Report on

PRINCIPLES, COMPUTATIONAL TOOLS AND CASE STUDIES WITH DATA SCIENCE AND MACHINE LEARNING

By

ADITYA RATHOR (202100286)

In partial fulfillment of requirements for the award of degree in
Bachelor of Technology in Computer Science and Engineering
(2023)

SMIT SIKKIM MANIPAL UNIVERSITY
SIKKIM MANIPAL INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SIKKIM MANIPAL INSTITUTE OF TECHNOLOGY
(A constituent college of Sikkim Manipal University)
MAJITAR, RANGPO, EAST SIKKIM – 737136

LIST OF CONTENTS

Content **Page No.**

ABSTRACT

This paper delves into the integral aspects of Data Science and Machine Learning, focusing on foundational principles, computational tools, and enlightening case studies. It underscores the significance of principles like data preprocessing, feature engineering, and model selection, highlighting their pivotal role in generating accurate insights. Ethical considerations and bias mitigation strategies are also explored in the context of responsible data-driven decision-making.

The paper then shifts its focus to essential computational tools that drive the effectiveness of these disciplines. Programming languages such as Python and R, along with libraries like TensorFlow and scikit-learn, empower practitioners to manipulate data and construct intricate models. The role of cloud computing platforms in facilitating scalable and efficient processing is also acknowledged. Furthermore, the paper presents a range of compelling case studies spanning diverse domains, revealing the practical applications of Data Science and Machine Learning. From healthcare diagnostics to financial analysis and beyond, these cases demonstrate how predictive algorithms revolutionize industries, optimize strategies, and enhance user experiences.

In conclusion, this paper offers a comprehensive understanding of the symbiotic relationship between principles, tools, and case studies in the realm of Data Science and Machine Learning. It serves as a roadmap for aspiring professionals to navigate these dynamic fields successfully.

MANIPAL UNIVERSITY
JAIPUR
(University under Section 2(f) of the UGC Act)

nirf
RANKINGS 2023
Engineering: 76 | Management: 79 | Law: 29
University: 101-150 (Rank-Band)

A+ NAAC
NAAC A+ Grade
with 3.28 Score

International Summer School - Manipal University Jaipur (ISSMUJ 2023)
26 June – 25 July, 2023

Certificate of Participation

This is to certify that **Aditya Rathor** from **Sikkim Manipal Institute of Technology** has successfully completed the International Summer School-Manipal University Jaipur **(ISSMUJ 2023)** course (equivalent to 03 credits) on **Principles, Computational tools and Case studies with Data Science and Machine Learning** at Manipal University Jaipur.

Organised By

Directorate of International Collaboration (DoIC)

Prof. Santosh Patil
Director, DoIC

Mr. Mohit Jain
Convener

Dr. Sonal Sidana
Co-Convener

CHAPTER 1

INTRODUCTION

- **Who is Data Scientist**

  Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decision

- **What is Big Data**

  Big Data is any data that is expensive to manage and hard to extract value from

  • Volume

  • The size of the data

  • Velocity

  • The latency of data processing relative to the growing demand for interactivity

  • Variety and Complexity

  • the diversity of sources, formats, quality, structures

- **Why Collect Data**

  1. It helps you **learn more about your customers**.
  2. It enables you to **discover trends** in the way people change their opinions and behaviour over time or in different circumstances.
  3. It lets you **segment your audience** into different customer groups and direct different marketing strategies at each of the groups based on their individual needs.
  4. It facilitates **decision making** and improves the quality of decisions made.
  5. It helps **resolve issues and improve the quality** of your product or service based on the feedback obtained.

- **Difference Between Quantitative Data & Qualitative Data**

Quantitative	Qualitative
• Things that are measurable and can be **expressed in numbers** or figures, or using other values that express quantity	• Qualitative data is usually **not easily measurable** as quantitative and can be gained through observation or open-ended survey or interview questions.
• Quantitative data is usually expressed in **numerical form** and can represent size, length, duration, amount, price, and so on	• As quantitative data collection methods usually are rather concerned **with words, sounds, thoughts, feelings, and other non-quantifiable data**, it allows a greater depth of understanding
• Quantitative data is most likely to provide answers to questions such as **who? when? where? what? and how many**	• Qualitative research is most likely to provide answers to questions such as **"why?" and "how?"**
• Quantitative survey questions are in most cases **closed-ended**, thus making the answers easily transformable into numbers, charts, graphs, and tables	• Qualitative data collection methods are most likely to consist of **open-ended** questions and descriptive answers and little or no numerical value

