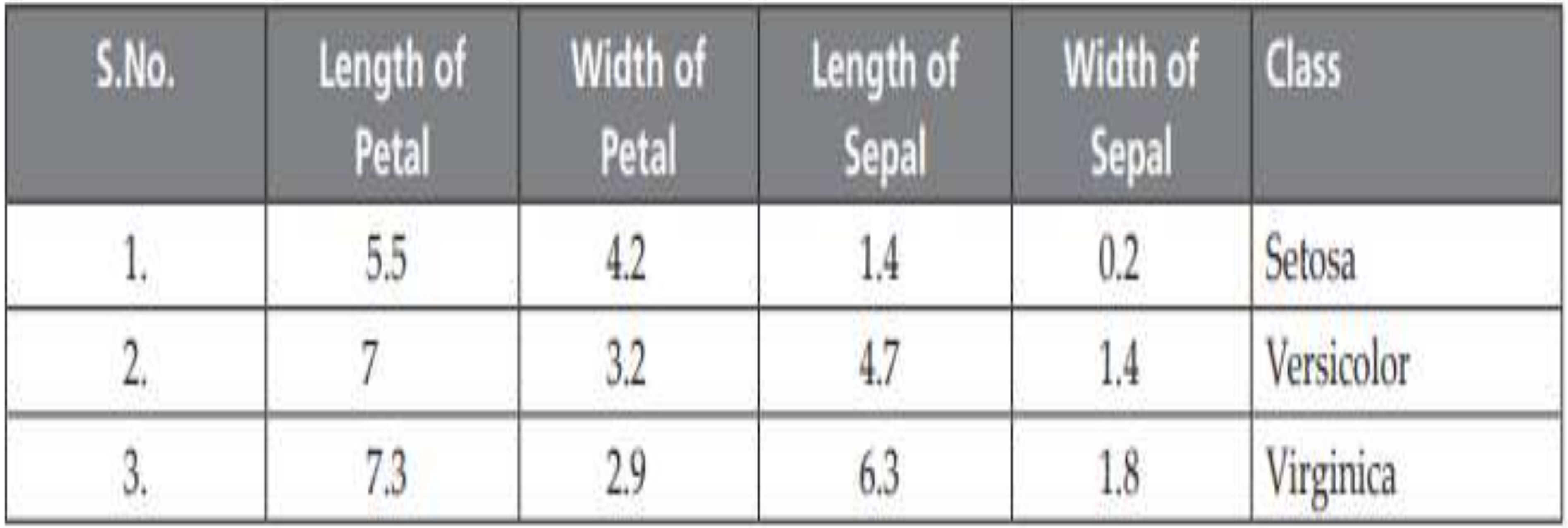
- Data is **a raw fact**.
- Normally, data is represented in the form of a **table**.
- Data also can be referred to as a data point, sample, or an example.
- Each row of the table represents a **data point**.
- **Features are attributes** or characteristics of an object.
- Normally, the **columns** of the table are attributes.
- Out of all attributes, one attribute is important and is called a label.
- **Label is the feature that we aim to predict**.
- Thus, there are two types of data – **labelled and**

TYPES OF MACHINE LEARNING

Labelled Data

- To illustrate labelled data, let us take one example dataset called Iris flower dataset or Fisher’s Iris dataset.
- The dataset has 50 samples of Iris – with four attributes, length and width of sepals and petals.
- The target variable is called class.
- There are three classes – Iris setosa, Iris virginica, and Iris versicolor.
- The partial data of Iris dataset is shown in Table 1.1.

TYPES OF MACHINE LEARNING



S.No.	Length of Petal	Width of Petal	Length of Sepal	Width of Sepal	Class
1.	5.5	4.2	1.4	0.2	Setosa
2.	7	3.2	4.7	1.4	Versicolor
3.	7.3	2.9	6.3	1.8	Virginica

TYPES OF MACHINE LEARNING

A dataset need not be always numbers. It can be images or video frames. Deep neural networks can handle images with labels. In the following Figure 1.6, the deep neural network takes images of dogs and cats with labels for classification.

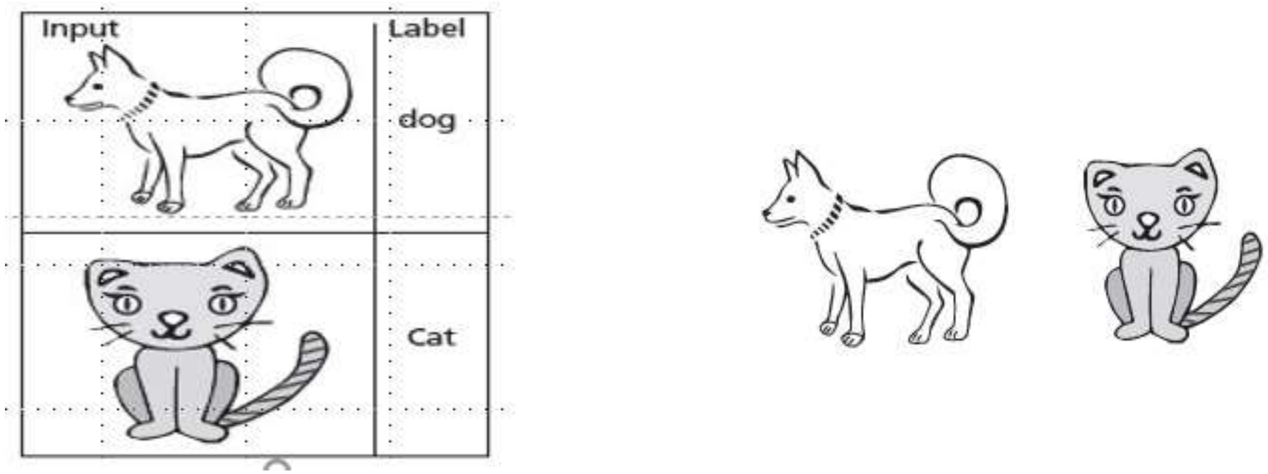


Figure 1.6: (a) Labelled Dataset (b) Unlabelled Dataset

In unlabelled data, there are no labels in the dataset.

TYPES OF MACHINE LEARNING

Supervised Learning

- Supervised algorithms **use labelled dataset**.
- As the name suggests, there is a **supervisor** or teacher **component** in supervised learning.
- A **supervisor provides labelled data** so that the model is constructed and generates test data.
- In supervised learning algorithms, learning takes place in two stages

TYPES OF MACHINE LEARNING

Supervised Learning

- In layman terms, during the first stage, the teacher communicates the information to the student that the student is supposed to master.
- The student receives the information and understands it.
- During this stage, the teacher has no knowledge of whether the information is grasped by the student.
- This leads to the second stage of learning.
- The teacher then asks the student a set of questions to find out how much information has been grasped by the student.
- Based on these questions the student is tested, and the teacher informs the student about his assessment.
- This kind of learning is typically called supervised learning.
- Supervised learning has two methods:

1. Classification **2. Regression**

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

- Classification is a supervised learning method.
- The input attributes of the classification algorithms are called **independent variables**.
- The target attribute is called label or **dependent variable**.
- The relationship between the input and target variable is represented in the form of a structure which is called a classification model.
- So, the focus of classification is to predict the ‘label’ that is in a discrete form (a value from the set of finite values).

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

- An example is shown in Figure 1.7 where a classification algorithm takes a set of labelled data images such as dogs and cats to construct a model that can later be used to classify an unknown test image data.

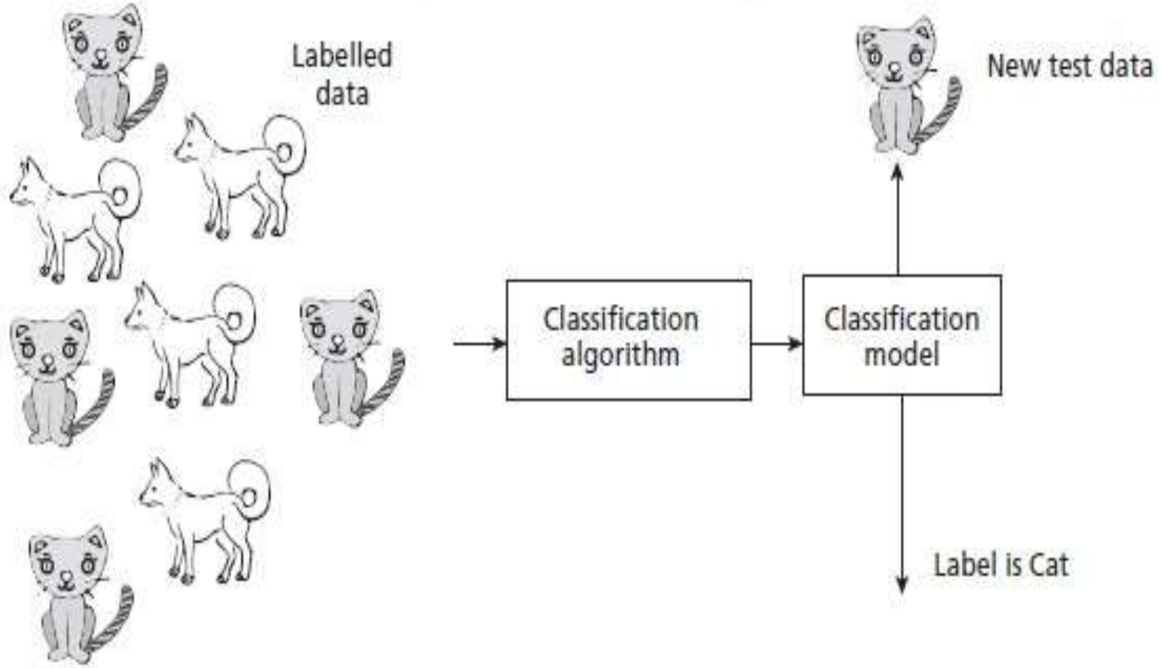


Figure 1.7: An Example Classification System

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

- In classification, learning takes place in two stages.
- During the first stage, called training stage, the learning algorithm takes a labelled dataset and starts learning.
- After the training set, samples are processed and the model is generated.
- In the second stage, the constructed model is tested with test or unknown sample and assigned a label.
- This is the classification process.

This is illustrated in the above Figure 1.7. Initially, the classification learning algorithm learns with the collection of labelled data and constructs the model. Then, a test case is selected, and the model assigns a label.

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

- In the case of Iris dataset, if the test is given as (6.3, 2.9, 5.6, 1.8, ?), the classification will generate the label for this.
- This is called classification.
- One of the examples of classification is – Image recognition, which includes classification of diseases like cancer, classification of plants, etc.

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

- The classification models can be categorized based on the implementation technology like decision trees, probabilistic methods, distance measures, and soft computing methods.
- Classification models can also be classified as generative models and discriminative models.
- Generative models deal with the process of data generation and its distribution.
- Probabilistic models are examples of generative models.
- Discriminative models do not care about the generation of data. Instead, they simply concentrate on classifying the

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

Some of the key algorithms of classification are:

- Decision Tree
- Random Forest
- Support Vector Machines
- Naïve Bayes
- Artificial Neural Network and Deep Learning networks like CNN

TYPES OF MACHINE LEARNING

Supervised Learning-Regression Models

- Regression models, unlike classification algorithms, predict continuous variables like price. In other words, it is a number.
- A fitted regression model is shown in Figure 1.8 for a dataset that represent weeks input x and product sales y

TYPES OF MACHINE LEARNING

Supervised Learning-Regression Models

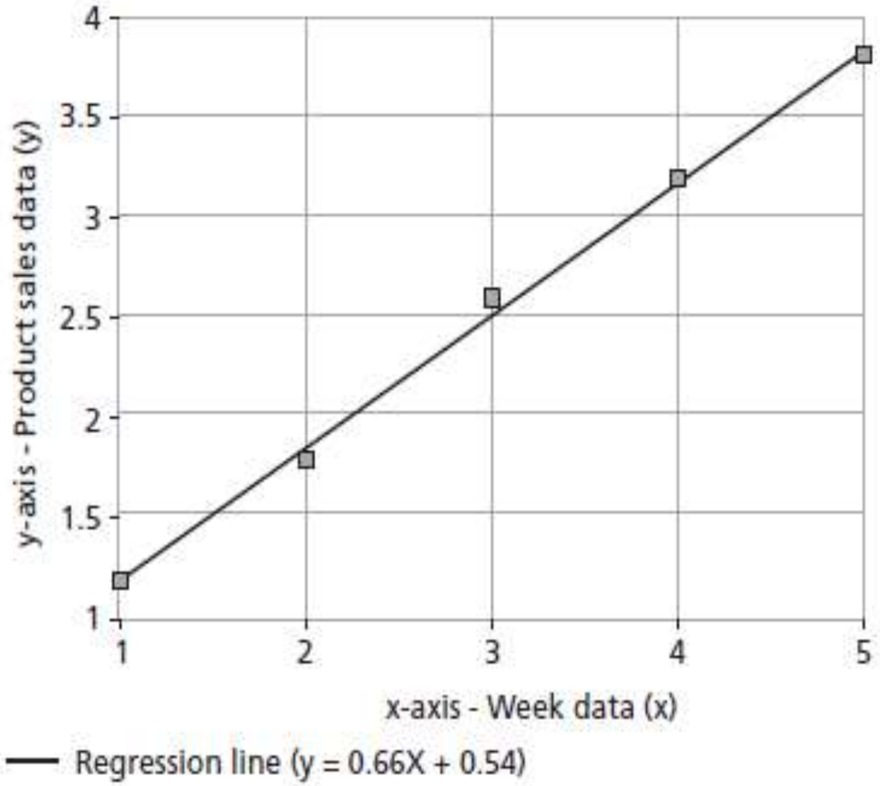


Figure 1.8: A Regression Model of the Form $y = ax + b$

product sales = 0.66 × Week + 0.54.

- ➢ The regression model takes input x and generates a model in the form of a fitted line of the form y = f(x).
- ➢ Here, x is the independent variable that may be one or more attributes and y is the dependent variable.
- ➢ In Figure 1.8, linear regression takes the training set and tries to fit it with a line – product sales = 0.66 × Week +

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

- Here, 0.66 and 0.54 are all regression coefficients that are learnt from data.
- The advantage of this model is that prediction for product sales (y) can be made for unknown week data (x).
- For example, the prediction for unknown eighth week can be made by substituting x as 8 in that regression formula to get y.

TYPES OF MACHINE LEARNING

Supervised Learning-Classification

- Both regression and classification models are supervised algorithms.
- Both have a supervisor and the concepts of training and testing are applicable to both.
- What is the difference between classification and regression models?
- The main difference is that regression models predict continuous variables such as product price, while classification concentrates on assigning labels such as class.

Unsupervised Learning

- The second kind of learning is by self-instruction.
- As the name suggests, there are no supervisor or teacher components. In the absence of a supervisor or teacher, self-instruction is the most common kind of learning process.
- This process of self-instruction is based on the concept of trial and error.
- Here, the program is supplied with objects, but no labels are defined.

TYPES OF MACHINE LEARNING

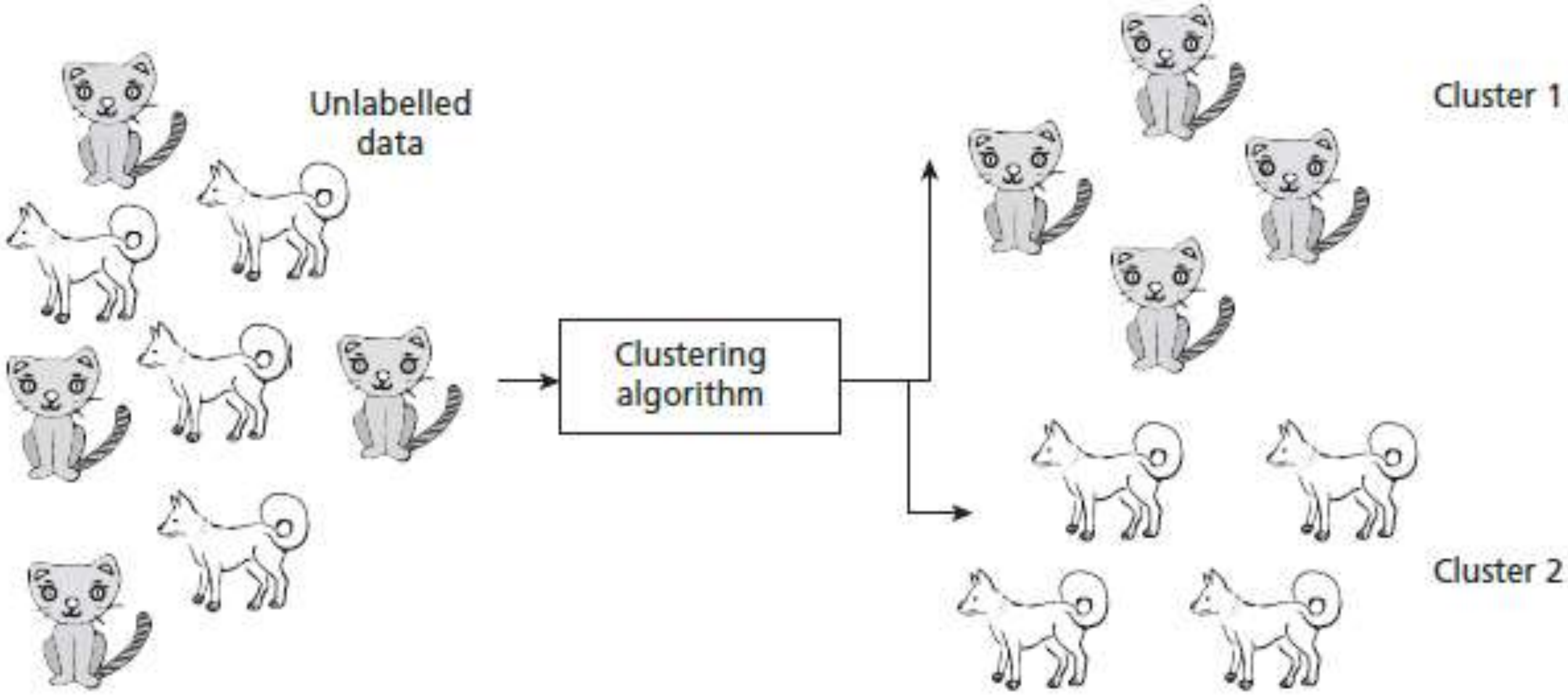
Unsupervised Learning

- The algorithm itself observes the examples and recognizes patterns based on the principles of grouping.
- Grouping is done in ways that similar objects form the same group.
- Cluster analysis and Dimensional reduction algorithms are examples of unsupervised algorithms.

TYPES OF MACHINE LEARNING

Unsupervised Learning-Cluster Analysis

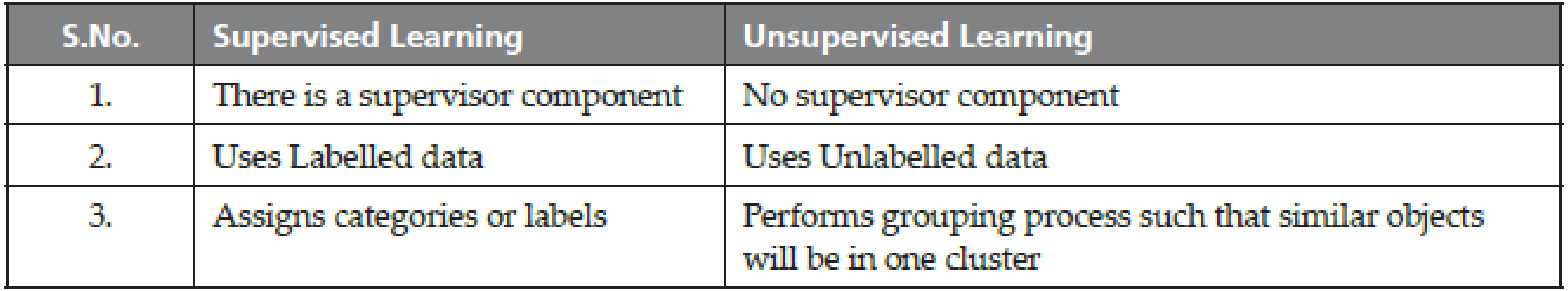
- Cluster analysis is an example of unsupervised learning.
- It aims to group objects into disjoint clusters or groups.
- Cluster analysis clusters objects based on its attributes.
- All the data objects of the partitions are similar in some aspect and vary from the data objects in the other partitions significantly.



- Some of the key clustering algorithms are:
  - k-means algorithm
  - Hierarchical algorithms

Dimensionality Reduction

- Dimensionality reduction algorithms are examples of unsupervised algorithms.
- It takes a **higher dimension data as input and outputs the data in lower dimension** by taking advantage of the variance of the data.
- It is a task of reducing the dataset with few features without losing the generality.



S.No.	Supervised Learning	Unsupervised Learning
1.	There is a supervisor component	No supervisor component
2.	Uses Labelled data	Uses Unlabelled data
3.	Assigns categories or labels	Performs grouping process such that similar objects will be in one cluster

Differences between Supervised and Unsupervised Learning

3. Semi-supervised Learning

- There are circumstances where the dataset has a huge collection of unlabelled data and some labelled data.
- Labelling is a costly process and difficult to perform by the humans.
- Semi-supervised algorithms use **unlabelled data by assigning a pseudo-label.** Then, the labelled and pseudo-labelled dataset can be combined.

4. Reinforcement Learning

- Reinforcement learning **mimics human beings**. Like human beings use ears and eyes to perceive the world and take actions, reinforcement learning allows the agent to interact with the environment to get rewards.
- The agent can be human, animal, robot, or any independent program.
- The rewards enable the agent to gain experience. The agent aims to maximize the reward.
- The **reward can be positive or negative** (Punishment). When the rewards are more, the behavior gets reinforced and learning becomes possible.

CHALLENGES OF MACHINE LEARNING

1.Problems – Machine learning can deal with the 'well-posed' problems where specifications are complete and available. Computers cannot solve 'ill-posed' problems.

2.Huge data – This is a primary requirement of machine learning. **Availability of a quality data is a challenge.** A quality data means it should be large and should not have data problems such as missing data or incorrect data.

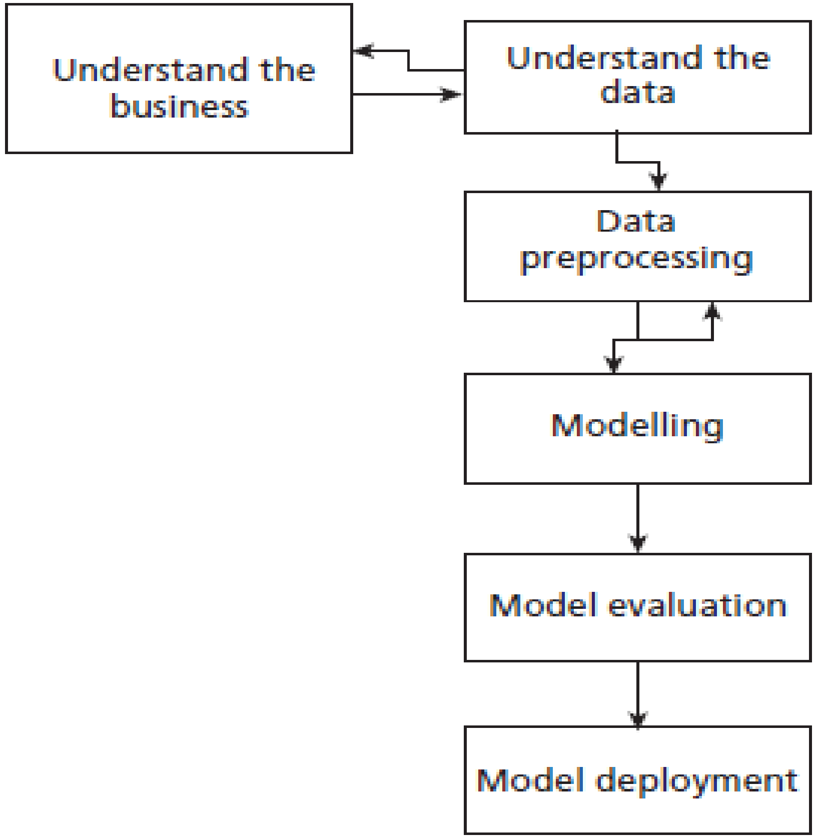
3.High computation power – With the availability of Big Data, the computational resource requirement has also increased. Systems with ***Graphics Processing Unit*** **(GPU) or even** ***Tensor Processing Unit*** **(TPU)** are required to execute machine learning algorithms.

4.Complexity of the algorithms – The **selection** of algorithms, describing the algorithms, application of algorithms to solve machine learning task, and comparison of algorithms have become necessary for machine learning or data scientists now.

5. Bias/Variance – Variance is the error of the model. This leads to a problem called bias/ variance tradeoff. **A model that fits the training data correctly but fails for test data**, in general lacks generalization, is called ***overfitting***. The reverse problem is called ***underfitting*** where the model fails for training data but has good generalization. **Overfitting and underfitting** are great challenges for machine learning algorithms.

MACHINE LEARNING PROCESS

1. **Understanding the business** – This step involves understanding the **objectives and requirements of the business** organization. Generally, a single data mining algorithm is enough for giving the solution.
2. **Understanding the data** – It involves the steps like **data collection, study of the characteristics of the data, formulation of hypothesis, and matching of patterns** to the selected hypothesis.
3. **Preparation of data** – This step involves producing the final dataset by **cleaning** the raw data and preparation of data for the data mining process.



4. **Modelling** – This step plays a role in the application of data mining algorithm for the data to **obtain a model or pattern.**

5. **Evaluate** – This step involves the evaluation of the data mining results using **statistical analysis and visualization methods**. The performance of the classifier is determined by evaluating the **accuracy** of the classifier.

6. **Deployment** – This step involves the **deployment of results** of the data mining algorithm to improve the existing process or for a new situation.

Some applications are listed below:

1. **Sentiment analysis** – This is an application of ***natural language processing (NLP)*** where the words of documents are converted to sentiments like happy, sad, and angry which are captured by emoticons effectively. For movie reviews or product reviews, five stars or one star are automatically attached using sentiment analysis programs.
2. **Recommendation systems** – These are systems that make personalized purchases possible. For example, Amazon recommends users to find related books or books bought by people who have the same taste like you, and Netflix suggests shows or related movies of your taste. The recommendation systems are based on ***machine learning***.
3. **Voice assistants** – Products like Amazon Alexa, Microsoft Cortana, Apple Siri, and Google Assistant are all examples of voice assistants. They take speech commands and perform tasks. These ***chatbots*** are the result of machine learning technologies.
4. **Technologies** like Google Maps and those used by Uber are all examples of machine learning which offer to locate and navigate shortest paths to reduce time.

What is Data ?

All facts are data. In computer systems, bits encode facts present in numbers, text, images, audio, and video.

Data is available in different data sources like flat files, databases, or data warehouses.

It can either be an **operational data** or a **non-operational data**.

Data by itself is **meaningless**. It has to be processed to generate any information. A string of bytes is meaningless. Only when a label is attached like height of students of a class, the data becomes **meaningful**.

Processed data is called **information** that includes patterns, associations, or relationships among data.

Elements of Big Data

Small Data

Big Data

Big Data can be characterized as follows :

1. **Volume** – Since there is a reduction in the cost of storing devices, there has been a tremendous growth of data. Small traditional data is measured in terms of gigabytes (GB) and terabytes (TB), but Big Data is measured in terms of petabytes (PB) and exabytes (EB).
2. **Velocity** – The fast arrival speed of data and its increase in data volume is noted as velocity. The availability of IoT devices and Internet power ensures that the data is arriving at a faster rate.

3. Variety – The variety of Big Data includes:

a. ***Form*** – There are many forms of data. Data types range from text, graph, audio, video, to maps.

b. ***Function*** – These are data from various sources like human conversations, transaction records, and old archive data.

c. ***Source of data*** – This is the third aspect of variety. There are many sources of data. Broadly, the data source can be classified as open/public data, social media data and multimodal data.

4. **Veracity of data** – Veracity of data deals with aspects like conformity to the facts, truthfulness, believability, and confidence in data. There may be many sources of error such as technical errors, typographical errors, and human errors.

5. **Validity** – Validity is the accuracy of the data for taking decisions or for any other goals that are needed by the given problem.

6. **Value** – Value is the characteristic of big data that indicates the value of the information that is extracted from the data and its influence on the decisions that are taken based on it.

Types of Data

1. **Structured Data**

Data is stored in an organized manner such as a database where it is available in the form of a table. The data can also be retrieved in an organized manner using tools like SQL.

The structured data frequently encountered in machine learning are listed below:

- ○**Record Data** A dataset is a collection of measurements taken from a process. We have a collection of objects in a dataset and each object has a set of measurements. The measurements can be arranged in the form of a matrix. Rows in the matrix represent an object and can be called as entities, cases, or records. The columns of the dataset are called attributes, features, or fields.
- ○**Data Matrix** It is a variation of the record type because it consists of numeric attributes. The standard matrix operations can be applied on these data.

○**Graph Data** It involves the **relationships among objects**. For example, a web page can refer to another web page. This can be modeled as a graph.

○**Ordered Data** Ordered data objects involve attributes that have an **implicit order among them**. The examples of ordered data are:

- ***Temporal data*** – It is the data whose attributes are **associated with time**. For example, the customer purchasing patterns during festival time is sequential data. Time series data is a special type of sequence data where the data is a series of measurements over time.
- ***Sequence data*** – It is like sequential data but does not have time stamps. This data involves the **sequence of words or letters**. For example, DNA data is a sequence of four characters – A T G C.
- ***Spatial data*** – It has attributes such as **positions or areas.** For example, maps are spatial data where the points are related by location.

2. **Unstructured Data**

Unstructured data includes video, image, and audio. It also includes textual documents, programs, and blog data. It is estimated that 80% of the data are unstructured data.

3. **Semi-Structured Data**

Semi-structured data are partially structured and partially unstructured. These include data like XML/JSON data, RSS feeds, and hierarchical data.

Data Storage and Representation

Flat Files

These are the simplest and most commonly available data source. It is also the cheapest way of organizing the data.

These flat files are the files where data is stored in plain ASCII or EBCDIC format.

Minor changes of data in flat files affect the results of the data mining algorithms.

Hence, flat file is suitable only for storing small dataset and not desirable if the dataset becomes larger.

Formats : CSV Files, TSV Files

Data Storage and Representation

Database System It normally consists of database files and a database management system (DBMS).

A transactional database is a collection of transactional records.

Time-series database stores time related information like log files where data is associated with a time stamp.

Spatial databases contain spatial information in a raster or vector format. Raster formats are either bitmaps or pixel maps. For example, images can be stored as a raster data. On the other hand, the vector format can be used to store maps as maps use basic geometric primitives like points, lines, polygons and so forth

Data Storage and Representation

World Wide Web (WWW) It provides a diverse, worldwide online information source. The objective of data mining algorithms is to mine interesting patterns of information present in WWW.

