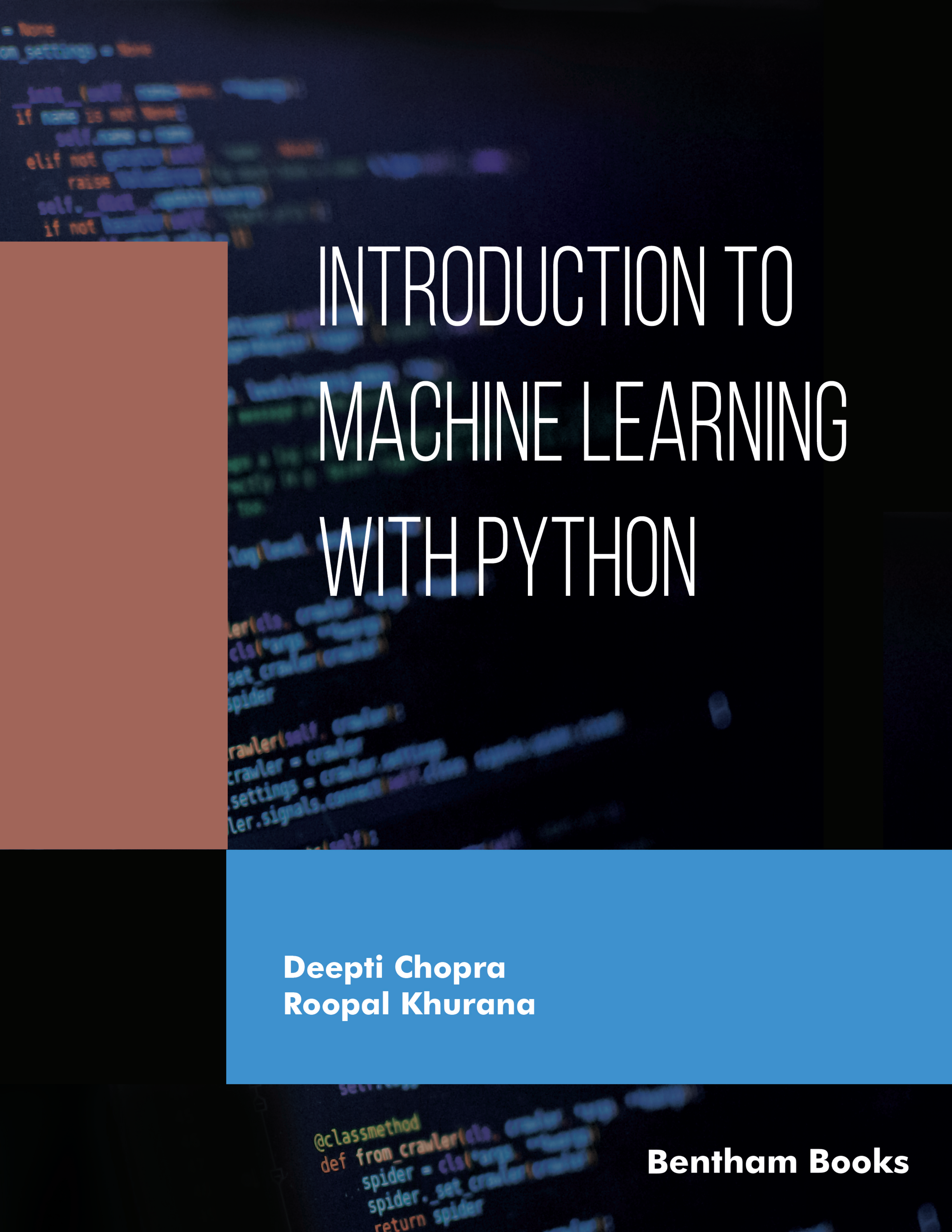

INTRODUCTION TO MACHINE LEARNING WITH PYTHON

Deepti Chopra
Roopal Khurana

Bentham Books

Introduction to Machine Learning with Python

Authored By

Deepti Chopra
Jagan Institute of Management Studies,
Sector 5, Rohini, Delhi-110085,
India

&

Roopal Khurana
Railtel Corporation of India Ltd,
IT Park, Shastri Park,
Delhi-110053,
India

Introduction to Machine Learning with Python

Authors: Deepti Chopra and Roopal Khurana

ISBN (Online): 978-981-5124-42-2

ISBN (Print): 978-981-5124-43-9

ISBN (Paperback): 978-981-5124-44-6

First published in 2023.

need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.