CHAPTER 2

Regression And Classification

DECISION TREE

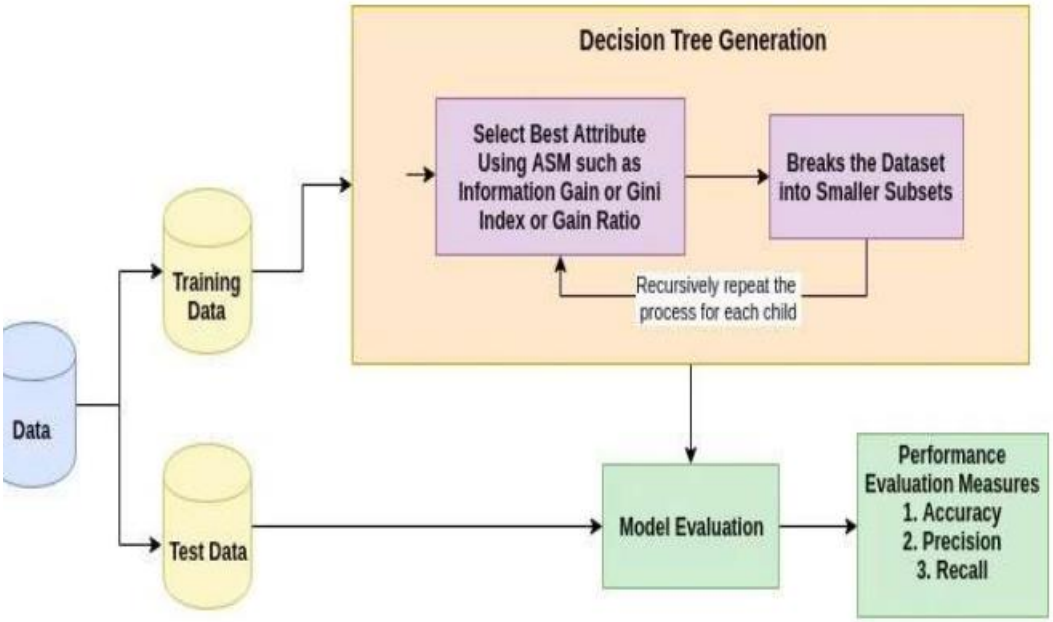
- Decision Tree Mining is a type technique that is used to build Classification Models. It builds classification models in the form of a tree-like structure, just like its name. This type of mining belongs to supervised class learning.
- In supervised learning, the target result is already known. Decision trees can be used for both categorical and numerical data. The categorical data represent gender, marital status, etc. while the numerical data represent age, temperature, etc.

EXAMPLE

How does Decision Tree Algorithm work

The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures (ASM) to split the records.
2. Make that attribute a decision node and breaks the dataset into smaller subsets.
3. Start tree building by repeating this process recursively for each child until one of the conditions will match:
   1. All the tuples belong to the same attribute value.
   2. There are no more remaining attributes.
   3. There are no more instances



CHAPTER 3

CLUSTERING

- **Classification of Machine Learning**

  At a broad level, machine learning can be classified into three types:

  1. Supervised Learning
  2. Unsupervised Learning
  3. Reinforcement Learning

- **SUPERVISED LEARNING**

  - Supervised learning is a type of machine learning method in which we provide sample labelled data to the machine learning system in order to train it, and on that basis, it predicts the output.
  - The system creates a model using labelled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.
  - The goal of supervised learning is to **map input data with the output data.**

  Supervised learning can be grouped further in two categories of algorithms:

  - Classification
  - Regression

Difference Between Classification and Regression:

	Classification	Regression
Description	A regression model seeks to predict a continuous quantity.	A classification model seeks to predict some class label.
Type of Algorithm	Supervised Learning Algorithm	Supervised Learning Algorithm
Type of response variable	Continuous	Categorial
How to assess model fit	Root mean squared error	Percentage of correct classifications

- **UNSUPERVISED LEARNING**

  - Unsupervised learning is a learning method in which a machine learns without any supervision.
  - The training is provided to the machine with the set of data that has not been labelled, classified, or categorized, and the algorithm needs to act on that data without any supervision
  - The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

It can be further classifieds into two categories of algorithms:

- Cluster
- Association

CLUSTERING V/S CLASSIFICATION

CLUSTERING	CLASSIFICATION
• Clustering is an unsupervised learning approach where grouping is done on similarities basis.	• Classification is a supervised learning approach where a specific label is provided to the machine to classify new observations.
• It does not use a training dataset.	• It uses a training dataset.
• objective is to group a set of objects to find whether there is any relationship between them.	• objective is to find which class a new object belongs to form the set of predefined classes.
• less complex as compared to clustering.	• more complex as compared to clustering.