XML (eXtensible Markup Language) It is both human and machine interpretable data format that can be used to represent data that needs to be shared across the platforms.

Data Stream It is dynamic data, which flows in and out of the observing environment. Typical characteristics of data stream are huge volume of data, dynamic, fixed order movement, and real-time constraints.

RSS (Really Simple Syndication) It is a format for sharing instant feeds across services.

JSON (JavaScript Object Notation) It is another useful data interchange format that is often used for many machine learning algorithms.

Data analysis is an activity that takes the data and generates useful information and insights for assisting the organizations.

There are four types of data analytics:

1. **Descriptive analytics** is about **describing the main features of the data**. After data collection is done, descriptive analytics deals with the collected data and quantifies it. It is often stated that analytics is essentially statistics.
2. **Diagnostic Analytics** deals with the question – 'Why?'. This is also known as **causal analysis**, as it aims to find out the **cause and effect of the events**. For example, if a product is not selling, diagnostic analytics aims to find out the reason.

3. **Predictive Analytics** deals with the **future**. It deals with the question – 'What will happen in future given this data?'. This involves the application of algorithms to identify the patterns to predict the future.

4. **Prescriptive Analytics** is about the **finding the best course of action for the business organizations**. Prescriptive analytics goes beyond prediction and **helps in decision making** by giving a set of actions. It helps the organizations to plan better for the future and to mitigate the risks that are involved.

BIG DATA ANALYSIS FRAMEWORK

Big data framework is a layered architecture.

A 4-layer architecture has the following layers:

1. Data connection layer
2. Data management layer
3. Data analytics layer
4. Presentation layer

BIG DATA ANALYSIS FRAMEWORK

Data Connection Layer: It has data ingestion mechanisms and data connectors. Data ingestion means **taking raw data and importing it into appropriate data structures.** It performs the tasks of ETL process.

Data Management Layer: It performs **preprocessing of data**. The purpose of this layer is to allow parallel execution of queries, and read, write and data management tasks.

Data Analytic Layer: It has many functionalities such as **statistical tests, machine learning algorithms to understand, and construction of machine learning models**. This layer implements many model validation mechanisms too. The processing is done as *Cloud Computing, Grid Computing and H-Computing*:

BIG DATA ANALYSIS FRAMEWORK

- *Cloud computing* is an emerging technology which is basically a business service model or simply called as pay-per-usage model. The term ‘Cloud’ refers to the Internet that provides sharing of processing power, applications, storage and services. It offers different kinds of services such as *Iaas, Paas, and SaaS.*

- *Grid Computing* Grid Computing is a parallel and distributed computing framework consisting of a network of computers offering a super computing service as a single virtual supercomputer.

- *H-Computing (High Performance Computing or HPC)* It enables to perform complex tasks at high speed. It aggregates computing power in such a way that provides much higher performance to solve complex problems in science, engineering, research or business.

Presentation Layer: It has mechanisms such as dashboards, and applications that **display the results of analytical engines and machine learning**

BIG DATA ANALYSIS FRAMEWORK

Big Data Processing Cycle involves data management that consists of the following steps.

1. Data Collection
2. Data Preprocessing
3. Application of Machine Learning Algorithms
4. Interpretation of results and visualization of machine learning algorithm

DATA COLLECTION

The first task of gathering datasets are the **collection of data.**

It is often estimated that most of the time is spent for collection of good quality data.

'Good data' is one that has the following properties:

1. **Timeliness** – The data should be relevant and **not stale or obsolete data**.
2. **Relevancy** – The data should be **relevant and ready** for the machine learning or data mining algorithms. All the necessary information should be available and there should be no bias in the data.
3. **Knowledge about the data** – The data should be **understandable and interpretable,** and should be self-sufficient for the required application as desired by the domain knowledge engineer.

DATA COLLECTION

Broadly, the data source can be classified as *open/public data, social media data and multimodal data*.

1. **Open or public data source** – It is a data source that does not have any stringent copyright rules or restrictions. Its data can be primarily used for many purposes.
2. **Social media** – It is the data that is generated by various social media platforms like Twitter, Facebook, YouTube, and Instagram. An enormous amount of data is generated by these platforms.
3. **Multimodal data** – It includes data that involves many modes such as text, video, audio and mixed types.

DATA PREPROCESSING

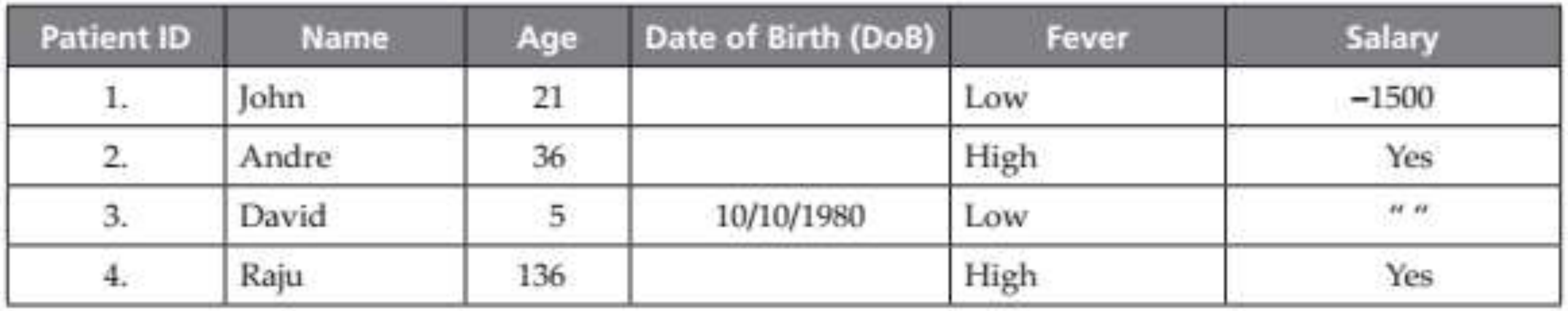
In real world, the available data is

- Incomplete data • Inaccurate data • Outlier data • Data with missing values • Data with inconsistent values • Duplicate data

Data preprocessing improves the quality of the data mining techniques. The raw data must be preprocessed to give accurate results.

The process of detection and removal of errors in data is called **data cleaning**.

DATA PREPROCESSING



Patient ID	Name	Age	Date of Birth (DoB)	Fever	Salary
1.	John	21		Low	−1500
2.	Andre	36		High	Yes
3.	David	5	10/10/1980	Low	" "
4.	Raju	136		High	Yes

Illustration of 'Bad' Data

It can be observed that data like Salary = ' ' is ***incomplete data***.

The DoB of patients, John, Andre, and Raju, is the ***missing data***.

The age of David .This is called ***inconsistent data***. Inconsistent data occurs due to problems in conversions, inconsistent formats, and difference in units.

Salary for John. It cannot be less than '0'. It is an instance of ***noisy data***.

Outliers are data that exhibit the characteristics that are different from other data and have very unusual values. It is often required to distinguish between noise and outlier data. The age of Raju .

DATA PREPROCESSING

Missing Data Analysis:

The primary data cleaning process is missing data analysis.

Data cleaning routines attempt to fill up the missing values, smoothen the noise while identifying the outliers and correct the inconsistencies of the data

DATA PREPROCESSING

The procedures that are given below can solve the problem of missing data:

1. **Ignore the tuple** – A tuple with missing data, especially the class label, is ignored. This method is not effective when the percentage of the missing values increases.
2. **Fill in the values manually** – Here, the domain expert can analyse the data tables and carry out the analysis and fill in the values manually. But, this is time consuming and may not be feasible for larger sets.
3. **A global constant can be used to fill in the missing attributes.** The missing values may be ’Unknown’ or be ’Infinity’. But, some data mining results may give spurious results by analysing these labels.
4. **The attribute value may be filled by the attribute value.** Say, the average income can replace a missing value.
5. **Use the attribute mean for all samples belonging to the same class.** Here, the average value replaces the missing values of all tuples that fall in this group.
6. **Use the most possible value to fill in the missing value.** The most probable value can be obtained from other methods like classification and decision tree prediction.

Removal of Noisy or Outlier Data:

Noise is a random error or variance in a measured value. It can be removed by using **binning**, which is a method where the given data values are sorted and distributed into equal frequency bins.

The bins are also called as ***buckets***.

The binning method then uses the neighbor values to smooth the noisy data.

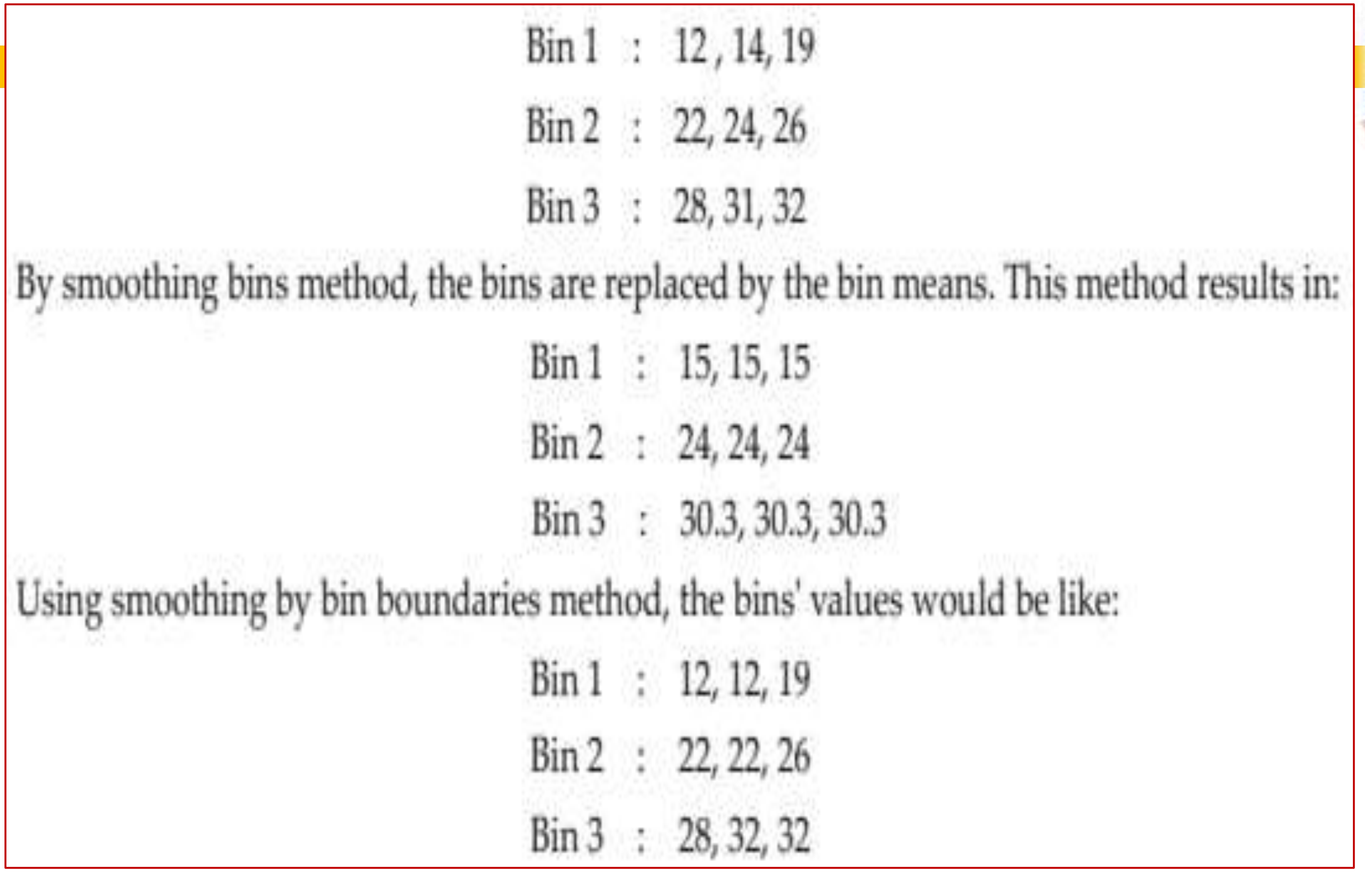
Some of the techniques commonly used are 'smoothing by means' where the mean of the bin removes the values of the bins, 'smoothing by bin medians' where the bin median replaces the bin values, and 'smoothing by bin boundaries' where the bin value is replaced by the closest bin boundary. The maximum and minimum values are called bin boundaries.

Binning methods may be used as a discretization technique.

DATA PREPROCESSING

Example 2.1: Consider the following set: S = {12, 14, 19, 22, 24, 26, 28, 31, 32}. Apply various binning techniques and show the result.

Solution: By equal-frequency bin method, the data should be distributed across bins. Let us assume the bins of size 3, then the above data is distributed across the bins as shown:



Bin 1 : 12 , 14, 19

Bin 2 : 22, 24, 26

Bin 3 : 28, 31, 32

By smoothing bins method, the bins are replaced by the bin means. This method results in:

Bin 1 : 15, 15, 15

Bin 2 : 24, 24, 24

Bin 3 : 30.3, 30.3, 30.3

Using smoothing by bin boundaries method, the bins' values would be like:

Bin 1 : 12, 12, 19

Bin 2 : 22, 22, 26

Bin 3 : 28, 32, 32

As per the method, the minimum and maximum values of the bin are determined, and it serves as bin boundary and does not change. Rest of the values are transformed to the nearest value. It can be observed in Bin 1, the middle value 14 is compared with the boundary values 12 and 19 and changed to the closest value, that is 12. This process is repeated for all bins.

DATA PREPROCESSING

Data Integration and Data Transformations:

Data integration involves routines that **merge data from multiple sources into a single data source**. So, this may lead to redundant data. The main goal of data integration is to detect and remove redundancies that arise from integration.

Data transformation routines perform operations like ***normalization*** to **improve the performance** of the data mining algorithms. technique. In normalization, the attribute values are **scaled to fit in a range** (say 0-1) to improve the performance of the data mining algorithm. Often, in neural networks, these techniques are used.

Some of the normalization procedures used are:

1. Min-Max   2. z-Score

DATA PREPROCESSING

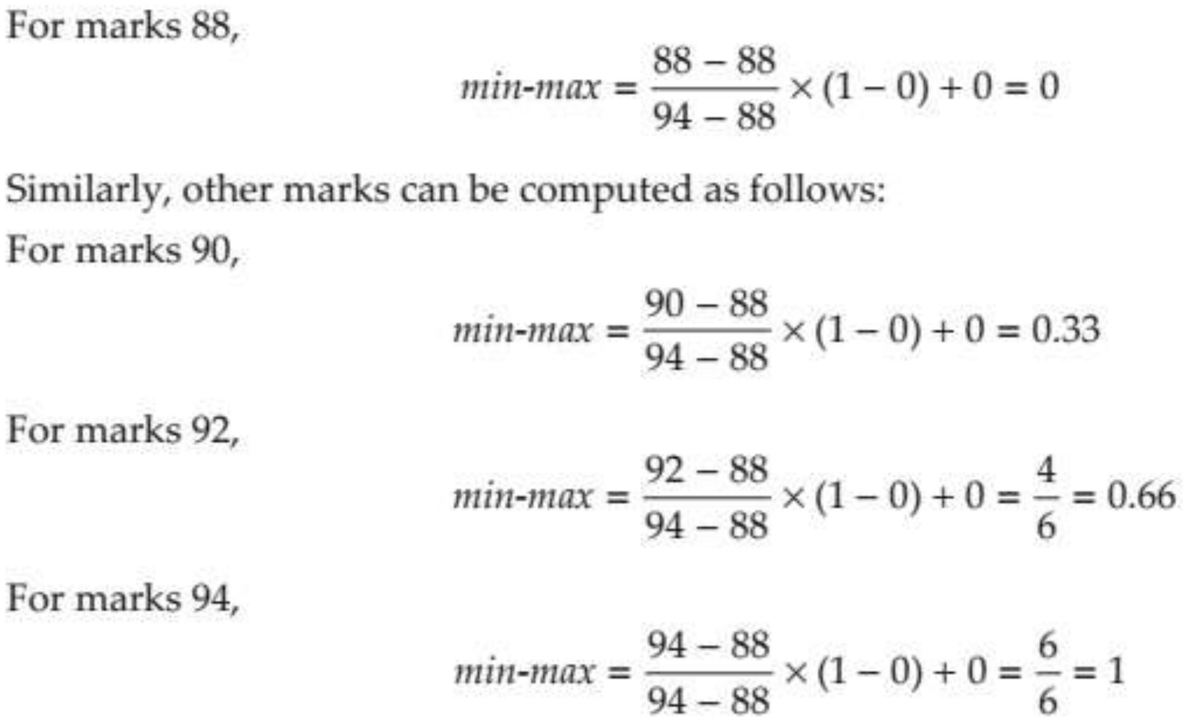
Min-Max Procedure It is a normalization technique where each variable V is **normalized by its difference with the minimum value divided by the range to a new range,** say 0–1. Often, neural networks require this kind of normalization. The formula to implement this normalization is given as:

Here max-min is the range. Min and max are the minimum and maximum of the given data, new max and new min are the minimum and maximum of the target range, say 0 and 1.

$$min\text{-}max = \frac{V - min}{max - min} \times (new\ max - new\ min) + new\ min$$

Example 2.2: Consider the set: V = {88, 90, 92, 94}. Apply Min-Max procedure and map the marks to a new range 0–1.

Solution: The minimum of the list V is 88 and maximum is 94. The new min and new max are 0 and 1, respectively. The mapping can be done using max-min as:



For marks 88,

$$\textit{min-max} = \frac{88-88}{94-88} \times (1-0) + 0 = 0$$

Similarly, other marks can be computed as follows:

For marks 90,

$$\textit{min-max} = \frac{90-88}{94-88} \times (1-0) + 0 = 0.33$$

For marks 92,

$$\textit{min-max} = \frac{92-88}{94-88} \times (1-0) + 0 = \frac{4}{6} = 0.66$$

For marks 94,

$$\textit{min-max} = \frac{94-88}{94-88} \times (1-0) + 0 = \frac{6}{6} = 1$$

So, it can be observed that the marks {88, 90, 92, 94} are mapped to the new range {0, 0.33, 0.66, 1}. Thus, the Min-Max normalization range is between 0 and 1.

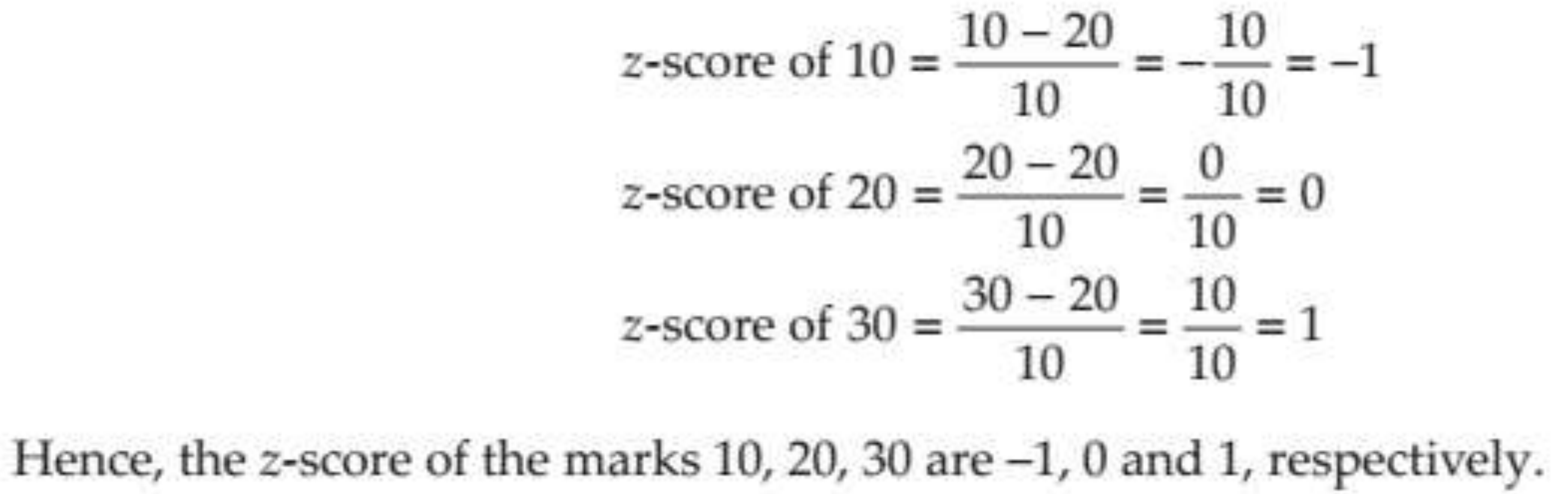
z-Score Normalization This procedure works by taking the difference between the field value and mean value, and by scaling this difference by standard deviation of the attribute.

$$V* = V - \mu/\sigma$$

Here, σ is the standard deviati n of the list V.

Example 2.3: Consider the mark list V = {10, 20, 30}, convert the marks to z-score.

Solution: The mean and Sample Standard deviation (s) values of the list V are 20 and 10, respectively. So the z-scores of these marks are calculated using ***V**** as:



$$\text{z-score of } 10 = \frac{10-20}{10} = -\frac{10}{10} = -1$$

$$\text{z-score of } 20 = \frac{20-20}{10} = \frac{0}{10} = 0$$

$$\text{z-score of } 30 = \frac{30-20}{10} = \frac{10}{10} = 1$$

Hence, the *z*-score of the marks 10, 20, 30 are –1, 0 and 1, respectively.

What is the use of z-scores?

z-scores are used to detect outlier detection. If the data value z-score function is either less than -3 or greater than +3, then it is possibly an outlier.

DATA PREPROCESSING

Data Reduction:

Data reduction reduces data size but produces the same results. There are different ways in which data reduction can be carried out such as data aggregation, feature selection, and dimensionality reduction.

DESCRIPTIVE STATISTICS

Descriptive statistics is a branch of statistics that does dataset summarization. It is used to summarize and describe data. Descriptive statistics are just descriptive and do not go beyond that.

Data visualization is a branch of study that is useful for investigating the given data.

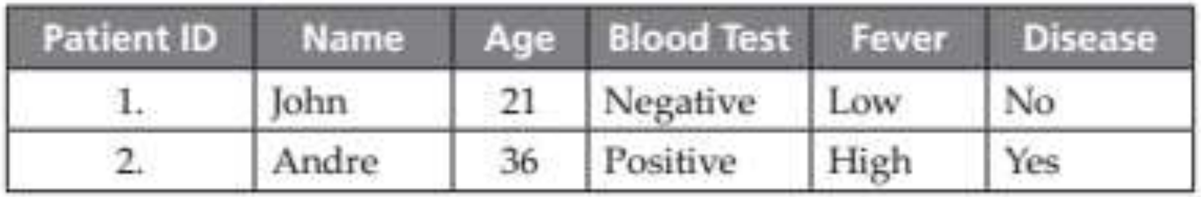
Descriptive analytics and data visualization techniques help to **understand the nature of the data**, which further helps to determine the kinds of machine learning or data mining tasks that can be applied to the data. This step is often known as **Exploratory Data Analysis (EDA)**.

Dataset and Data Types

A dataset can be assumed to be a collection of data objects.

The data objects may be records, points, vectors, patterns, events, cases, samples or observations. These records contain many attributes.

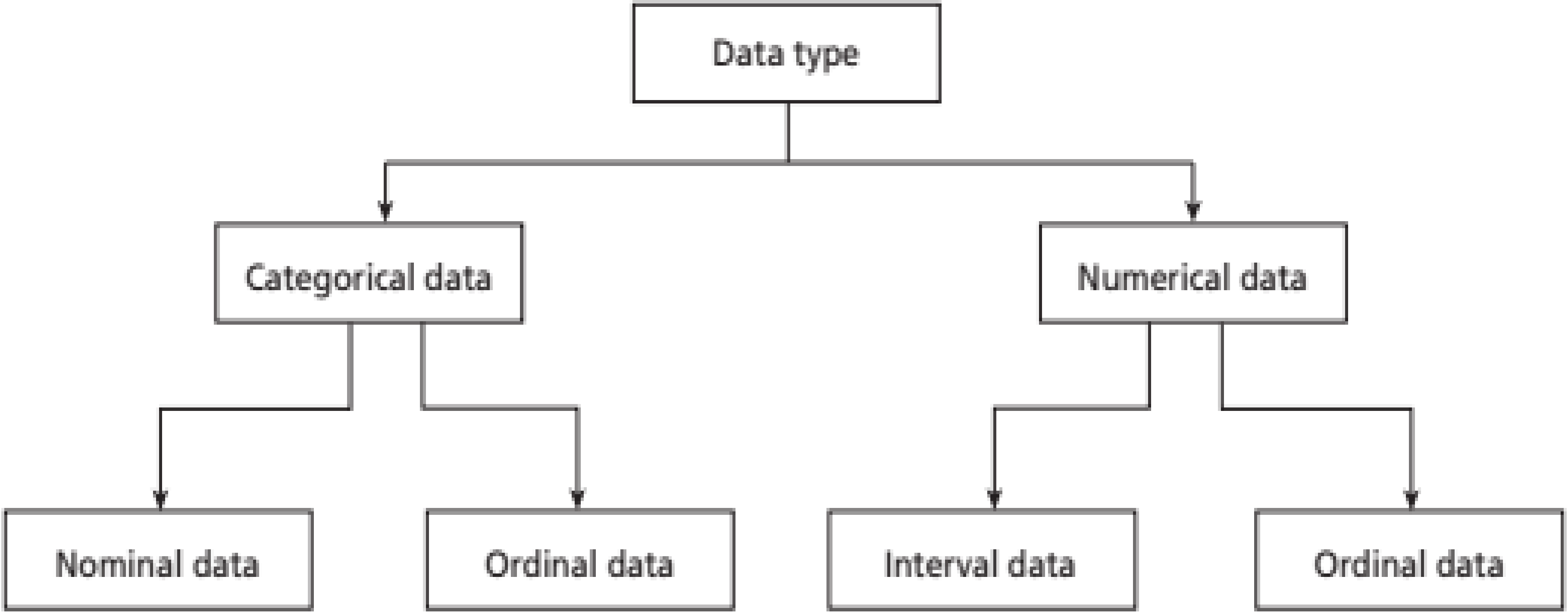
An attribute can be defined as the property or characteristics of an object. For example, consider the following database shown in sample .



Patient ID	Name	Age	Blood Test	Fever	Disease
1.	John	21	Negative	Low	No
2.	Andre	36	Positive	High	Yes

Every attribute should be associated with a value. This process is called **measurement**.

The type of attribute determines the data types, often referred to as measurement scale types



Data type

Categorical data

Numerical data

Nominal data

Ordinal data

Interval data

Ordinal data

Types of Data

Broadly, data can be classified into two types:

1. Categorical or qualitative data
2. Numerical or quantitative data

Categorical or Qualitative Data The categorical data can be divided into two types. They are **nominal type** and **ordinal type**.

❑**Nominal Data :** Nominal data are **symbols and cannot be processed like a number.**

Nominal data type provides only information but **has no ordering among data**. Only operations like (=, ≠) are meaningful for these data. For example, the patient ID can be checked for equality and nothing else.

❑**Ordinal Data** – It provides **enough information and has natural order**. For example, Fever = {Low, Medium, High} is an ordinal data

Numeric or Qualitative Data **:**It can be divided into two categories. They are ***interval type*** and ***ratio type***.

Interval Data – Interval data is a **numeric data for which the differences between values are meaningful**.

For example, there is a difference between 30 degree and 40 degree. Only the permissible operations are + and -.

Ratio Data – For ratio data, **both differences and ratio are meaningful**. The difference between the ratio and interval data is the position of zero in the scale.

For example, take the Centigrade-Fahrenheit conversion. The zeroes of both scales do not match. Hence, these are interval data.

Another way of classifying the data is to classify it as:

Discrete Data : This kind of data is recorded as integers.

Eg : response of the survey can be discrete data.

Continuous Data : It can be fitted into a range and includes decimal point.

Eg: age is a continuous data

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

Univariate analysis is the simplest form of statistical analysis.

As the name indicates, the dataset has only one variable.

A variable can be called as a **category**.

Univariate does not deal with cause or relationships.

The aim of univariate analysis is to **describe data and find patterns.**

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

1.Data Visualization:

To understand data, graph visualization is must.

Data visualization helps to understand data.

It helps to present information and data to customers.

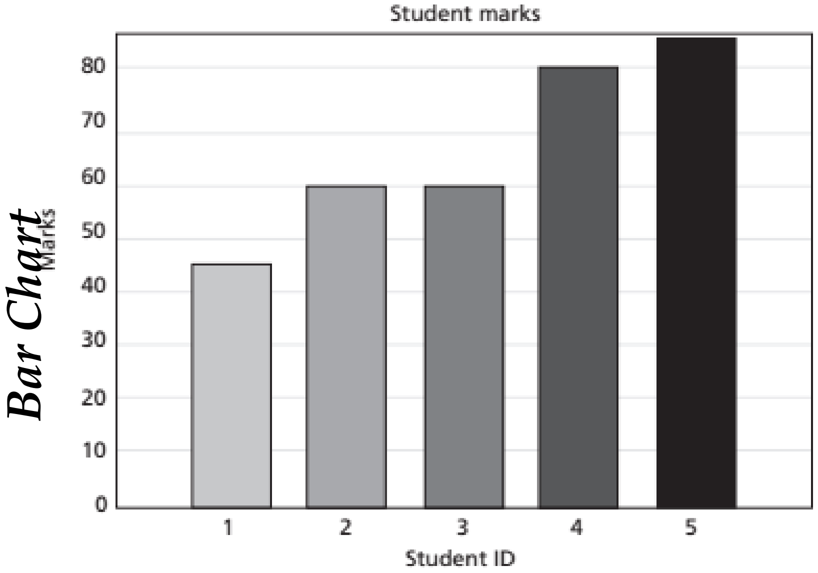
Some of the graphs that are used in univariate data analysis are *bar charts, histograms, frequency polygons and pie charts*.

Bar Chart A Bar chart is used to **display the frequency distribution for variables.**

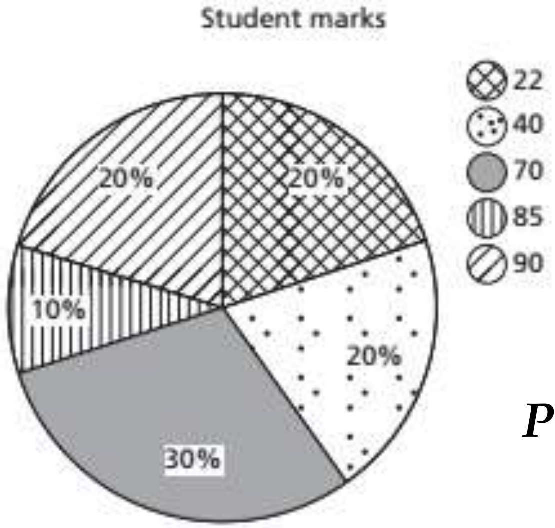
Bar charts are used to illustrate **discrete data.** The charts can also help to explain the **counts of nominal data**.