3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Pte. Ltd.
80 Robinson Road #02-00
Singapore 068898
Singapore
Email: subscriptions@benthamscience.net

CONTENTS

FOREWORD i

PREFACE ii
- **CONSENT FOR PUBLICATION** ii
- **CONFLICT OF INTEREST** ii
- **ACKNOWLEDGEMENT** ii

CHAPTER 1 INTRODUCTION TO PYTHON 1
- **INTRODUCTION** 1
  - Web Development 1
  - Game Development 2
  - Artificial Intelligence and Machine Learning 2
  - Desktop GUI 2
- **SETTING UP PYTHON ENVIRONMENT** 2
  - Steps Involved In Installing Python On Windows Include The Following: 3
  - Steps involved in installing Python on Macintosh include the following 3
  - Setting Up Path 3
  - Setting Up Path In The Unix/linux 3
- **WHY PYTHON FOR DATA SCIENCE?** 4
- **ECOSYSTEM FOR PYTHON MACHINE LEARNING** 5
- **ESSENTIAL TOOLS AND LIBRARIES** 6
  - Jupyter Notebook 6
    - *Pip Install Jupiter* 7
  - NumPy 7
  - Pandas 8
  - Scikit-learn 9
  - SciPy 9
  - Matplotlib 10
  - Mglearn 10
- **PYTHON CODES** 10
- **CONCLUSION** 13
- **EXERCISES** 13
- **REFERENCES** 14

CHAPTER 2 INTRODUCTION TO MACHINE LEARNING 15
- **INTRODUCTION** 15
- **DESIGN A LEARNING SYSTEM** 16
  - Selection Of Training Set 17
  - Selection Of Target Function 17
  - Selection Of A Function Approximation Algorithm 17
- **PERSPECTIVE AND ISSUES IN MACHINE LEARNING** 18
  - Issues In Machine Learning 19
    - *Quality of Data* 19
    - *Improve the Quality of Training* 20
    - *Overfitting the Training Data* 20
    - *Machine Learning Involves A Complex Process* 20
    - *Insufficient training data* 20
  - Feasibility of Learning An Unknown Target Function 20
    - *Collection of Data* 21
    - *Pre-processing of Data* 21
    - *Finding The Model That Will Be Best For The Data* 22

Training and Testing Of The Developed Model Evaluation 22
In Sample Error and Out of Sample Error 22
APPLICATIONS OF MACHINE LEARNING 23
Virtual Personal Assistants 23
Traffic Prediction 23
Online Transportation Networks 23
Video Surveillance System 24
Social Media Services 24
People you May Know 24
Face Recognition 24
Similar Pins 24
Sentiment Analysis 25
Email Spam and Malware Filtering 25
Online Customer Support 25
Result Refinement of a Search Engine 25
Product Recommendations 26
Online Fraud Detection 26
Online Gaming 26
Financial Services 26
Healthcare 26
Oil and Gas 27
Self-driving Cars 27
Automatic Text Translation 27
Dynamic Pricing 27
Classification of News 27
Information Retrieval 28
Robot Control 28
CONCLUSION 28
EXERCISES 28
REFERENCES 28

CHAPTER 3 LINEAR REGRESSION AND LOGISTIC REGRESSION 30
INTRODUCTION 30
LINEAR REGRESSION 30
Linear Regression In One Variable 32
Linear Regression In Multiple Variables 32
Overfitting and Regularization In Linear Regression 33
GRADIENT DESCENT 33
POLYNOMIAL REGRESSION 34
Features of Polynomial Regression 34
LOGISTIC REGRESSION 35
Overfitting and Regularisation in Logistic Regression 35
BINARY CLASSIFICATION AND MULTI-CLASS CLASSIFICATION 35
Binary Classification Tests 38
Classification Accuracy 38
Error Rate 38
Sensitivity 38
Specificity 38
PYTHON CODES 39
CONCLUSION 57
EXERCISES 57

REFERENCES 57

CHAPTER 4 SUPPORT VECTOR MACHINE 58
INTRODUCTION 58
SUPPORT VECTOR CLASSIFICATION 58
The Maximal Margin Classifier 61
Soft Margin Optimization 62
Linear Programming Support Vector Machines 63
SUPPORT VECTOR REGRESSION 63
Kernel Ridge Regression 64
Gaussian Processes 64
APPLICATIONS OF SUPPORT VECTOR MACHINE 64
Text Categorisation 65
Image Recognition 65
Bioinformatics 65
PYTHON CODE 65
CONCLUSION 72
EXERCISES 73
REFERENCES 73

CHAPTER 5 DECISION TREES 74
INTRODUCTION 74
REGRESSION TREES 75
STOPPING CRITERION AND PRUNING LOSS FUNCTIONS IN DECISION TREE 75
CATEGORICAL ATTRIBUTES, MULTIWAY SPLITS AND MISSING VALUES IN DECISION TREES 76
ISSUES IN DECISION TREE LEARNING 76
Preventing Overfitting of Data 76
Incorporating Continuous Valued Attributes 77
Other Measures for Attributes Selection 78
Handling Missing Values 79
Handling of Attributes with Differing Costs 79
INSTABILITY IN DECISION TREES 80
PYTHON CODE 80
CONCLUSION 81
EXERCISES 81
REFERENCES 81