- **RIENFORCEMENT LEARNING**

  - The idea behind Reinforcement Learning is that an **agent** (an AI) will learn from the **environment** by interacting with it (through trial and

error) and receiving rewards (negative or positive) as feedback for performing **actions.**

Reinforcement learning has four essential elements:

1. **Agent** The program you train, with the aim of doing a job you specify.
2. **Environment** The world, real or virtual, in which the agent performs actions.
3. **Action** A move made by the agent, which causes a status change in the environment.
4. **Rewards** The evaluation of an action, which can be positive or negative

Correlation and Regression

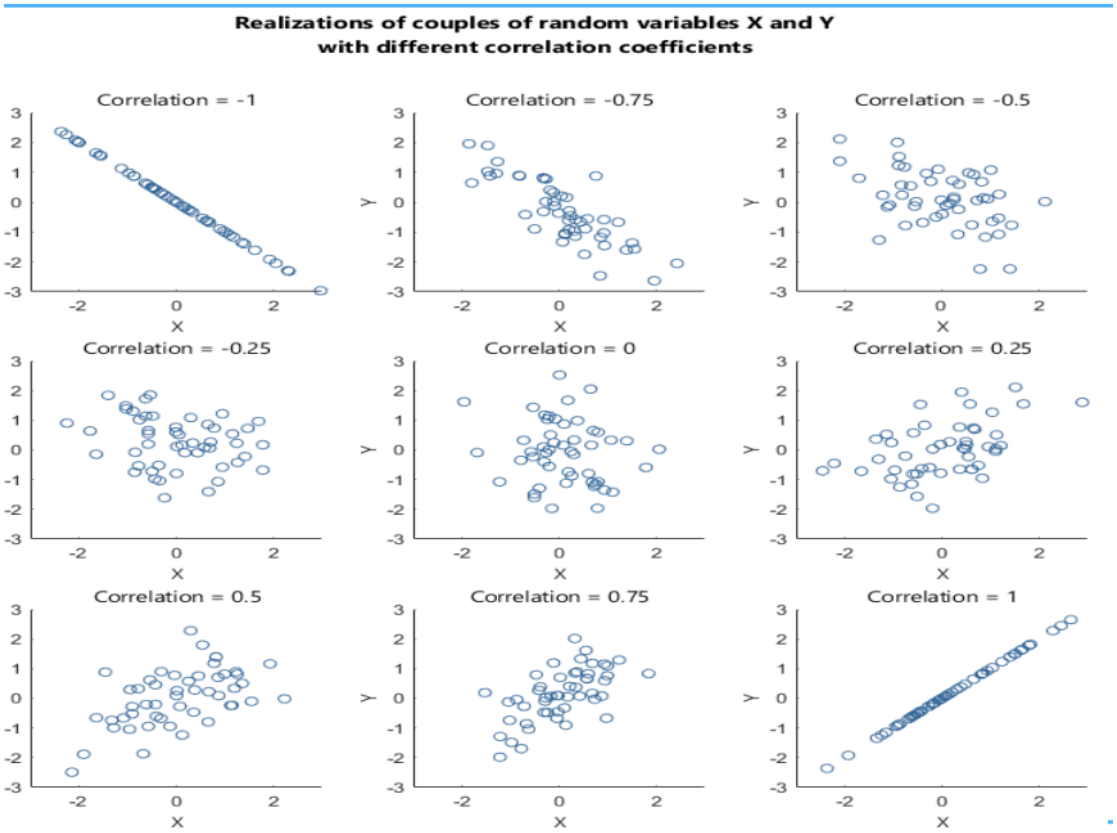
- Correlation and regression are the two most commonly used techniques for investigating the relationship between quantitative variables.
- Correlation is used to give the strength of relationship between the variables whereas linear regression uses an equation to express this relationship.
- Correlation and regression are **statistical measurements** that are used to give a relationship between two variables.
- For example, suppose a person is driving an expensive car then it is assumed that she must be financially well.
- To **numerically quantify this relationship**, correlation and regression are used.

Correlation Coefficient

The correlation coefficient, **r**, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1.

When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables.