The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown

UNIVARIATE DATA ANALYSIS AND VISUALIZATION



Bar Chart



Pie Chart

Pie Chart These are equally helpful in illustrating the univariate data. The percentage frequency distribution of students' marks {22, 22, 40, 40, 70, 70, 70, 85, 90, 90} is below in Figure 2.18.

It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, 2/10 × 100 = 20% space in a pie of 100% is allotted for marks 22 in Figure 2.18.

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

Histogram : shows frequency distributions. The histogram for students' marks {45, 60, 60, 80, 85} in the group range of 0-25, 26-50, 51-75, 76-100 is given below in Figure 2.19. One can visually inspect from Figure 2.19 that the number of students in the range 76-100 is 2.

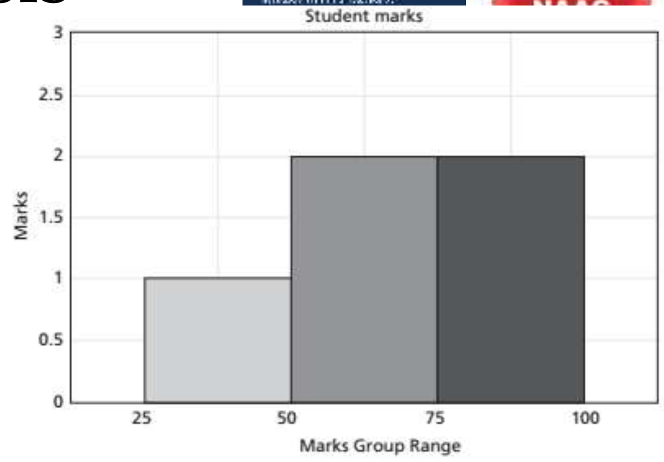


Figure 2.19: *Sample Histogram of English Marks*

Dot Plots similar to bar charts. The dot plot of English marks for five students with ID as {1, 2, 3, 4, 5} and marks {45, 60, 60, 80, 85} is given in Figure 2.20. The advantage is that by visual inspection one can find out who got more marks.

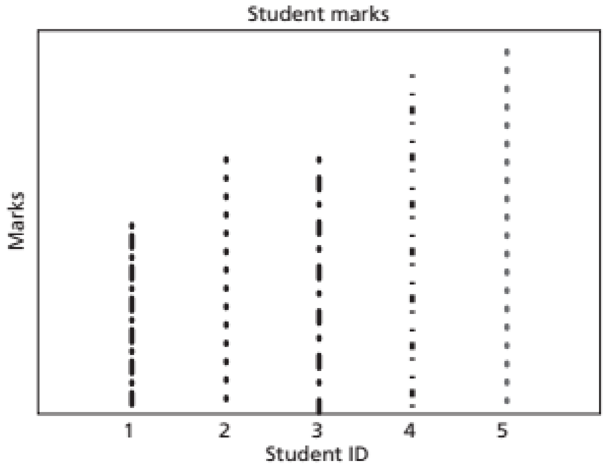


Figure 2.20: *Dot Plots*

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

2. Central Tendency

A condensation or summary of the data is necessary. This makes the data analysis easy and simple.

One such summary is called central tendency.

Thus, central tendency can explain the characteristics of data and that further helps in comparison.

Mass data have tendency to concentrate at certain values, normally in the central location.

It is called measure of central tendency (or averages).

This represents the first order of measures.

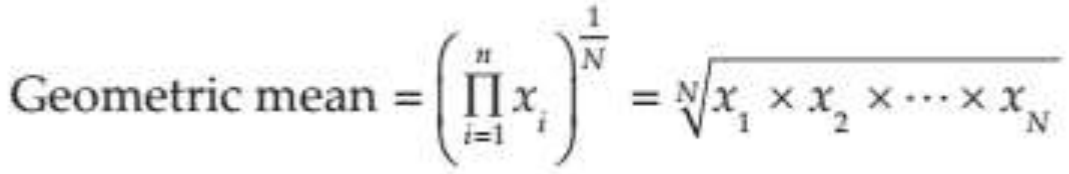
Popular measures are mean, median and mode.

Mean – Arithmetic average (or mean) is a measure of central tendency that represents the ‘center’ of the dataset.

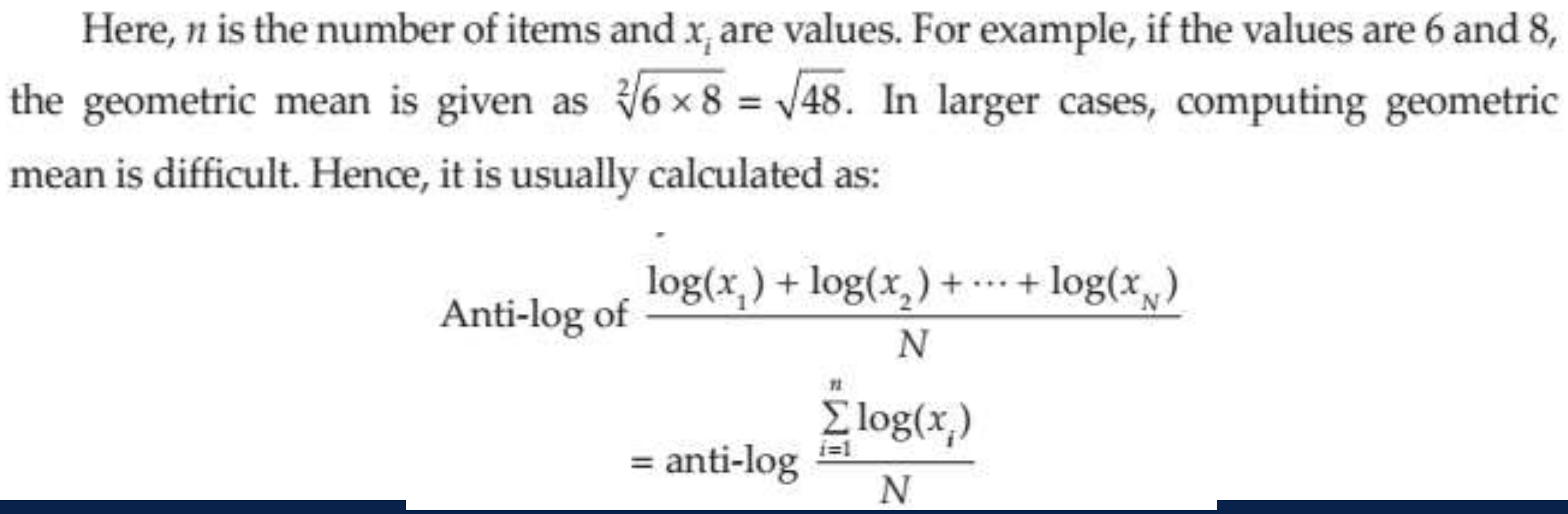
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Weighted mean – Unlike arithmetic mean that gives the weightage of all items equally, **weighted mean gives different importance to all items** as the item importance varies. Hence, different weightage can be given to items.

Geometric mean – Let $x_1, x_2, \ldots, x_N$ be a set of 'N' values or observations. Geometric mean is the N^{th} root of the product of N items. The formula for computing geometric mean is given as follows:



$$\text{Geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{N}} = \sqrt[N]{x_1 \times x_2 \times \cdots \times x_N}$$



Here, n is the number of items and x_i are values. For example, if the values are 6 and 8, the geometric mean is given as $\sqrt[2]{6 \times 8} = \sqrt{48}$. In larger cases, computing geometric mean is difficult. Hence, it is usually calculated as:

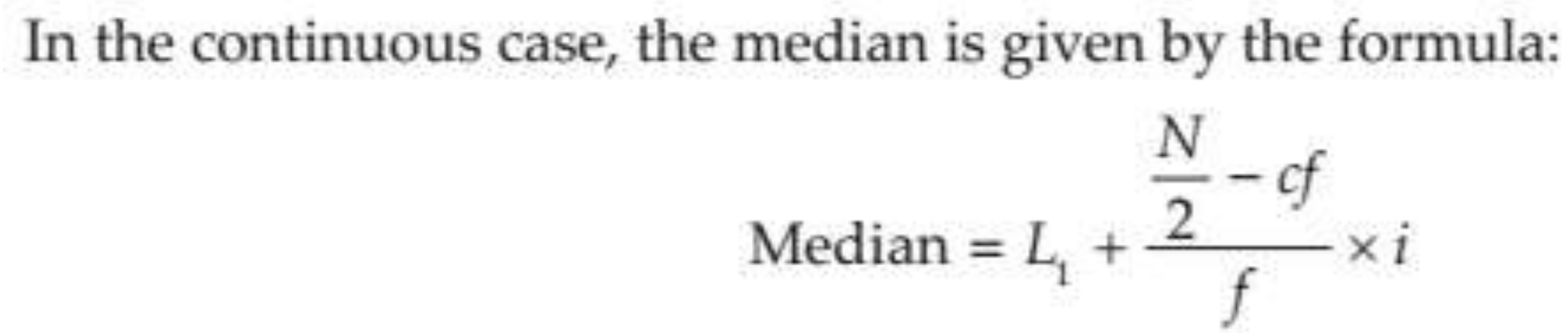
$$\text{Anti-log of } \frac{\log(x_1) + \log(x_2) + \cdots + \log(x_N)}{N}$$

$$= \text{anti-log } \frac{\sum_{i=1}^{n} \log(x_i)}{N}$$

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

Median – The middle value in the distribution is called **median**.

- If the total number of items in the distribution is ***odd,*** then the middle value is called median. If the numbers are ***even,*** then the average value of two items in the centre is the median.



In the continuous case, the median is given by the formula:

$$\text{Median} = L_1 + \frac{\frac{N}{2} - cf}{f} \times i$$

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

Mode – Mode is the value that occurs more frequently in the dataset.

In other words, the value that has the highest frequency is called mode.

Mode is only for discrete data and is not applicable for continuous data as there are no repeated values in continuous data.

3. Dispersion

- The spreadout of a set of data around the central tendency (mean, median or mode) is called **dispersion**.
- It shows how data are spread and how different they are from one another,
- Dispersion is represented by various ways such as ***range, variance, standard deviation, and standard error***.

Range

- Range is the difference between the maximum and minimum of values of the given list of data.

Standard Deviation

- The mean does not convey much more than a middle point.

For example, the following datasets {10, 20, 30} and {10, 50, 0} both have a mean of 20. The difference between these two sets is the spread of data.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Here, N is the size of the population, x_i is observation or value from the population and μ is the population mean. Often, N – 1 is used instead of N in the denominator.

Quartiles and Inter Quartile Range

- Quartiles are values that divide a series into four equal parts.
- The interquartile range (IQR) is the difference between the first and third quartiles.
- It's a measure of variability used with the median It is sometimes convenient to subdivide the dataset using coordinates. Percentiles are about data that are less than the coordinates by some percentage of the total value.
- kth percentile is the property that the k% of the data lies at or below X_i.

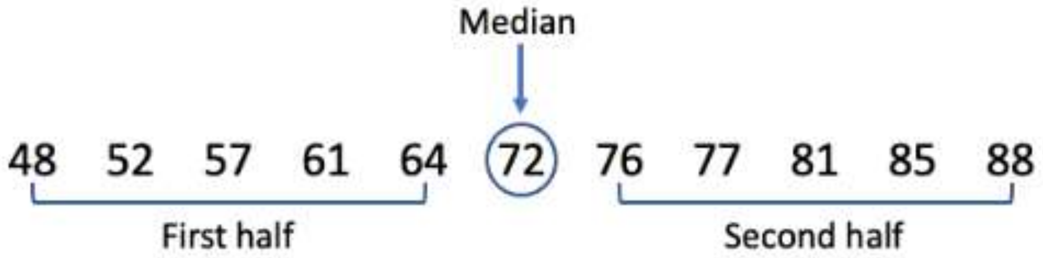
For example, median is 50th percentile and can be denoted as $Q_{0.50}$. The 25th percentile is called **first quartile (Q_1)** and the 75th percentile is called **third quartile (Q_3).**

Step 1: Order your values from low to high.

48 52 57 61 64 72 76 77 81 85 88

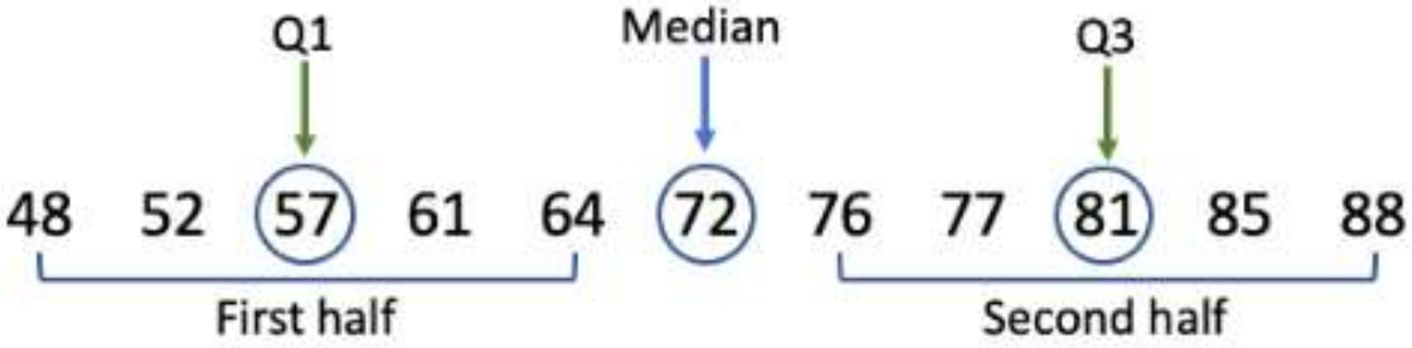
Step 2: Locate the median, and then separate the values below it from the values above it.

In an odd-numbered data set, the median is the number in the middle of the list. The median itself is excluded from both halves: one half contains all values below the median, and the other contains all the values above it.



Step 3: Find Q1 and Q3.

Q1 is the median of the first half and Q3 is the median of the second half. Since each of these halves have an odd-numbered size, there is only one value in the middle of each half.



Q1 Median Q3

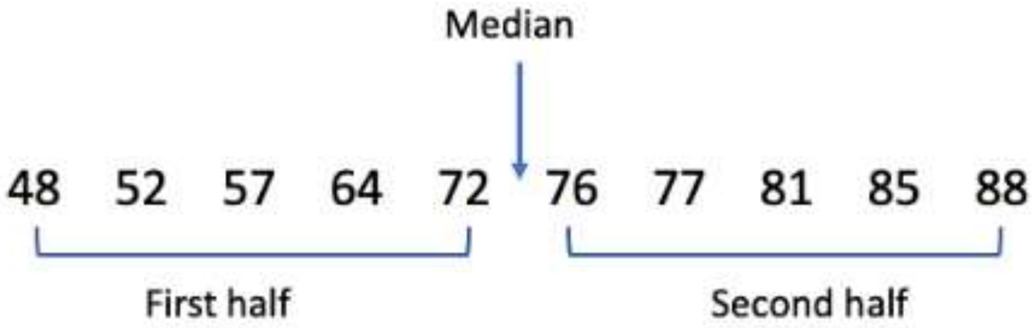
48 52 57 61 64 72 76 77 81 85 88

First half Second half

Step 4: Calculate the interquartile range.

$$IQR = Q3 \; - Q1$$
$$IQR = 81 \; - 57 = 24$$

48 52 57 64 72 76 77 81 85 88

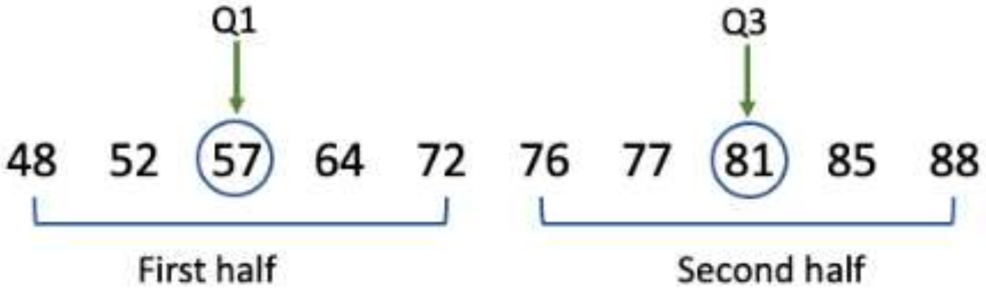


Median

48 52 57 64 72 76 77 81 85 88

First half

Second half



Q1

Q3

48 52 57 64 72 76 77 81 85 88

First half

Second half

$$IQR = Q3\ - Q1$$

$$IQR = 81\ - 57 = 24$$

Example 2.4: For patients' age list {12, 14, 19, 22, 24, 26, 28, 31, 34}, find the IQR.

Solution: The median is in the fifth position. In this case, 24 is the median. The first quartile is median of the scores below the mean i.e., {12, 14, 19, 22}. Hence, it's the median of the list below 24. In this case, the median is the average of the second and third values, that is, $Q_{0.25}$ = 16.5. Similarly, the third quartile is the median of the values above the median, that is {26, 28, 31, 34}. So, $Q_{0.75}$ is the average of the seventh and eighth score. In this case, it is 28 + 31/2 = 59/2 = 29.5.

Hence, the IQR is:

$$\begin{aligned} &= Q_{0.75} - Q_{0.25} \\ &= 29.5 - 16.5 = 13 \end{aligned}$$

The half of IQR is called semi-quartile range. The Semi Inter Quartile Range (SIQR) is given as:

$$\begin{aligned} \text{SIQR} &= \frac{1}{2} \times \text{IQR} \\ &= \frac{1}{2} \times 13 = 6.5 \end{aligned}$$

Five-point Summary and Box Plots

- The median, quartiles Q1 and Q3, and minimum and maximum written in the order < Minimum, Q1, Median, Q3, Maximum > is known as **five-point summary**.
- Box plots are suitable for continuous variables and a nominal variable.
- Box plots can be used to illustrate data distributions and summary of data. It is the popular way for plotting five number summaries. A Box plot is also known as a **Box and whisker plot**.

Example 2.5: Find the 5-point summary of the list {13, 11, 2, 3, 4, 8, 9}.

Solution: The minimum is 2 and the maximum is 13. The Q1, Q2 and Q3 are 3, 8 and 11, respectively. Hence, 5-point summary is {2, 3, 8, 11, 13}, that is, {minimum, Q1, median, Q3, maximum}. Box plots are useful for describing 5-point summary. The Box plot for the set is given in Figure 2.21.

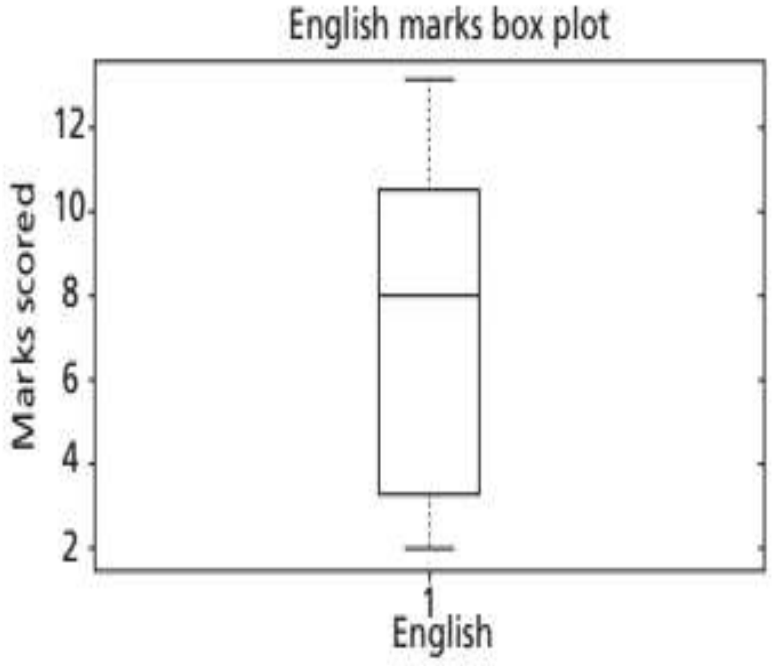


Figure 2.21: *Box Plot for English Marks*

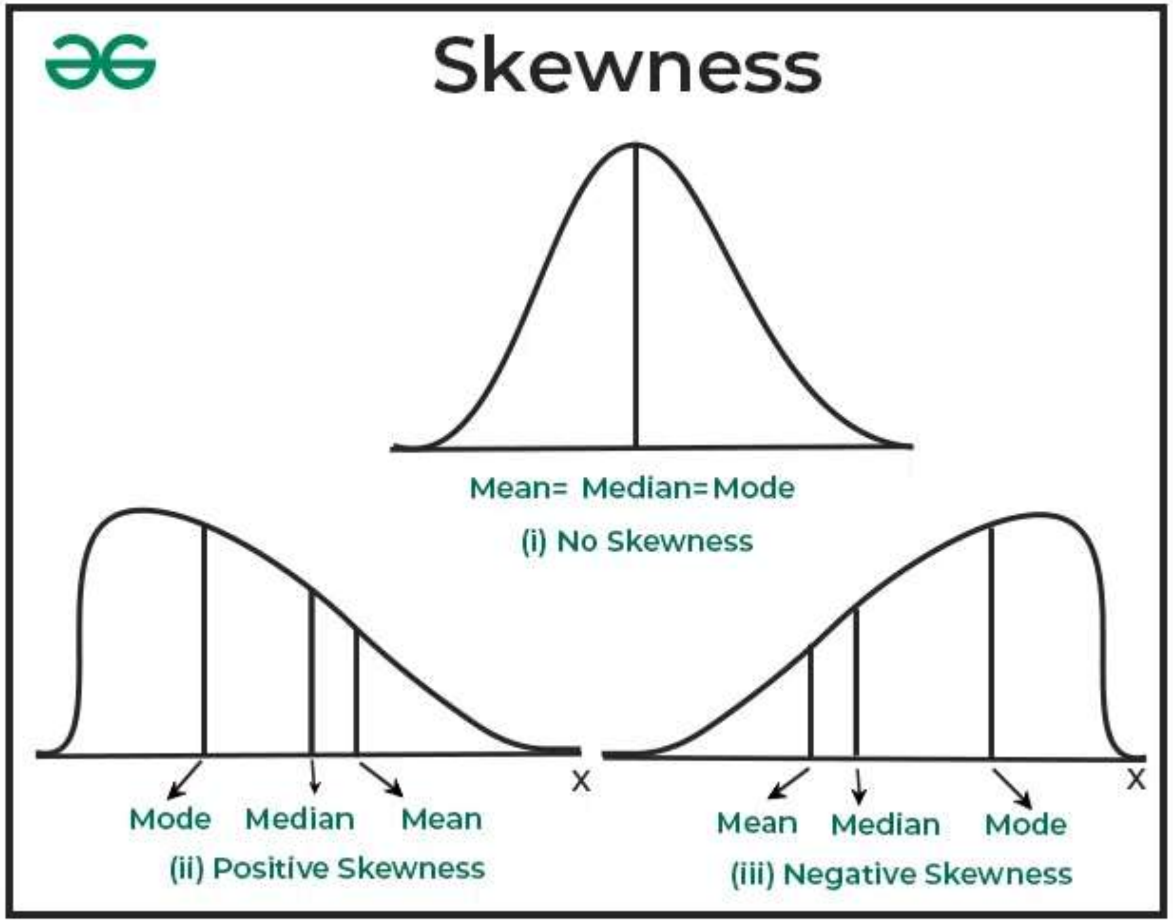
UNIVARIATE DATA ANALYSIS AND VISUALIZATION

4. Shape

- Skewness and Kurtosis (called moments) indicate the symmetry/asymmetry and peak location of the dataset.

Skewness

- The measures of direction and degree of symmetry are called **measures of third order**.
- Ideally, skewness should be zero as in ideal normal distribution. More often, the given dataset may not have perfect symmetry (consider the following Figure 2.22).


Skewness
Mean= Median=Mode
(i) No Skewness
X
Mode
Median
Mean
(ii) Positive Skewness
X
Mean
Median
Mode
(iii) Negative Skewness

UNIVARIATE DATA ANALYSIS AND VISUALIZATION

- The relationship between skew and the relative size of the mean and median can be summarized by a convenient numerical skew index known as Pearson 2 skewness coefficient.

$$\frac{3 \times (\mu - median)}{\sigma}$$

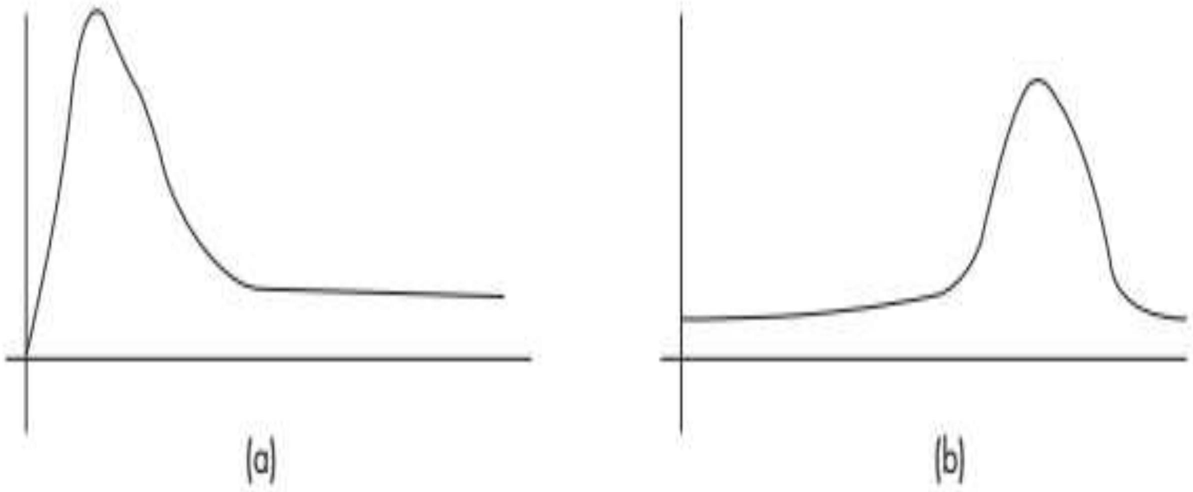


Figure 2.22: *(a) Positive Skewed and (b) Negative Skewed Data*

Kurtosis

- Kurtosis is used to find the **presence of outliers** in our data. It gives us the total degree of outliers present.
- Kurtosis is the measure of whether the data is heavy tailed or light tailed relative to normal distribution.
- It can be observed that normal distribution has bell-shaped curve with no long tails. Low kurtosis tends to have light tails. The implication is that there is no outlier data.
- Let x_1, x_2, …, x_N be a set of 'N' values or observations. Then, kurtosis is measured using the formula given below:

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^4 / N}{\sigma^4}$$

Thank You
End of Presentation

ATME College of Engineering

Machine Learning – BCS602

ASHWINI p.,
Assistant Professor
Dept. of Computer Science & Engineering
ATMECE, Mysuru

Course outcomes

At the end of the course, the student will be able to:

- **Describe** the machine learning techniques, their types and data analysis framework.
- **Apply** mathematical concepts for feature engineering and perform dimensionality reduction to enhance model performance.
- **Develop** similarity-based learning models and regression models for solving classification and prediction tasks.
- **Build** probabilistic learning models and design neural network models using perceptrons and multilayer architectures
- **Utilize** clustering algorithms to identify patterns in data and implement reinforcement learning techniques.

Overview of syllabus

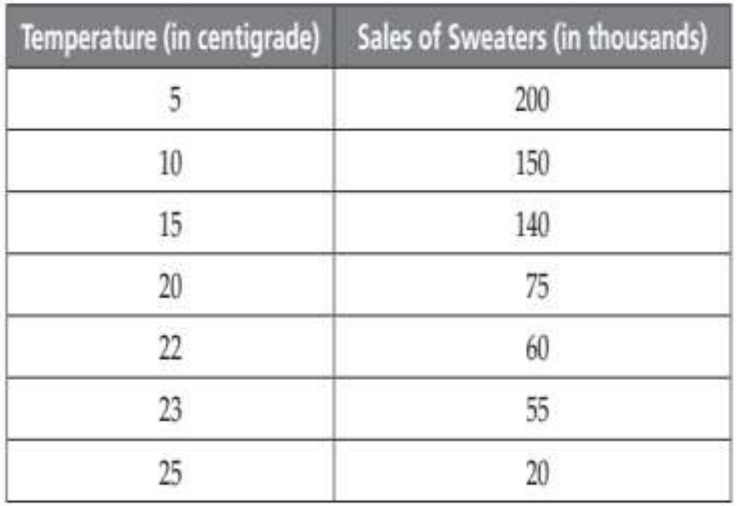
Module-2
Understanding Data – 2: Bivariate Data and Multivariate Data, Multivariate Statistics, Essential Mathematics for Multivariate Data, Feature Engineering and Dimensionality Reduction Techniques. **Basic Learning Theory:** Design of Learning System, Introduction to Concept of Learning, Modelling in Machine Learning. **Chapter-2 (2.6-2.8, 2.10), Chapter-3 (3.3, 3.4, 3.6)**

MODULE-2

UNDERSTANDING DATA

BIVARIATE DATA AND MULTIVARIATE DATA

- Bivariate Data involves **two variables**.
- Bivariate data deals with causes of relationships.
- The aim is to find **relationships among data**.
- Consider the following Table 2.6, with data of the temperature in a shop and sales of sweaters.
- Figure 2.23 and 2.24 shows scatter plot and line chart for the Table 2.6



Temperature (in centigrade)	Sales of Sweaters (in thousands)
5	200
10	150
15	140
20	75
22	60
23	55
25	20

Table 2.6: *Temperature in a Shop and Sales Data*

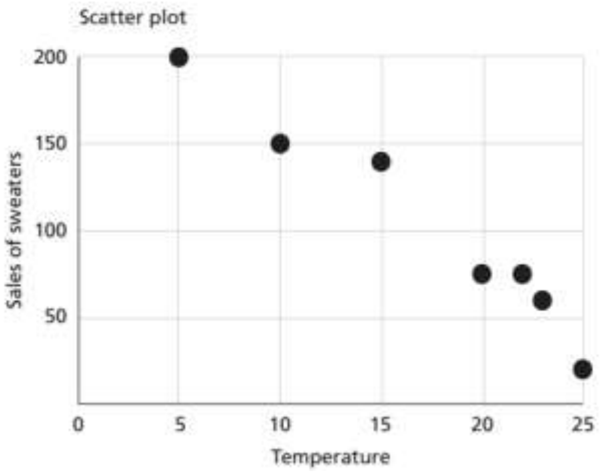


Figure 2.23: *Scatter Plot*

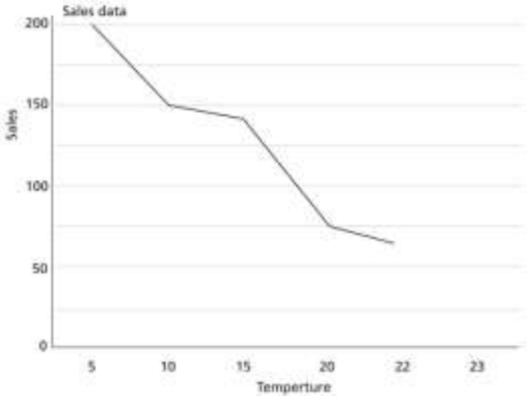


Figure 2.24: *Line Chart*

BIVARIATE DATA AND MULTIVARIATE DATA

1. **Bivariate Statistics:** *Covariance* and *Correlation* are examples of bivariate statistics.