Thus, correlation can be positive (direct correlation), negative (indirect correlation), or zero.



Correlation Formula (Pearson Correlation Coefficient)

Correlation shows the relation between two variables. Correlation coefficient shows the measure of correlation. To compare two datasets, we use the correlation formulas

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where n = Quantity of Information

Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value

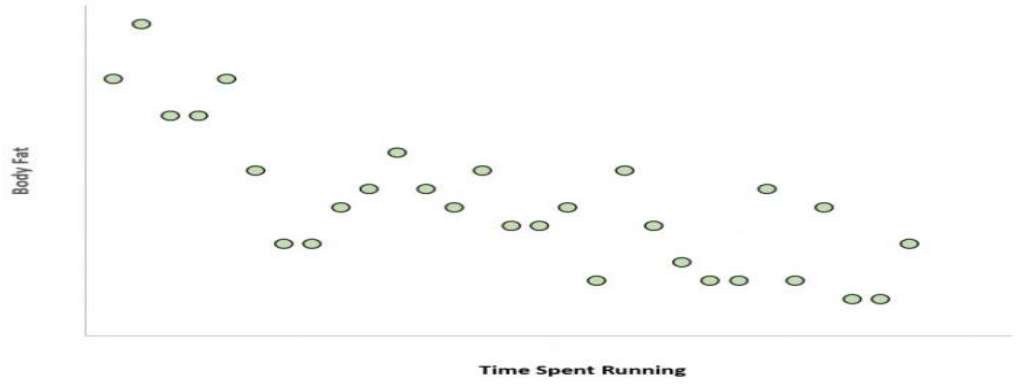
Σxy = Sum of the Product of first & Second Value

Σx^2 = Sum of the Squares of the First Value

Σy^2 = Sum of the Squares of the Second Value

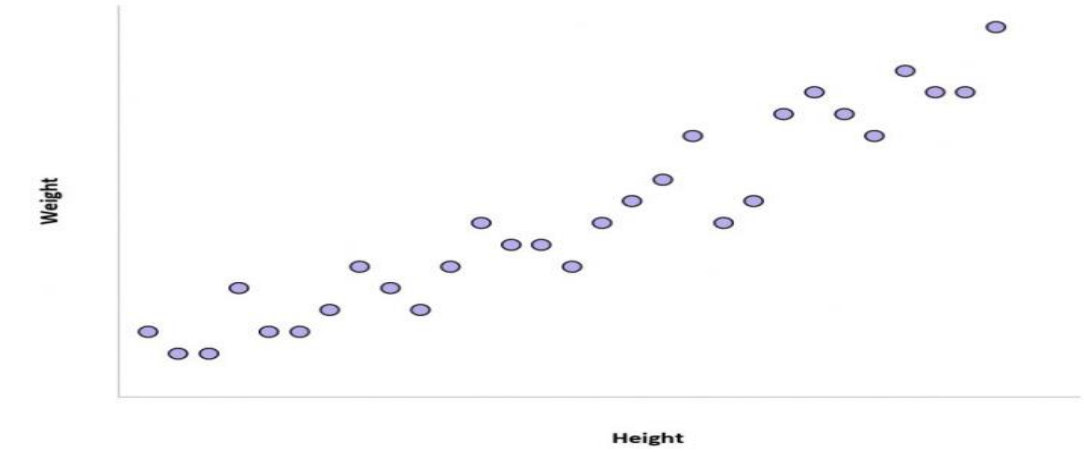
Example : Time Spent Running vs. Body Fat (Negative Correlation)

The more time an individual spends running, the lower their body fat tends to be. In other words, the variable running time and the variable body fat have a negative correlation. As time spent running increases, body fat decreases. If we created a scatterplot of time spent running vs. body fat, it may look something like this:



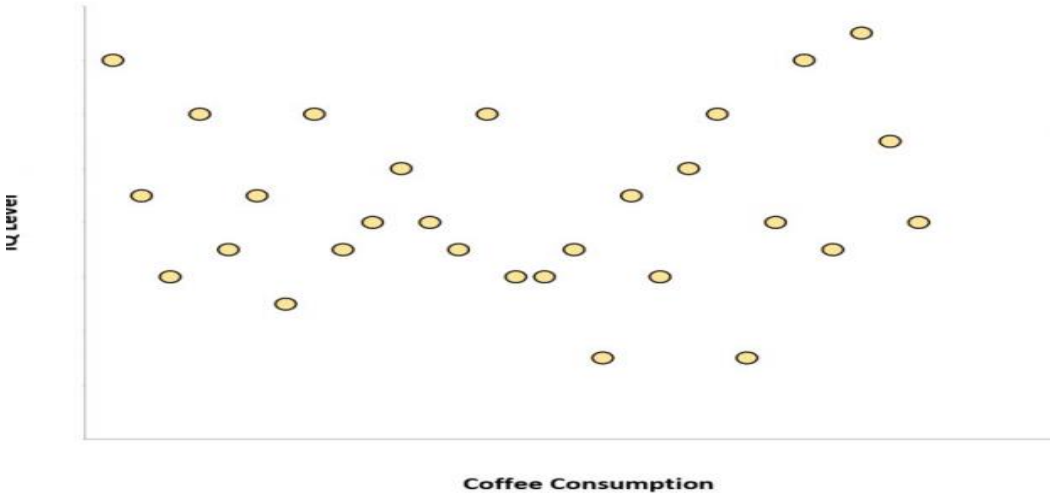
Example : Height vs. Weight (Positive Correlation)

The correlation between the height of an individual and their weight tends to be positive. In other words, individuals who are taller also tend to weigh more. If we created a scatterplot of height vs. weight, it may look something like this:



Example : Coffee Consumption vs. Intelligence (No correlation)

The amount of coffee that individuals consume and their IQ level has a correlation of zero. In other words, knowing how much coffee an individual drinks doesn't give us an idea of what their IQ level might be.



CORRELATION V/S REGRESSION

	Correlation	Linear Regression
Purpose	Description, Inferential Statistics	Prediction, Design Experiment
Statistic	**r**	r, R^2, R^2 -adjust
Variables	Paired	2, 3, more
Variables	No differentiation between the variables.	1 or more independent x's, 1 dependent: y = f(x)
Fits a line thru data	Implicitly	Explicitly: y = a+ bx
Cause & Effect	Does not address	Attempts to show

Simple Linear Regression-Python Implementation

```
import numpy as np
import matplotlib.pyplot as plt
def estimate_coef(x, y):
 # number of observations/points
 n = np.size(x)
# mean of x and y vector
 m_x = np.mean(x)
 m_y = np.mean(y)
 # calculating cross-deviation and deviation about x
 SS_xy = np.sum(y*x) - n*m_y*m_x
 SS_xx = np.sum(x*x) - n*m_x*m_x
 # calculating regression coefficients
 b_1 = SS_xy / SS_xx
```

```
 b_0 = m_y - b_1*m_x
 return (b_0, b_1)
def plot_regression_line(x, y, b):
 # plotting the actual points as scatter plot
 plt.scatter(x, y, color = "r",
 marker = "o", s = 35)
# predicted response vector
 y_pred = b[0] + b[1]*x
# plotting the regression line
 plt.plot(x, y_pred, color = "b")
# putting labels
 plt.xlabel('independent variable-x')
 plt.ylabel('dependent variable-y')
 # function to show plot
 plt.show()
def regression():
 # observations / data
 x = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
 y = np.array([1, 3, 2, 5, 7, 8, 8, 9, 10, 12, 13])
# estimating coefficients
 b = estimate_coef(x, y)
 print("Estimated coefficients:\nb_0 = {} \nb_1 = {}".format(b[0], b[1]))
# plotting regression line
 plot_regression_line(x, y, b)
regression()
```

INFORMATION GAIN

- This method is the main method that is used to build decision trees. It reduces the information that is required to classify the tuples. It reduces the number of tests that are needed to classify the given tuple. The attribute with the highest information gain is selected.
- The original information needed for classification of a tuple in dataset D is given by:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where p is the probability that the tuple belongs to class C. The information is encoded in bits, therefore, log to the base 2 is used. E(s) represents the average amount of information required to find out the class label of dataset D. This information gain is also called **Entropy**.

The information required for exact classification after portioning is given by the formula:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

Where P (c) is the weight of partition. This information represents the information needed to classify the dataset D on portioning by X.

Information gain is the difference between the original and expected information that is required to classify the tuples of dataset D.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Gain is the reduction of information that is required by knowing the value of x. The attribute with the highest information gain is chosen as “best”.