Covariance

- Covariance is an indicator of the extent to which **2 random variables are dependent** on each other.
- Covariance implies whether the **two variables are directly or inversely proportional.**
- A higher number denotes higher dependency.
- Correlation is a statistical measure that indicates **how strongly two variables are related.**
- The value of covariance lies in the range of -∞ and +∞.

Covariance

- It is a **measure of joint probability of random variables**, say X and Y.
- Generally, random variables are represented in capital letters
- It is defined as covariance(X, Y) or COV(X, Y) and is used to measure the variance between two dimensions.
- The formula for finding co-variance for specific x, and y are:

$$\operatorname{cov}(X,Y)=\frac{1}{N}\sum_{i=1}^{N}(x_i-E(X))(y_i-E(Y))$$

- Here, ***xi* and *yi*** are **data values** from X and Y. **E(X) and E(Y)** are the **mean values** of x_i and y_i. **N** is the **number of given data**.
- Also, the COV(X, Y) is same as COV(Y, X).

Example 2.6: Find the covariance of data X = {1, 2, 3, 4, 5} and Y = {1, 4, 9, 16, 25}.

Solution: Mean(X) = E(X) = 15/5 = 3,

Mean(Y) = E(Y) = 55/5 = 11.

The covariance is computed using COV(X, Y) as:

$$\frac{(1-3)(1-11)+(2-3)(4-11)+(3-30)(9-11)+(4-3)(16-11)+(5-3)(25-11)}{5} = 12$$

The covariance between X and Y is 12.

It can be normalized to a value between -1 and +1.

This is done by dividing it by the correlation of variables.

Correlation

- Correlation is a statistical concept **determining the relationship** potency of two numerical variables.
- While deducing the relation between variables, we conclude the change in **one variable that impacts a difference in another**.
- The correlation indicates the **relationship between dimensions** using its sign.
- The sign is more important than the actual value.

Correlation

1. If the value is positive, it indicates that the **dimensions increase together.**
2. If the value **is negative**, it indicates that **while one-dimension increases, the other dimension decreases.**
3. If the value is **zero**, then it indicates that **both the dimensions are independent of each other.**
4. If the **dimensions are correlated**, then it is better to **remove one dimension as it is a redundant dimension.**
5. If the given attributes are X = (x1, x2, …, xN) and Y = (y1, y2, …, yN), then the Pearson correlation coefficient, that is denoted as r, is given as:

$$r = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

where, σ_X, σ_Y are the standard deviations of X and Y.

Example 2.7: Find the correlation coefficient of data X = {1, 2, 3, 4, 5} and Y = {1, 4, 9, 16, 25}.

Solution:

The mean values of X and Y are 15/5 = 3 and 55/5 = 11.

The standard deviations of X and Y are 1.41 and 8.6486, respectively.

Therefore, the correlation coefficient is given as ratio of covariance (12 from the previous problem 2.6) standard deviation of x and y as per the above equation as

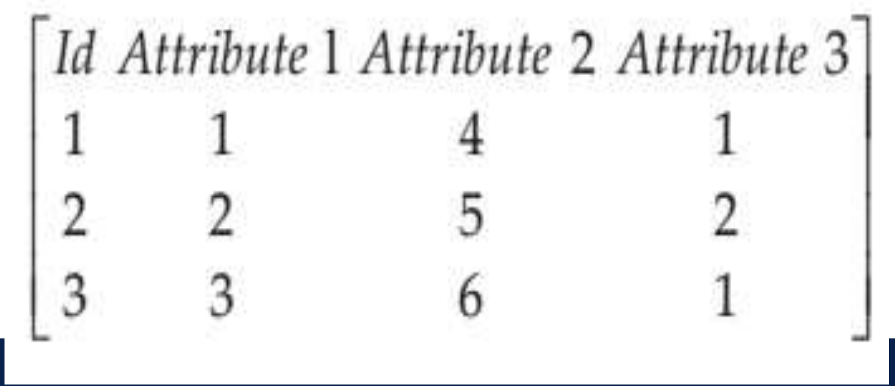
$$r = \frac{12}{1.41 \times 8.6486} \approx 0.984$$

2. Multivariate Statistics:

- In machine learning, almost all datasets are multivariable.
- Multivariate data is the analysis of more than two observable variables, and often, thousands of multiple measurements need to be conducted for one or more subjects.
- The multivariate data is like bivariate data but may have more than two dependant variables.
- Some of the multivariate analysis are regression analysis, principal component analysis, and path analysis.

BIVARIATE DATA AND MULTIVARIATE DATA

- The mean of multivariate data is a **mean vector** and the mean of the shown three attributes is given as (2, 7.5, 1.33).
- The variance of multivariate data becomes the **covariance matrix**.
- The mean vector is called **centroid** and variance is called **dispersion matrix** *(Will be discussed later)*.
- Multivariate data has three or more variables.



$$\begin{bmatrix} Id & Attribute\ 1 & Attribute\ 2 & Attribute\ 3 \\ 1 & 1 & 4 & 1 \\ 2 & 2 & 5 & 2 \\ 3 & 3 & 6 & 1 \end{bmatrix}$$

BIVARIATE DATA AND MULTIVARIATE DATA

Heatmap

- Heatmap is a graphical representation of 2D matrix.
- It takes a matrix as input and colours it. The darker colours indicate very large values and lighter colours indicate smaller values.
- The advantage of this method is that humans perceive colours well. So, by colour shaping, larger values can be perceived well.
- For example, in vehicle traffic data, heavy traffic regions can be differentiated from low traffic regions through heatmap.
- In Figure 2.25, patient data highlighting weight and health status is plotted. Here, X-axis is weights and Y-axis is patient counts. The dark colour regions highlight patients' weights vs patient counts in health status.

BIVARIATE DATA AND MULTIVARIATE DATA

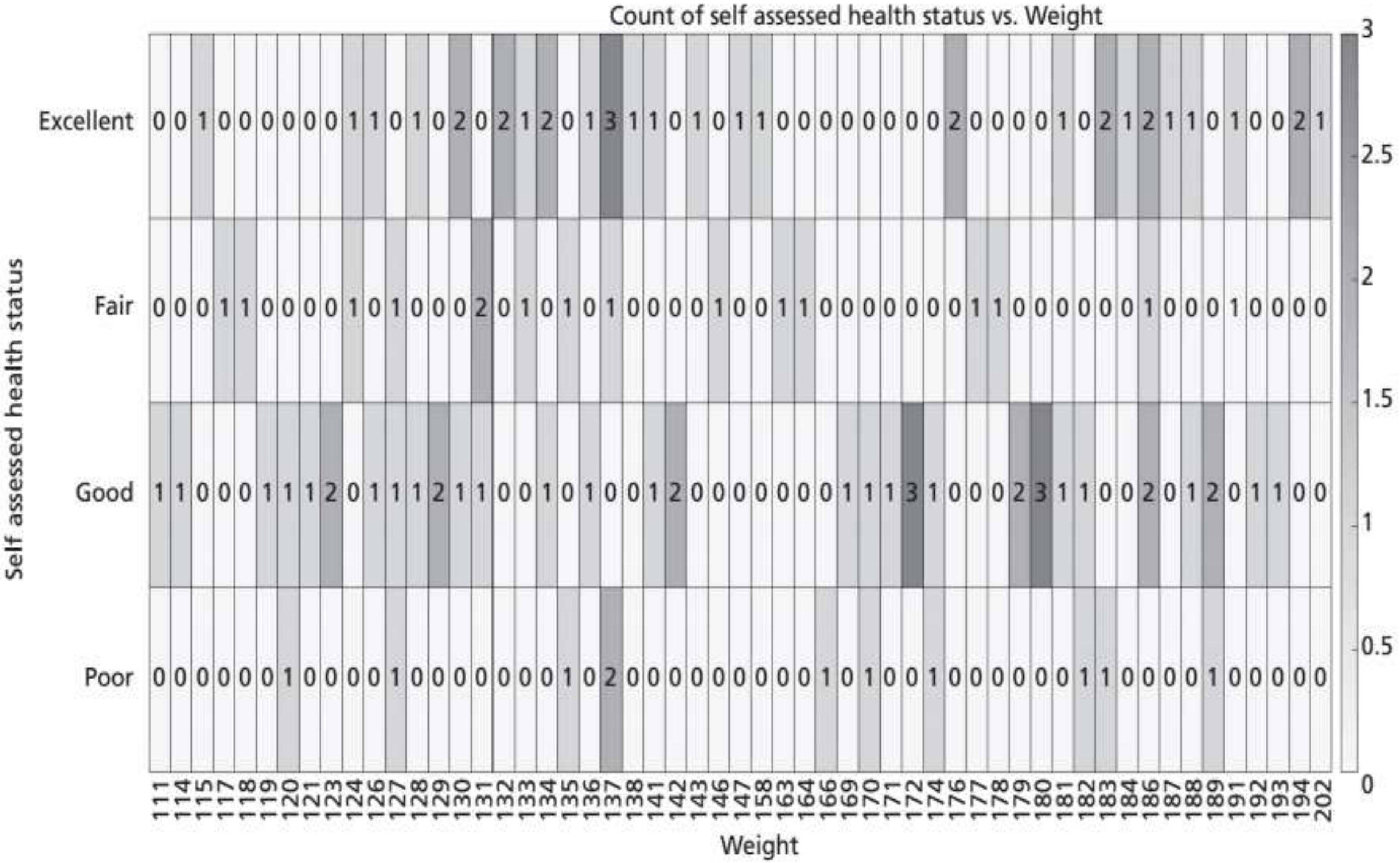


Figure 2.25: *Heatmap for Patient Data*

BIVARIATE DATA AND MULTIVARIATE DATA

Pairplot

- Pairplot or scatter matrix is a data visual technique for multivariate data. A scatter matrix consists of several pair-wise scatter plots of variables of the multivariate data. All the results are presented in a matrix format.
- By visual examination of the chart, one can easily find relationships among the variables such as correlation between the variables.
- A random matrix of three columns is chosen and the relationships of the columns is plotted as a pairplot (or scattermatrix) as shown below in Figure 2.26.

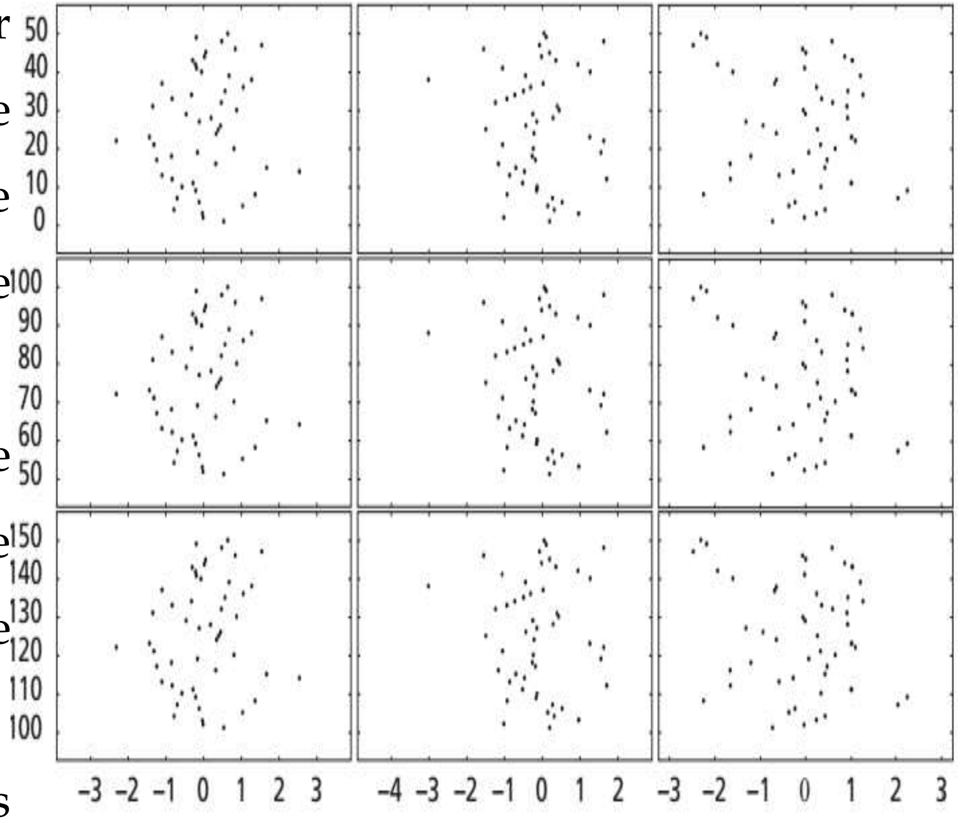


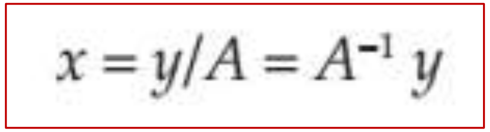
***Figure 2.26:** Pairplot for Random Data*

ESSENTIAL MATHEMATICS FOR MULTIVARIATE DATA

- Machine learning involves many mathematical concepts from the domain of ***Linear algebra, Statistics, Probability and Information theory***.
- Here we discuss important aspects of linear algebra and probability.
- **'Linear Algebra'** is a branch of mathematics that is central for many scientific applications and other mathematical subjects.
- Linear algebra deals with *linear equations, vectors, matrices, vector spaces and transformations*. These are the driving forces of machine learning and machine learning cannot exist without these data types.

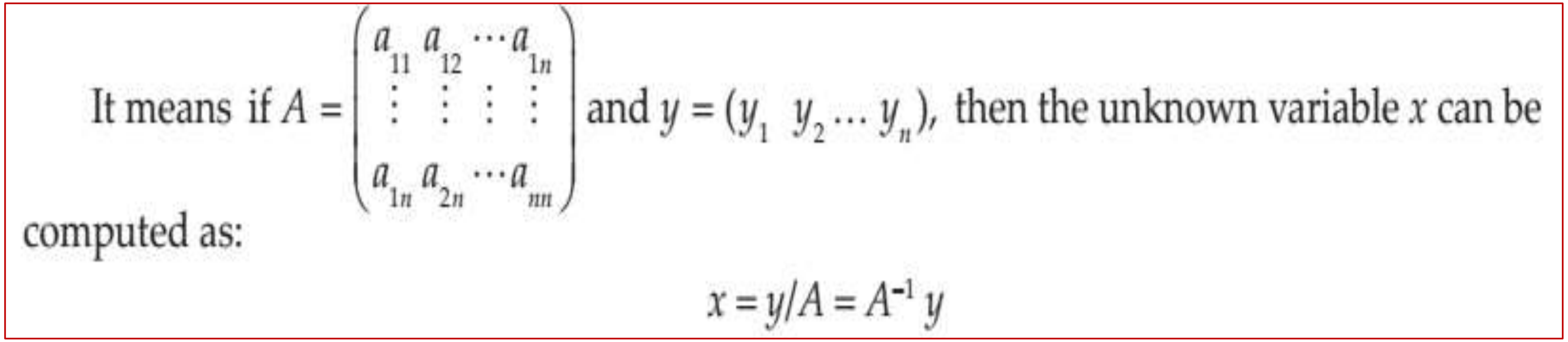
1. Linear Systems and Gaussian Elimination for Multivariate Data:

- A linear system of equations is a group of equations with unknown variables.
- Let $Ax = y$, then the solution x is given as:



$$x = y/A = A^{-1}\,y$$

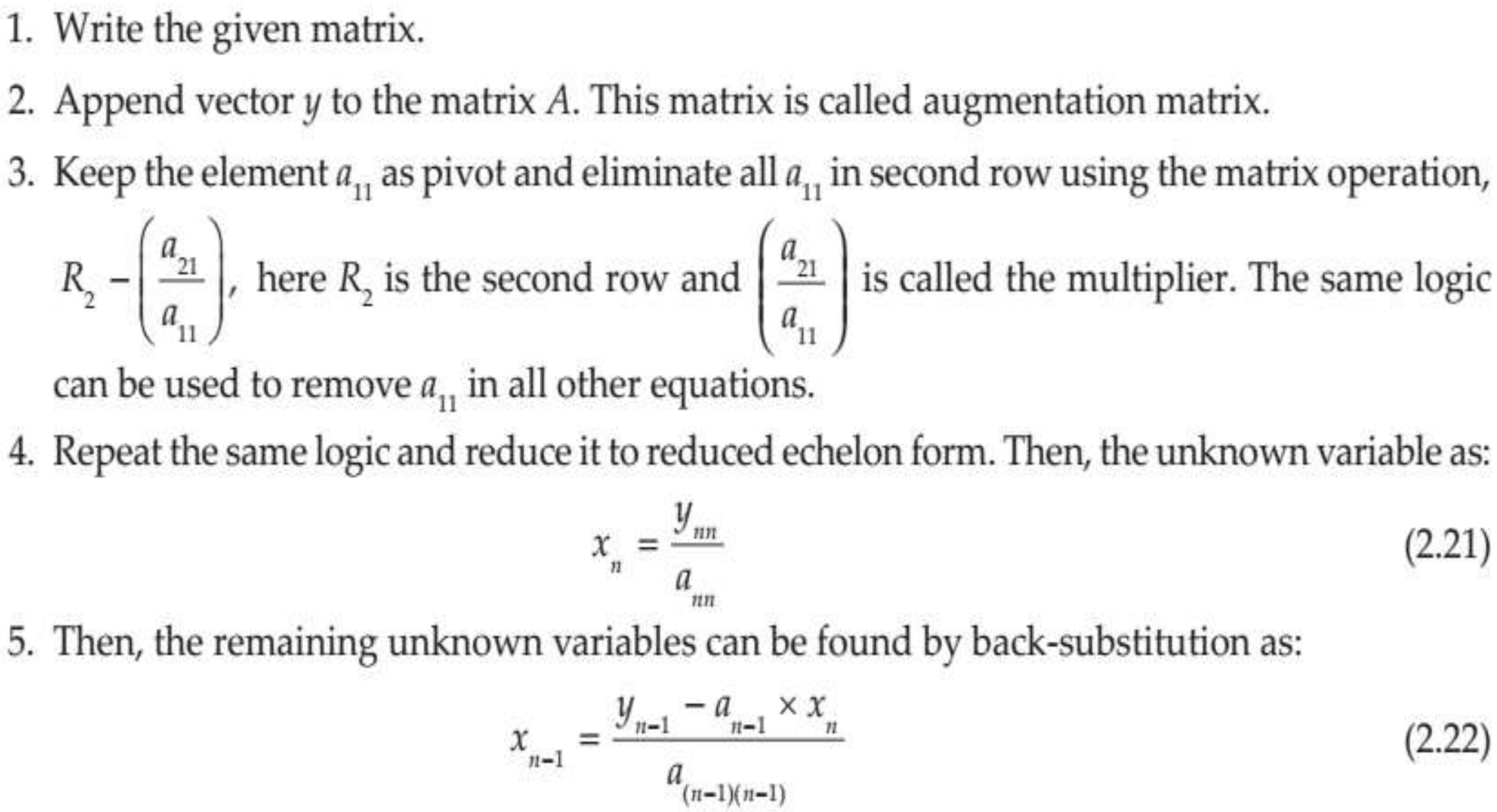
- This is true if y is not zero and A is not zero. The logic can be extended for N-set of equations with 'n' unknown variables.



It means if $A = \begin{pmatrix} a_{11} & a_{12} & \cdots a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1n} & a_{2n} & \cdots a_{nn} \end{pmatrix}$ and $y = (y_1 \;\; y_2 \ldots y_n)$, then the unknown variable x can be computed as:

$$x = y/A = A^{-1}\,y$$

- If there is a unique solution, then the system is called **consistent independent**.
- If there are various solutions, then the system is called **consistent dependant**.
- If there are no solutions and if the equations are contradictory, then the system is called **inconsistent**.
- For solving large number of system of equations, ***Gaussian elimination*** can be used. The procedure for applying Gaussian elimination is given as follows:



1. Write the given matrix.
2. Append vector y to the matrix A. This matrix is called augmentation matrix.
3. Keep the element a_{11} as pivot and eliminate all a_{11} in second row using the matrix operation, $R_2 - \left(\frac{a_{21}}{a_{11}}\right)$, here R_2 is the second row and $\left(\frac{a_{21}}{a_{11}}\right)$ is called the multiplier. The same logic can be used to remove a_{11} in all other equations.
4. Repeat the same logic and reduce it to reduced echelon form. Then, the unknown variable as:

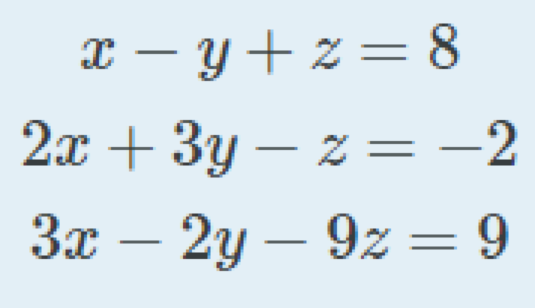
$$x_n = \frac{y_{nn}}{a_{nn}} \tag{2.21}$$

5. Then, the remaining unknown variables can be found by back-substitution as:

$$x_{n-1} = \frac{y_{n-1} - a_{n-1} \times x_n}{a_{(n-1)(n-1)}} \tag{2.22}$$

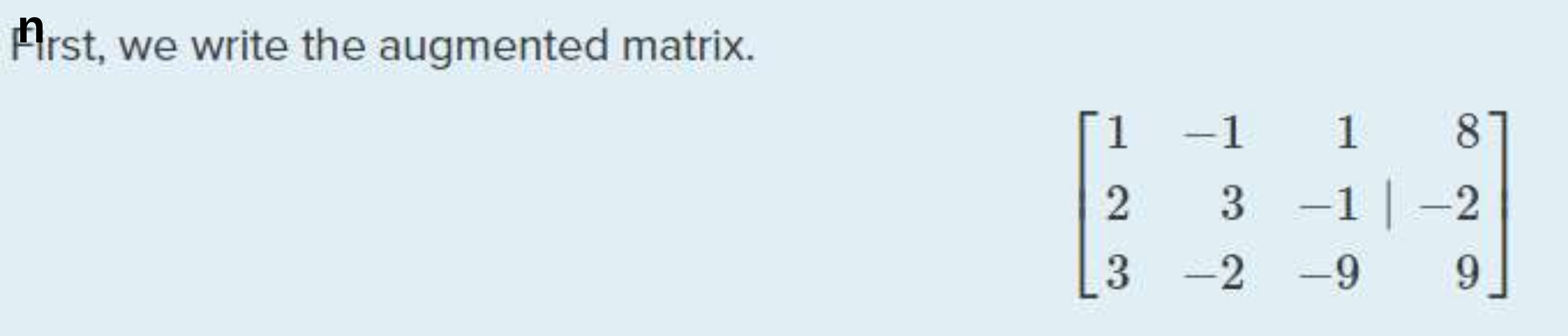
- To facilitate the application of Gaussian elimination method, the following row operations are applied:
  1. Swapping the rows
  2. Multiplying or dividing a row by a constant
  3. Replacing a row by adding or subtracting a multiple of another row to it These concepts are illustrated in Example 2.8.

Example 2.8: Solve the following set of equations using Gaussian Elimination method.



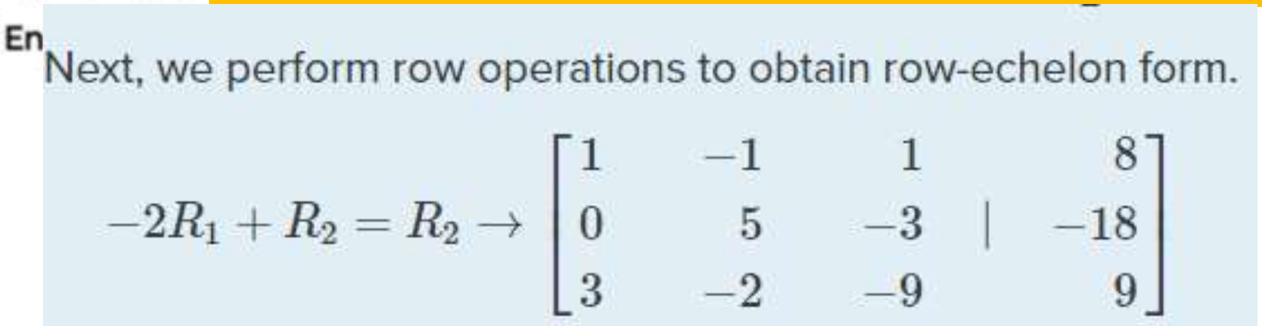
$$\begin{aligned} x - y + z &= 8 \\ 2x + 3y - z &= -2 \\ 3x - 2y - 9z &= 9 \end{aligned}$$

Solution



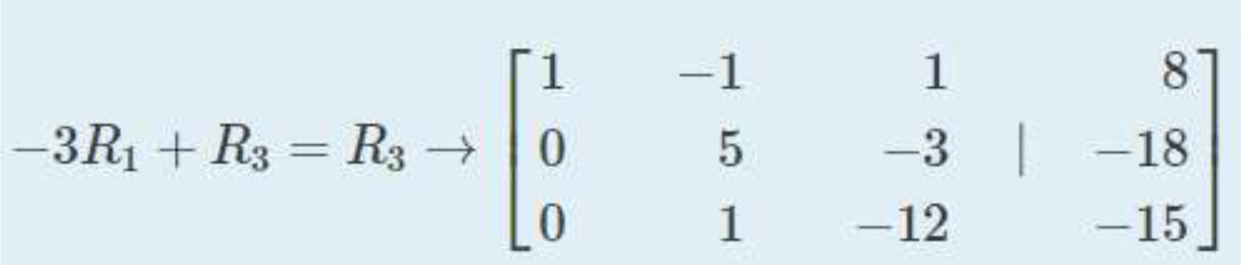
First, we write the augmented matrix.

$$\left[\begin{array}{ccc|c} 1 & -1 & 1 & 8 \\ 2 & 3 & -1 & -2 \\ 3 & -2 & -9 & 9 \end{array}\right]$$

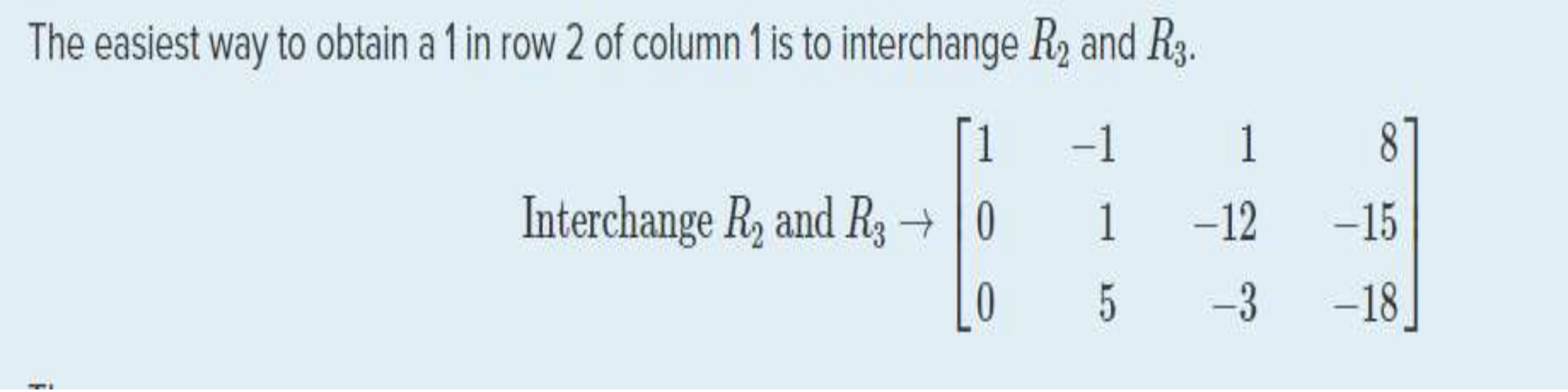


Next, we perform row operations to obtain row-echelon form.

$$-2R_1 + R_2 = R_2 \rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 1 & 8 \\ 0 & 5 & -3 & -18 \\ 3 & -2 & -9 & 9 \end{array}\right]$$

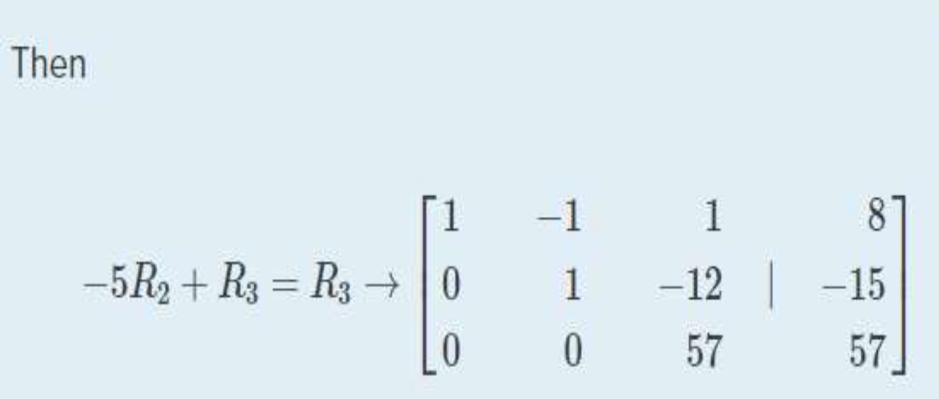


$$-3R_1 + R_3 = R_3 \rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 1 & 8 \\ 0 & 5 & -3 & -18 \\ 0 & 1 & -12 & -15 \end{array}\right]$$



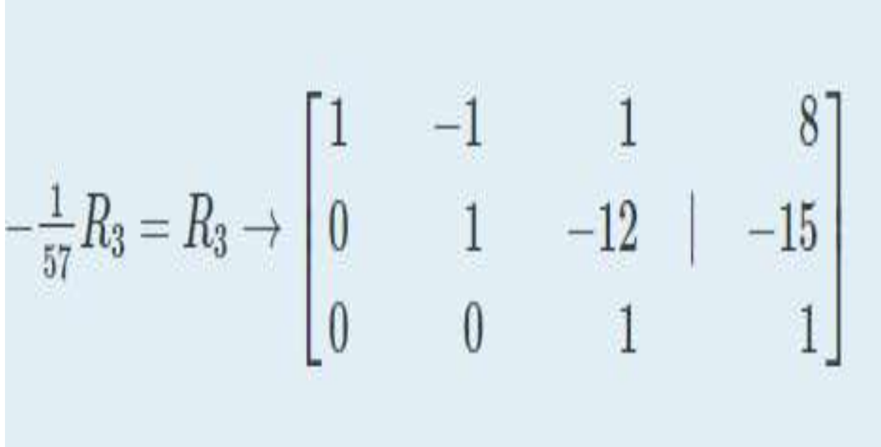
The easiest way to obtain a 1 in row 2 of column 1 is to interchange R_2 and R_3.

$$\text{Interchange } R_2 \text{ and } R_3 \rightarrow \begin{bmatrix} 1 & -1 & 1 & 8 \\ 0 & 1 & -12 & -15 \\ 0 & 5 & -3 & -18 \end{bmatrix}$$

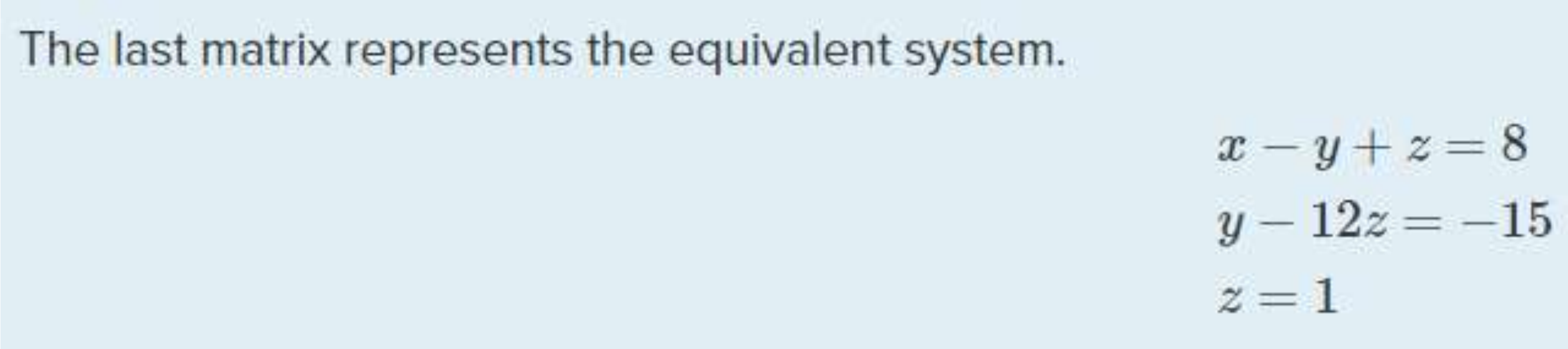


Then

$$-5R_2 + R_3 = R_3 \rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 1 & 8 \\ 0 & 1 & -12 & -15 \\ 0 & 0 & 57 & 57 \end{array}\right]$$

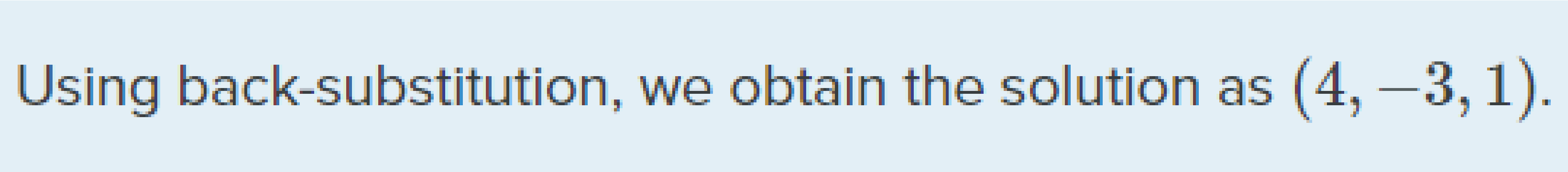


$$-\frac{1}{57}R_3 = R_3 \rightarrow \left[\begin{array}{ccc|c} 1 & -1 & 1 & 8 \\ 0 & 1 & -12 & -15 \\ 0 & 0 & 1 & 1 \end{array}\right]$$



The last matrix represents the equivalent system.

$$\begin{aligned} x - y + z &= 8 \\ y - 12z &= -15 \\ z &= 1 \end{aligned}$$



Using back-substitution, we obtain the solution as $(4, -3, 1)$.