CHAPTER 4

ARTIFICIAL NEURAL NETWORK

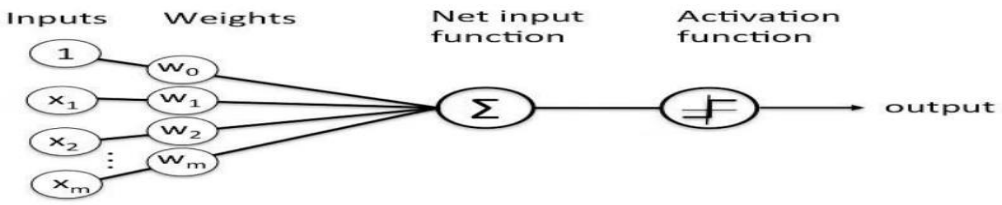
The artificial neuron has the following characteristics:

- A neuron is a mathematical function modelled on the working of biological neurons.
- It is an elementary unit in an artificial neural network.
- One or more inputs are separately weighted.
- Inputs are summed and passed through a nonlinear function to produce output.
- Every neuron holds an internal state called activation signal.
- Every neuron holds an internal state called activation signal.
- Every neuron is connected to another neuron via connection link.

A neural network without an activation function is essentially just a linear regression model. The activation function does the non-linear transformation to the input making it capable to learn and perform more complex tasks.

Perceptron

Perceptron was introduced by Frank Rosenblatt in 1957. He proposed a Perceptron learning rule based on the original MCP neuron. A Perceptron is an algorithm for **supervised learning of binary classifiers**. This algorithm enables neurons to learn and processes elements in the training set one at a time.



Types of Perceptron:

1. Single layer: Single layer perceptron can learn only linearly separable patterns.
2. Multilayer: Multilayer perceptron can learn about two or more layers having a greater processing power.

   The Perceptron algorithm learns the weights for the input signals in order to draw a decision boundary.

   The perceptron is a simplified model of the real neuron that attempts to imitate it by the following process: it takes the **input signals**, let's call them **x1, x2, …, xn**, computes a weighted sum z of those inputs, then passes it through a threshold function ϕ and outputs the result.

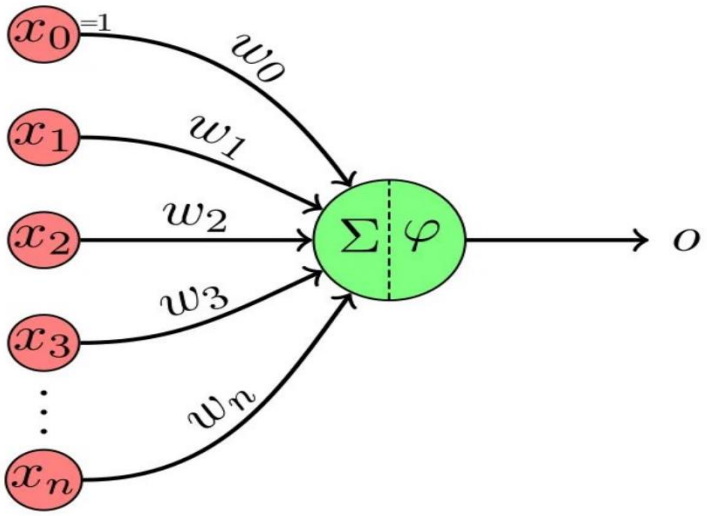
$$z = \sum_{i=1}^{n} x_i w_i$$

$$\phi(z) = \begin{cases} 0 & z \leq w_0 \\ 1 & z > w_0 \end{cases}$$

But having w0 as a threshold is the same thing as adding w0 to the sum as bias and having instead a threshold of 0. That is, we consider an additional input signal x0 that is always set to 1.

$$z = \sum_{i=0}^{n} x_i w_i$$

$$\phi(z) = \begin{cases} 0 & z \leq 0 \\ 1 & z > 0 \end{cases}$$



The bias term is an adjustable, numerical term added to a perceptron's weighted sum of inputs and weights that can increase classification model accuracy.

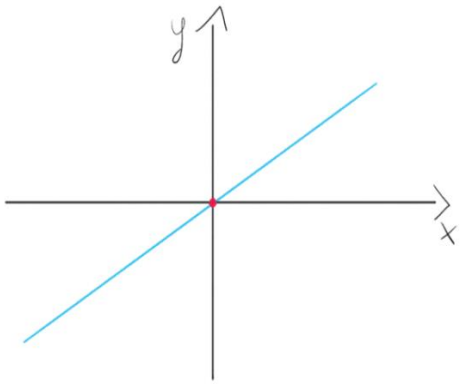
In case i=1 and x_0=1. As a result, such Neural Network is actually a linear regression model:

$$y = x_1 w_1 + w_0$$

Now the crucial part. To understand why we need bias neuron, let's see what happens when there is no bias input at all. It means that there will be only one input x_1 and nothing more:

$$Y = x_1 w_1$$

Such a model is not very flexible. It means that the line needs to go through the point (0, 0). A Slope of the line may change; however, it is tied to the coordinate system's origin. Take a look at this visualization