Row Echelon Form

A matrix is in Row Echelon form if it has the following properties:

- Any row consisting entirely of zeros occurs at the bottom of the matrix.
- For each row that does not contain entirely zeros, the first non-zero entry is 1 (called a leading 1).
- For two successive (non-zero) rows, the leading 1 in the higher row is further left than the leading one in the lower row.

For reduced row echelon form, the leading 1 of every row contains 0 below and above its in that column. Below is an example of row-echelon form:

$$\begin{bmatrix} 1 & 2 & -1 & 4 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

BACK

2. Matrix Decompositions:

- It is often necessary to reduce a matrix to its constituent parts so that complex matrix operations can be performed. These methods are also known as **matrix factorization methods**.
- The most popular matrix decomposition is called **eigen decomposition**. It is a way of reducing the matrix into eigen values and eigen vectors. Then, the matrix A can be decomposed as:

$$\boxed{A = Q\Lambda Q^T}$$

where, Q is the matrix of eigen vectors, Λ is the diagonal matrix and Q^T is the transpose of matrix Q.

LU Decomposition

- One of the simplest matrix decompositions is **LU decomposition** where the matrix A can be decomposed matrices:

$$A = LU$$

  - Here, L is the lower triangular matrix and U is the upper triangular matrix. The decomposition can be done using ***Gaussian elimination method***.
  - First, an identity matrix is augmented to the given matrix. Then, row operations and Gaussian elimination is applied to reduce the given matrix to get matrices L and U.
  - Example 2.9 illustrates the application of Gaussian elimination to get LU.

Example 2.9: Find LU decomposition of the given matrix: $A = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 3 & 2 \\ 3 & 4 & 2 \end{pmatrix}$

Solution: First, augment an identity matrix and apply Gaussian elimination. The steps are as shown in:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 3 & 3 & 2 \\ 3 & 4 & 2 \end{bmatrix} \qquad \boxed{\text{Initial Matrix}}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 3 & 4 & 2 \end{bmatrix} \qquad \boxed{R_2 = R_2 - 3R_1}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & -2 & -10 \end{bmatrix} \qquad \boxed{R_3 = R_3 - 3R_1}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & \frac{2}{3} & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & 0 & \frac{-10}{3} \end{bmatrix} \qquad \boxed{R_3 = R_3 - \frac{2}{3}R_2}$$

Now, it can be observed that the first matrix is *L* as it is the lower triangular matrix whose values are the determiners used in the reduction of equations above such as 3, 3 and 2/3. The second matrix is *U*, the upper triangular matrix whose values are the values of the reduced matrix because of Gaussian elimination.

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & \frac{2}{3} & 1 \end{pmatrix} \text{ and } U = \begin{pmatrix} 1 & 2 & 4 \\ 0 & -3 & -10 \\ 0 & 0 & -\frac{10}{3} \end{pmatrix}.$$

It can be cross verified that the multiplication of LU yields the original matrix A.

3. Machine Learning and Importance of Probability and Statistics:

- Machine learning is linked with statistics and probability.
- Like linear algebra, **statistics** is the heart of machine learning.
- The importance of statistics needs to be stressed as without statistics; analysis of data is difficult.
- **Probability** is especially important for machine learning.
- Any data can be assumed to be generated by a probability distribution.

3.1 Probability Distributions

- A probability distribution of a variable, say *X*, summarizes the probability associated with X’s events.
- Distribution is a function that describes the relationship between the observations in a sample space.
- Probability distributions are of two types:

  1. Discrete probability distribution
  2. Continuous probability distribution

- The relationships between the events for a continuous random variable and their probabilities is called a **continuous probability distribution**. It is summarized as **Probability Density Function (PDF)**. The plot of PDF shows the shape of the distribution.
- **Cumulative Distributive Function (CDF)** computes the probability of an observation ≤ value. Both PDF and CDF are continuous values.
- The discrete equivalent of PDF in discrete distribution is called **Probability Mass Function (PMF)**.

3.1.1 Continuous Probability Distributions

- Continuous Probability Distributions categories are Normal, Rectangular, and Exponential distributions.

Normal Distribution

- Normal distribution is a continuous probability distribution. This is also known as **gaussian distribution** or **bell-shaped curve distribution**.
- The shape of this distribution is a typical bell-shaped curve.
- The heights of the students, blood pressure of a population, and marks scored in a class can be approximated using normal distribution.
- PDF of the normal distribution is given as:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e$$

Here, μ is mean and σ is the standard deviation. Normal distribution is characterized by two parameters – mean and variance.

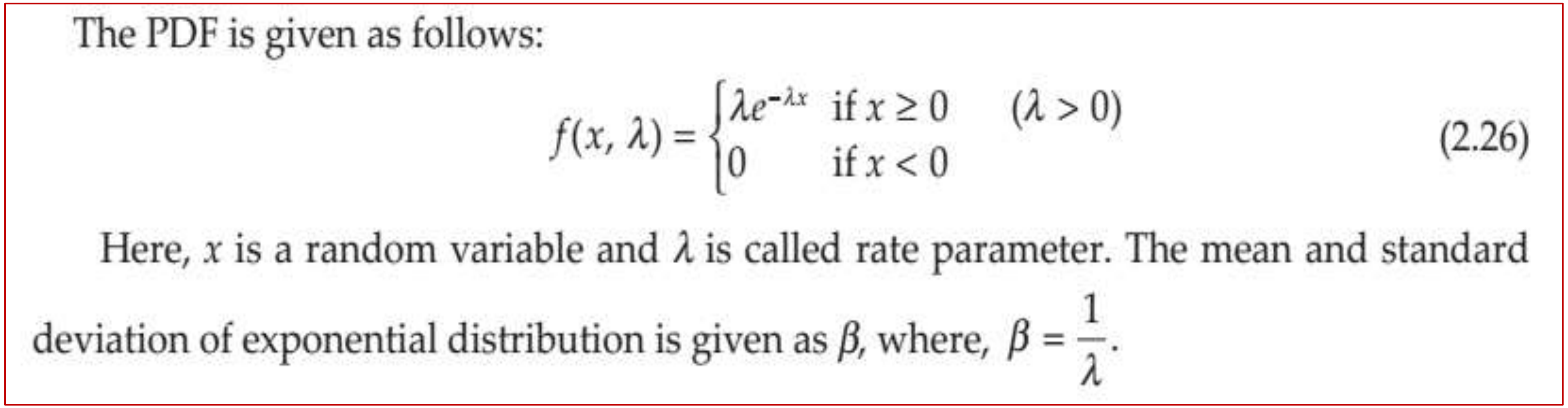
Rectangular Distribution

- This is also known as uniform distribution. It has equal probabilities for all values in the range *a*, *b*.
- The uniform distribution is given as follows:

$$P(X = x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{Otherwise} \end{cases}$$

Exponential Distribution

- This is a continuous uniform distribution.
- This probability distribution is used to describe the time between events in a Poisson process.
- Exponential distribution is another special case of Gamma distribution with a fixed parameter of 1.
- This distribution is helpful in modelling of time until an event occurs.



The PDF is given as follows:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \quad (\lambda > 0) \\ 0 & \text{if } x < 0 \end{cases} \tag{2.26}$$

Here, x is a random variable and λ is called rate parameter. The mean and standard deviation of exponential distribution is given as β, where, $\beta = \frac{1}{\lambda}$.

3.1.2 Discrete Probability Distributions

- Binomial, Poisson, and Bernoulli distributions fall under this category.

Binomial Distribution

- Binomial distribution is another distribution that is often encountered in machine learning. It has only two outcomes: success or failure. This is also called **Bernoulli trial**.
- The objective of this distribution is to find probability of getting success k out of n trials. The way to get success out of k out of n number of trials is given as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- The binomial distribution function is given as follows, where *p* is the probability of success and probability of failure is (1 - *p*). The probability of success in a certain number of trials is given as:

$$\boxed{p^k(1-p)^{n-k} \text{ or } p^k\, q^{n-k}}$$

- Combining both, one gets PDF of binomial distribution as:

$$\boxed{\binom{n}{k} p^k(1-p)^{n-k}}$$

- Here, *p* is the probability of each choice, *k* is the number of choices, and *n* is the total number of choices. The mean of binomial distribution is given below: $\boxed{\mu = n \times p}$

- And the variance is given as: $\boxed{\sigma^2 = np(1-p)}$

- Hence, the standard deviation is given as: $\boxed{\sigma = \sqrt{np(1-p)}}$

Poisson Distribution

- It is another important distribution that is quite useful.
- Given an interval of time, this distribution is used to model the probability of a given number of events k.
- The mean rule λ is inclusive of previous events.
- Some of the examples of Poisson distribution are number of emails received, number of customers visiting a shop and the number of phone calls received by the office.
- The PDF of Poisson distribution is given as follows:

3.2.1 Parametric Density Estimation

- It assumes that the data is from a known probabilistic distribution and can be estimated as $p(x \mid \Theta)$, where, Θ is the parameter.
- *Maximum likelihood function* is a parametric estimation method.

Maximum Likelihood Estimation

- For a sample of observations, one can estimate the probability distribution. This is called **density estimation**. Maximum Likelihood Estimation (MLE) is a probabilistic framework that can be used for density estimation.
- This involves formulating a function called **likelihood function** which is the conditional probability of observing the observed samples and distribution function with its parameters.

- For example, consider a joint probability $p(X; \theta)$, where, $X = \{x1, x2, \ldots, xn\}$
- The likelihood of observing the data is given as a function $L(X; \theta)$.
- The objective of MLE is to maximize this function as *max* $L(X; \theta)$.
- The joint probability of this problem can be restated as: $\boxed{\prod_{i=1}^{n} p(x_i;\theta).}$
- The computation of the above formula is unstable and the hence the problem is restated as maximum of log conditional probability given θ. This is given as:

$$\boxed{\sum_{i=1}^{n} \log p(x_i;\theta)}$$

- Instead of maximizing, one can minimize this function as: $\boxed{min = -\sum_{i=1}^{n} \log p(x_i;\theta)}$

FEATURE ENGINEERING AND DIMENSIONALITY REDUCTION TECHNIQUES

- Features are attributes.
- Feature engineering is about determining the subset of features that form an important part of the input that improves the performance of the model, be it classification or any other model in machine learning.
- Feature engineering deals with two problems – **Feature Transformation** and **Feature Selection**.
  - Feature transformation is extraction of features and creating new features that may be helpful in increasing performance. For example, the height and weight may give a new attribute called Body Mass Index (BMI).
  - Feature subset selection is another important aspect of feature engineering that focuses on selection of features to reduce the time but not at the cost of reliability.
- The subset selection reduces the dataset size by removing irrelevant features and constructs a minimum set of attributes for machine learning.

- The features can be removed based on two aspects:

1. Feature relevancy – Some features contribute more for classification than other features. For example, a mole on the face can help in face detection than common features like nose. In simple words, the features should be relevant. The relevancy of the features can be determined based on information measures such as mutual information, correlation based features like correlation coefficient and distance measures.

2. Feature redundancy – Some features are redundant. For example, when a database table has a field called Date of birth, then age field is not relevant as age can be computed easily from date of birth. This helps in removing the column age that leads to reduction of dimension one.

1. Stepwise Forward Selection

- This procedure starts with an empty set of attributes. Every time, an attribute is tested for statistical significance for best quality and is added to the reduced set. This process is continued till a good reduced set of attributes is obtained.

2. Stepwise Backward Elimination

- This procedure starts with a complete set of attributes. At every stage, the procedure removes the worst attribute from the set, leading to the reduced set.

3. Principal Component Analysis

- The idea of the principal component analysis (PCA) or KL transform is to transform a given set of measurements to a new set of features so that the features exhibit high information packing properties.
- This leads to a reduced and compact set of features.
- Consider a group of random vectors of the form:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- The mean vector of the set of random vectors is defined as:

$$m_x = E\{x\}$$

Example 2.12: Let the data points be $\binom{2}{6}$ and $\binom{1}{7}$. Apply PCA and find the transformed data. Again, apply the inverse and prove that PCA works.

Solution: One can combine two vectors into a matrix as follows:

The mean vector can be computed as Eq. (1) as follows:

$$\mu = \begin{pmatrix} \frac{2+1}{2} \\ \frac{6+7}{2} \end{pmatrix} = \begin{pmatrix} 1.5 \\ 6.5 \end{pmatrix}$$

As part of PCA, the mean must be subtracted from the data to get the adjusted data:

$$x_1 = \begin{pmatrix} 2-1.5 \\ 6-6.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$$

$$x_2 = \begin{pmatrix} 1-1.5 \\ 7-6.5 \end{pmatrix} = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}$$

One can find the covariance for these data vectors. The covariance can be obtained using Eq. (2):

$$m_1 = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix}$$

$$m_2 = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix} \begin{pmatrix} -0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix}$$

The final covariance matrix is obtained by adding these two matrices as:

$$C = \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

The eigen values and eigen vectors of matrix C can be obtained as $\lambda 1 = 1$, $\lambda 1 = 0$. The eigen vectors are $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The matrix A can be obtained by packing the eigen vector of these eigen values (after sorting it) of matrix C. For this problem, $A = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$. The transpose of A, $A^T = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$ is also the same matrix as it is an orthogonal matrix. The matrix can be normalized by diving each elements of the vector, by the norm of the vector to get:

$$A = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

One can check that the PCA matrix A is orthogonal. A matrix is orthogonal is $A^{-1} = A$ and $AA^{-1} = I$.

G HJU

$$AA^T = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The transformed matrix y using Eq. (3) is given as: $y = A \times (x - m)$

Recollect that (x-m) is the adjusted matrix.

$$y = A(x-m) = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left(\textit{for convenience } 0.5 = \frac{1}{2} \right)$$

$$= \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix}$$

One can check the original matrix can be retrieved from this matrix as:

$$\{(A)^T \times y\} + m$$

$$x = A^T y + m = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1.5 \\ 6.5 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 1.5 \\ 6.5 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 6 & 7 \end{pmatrix}$$

Therefore, one can infer the original is obtained without any loss of information.

4. Linear Discriminant Analysis

- Linear Discriminant Analysis (LDA) is also a feature reduction technique like PCA. The focus of LDA is to project higher dimension data to a line (lower dimension data).
- LDA is also used to classify the data. Let there be two classes, c1 and c2. Let m1 and m2 be the mean of the patterns of two classes. The mean of the class c1 and c2 can be computed as:

$$\mu_1 = \frac{1}{N_1}\sum_{x_i \in c_1}^{n} x_i \text{ and } \mu_2 = \frac{1}{N_2}\sum_{x_i \in c_2}^{n} x_i$$

- The aim of LDA is to optimize the function:

$$J(V) = \frac{V^T \sigma_B V}{V^T \sigma_W V}$$

where, V is the linear projection and σ_B and σ_W are class scatter matrix and within scatter matrix, respectively. For the two-class problem, these matrices are given as:

$$\sigma_B = N_1(\mu_1 - \mu)(\mu_1 - \mu)^T + N_2(\mu_2 - \mu)(\mu_2 - \mu)^T$$
$$\sigma_W = \sum_{x_i \in c_1}(x_i - \mu_1)(x_i - \mu_1)^T + \sum_{x_i \in c_2}(x_i - \mu_2)(x_i - \mu_2)^T$$

- The maximization of J(V) should satisfy the equation:

$$\sigma_B V = \lambda \sigma_W V \text{ or } \sigma_W^{-1}\sigma_B V = \lambda V$$

5. Singular Value Decomposition

- Singular Value Decomposition (SVD) is another useful decomposition technique. Let A be the matrix, then the matrix A can be decomposed as:

$$\boxed{A = USV^T}$$

Here, A is the given matrix of dimension $m \times n$, U is the orthogonal matrix whose dimension is $m \times n$, S is the diagonal matrix of dimension $n \times n$, and V is the orthogonal matrix.

- The procedure for finding decomposition matrix is given as follows:

1. For a given matrix, find AA^T

2. Find eigen values of AA^T

3. Sort the eigen values in a descending order. Pack the eigen vectors as a matrix U.

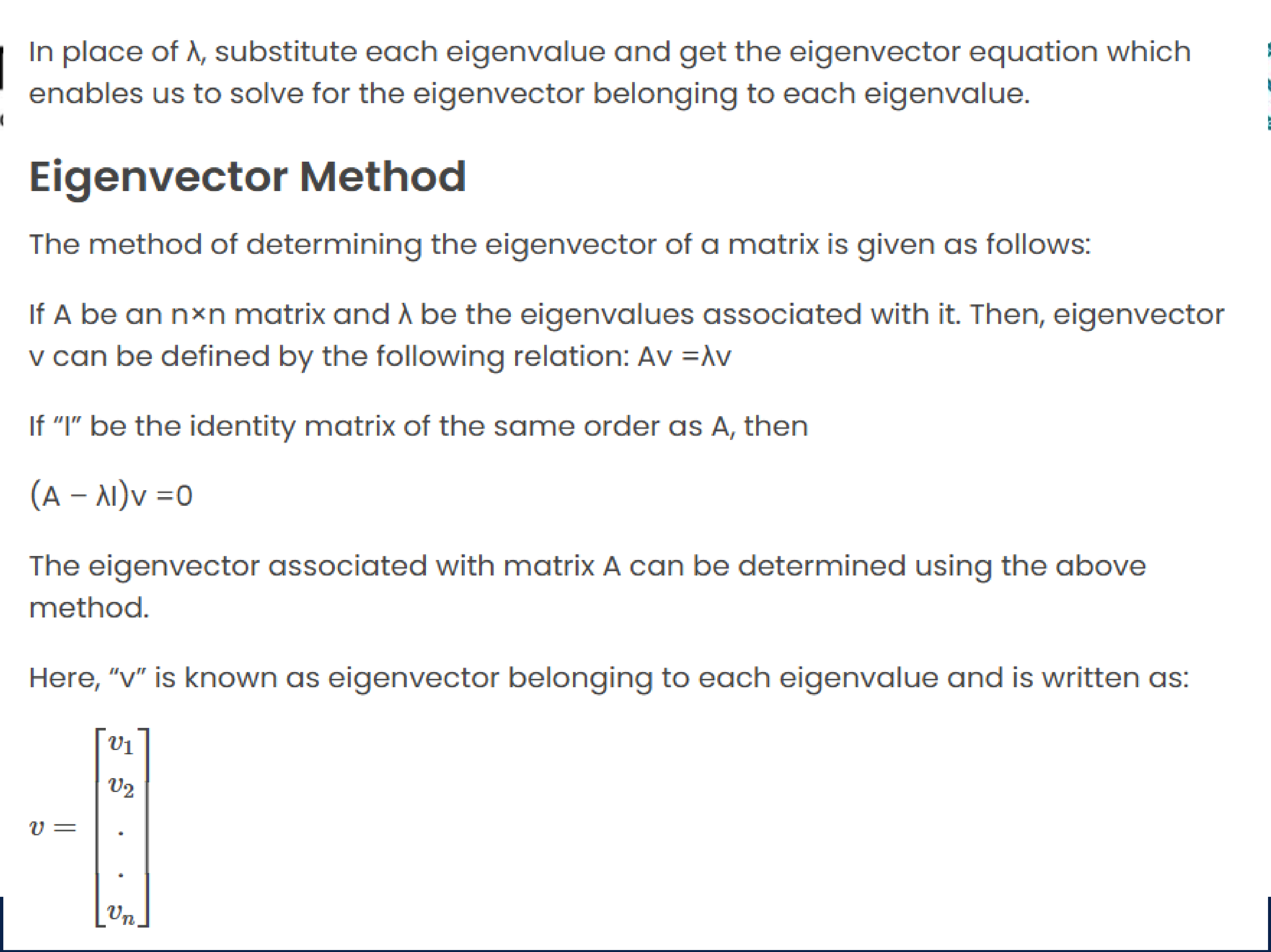
4. Arrange the square root of the eigen values in diagonal. This matrix is diagonal matrix, S.

5. Find eigen values and eigen vectors for $A^T A$. Find the eigen value and pack the eigen vector as a matrix called V.

Thus, $A = USV^T$. Here, U and V are orthogonal matrices. The columns of U and V are left and right singular values, respectively. SVD is useful in compression, as one can decide to retain only a certain component instead of the original matrix A as:

$$a_{ij} = \sum_{k=1}^{n} u_{ik} s_k v_{jk}$$

Based on the choice of retention, the compression can be controlled.



In place of λ, substitute each eigenvalue and get the eigenvector equation which enables us to solve for the eigenvector belonging to each eigenvalue.

Eigenvector Method

The method of determining the eigenvector of a matrix is given as follows:

If A be an n×n matrix and λ be the eigenvalues associated with it. Then, eigenvector v can be defined by the following relation: Av =λv

If "I" be the identity matrix of the same order as A, then

(A − λI)v =0

The eigenvector associated with matrix A can be determined using the above method.

Here, "v" is known as eigenvector belonging to each eigenvalue and is written as:

$$v = \begin{bmatrix} v_1 \\ v_2 \\ . \\ . \\ v_n \end{bmatrix}$$

Example 2.13: Find SVD of the matrix: $A = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix}$

Solution: The first step is to compute: $AA^T = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix}\begin{pmatrix} 1 & 4 \\ 2 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 22 \\ 22 & 97 \end{pmatrix}$

The eigen value and eigen vector of this matrix can be calculated to get U. The eigen values of this matrix are 0.0098 and 101.9902.

The eigen vectors of this matrix are:

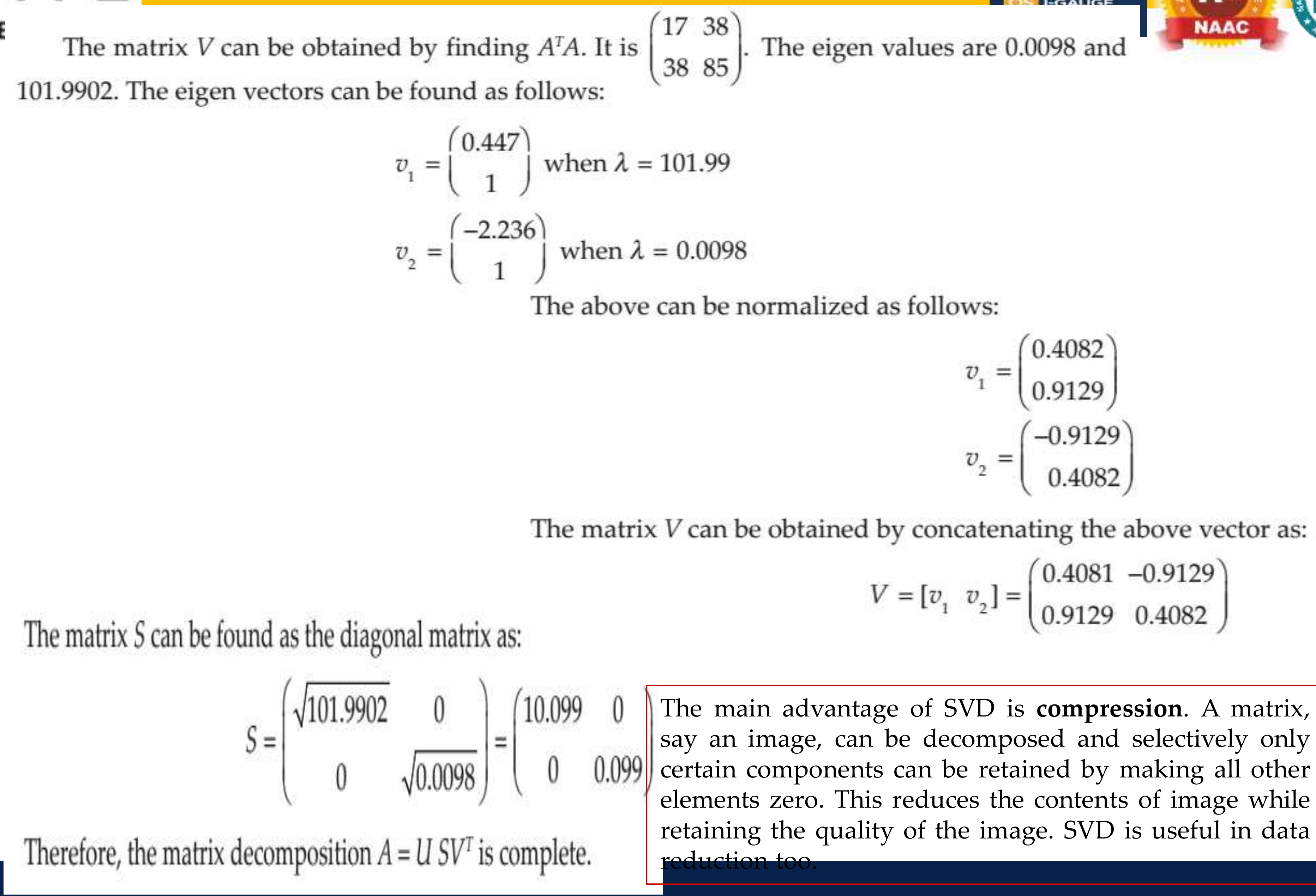
$$u_1 = \begin{pmatrix} 0.2268 \\ 1 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} -4.4086 \\ 1 \end{pmatrix}$$

These vectors are normalized to get the vectors respectively as: $u_1 = \begin{pmatrix} 0.2212 \\ 0.9752 \end{pmatrix}$ $u_2 = \begin{pmatrix} -0.9752 \\ 0.2212 \end{pmatrix}$

The matrix U can be obtained by concatenating the above vector as:

$$U = [u_1, u_2] = \begin{pmatrix} 0.2212 & -0.9752 \\ 0.9752 & 0.2212 \end{pmatrix}$$



The matrix V can be obtained by finding A^TA. It is $\begin{pmatrix} 17 & 38 \\ 38 & 85 \end{pmatrix}$. The eigen values are 0.0098 and 101.9902. The eigen vectors can be found as follows:

$$v_1 = \begin{pmatrix} 0.447 \\ 1 \end{pmatrix} \text{ when } \lambda = 101.99$$

$$v_2 = \begin{pmatrix} -2.236 \\ 1 \end{pmatrix} \text{ when } \lambda = 0.0098$$

The above can be normalized as follows:

$$v_1 = \begin{pmatrix} 0.4082 \\ 0.9129 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} -0.9129 \\ 0.4082 \end{pmatrix}$$

The matrix V can be obtained by concatenating the above vector as:

$$V = [v_1 \;\; v_2] = \begin{pmatrix} 0.4081 & -0.9129 \\ 0.9129 & 0.4082 \end{pmatrix}$$

The matrix S can be found as the diagonal matrix as:

$$S = \begin{pmatrix} \sqrt{101.9902} & 0 \\ 0 & \sqrt{0.0098} \end{pmatrix} = \begin{pmatrix} 10.099 & 0 \\ 0 & 0.099 \end{pmatrix}$$

Therefore, the matrix decomposition $A = U\,SV^T$ is complete.

The main advantage of SVD is **compression**. A matrix, say an image, can be decomposed and selectively only certain components can be retained by making all other elements zero. This reduces the contents of image while retaining the quality of the image. SVD is useful in data reduction too.

Chapter 2
Basics of Learning Theory

3.3 DESIGN OF A LEARNING SYSTEM

Training Experience

Let us consider designing of a chess game. In direct experience, individual board states and correct moves of the chess game are given directly.

In indirect system, the move sequences and results are only given.

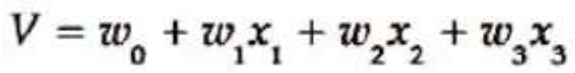
The training experience also depends on the presence of a supervisor who can label all valid moves for a board state.

In the absence of a supervisor, the game agent plays against itself and learns the good moves, if the training samples cover all scenarios, or in other words, distributed enough for performance computation. If the training samples and testing samples have the same distribution, the results would be good.

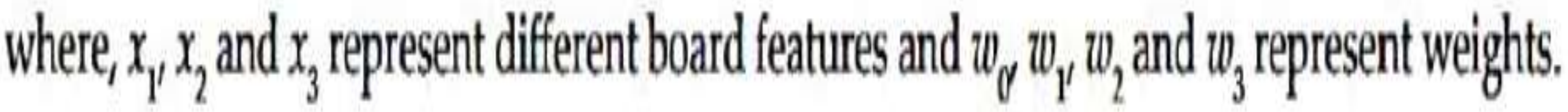
Determine the Target Function

The representation of knowledge may be a table, collection of rules or a neural network.

The linear combination of these factors can be coined as:



$$V = w_0 + w_1x_1 + w_2x_2 + w_3x_3$$



where, x_1, x_2 and x_3 represent different board features and w_0, w_1, w_2 and w_3 represent weights.

Choosing an Approximation Algorithm for the Target Function

$$E \equiv \sum_{\substack{\textit{Training}\\ \textit{Samples}}} \left[V_{train}(b) - \hat{V}(b)\right]^{2}$$

Here, b is the sample and $\hat{V}(b)$ the predicted hypothesis. The approximation is carried out as:

Computing the error as the difference between trained and expected hypothesis. Let error
be error(b).

Then, for every board feature x, the weights are updated as:

$$w_i = w_i + \mu \times error(b) \times x_i$$

Here, μ is the constant that moderates the size of the weight update.

Thus, the learning system has the following components:

A Performance system to allow the game to play against itself.

A Critic system to generate the samples.

A Generalizer system to generate a hypothesis based on samples.

An Experimenter system to generate a new system based on the currently learnt function.

This is sent as input to the performance system.

3.4 INTRODUCTION TO CONCEPT LEARNING

Formally, Concept learning is defined as—"Given a set of hypotheses, the learner searches through the hypothesis space to identify the best hypothesis that matches the target concept”.

Consider the following set of training instances shown in Table 3.1.

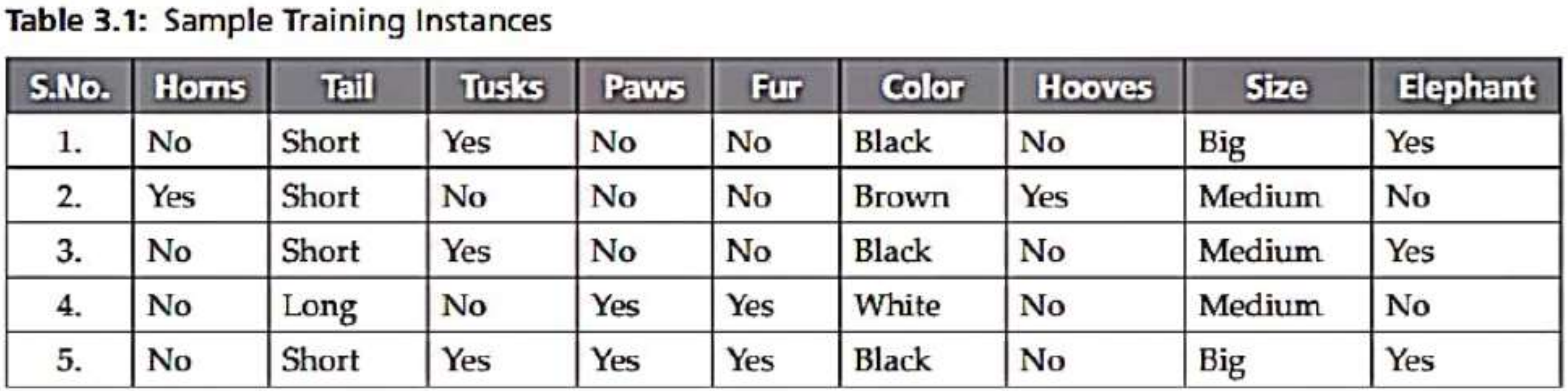
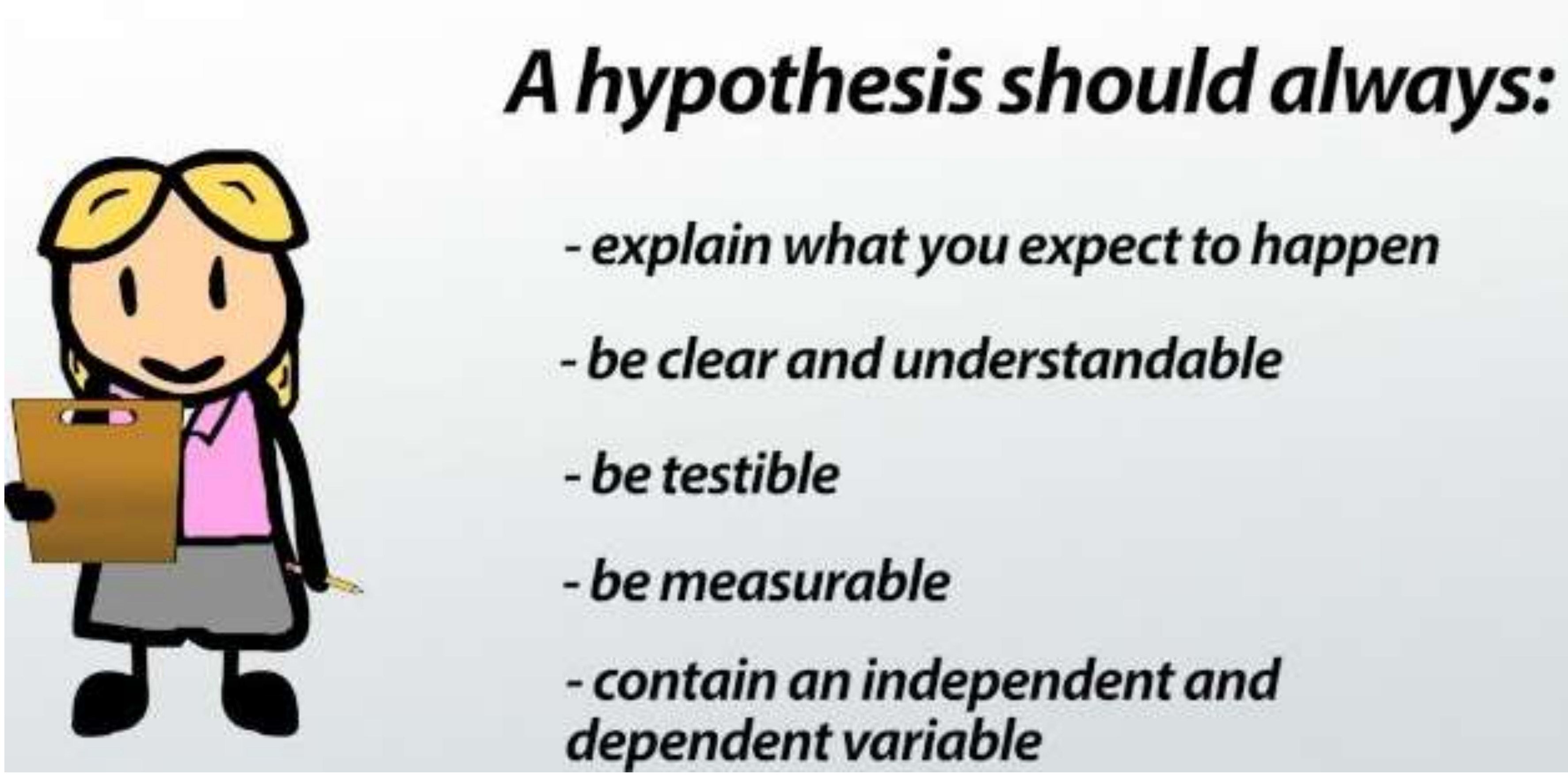


Table 3.1: Sample Training Instances

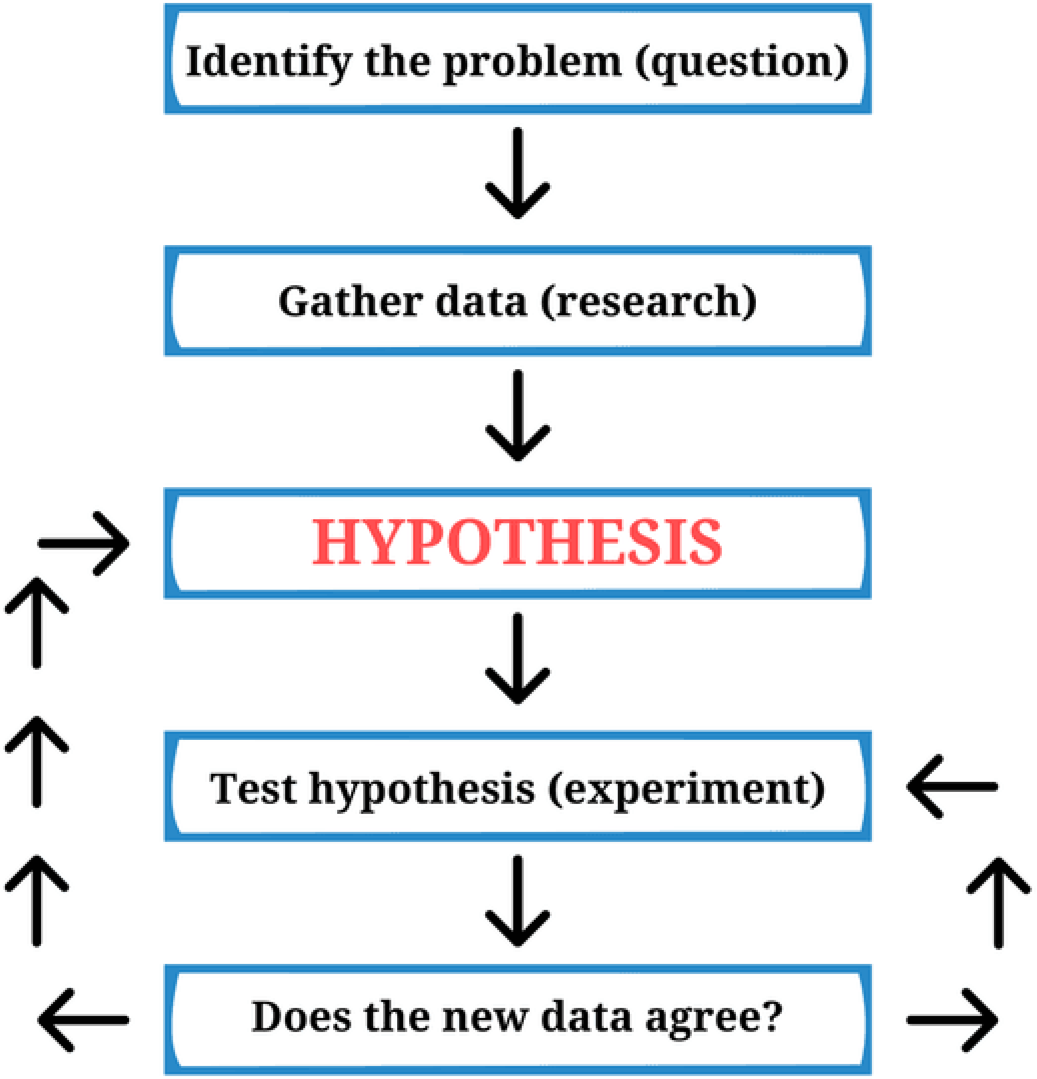
S.No.	Horns	Tail	Tusks	Paws	Fur	Color	Hooves	Size	Elephant
1.	No	Short	Yes	No	No	Black	No	Big	Yes
2.	Yes	Short	No	No	No	Brown	Yes	Medium	No
3.	No	Short	Yes	No	No	Black	No	Medium	Yes
4.	No	Long	No	Yes	Yes	White	No	Medium	No
5.	No	Short	Yes	Yes	Yes	Black	No	Big	Yes

3.4.1 Representation of a Hypothesis



A hypothesis should always:

- explain what you expect to happen
- be clear and understandable
- be testible
- be measurable
- contain an independent and dependent variable



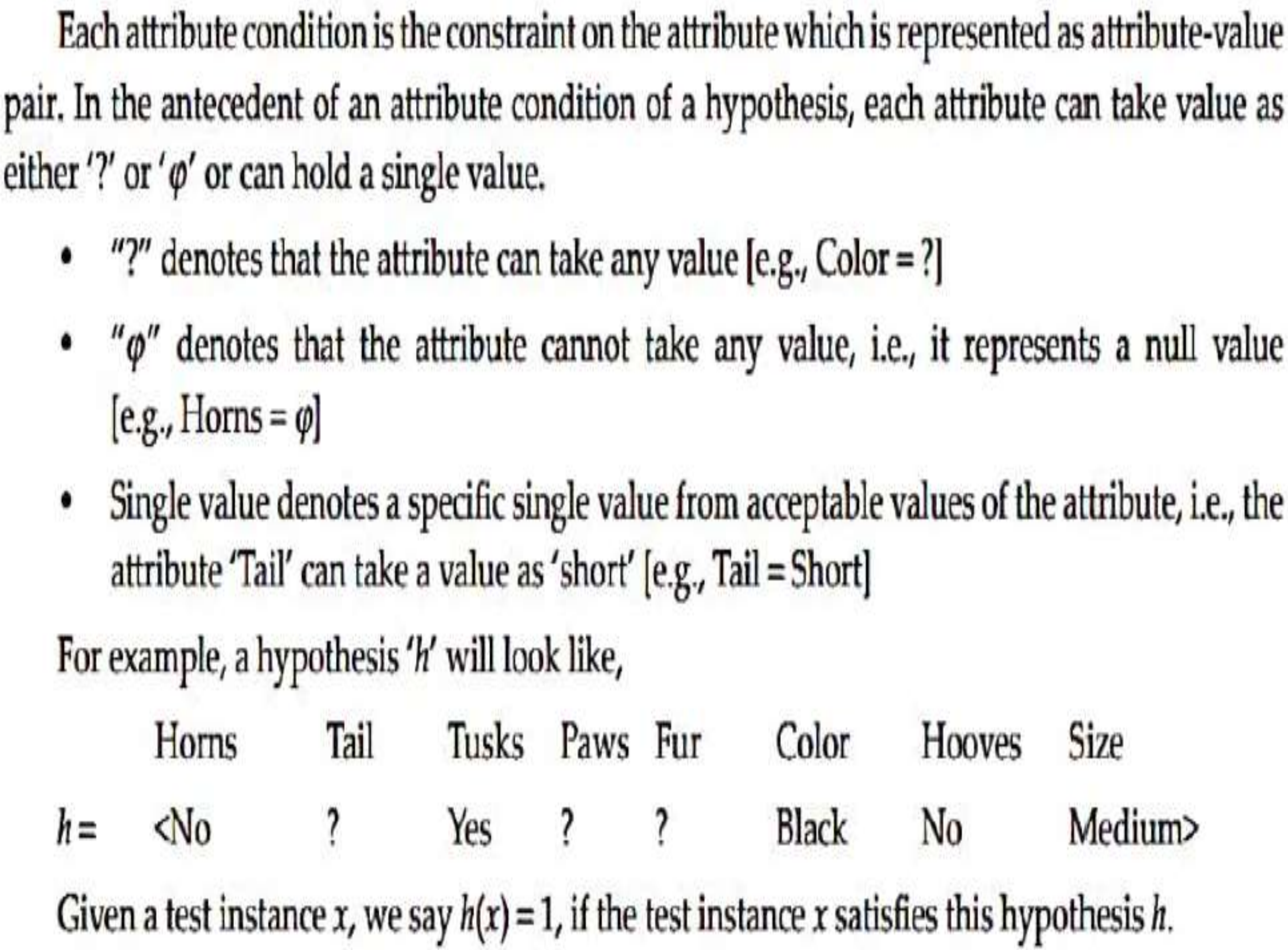
Identify the problem (question)

Gather data (research)

HYPOTHESIS

Test hypothesis (experiment)

Does the new data agree?



Each attribute condition is the constraint on the attribute which is represented as attribute-value pair. In the antecedent of an attribute condition of a hypothesis, each attribute can take value as either '?' or 'φ' or can hold a single value.

- "?" denotes that the attribute can take any value [e.g., Color = ?]
- "φ" denotes that the attribute cannot take any value, i.e., it represents a null value [e.g., Horns = φ]
- Single value denotes a specific single value from acceptable values of the attribute, i.e., the attribute 'Tail' can take a value as 'short' [e.g., Tail = Short]

For example, a hypothesis '*h*' will look like,

	Horns	Tail	Tusks	Paws	Fur	Color	Hooves	Size
h =	<No	?	Yes	?	?	Black	No	Medium>

Given a test instance *x*, we say $h(x) = 1$, if the test instance *x* satisfies this hypothesis *h*.

Hypothesis space is the set of all possible hypotheses that approximates the target function f. In other words, the set of all possible approximations of the target function can be defined as hypothesis space. From this set of hypotheses in the hypothesis space, a machine learning algorithm would determine the best possible hypothesis that would best describe the target function or best fit the outputs. Generally, a hypothesis representation language represents a larger hypothesis space. Every machine learning algorithm would represent the hypothesis space in a different manner about the function that maps the input variables to output variables.

For example, a regression algorithm represents the hypothesis space as a linear function whereas a decision tree algorithm represents the hypothesis space as a tree. The set of hypotheses that can be generated by a learning algorithm can be further reduced by specifying a language bias.

The subset of hypothesis space that is consistent with all-observed training instances is called as Version Space. Version space represents the only hypotheses that are used for the classification.

For example, each of the attribute given in the Table 3.1 has the following possible set of values.

Heuristic search is a search strategy that finds an optimized hypothesis/solution to a problem by iteratively improving the hypothesis/solution based on a given heuristic function or a cost measure. Heuristic search methods will generate a possible hypothesis that can be a solution in the hypothesis space or a path from the initial state.

This hypothesis will be tested with the target function or the goal state to see if it is a real solution. If the tested hypothesis is a real solution, then it will be selected. This method generally increases the efficiency because it is guaranteed to find a better hypothesis but may not be the best hypothesis.

It is useful for solving tough problems which could not solved by any other method. The typical example problem solved by heuristic search is the travelling salesman problem.

Several commonly used heuristic search methods are hill climbing methods, constraint satisfaction problems, best-first search, simulated-annealing, A* algorithm, and genetic algorithms.

3.4.4 Generalization and Specialization

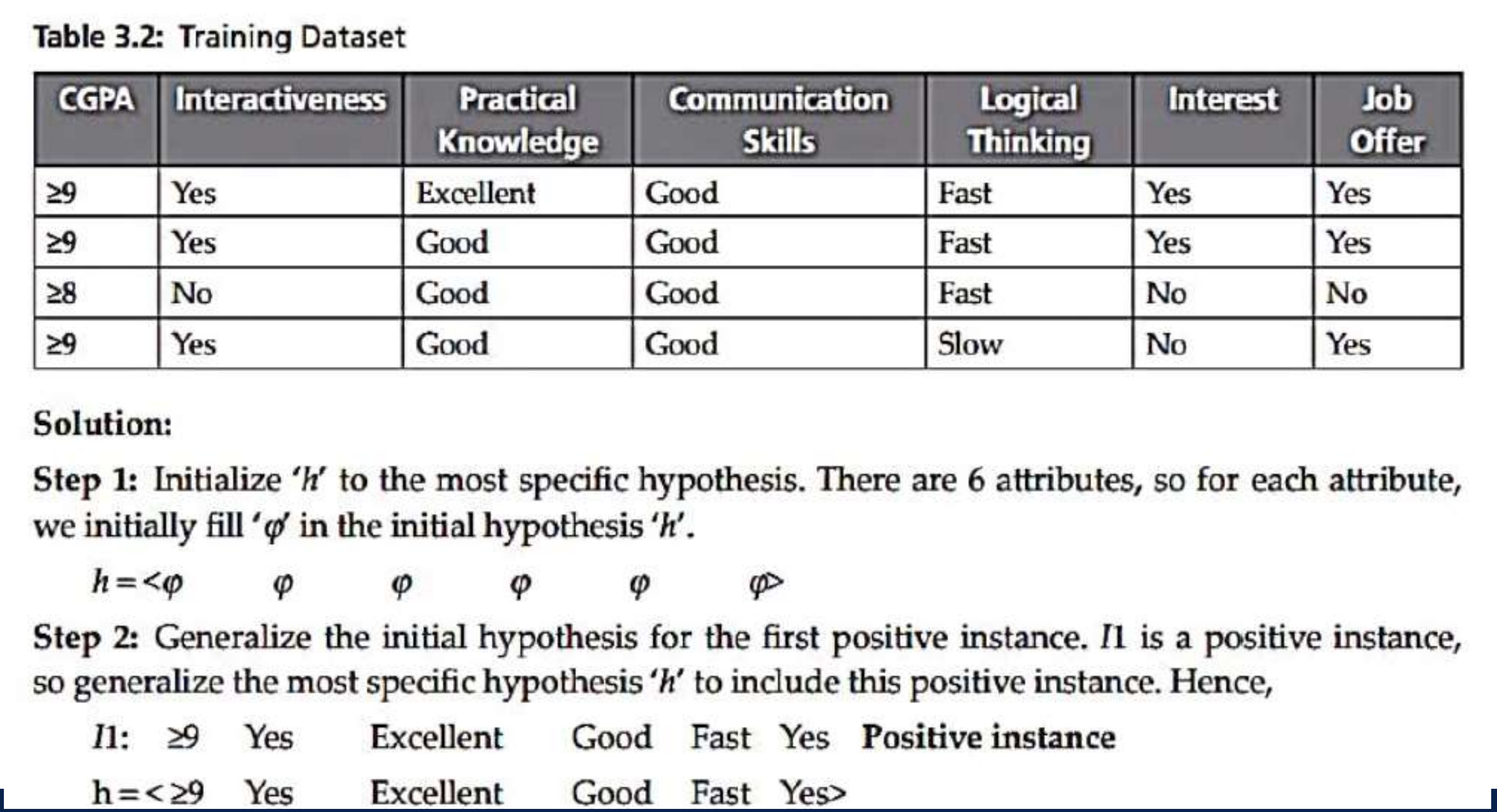


Table 3.2: Training Dataset

CGPA	Interactiveness	Practical Knowledge	Communication Skills	Logical Thinking	Interest	Job Offer
≥9	Yes	Excellent	Good	Fast	Yes	Yes
≥9	Yes	Good	Good	Fast	Yes	Yes
≥8	No	Good	Good	Fast	No	No
≥9	Yes	Good	Good	Slow	No	Yes

Solution:

Step 1: Initialize '*h*' to the most specific hypothesis. There are 6 attributes, so for each attribute, we initially fill '*φ*' in the initial hypothesis '*h*'.

$h = <\varphi \quad \varphi \quad \varphi \quad \varphi \quad \varphi \quad \varphi>$

Step 2: Generalize the initial hypothesis for the first positive instance. *I*1 is a positive instance, so generalize the most specific hypothesis '*h*' to include this positive instance. Hence,

*I*1: ≥9 Yes Excellent Good Fast Yes **Positive instance**

h = < ≥9 Yes Excellent Good Fast Yes>

Limitations of Find-S Algorithm

1.Find-S algorithm tries to find a hypothesis that is consistent with positive instances, ignoring all negative instances. As long as the training dataset is consistent, the hypothesis found by this algorithm may be consistent.

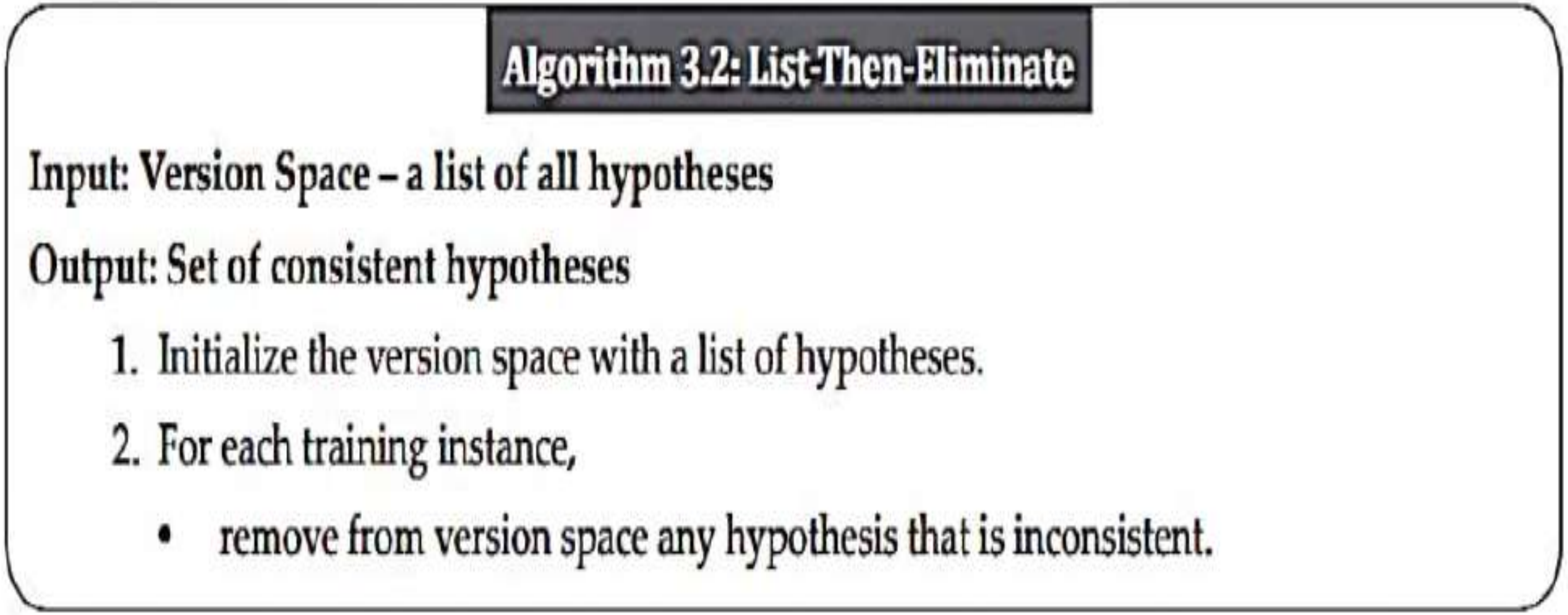
2.The algorithm finds only one unique hypothesis, wherein there may be many other hypotheses that are consistent with the training dataset.

3.Many times, the training dataset may contain some errors; hence such inconsistent data instances can mislead this algorithm in determining the consistent hypothesis since it ignores negative instances.

4.Hence, it is necessary to find the set of hypotheses that are consistent with the training data including the negative examples. To overcome the limitations of Find-S algorithm, Candidate Elimination algorithm was proposed to output the set of all hypotheses consistent with the training dataset.

3.4.6 Version Spaces

List-Then-Eliminate Algorithm



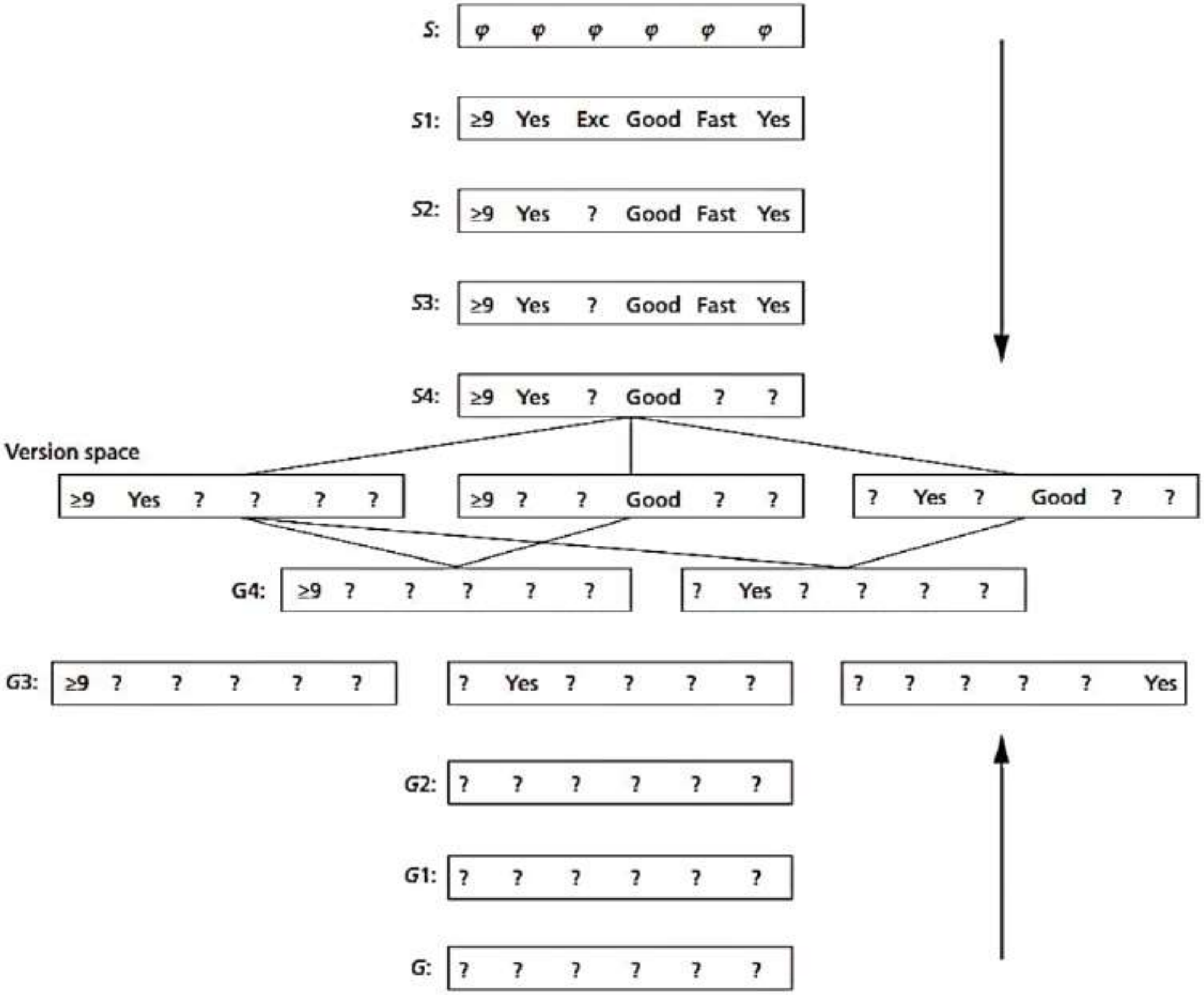
Algorithm 3.2: List-Then-Eliminate

Input: Version Space – a list of all hypotheses

Output: Set of consistent hypotheses

1. Initialize the version space with a list of hypotheses.
2. For each training instance,
   - remove from version space any hypothesis that is inconsistent.