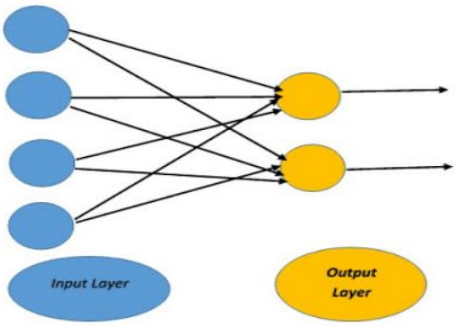


To gain more flexibility we need to get back to the original model with bias. It will equip us with weight w_0, not tied to any input. This weight allows the model to move up and down if it's needed to fit the data.

That's the reason why we need bias neurons in neural networks. Without these bias weights, our model has quite limited "movement" while searching through solution space.

Single Layer Perceptron:

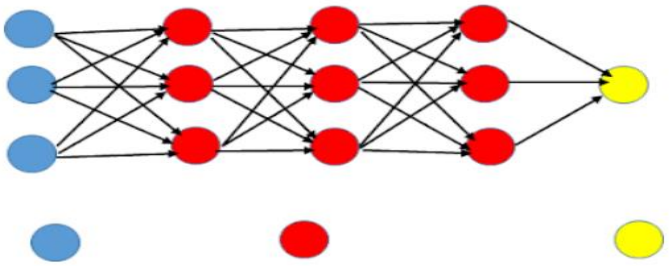
Single Layer Perceptron has just two layers of input and output. It only has single layer hence the name single layer perceptron. It does not contain Hidden Layers as that of Multilayer perceptron.



Input nodes are connected fully to a node or multiple nodes in the next layer. A node in the next layer takes a weighted sum of all its inputs.

Multi-layer Perceptron:

A multilayer perceptron is a type of feed-forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. The MLP network consists of input, output, and hidden layers. Each hidden layer consists of numerous perceptron's which are called hidden layers or hidden unit.



CHAPTER 5

DEEP LEARNING

What is Deep Learning?

Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions.

Why is deep learning important?

Artificial intelligence (AI) attempts to train computers to think and learn as humans do. Deep learning technology drives many AI applications used in everyday products, such as the following:

- Digital assistants
- Voice-activated television remotes
- Fraud detection
- Automatic facial recognition

What are the uses of deep learning?

Deep learning has several use cases in automotive, aerospace, manufacturing, electronics, medical research, and other fields. These are some examples of deep learning:

- Self-driving cars use deep learning models to automatically detect road signs and pedestrians.
- Defense systems use deep learning to automatically flag areas of interest in satellite images.

- Medical image analysis uses deep learning to automatically detect cancer cells for medical diagnosis.
- Factories use deep learning applications to automatically detect when people or objects are within an unsafe distance of machines.

Classification Of Neural Network:

1. **Shallow Neural Network:** It has only one hidden layer between the input and output.
2. **Deep Neural Network:** It has more than one layer.

Types Of Deep Learning Network:

1. **Feed-forward Neural Networks**: The feedforward neural network has an open loop but the feedback neural network has a closed loop. Input is more essential in a feedforward network system whereas the output is the most essential part of a feedback network system.
2. **Recurrent Neural Networks**: It is a type of Neural Network where the output from the previous step is fed as input to the current step which solved the issue with the help of a Hidden Layer. The main and most important feature of RNN is its Hidden state, which remembers some information about a sequence.
3. **Convolutional Neural Network**: A Convolutional Neural Network, also known as CNN or ConvNet, is a class of neural networks that specializes in processing data that has a grid-like topology, such as an image. A digital image is a binary representation of visual data. It contains a series of pixels arranged in a grid-like fashion that contains pixel values to denote how bright and what colour each pixel should be.
4. **Deep Reinforcement Learning:** It is a subfield of machine learning that combines reinforcement learning (RL) and deep learning. RL considers the problem of a computational agent learning to make decisions by trial and error.

   EXAMPLE: Self-Driving Cars

Difference Between Machine Learning & Deep Learning

Machine learning uses algorithms to parse data, learn from that data, and make informed decisions based on what it has learned.