Version Spaces and the Candidate Elimination Algorithm



S: φ φ φ φ φ φ

S1: ≥9 Yes Exc Good Fast Yes

S2: ≥9 Yes ? Good Fast Yes

S3: ≥9 Yes ? Good Fast Yes

S4: ≥9 Yes ? Good ? ?

Version space

≥9 Yes ? ? ? ?

≥9 ? ? Good ? ?

? Yes ? Good ? ?

G4: ≥9 ? ? ? ? ?

? Yes ? ? ? ?

G3: ≥9 ? ? ? ? ?

? Yes ? ? ? ?

? ? ? ? ? Yes

G2: ? ? ? ? ? ?

G1: ? ? ? ? ? ?

G: ? ? ? ? ? ?

Figure 3.2: Deriving the Version Space

3.6 MODELLING IN MACHINE LEARNING

A machine learning model is an abstraction of the training dataset that can perform a prediction on new data. Training the model means feeding instances to the machine learning algorithm.

Training datasets are used to fit and tune the model. After training a machine learning algorithm with the training data, a predictive model is generated as output to which a new data is fed to make predictions.

The process of modelling means training a machine learning algorithm with the training dataset, tuning it to increase performance, validating it and making predictions for a new unseen data. The major concern in machine learning is what model to select, how to train the model, time required to train, the dataset to be used, what performance to expect, and so on.

During prediction, an error occurs when the estimated output does not match with the true output. Training error, also called as in-sample error, results when applying the predicted model on the training data, while Test error also called as out-of-sample error is the average error when predicting on unseen observations. The error function or the loss function is the aggregation of the differences between the true values and the predicted values.

This loss function is defined as the Mean Squared Error (MSE), which is the average of the squared differences between the true values Y,and the predicted values f{X)) for an input value ‘X, . A smaller value of MSE denotes that the error is less and, therefore, the prediction is more accurate.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}[Y_i - f(X_i)]^2$$

Machine Learning Process

3.6.1 Model Selection and Model Evaluation

3.6.2 Re-sampling Methods

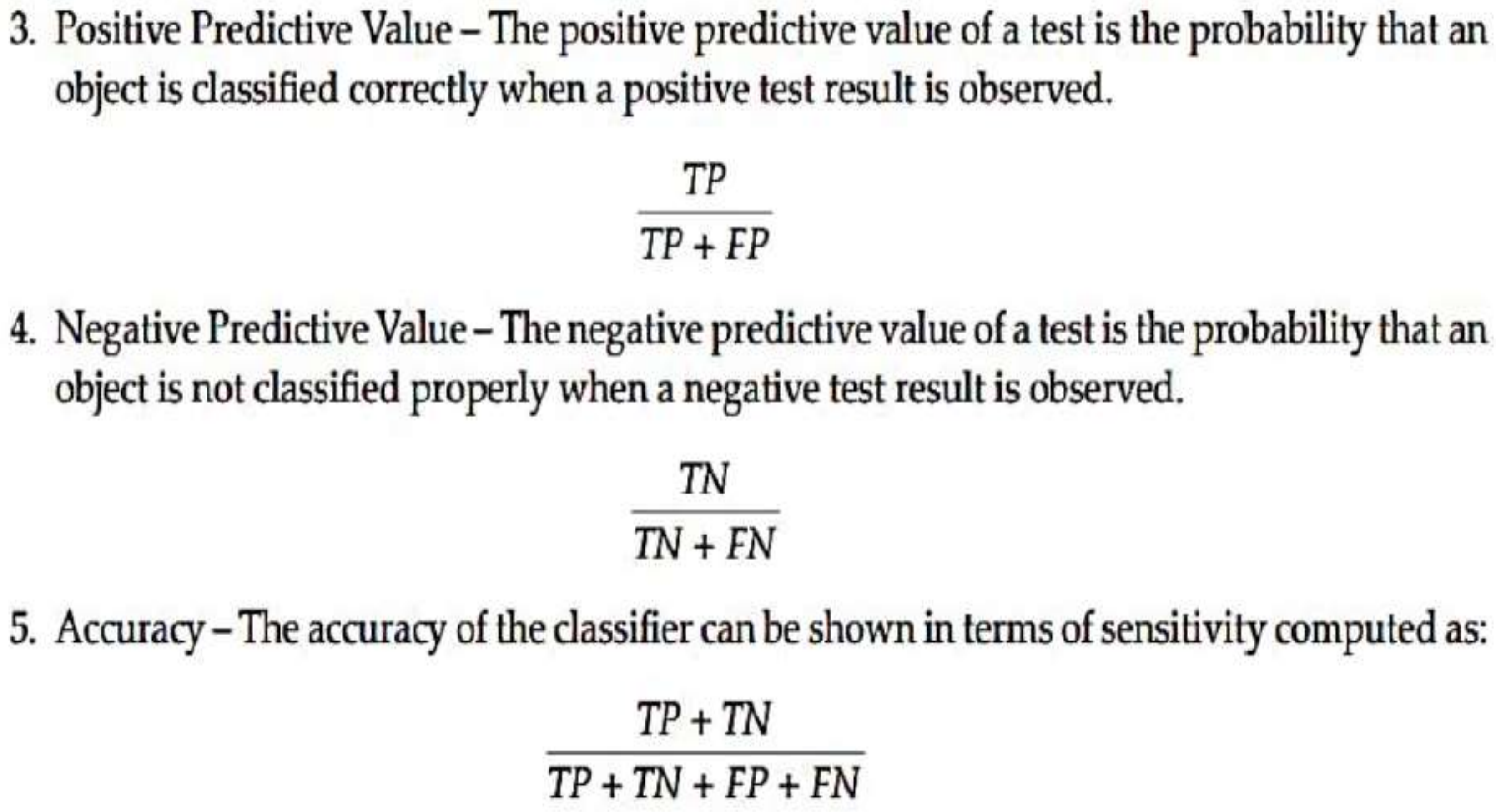
K-fold Cross-Validation

Leave-One-Out Cross-Validation (LOOCV)

Model Performance

In this table, True Positive (TP) = Number of cancer patients who are classified by the test correctly, True Negative (TN) = Number of normal patients who do not have cancer are correctly detected.

The two errors that are involved in this process is False Positive (FP) that is an alarm that indicates that the tests show positive when the patient has no disease and False Negative (FN) is another error that says a patient has cancer when tests says negative or normal. FP and FN are costly errors in this classification process.



3. Positive Predictive Value – The positive predictive value of a test is the probability that an object is classified correctly when a positive test result is observed.

$$\frac{TP}{TP + FP}$$

4. Negative Predictive Value – The negative predictive value of a test is the probability that an object is not classified properly when a negative test result is observed.

$$\frac{TN}{TN + FN}$$

5. Accuracy – The accuracy of the classifier can be shown in terms of sensitivity computed as:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Classifier Performance as Distance Measures The classifier performance can be computed as a distance measure also. The classifier accuracy can be plotted as a point. A point in the north-west is a better classifier. Euclid distance of two points of the two classifiers can give a performance measure. The value ranges from 0 to 1.

Visual Classifier Performance Receiver Operating Characteristic (ROC) curve and Precision-Recall curves indicate the performance of classifiers visually. ROC curves are visual means of checking the accuracy and comparison of classifiers. ROC is a plot of sensitivity (True Positive Rate) and the 1-specificity (False Positive Rate) for a given model.

A sample ROC curve is shown in Figure 3.6, where results of five classifiers are given. A is the ROC of an average classifier. The ideal classifier is E where the area under curve is 1.0. Theoretically, it can range from 0.9 to 1. The rest of the classifiers B, C, D are categorized based on area under curve as good, better and still better based on the area under curve values.

Instead of predicting the label of a classifier, one can predict the probabilities of the model. Probabilities allow some better evaluation by functions that are called scoring functions or scoring rules.

The area under curve (AUC) is one such score that can be used for classifier model evaluation. The integrated AUC is a measure of the model across threshold values. AUC indicates the accuracy of the model. A model is perfect if it has area under ROC curve as one. The AUC score 0 of a model indicates the wrong model.

The approximate area under precision-recall curve also indicates the power of the model across thresholds. A precision-recall curve is a plot of precision and recall for different threshold values.

This curve is useful if there is an imbalance in the classes where one class has more labels and other classes have less samples. ROC is used when there is no class imbalance and precision-recall curves are used when there is a moderate-to-large class imbalance.

where, L(D k) is the number of bits used to represent the predictions D based on the training set.

MDL can be expressed in terms of negative log-likelihood also as:

$$\mathrm{MDL} = -\log(p(\Theta)) - \log(p(y \mid x, \Theta))$$

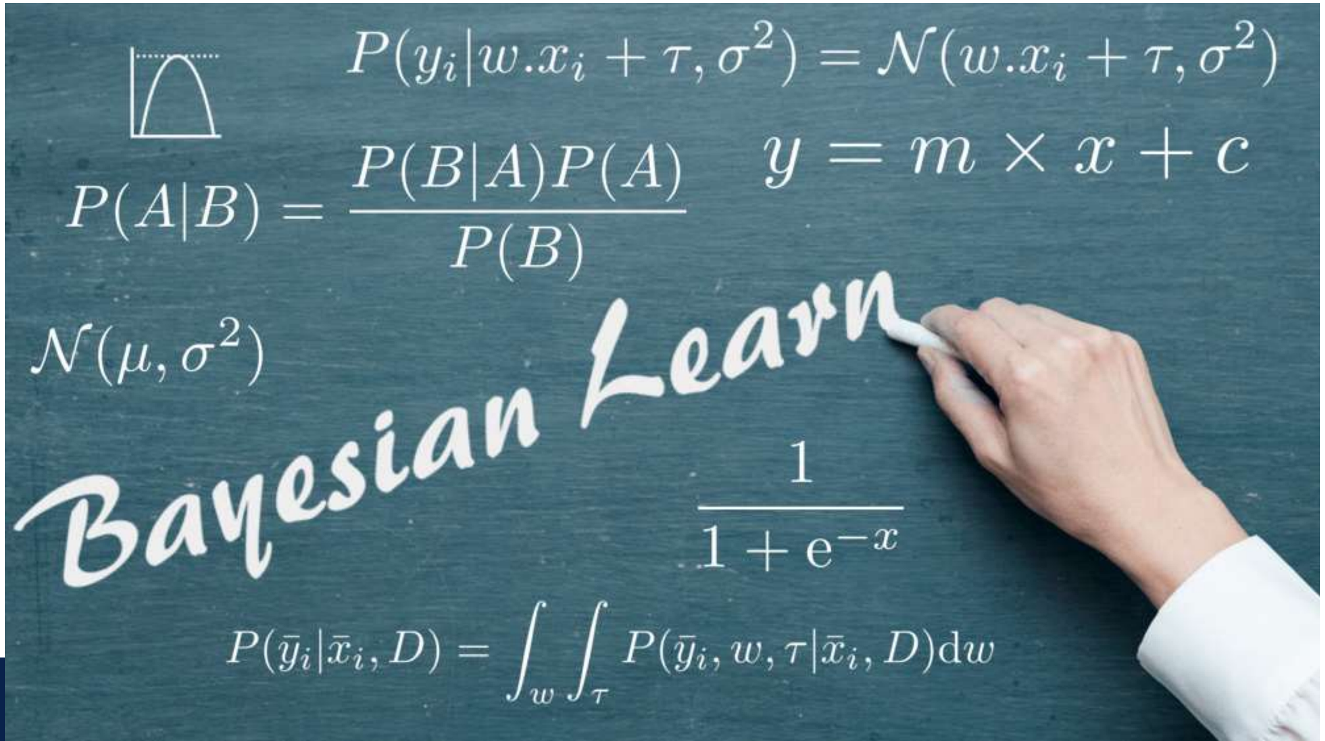
where, y is the target variable, x is the input and Θ is the model parameters.

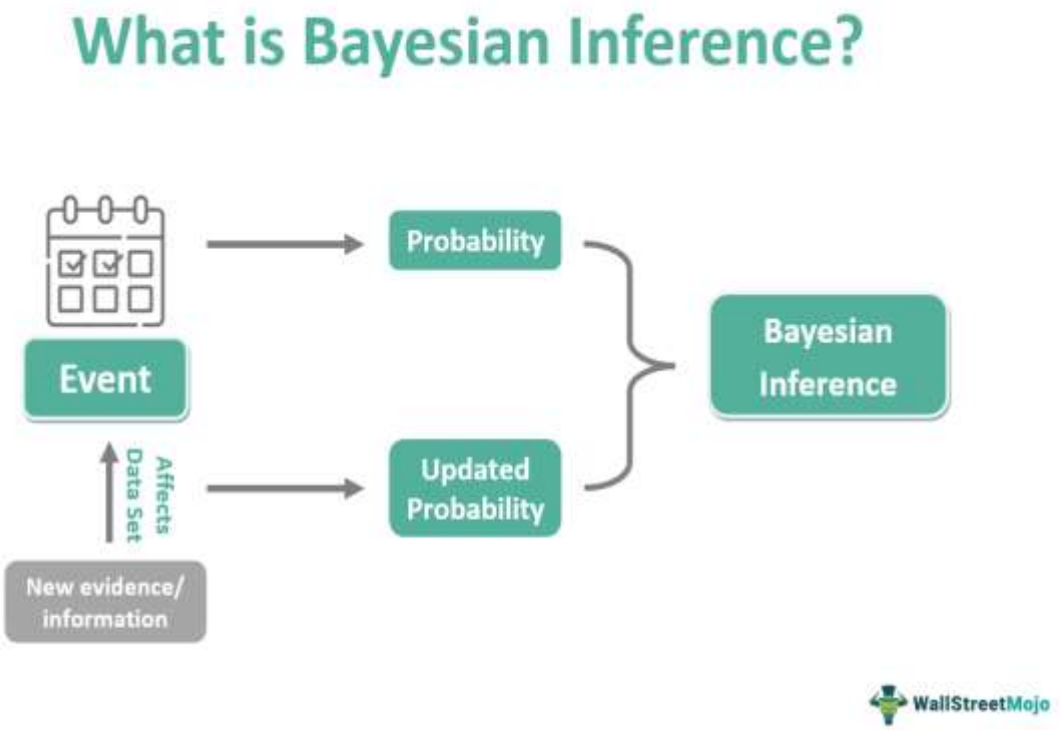
Thank You!

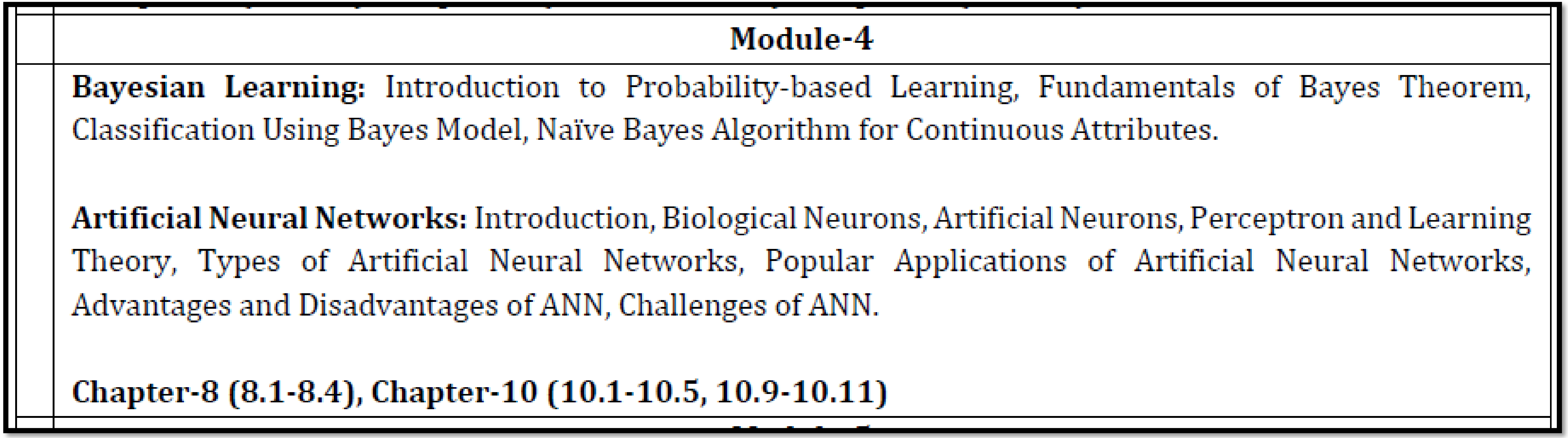
Module 4

Bayesian Learning

By
Ashwini P.,
Assistant Professor
Dept. of CSE
ATMECE, Mysuru


$P(y_i|w.x_i+\tau,\sigma^2)=\mathcal{N}(w.x_i+\tau,\sigma^2)$
$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$
$y=m\times x+c$
$\mathcal{N}(\mu,\sigma^2)$
Bayesian Learn
$\frac{1}{1+e^{-x}}$
$P(\bar{y}_i|\bar{x}_i,D)=\int_w\int_\tau P(\bar{y}_i,w,\tau|\bar{x}_i,D)dw$





	Module-4
	Bayesian Learning: Introduction to Probability-based Learning, Fundamentals of Bayes Theorem, Classification Using Bayes Model, Naïve Bayes Algorithm for Continuous Attributes. **Artificial Neural Networks:** Introduction, Biological Neurons, Artificial Neurons, Perceptron and Learning Theory, Types of Artificial Neural Networks, Popular Applications of Artificial Neural Networks, Advantages and Disadvantages of ANN, Challenges of ANN. **Chapter-8 (8.1-8.4), Chapter-10 (10.1-10.5, 10.9-10.11)**

Bayesian Learning is a learning method that describes and represents knowledge in an uncertain domain and provides a way to reason about this knowledge using probability measure. It uses Bayes theorem to infer the unknown parameters of a model.

Bayesian inference is useful in many applications which involve reasoning and diagnosis such as game theory, medicine, etc.

Bayesian inference is much more powerful in handling missing data and for estimating any uncertainty in predictions.

Probability-based learning is one of the most important practical learning methods which combines prior knowledge or prior probabilities with observed data. Probabilistic learning uses the concept of probability theory that describes how to model randomness, uncertainty, and noise to predict future events.

It is a tool for modelling large datasets and uses Bayes rule to infer unknown quantities, predict and learn from data. In a probabilistic model, randomness plays a major role which gives probability distribution a solution, while in a deterministic model there is no randomness and hence it exhibits the same initial conditions every time the model is run and is likely to get a single possible outcome as the solution.

Bayesian learning differs from probabilistic learning as it uses subjective probabilities (i.e., probability that is based on an individual's belief or interpretation about the outcome of an event and it can change over time) to infer parameters of a model. Two practical learning algorithms called Naive Bayes learning and Bayesian Belief Network (BBN) form the major part of Bayesian learning. These algorithms use prior probabilities and apply Bayes rule to infer useful information.

Naive Bayes Model relies on Bayes theorem that works on the principle of three kinds of probabilities called prior probability, likelihood probability, and posterior probability.

Prior Probability It is the general probability of an uncertain event before an observation is seen or some evidence is collected. It is the initial probability that is believed before any new information is collected.

Likelihood Probability Likelihood probability is the relative probability of the observation occurring for each class or the sampling density for the evidence given the hypothesis. It is stated as P (Evidence | Hypothesis), which denotes the likeliness of the occurrence of the evidence given the parameters.

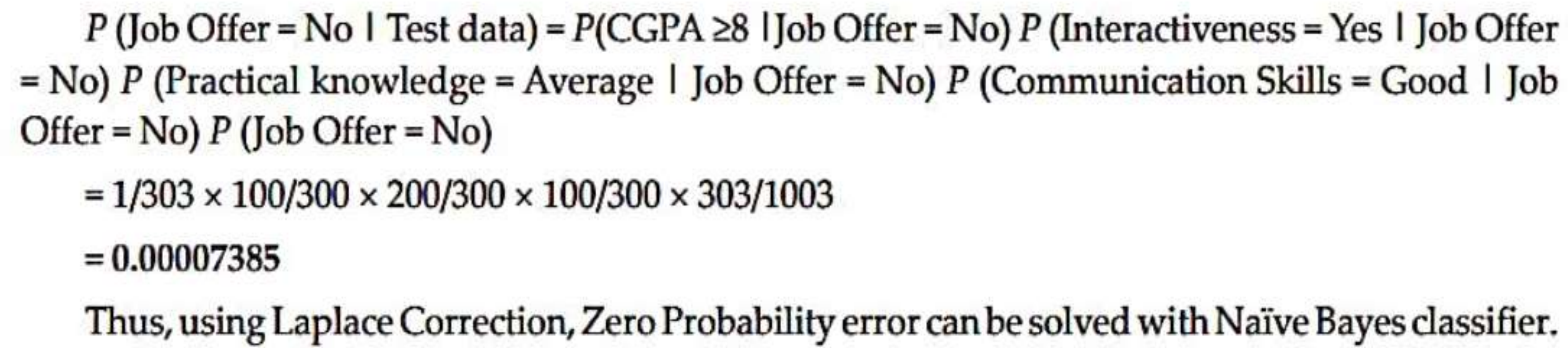
Posterior Probability It is the updated or revised probability of an event taking into account the observations from the training data. P (Hypothesis | Evidence) is the posterior distribution representing the belief about the hypothesis, given the evidence from the training data. Therefore, Posterior probability = prior probability + new evidence

Naive Bayes Classification models work on the principle of Bayes theorem. Bayes’ rule is a mathematical formula used to determine the posterior probability, given prior probabilities of events. Generally, Bayes theorem is used to select the most probable hypothesis from data, considering both prior knowledge and posterior distributions. It is based on the calculation of the posterior probability and is stated as:

P (Hypothesis h | Evidence E) where, Hypothesis k is the target class to be classified and Evidence E is the given test instance.

P (Hypothesis k| Evidence E) is calculated from the prior probability P (Hypothesis k), the likelihood probability P (Evidence E |Hypothesis k) and the marginal probability P (Evidence E). It can be written as:

$$P\,(\text{Hypothesis}\ h \mid \text{Evidence}\ E) = \frac{P(\text{Evidence}\ E \mid \text{Hypothesis}\ h)\, P(\text{Hypothesis}\ h)}{P(\text{Evidence}\ E)}$$

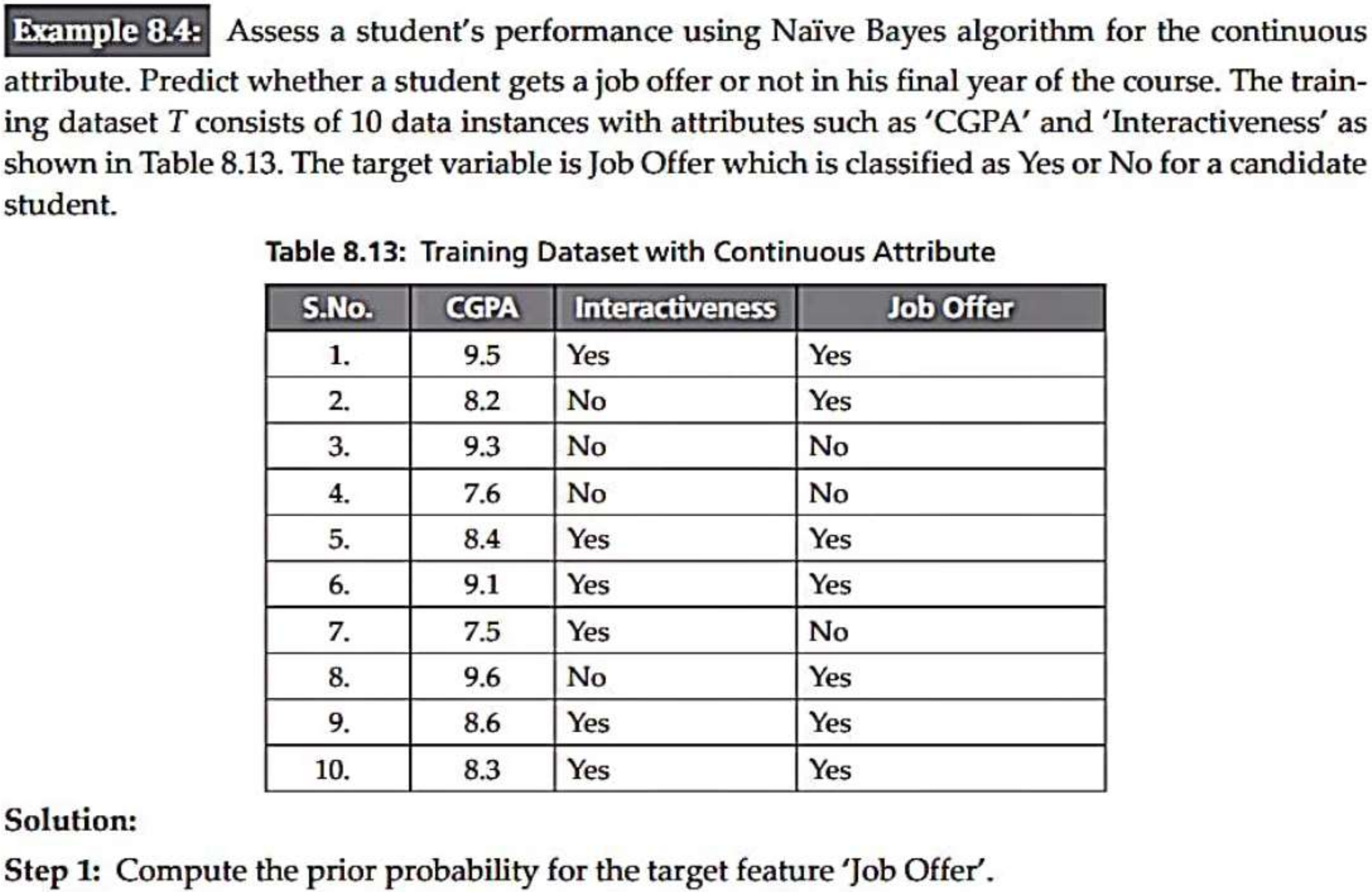


P (Job Offer = No | Test data) = *P*(CGPA $\geq$8 |Job Offer = No) *P* (Interactiveness = Yes | Job Offer = No) *P* (Practical knowledge = Average | Job Offer = No) *P* (Communication Skills = Good | Job Offer = No) *P* (Job Offer = No)

= 1/303 × 100/300 × 200/300 × 100/300 × 303/1003

= 0.00007385

Thus, using Laplace Correction, Zero Probability error can be solved with Naïve Bayes classifier.



Example 8.4: Assess a student's performance using Naïve Bayes algorithm for the continuous attribute. Predict whether a student gets a job offer or not in his final year of the course. The training dataset *T* consists of 10 data instances with attributes such as 'CGPA' and 'Interactiveness' as shown in Table 8.13. The target variable is Job Offer which is classified as Yes or No for a candidate student.

Table 8.13: Training Dataset with Continuous Attribute

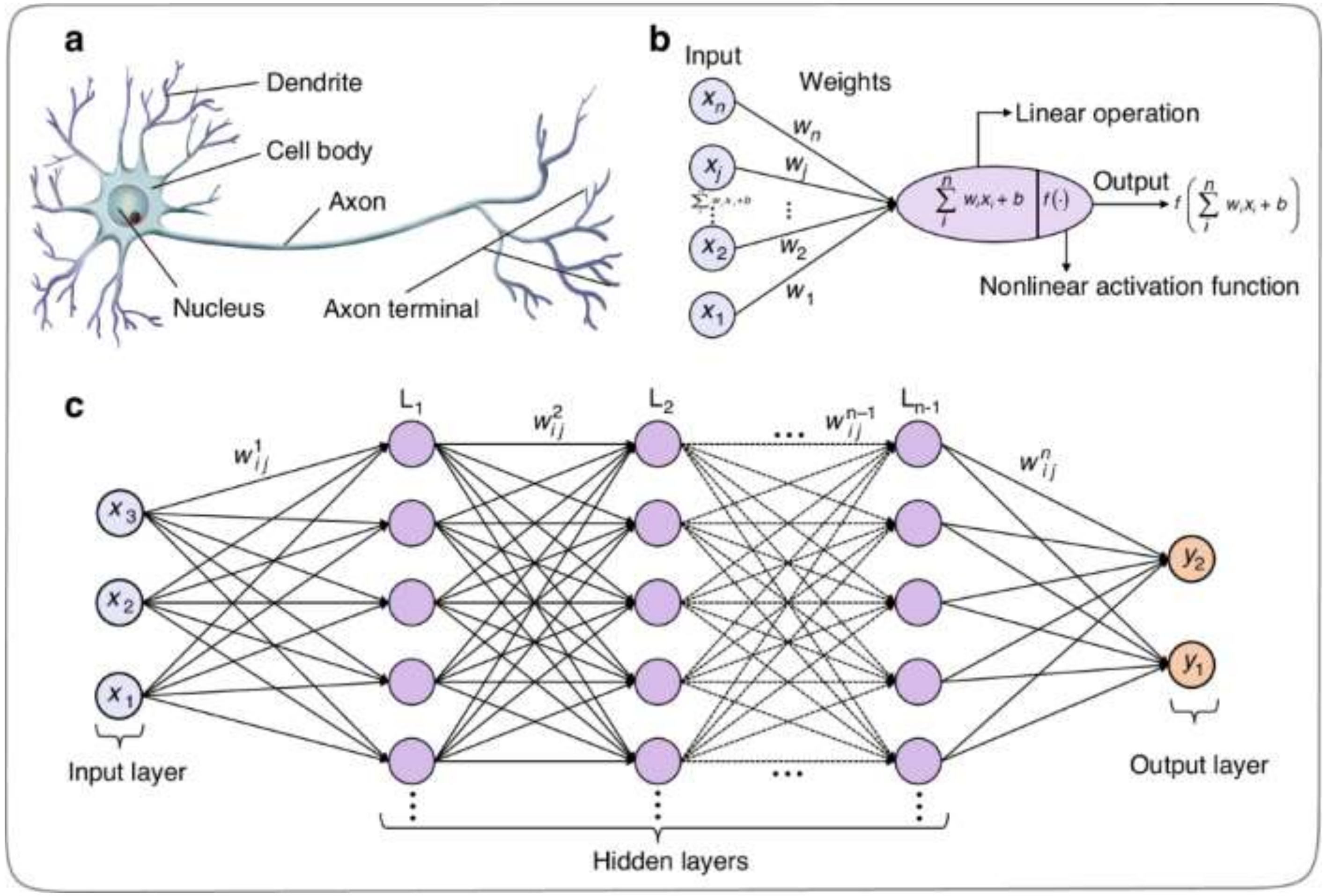
S.No.	CGPA	Interactiveness	Job Offer
1.	9.5	Yes	Yes
2.	8.2	No	Yes
3.	9.3	No	No
4.	7.6	No	No
5.	8.4	Yes	Yes
6.	9.1	Yes	Yes
7.	7.5	Yes	No
8.	9.6	No	Yes
9.	8.6	Yes	Yes
10.	8.3	Yes	Yes

Solution:

Step 1: Compute the prior probability for the target feature 'Job Offer'.