Deep learning structures algorithms in layers to create an “artificial neural network” that can learn and make intelligent decisions on its own. Deep learning is a subset of machine learning.

CHAPTER 6

TABLEAU

What is Data Visualization?

1. Visual Representation of Data.
2. For exploration, discovery, insight, ..
3. Interactive component provides more insight as compared to a static image.

Why do we need data visualization?

1. A visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows on a spreadsheet.
2. Since the purpose of data analysis is to gain insights, data is much more valuable when it is visualized.
3. Even if a data analyst can pull insights from data without visualization, it will be more difficult to communicate the meaning without visualization
4. Charts and graphs make communicating data findings easier even if you can identify the patterns without them.

Data visualization is the representation of data or information in a graph, chart, or other visual format.

It communicates relationships of the data with images.

The representation and presentation of data that exploits our visual perception abilities in order to amplify cognition

Taxonomy

- The representation of data is the way you decide to depict data through a choice of physical forms. Whether it is via a line, a bar, a circle, or any other visual variable, you are taking data as the raw material and creating a representation to best portray its attribute.
- The presentation of data goes beyond the representation of data and concerns how you integrate your data representation into the overall communicated work, including the choice of colours, annotations, and interactive features.
- Exploiting our visual perception abilities relates to the scientific understanding of how our eyes and brains process information most effectively, as we've just discussed. This is about harnessing our abilities with spatial reasoning, pattern recognition, and big-picture thinking.
- Amplify cognition is about maximizing how efficiently and effectively we are able to process the information into thoughts, insights, and knowledge. Ultimately, the objective of data visualization should be to make a reader or users feel like they have become better informed about a subject.

Six principles of graphical integrity

- Representations of numbers should match their true proportions.
- Labelling should be clear and detailed..
- Designs should not vary from some ulterior motive, but show only data variations
- Well known units are best when representing money.
- The number of dimensions represented should be the same as the number of dimensions in the data.
- Representations should not imply an unintended context.

CHAPTER 7

SUMMARY AND CONCLUSION

SUMMARY OF ACHIEVEMENT

successfully delved into the multifaceted realm of Data Science and Machine Learning, uncovering the essential pillars that drive these transformative disciplines. Mastery over foundational principles, including data preprocessing, feature engineering, and model selection, has enabled the creation of accurate and impactful insights. Proficiently navigated through a diverse set of computational tools, leveraging programming languages like Python and utilizing libraries such as TensorFlow and scikit-learn for constructing intricate models.

Furthermore, adeptly dissected real-world case studies spanning industries like healthcare, finance, and marketing, showcasing the practical application of predictive algorithms to revolutionize strategies and improve decision-making. Effectively grasped the symbiotic relationship between principles, tools, and case studies, demonstrating a comprehensive understanding of the dynamic interplay within Data Science and Machine Learning. These achievements have solidified the capacity to contribute to data-driven innovation and informed problem-solving across various domains.

MAIN DIFFICULTY ENCOUNTERED

1. **Complex Concepts and Algorithms**: Understanding intricate concepts like neural networks, ensemble methods, and deep learning architectures posed initial challenges. Overcoming this required investing extra time in thorough research and seeking clarification from experts.
2. **Algorithm Selection and Tuning:** Choosing the right algorithms and hyperparameters for specific tasks proved to be a trial-and-error process. It required experimentation and a deep understanding of the problem domain to achieve optimal performance.
3. **Keeping Pace with Advancements**: The rapid evolution of Data Science and Machine Learning presented the challenge of staying updated with new techniques, libraries, and best practices. Continuous learning and adaptation were necessary to remain effective in the field.

Despite these difficulties, each challenge presented an opportunity for growth. Overcoming them fostered a deeper understanding of the intricacies of Data Science and Machine Learning, ultimately enhancing problem-solving skills and the ability to contribute meaningfully to the field.

CONCLUSION

The odyssey through Data Science and Machine Learning culminates in a synthesis of principles, vital tools, and tangible impact. Challenges metamorphosed into growth opportunities, while real-world cases solidified transformative potential. Armed with this journey's insights, we stand prepared for ongoing evolution, poised to innovate and contribute meaningfully to dynamic landscapes.

CHAPTER 8
REFERENCES

REFERENCES

1. Dr. Deevesh Choudhary, MUJ
2. Dr. Shaumik Tiwari, UPES, Dehradun
3. Dr. Aprna Tripathi, MUJ