Chapter 10

Artificial Neural Networks



Artificial Neural Networks (ANNs) imitate human brain behaviour and the way in which learning happens in a human. The human brain constitutes a mass of neurons that are all connected as a network, which is actually a directed graph.

These neurons are the processing units which receive information, process it and then transmit this data to other neurons that allows humans to learn almost any task. ANN is a learning mechanism that models a human brain to solve any non-linear and complex problem.

Each neuron is modelled as a computing unit, or simply called as a node in ANN, that is capable of doing complex calculations. ANN is a system that consists of many such computing units operating in parallel that can learn from observations. They are widely used in developing artificial learning systems and have inspired researchers and industry in Machine Learning nowadays. Some typical applications of ANN in the field of computer science are Natural Language Processing (NLP), pattern recognition, face recognition, speech recognition, character recognition, text processing, stock prediction, computer vision, etc.

ANNs also have been considerably used in other engineering fields such as Chemical industry, Medicine, Robotics, Communications, Banking, and Marketing.

10.1 INTRODUCTION

The human nervous system has billions of neurons that are the processing units which make humans to perceive things, to hear, to see and to smell. The human nervous system works beautifully, making us understand who we are, what we do, where we are and everything in our surrounding. It makes us to remember, recognize and correlate things around us.

It is a learning system that consists of functional units called nerve cells, typically called as neurons.

The human nervous system is divided into two sections called the Central Nervous System (CNS) and the Peripheral Nervous System (PNS). The brain and the spinal cord constitute the CNS and the neurons inside and outside the CNS constitute the PNS. The neurons are basically classified into three types called sensory neurons, motor neurons and interneurons.

Sensory neurons get information from different parts of the body and bring it into the CNS, whereas motor neurons receive information from other neurons and transmit commands to the body parts.

The CNS consists of only interneurons which connect one neuron to another neuron by receiving information from one neuron and transmitting it to another. The basic functionality of a neuron is to receive information, process it and then transmit it to another neuron or to a body part.

10.2 BIOLOGICAL NEURONS

A typical biological neuron has four parts called dendrites, soma, axon and synapse. The body of the neuron is called as soma. Dendrites accept the input information and process it in the cell body called soma. A single neuron is connected by axons to around 10,000 neurons and through these axons the processed information is passed from one neuron to another neuron.

A neuron gets fired if the input information crosses a threshold value and transmits signals to another neuron through a synapse. A synapse gets fired with an electrical impulse called spikes which are transmitted to another neuron. A single neuron can receive synaptic inputs from one neuron or multiple neurons. These neurons form a network structure which processes input information and gives out a response. The simple structure of a biological neuron is shown in Figure 10.1.

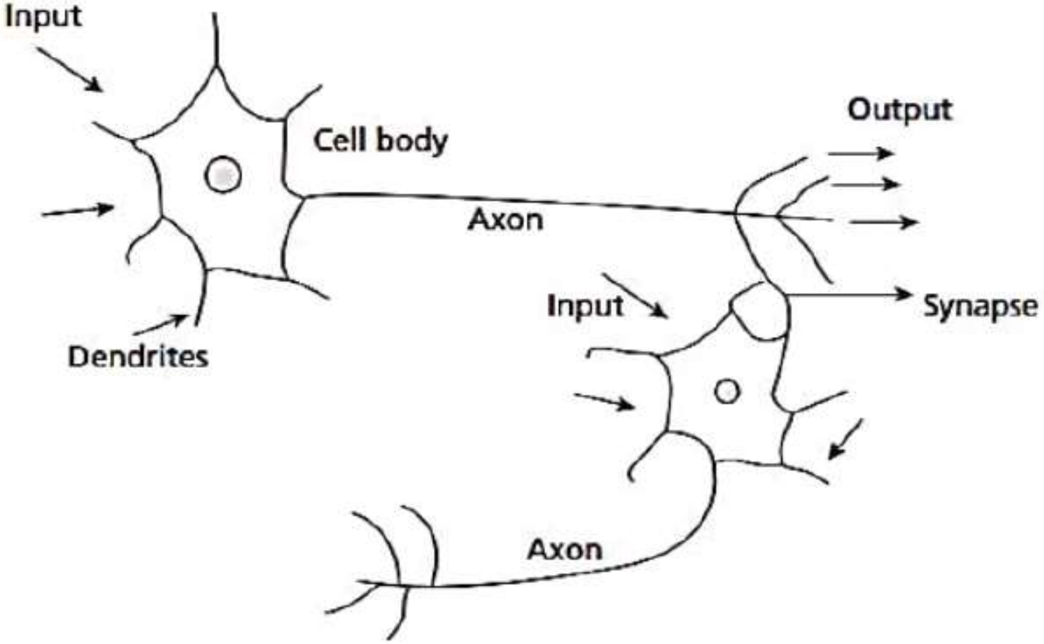
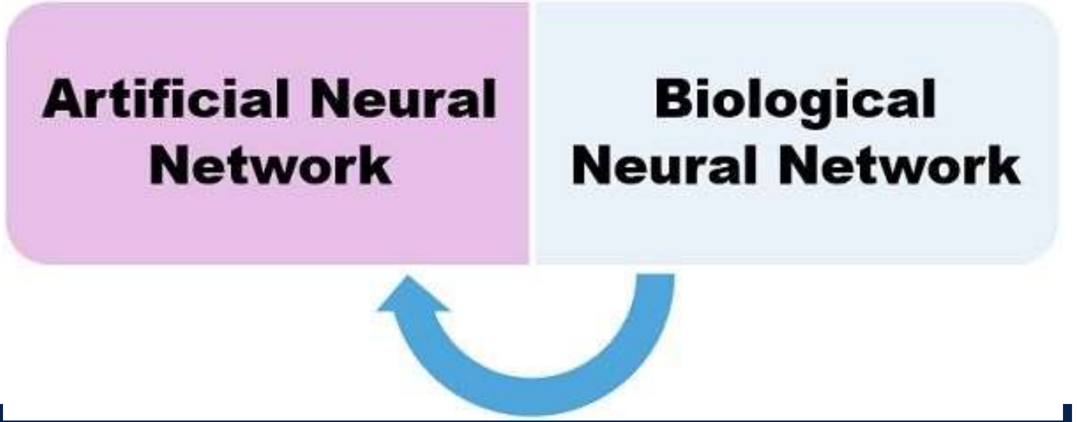
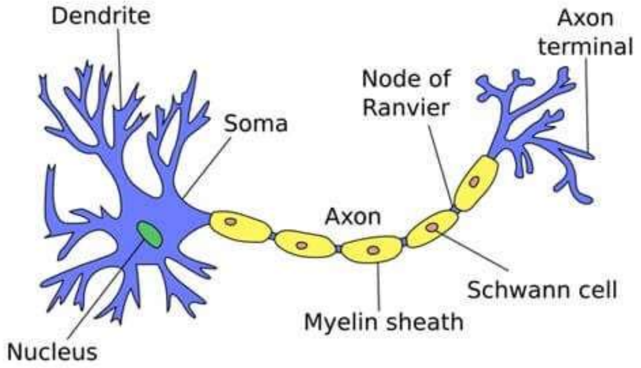
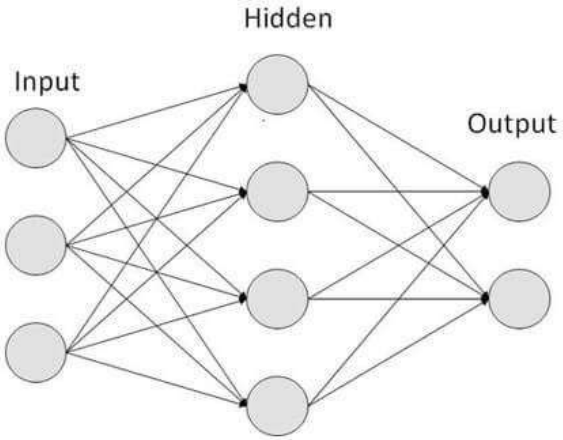


Figure 10.1: A Biological Neuron

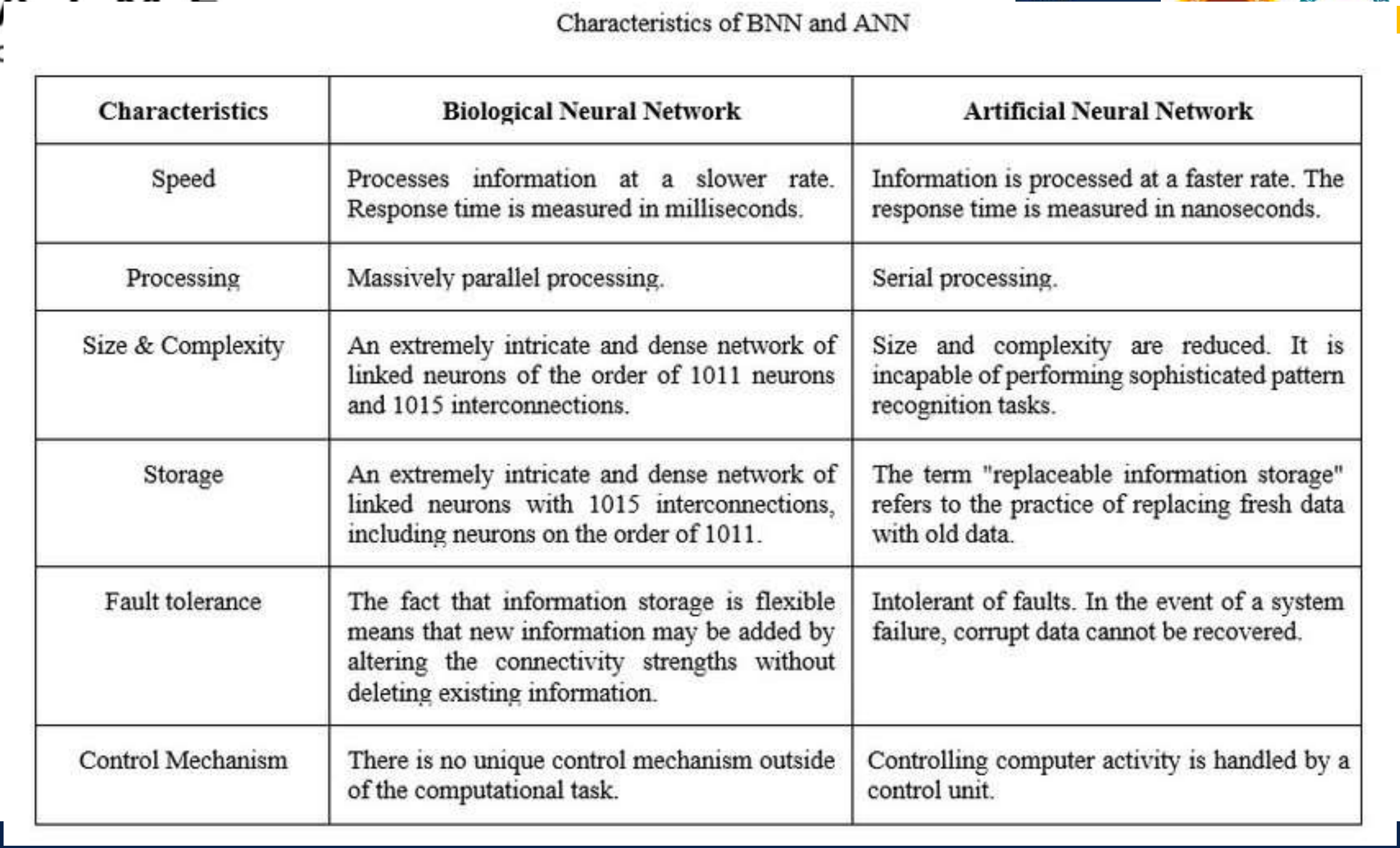
Biological Neuron	Artificial Neuron
Dendrites	Input
Cell Nucleus(Soma)	Node
Axon	Output
Synapse	Interconnections







🧠	🖥️
Biological neurons or nerve cells	Silicon transistors
200 billion neurons, 32 trillion interconnections.	1 billion bytes RAM, trillion of bytes on disk.
Neuron size: 10-6 m.	Single transistor size: 10-9m.
Energy consumption: 6-10 joules per operation per sec.	Energy consumption: 10-16 joules per operation per second.
Learning capability	Programming capability



Characteristics of BNN and ANN

Characteristics	Biological Neural Network	Artificial Neural Network
Speed	Processes information at a slower rate. Response time is measured in milliseconds.	Information is processed at a faster rate. The response time is measured in nanoseconds.
Processing	Massively parallel processing.	Serial processing.
Size & Complexity	An extremely intricate and dense network of linked neurons of the order of 1011 neurons and 1015 interconnections.	Size and complexity are reduced. It is incapable of performing sophisticated pattern recognition tasks.
Storage	An extremely intricate and dense network of linked neurons with 1015 interconnections, including neurons on the order of 1011.	The term "replaceable information storage" refers to the practice of replacing fresh data with old data.
Fault tolerance	The fact that information storage is flexible means that new information may be added by altering the connectivity strengths without deleting existing information.	Intolerant of faults. In the event of a system failure, corrupt data cannot be recovered.
Control Mechanism	There is no unique control mechanism outside of the computational task.	Controlling computer activity is handled by a control unit.

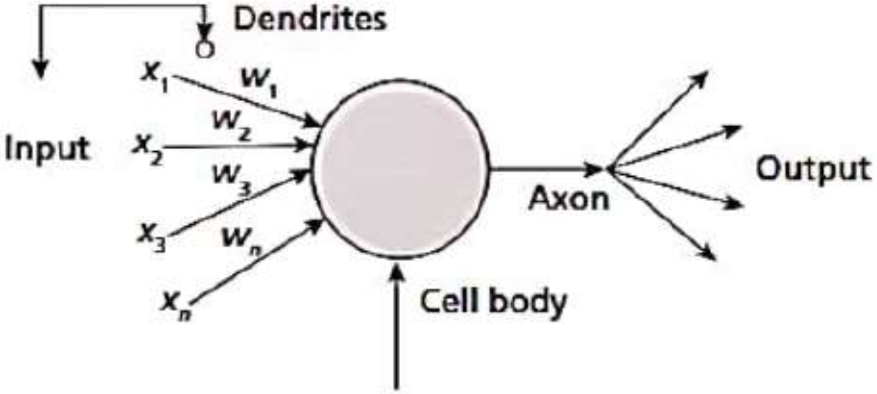


Figure 10.2: An Artificial Neuron

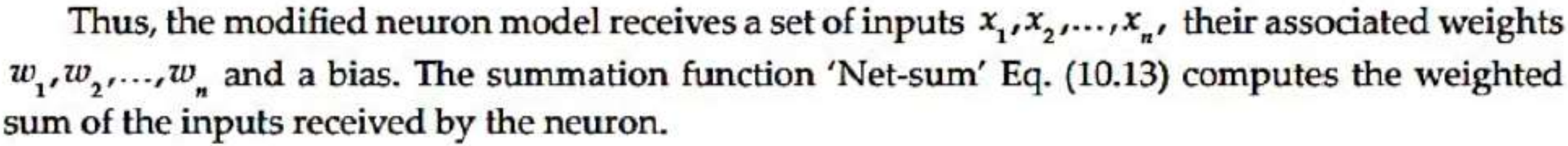
The first mathematical model of a biological neuron was designed by McCulloch & Pitts in 1943. It includes two steps:

1. It receives weighted inputs from other neurons

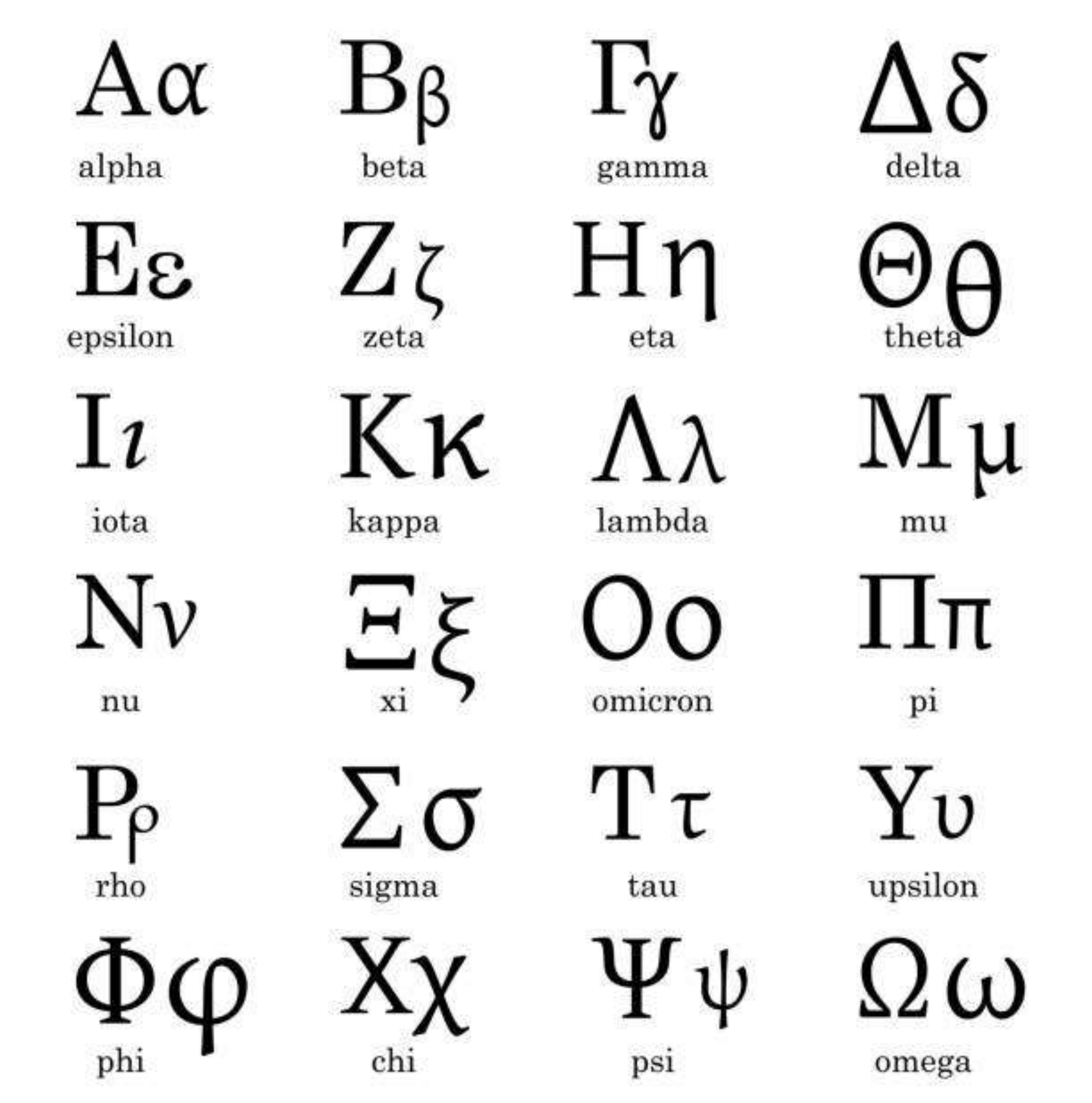
2. It operates with a threshold function or activation function. The received inputs are computed as a weighted sum which is given to the activation function and if the sum exceeds the threshold value the neuron gets fired. The mathematical model of a neuron is shown in Figure 10.3.

The neuron is the basic processing unit that receives a set of inputs x1,x2,,...,xn, and their associated weights w1,w2,...,wn. The Summation function ‘Net-sum’ Eq. (10.1) computes the weighted sum of the inputs received by the neuron.

$$\text{Net-sum} = \sum_{i=1}^{n} x_i w_i$$



Thus, the modified neuron model receives a set of inputs $x_1, x_2, \ldots, x_n$, their associated weights $w_1, w_2, \ldots, w_n$ and a bias. The summation function 'Net-sum' Eq. (10.13) computes the weighted sum of the inputs received by the neuron.



Αα	Ββ	Γγ	Δδ
alpha	beta	gamma	delta
Εε	Ζζ	Ηη	Θθ
epsilon	zeta	eta	theta
Ιι	Κκ	Λλ	Μμ
iota	kappa	lambda	mu
Νν	Ξξ	Οο	Ππ
nu	xi	omicron	pi
Ρρ	Σσ	Ττ	Υυ
rho	sigma	tau	upsilon
Φφ	Χχ	Ψψ	Ωω
phi	chi	psi	omega

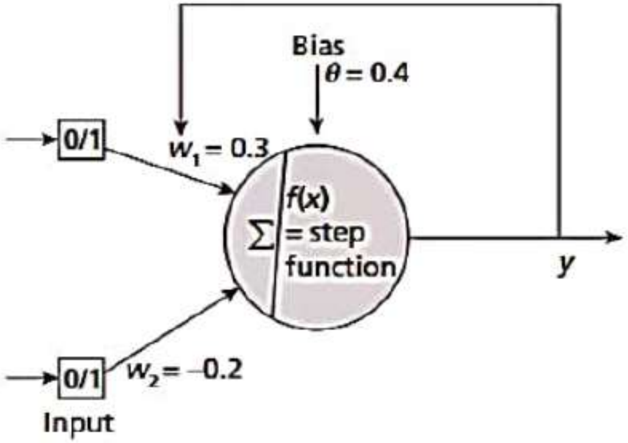


Figure 10.6: Perceptron for Boolean Function AND

Solution: Desired output for Boolean function AND is shown in Table 10.1.

Table 10.1: AND Truth Table

x_1	x_2	Y_{des}
0	0	0
0	1	0
1	0	0
1	1	1

For input (1, 0) the weights are updated as follows:

$$\Delta w_1 = \propto \times e(t) \times x_1 = 0.2 \times -1 \times 1 = -0.2$$

$$w_1 = w_1 + \Delta w_1 = 0.5 + \Delta w_1 = 0.5 - 0.2 = 0.3$$

$$\Delta w_2 = \propto \times e(t) \times x_2 = 0.2 \times -1 \times 0 = 0$$

$$w_2 = w_2 + \Delta w_2 = 0 + \Delta w_2 = 0 + 0 = 0$$

For input (1, 1), the weights are updated as follows:

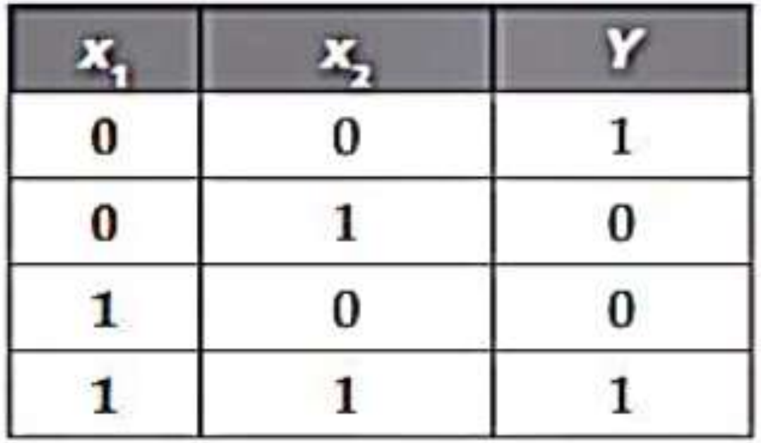
$$\Delta w_1 = \propto \times e(t) \times x_1 = 0.2 \times 1 \times 1 = 0.2$$

$$w_1 = w_1 + \Delta w_1 = 0.3 + \Delta w_1 = 0.3 + 0.2 = 0.5$$

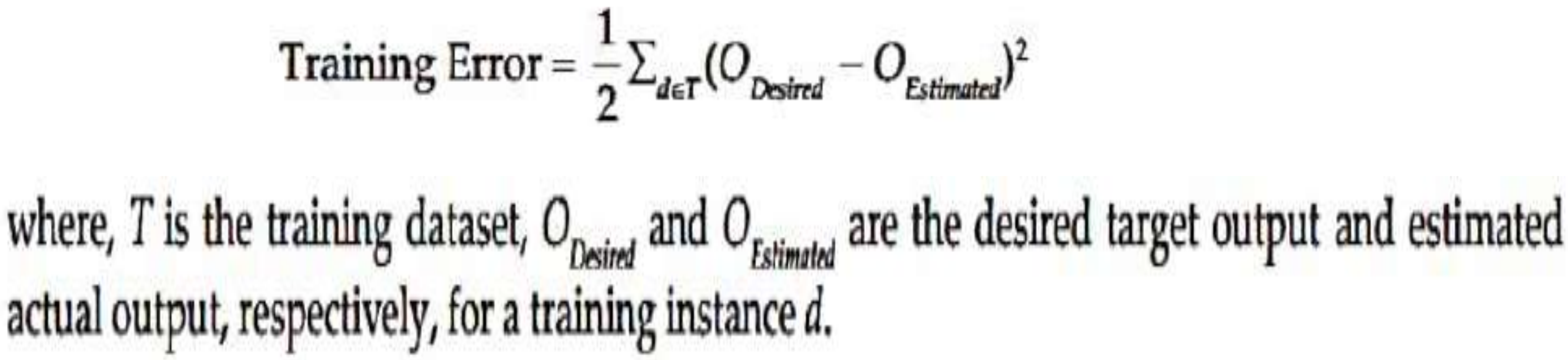
$$\Delta w_2 = \propto \times e(t) \times x_2 = 0.2 \times 1 \times 1 = 0.2$$

$$w_2 = w_2 + \Delta w_2 = 0 + \Delta w_2 = 0 + 0.2 = 0.2$$

Table 10.6: XOR Truth Table



x_1	x_2	Y
0	0	1
0	1	0
1	0	0
1	1	1



$$\text{Training Error} = \frac{1}{2}\sum_{d \in T}(O_{Desired} - O_{Estimated})^2$$

where, T is the training dataset, $O_{Desired}$ and $O_{Estimated}$ are the desired target output and estimated actual output, respectively, for a training instance d.

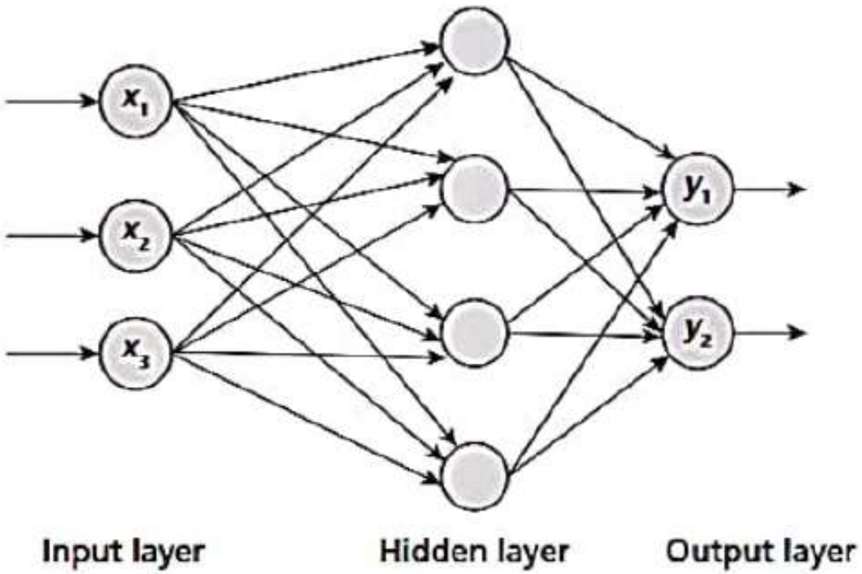


Figure 10.8: Model of a Fully Connected Neural Network

This ANN consists of multiple layers with one input layer, one output layer and one or more hidden layers. Every neuron in a layer is connected to all neurons in the next layer and thus they are fully connected.

The information flows in both the directions. In the forward direction, the inputs are multiplied by weights of neurons and forwarded to the activation function of the neuron and output is passed to the next layer.

If the output is incorrect, then in the backward direction, error is back propagated to adjust the weights and biases to get correct output. Thus, the network learns with the training data. This type of ANN is used in deep learning for complex classification, speech recognition, medical diagnosis, forecasting, etc. They are comparatively complex and slow. The model of an MLP is shown in Figure 10.9.

10.9 POPULAR APPLICATIONS OF ARTIFICIAL NEURAL NETWORKS

ANN learning mechanisms are used in many complex applications that involve modelling of non-linear processes. ANN is a useful model that can handle even noisy and incomplete data. They are used to model complex patterns, recognize patterns and solve prediction problems like humans in many areas such as:

1. Real-time applications: Face recognition, emotion detection, self-driving cars, navigation systems, routing systems, target tracking, vehicle scheduling, etc.
2. Business applications: Stock trading, sales forecasting, customer behaviour modelling, Market research and analysis, etc.
3. Banking and Finance: Credit and loan forecasting, fraud and risk evaluation, currency price prediction, real-estate appraisal, etc.
4. Education: Adaptive learning software, student performance modelling, etc.
5. Healthcare: Medical diagnosis or mapping symptoms to a medical case, image interpretation and pattern recognition, drug discovery, etc.
6. Other Engineering Applications: Robotics, aerospace, electronics, manufacturing, communications, chemical analysis, food research, etc.

1. An ANN requires processors with parallel processing capability to train the network running for many epochs. The function of each node requires a CPU capability which is difficult for very large networks with a large amount of data.

2. They work like a ‘black box’ and it is exceedingly difficult to understand their working in inner layers. Moreover, it is hard to understand the relationship between the representations learned at each layer.

3. The modelling with ANN is also extremely complicated and the development takes a much longer time.

4. Generally, neural networks require more data than traditional machine learning algorithms, and they do not perform well on small datasets.

5. They are also more computationally expensive than traditional learning techniques.

The major challenges while modelling a real-time application with ANNs are:

1.Training a neural network is the most challenging part of using this technique. Overfitting or underfitting issues may arise if datasets used for training are not correct. It is also hard to generalize to the real-world data when trained with some simulated data. Moreover, neural network models normally need a lot of training data to be robust and are usable for a real-time application.

2. Finding the weight and bias parameters for neural networks is also hard and it is difficult to calculate an optimal model.

THANK YOU!