CHAPTER 6 NEURAL NETWORK 83
INTRODUCTION 83
EARLY MODELS 83
PERCEPTRON LEARNING 84
BACKPROPAGATION 85
AN ILLUSTRATIVE EXAMPLE: FACE RECOGNITION 85
STOCHASTIC GRADIENT DESCENT 86
ADVANCED TOPICS IN ARTIFICIAL NEURAL NETWORK 87
Alternative Error Functions 87
Alternative Error Minimization Mechanism 87
Recurrent Networks 88
Dynamically Modifying Network Structures 88
PYTHON CODES 89
CONCLUSION 95

EXERCISES 95
REFERENCES 96

CHAPTER 7 SUPERVISED LEARNING 97
INTRODUCTION 97
USING STATISTICAL DECISION THEORY 97
Gaussian or Normal Distribution 98
Conditionally Independent Binary Components 99
LEARNING BELIEF NETWORKS 100
NEAREST-NEIGHBOUR METHODS 100
CONCLUSION 101
EXERCISES 102
REFERENCES 102

CHAPTER 8 UNSUPERVISED LEARNING 103
INTRODUCTION 103
CLUSTERING 103
K-means Clustering 104
Hierarchical Clustering 104
Principal Component Analysis (PCA) 104
PYTHON CODE 105
CONCLUSION 111
EXERCISES 111
REFERENCES 111

CHAPTER 9 THEORY OF GENERALISATION 113
INTRODUCTION 113
BOUNDING THE TESTING ERROR 113
VAPNIK CHERVONENKIS INEQUALITY 114
PROOF OF VC INEQUALITY 114
CONCLUSION 115
EXERCISES 115
REFERENCES 115

CHAPTER 10 BIAS AND FAIRNESS IN ML 116
INTRODUCTION 116
HOW TO DETECT BIAS? 116
HOW TO FIX BIASES OR ACHIEVE FAIRNESS IN ML? 118
CONFIDENCE INTERVALS 119
HYPOTHESIS TESTING 119
COMPARING LEARNING ALGORITHMS 120
CONCLUSION 121
EXERCISES 121
REFERENCES 121

APPENDIX 123
CONCLUSION 123

SUBJECT INDEX 124

FOREWORD

I take the opportunity to congratulate the authors, Dr. Deepti Chopra and Mr. Roopal Khurana who have written this book titled, **“Introduction to Machine Learning With Python”**.

The advancement in technology in the past decade has been due to the introduction of Machine Learning. Today, machine learning has escalated Artificial Intelligence Revolution, be it in Fraud Detection and Prevention, Self-driving cars, Recommendation Systems, Facial Recognition technology, *etc.*

Machine Learning is one of the approaches of Artificial Intelligence in which Machines become capable of drawing intelligent decisions like humans by learning from their past experiences. In classical methods of Artificial Intelligence, step-by-step instructions are provided to the machines to solve a problem. Machine learning combines classical methods of Artificial Intelligence with the knowledge of the past to gain human-like intelligence.

The authors of this book have given explanations on Machine Learning with Python from the basics to the advanced level so as to assist beginners in building a strong foundation and developing practical understanding.

Beginners with zero or little knowledge about Machine Learning can gain insight into this subject from this book. This book explains Machine Learning concepts using real-life examples implemented in Python.

After learning from this book, one will be able to apply concepts of Machine Learning to real-life problems.

I am sure readers will benefit from this book and gain a lot in the field of machine learning.

Happy Reading!!

Best regards,

Rajesh Pokhriyal | Scientist 'D'
Indian Computer Emergency Response Team (CERT-In)
Ministry of Electronics & IT
Electronics Niketan 6 CGO Complex Lodhi Road
New Delhi 110003

PREFACE

Machine learning has become part and parcel of day-to-day private/non-profit/business and government operations because of its ability to grasp automatically through past experiences without being explicitly programmed. Today, machine learning has conquered the entire industry due to its numerous applications ranging from digital marketing to space research. Today, it governs the industry in terms of building high-tech products, ranking web searches, building speech recognition systems, recommendation systems, *etc.* However, we have not yet developed fully operational machines that give judgments on their own like humans but it is not far away to reach that level. From this book, we intend to re-discover the core concepts of Machine learning paradigms along with numerous architectures and algorithms used in different paradigms. The book elaborates on various topics related to the implementation side using Python with real-life examples. The book can kickstart your career in the field of Machine Learning. It also provides the basic knowledge of Python which is a prerequisite of this course. We can say that this book is meant for neophyte users who wish to get acquainted with the implementation of machine learning using Python. The reader will be able to read well-explained examples and exercises and it will be an ideal choice for Machine Learning enthusiasts. The book presents detailed practice exercises for offering a comprehensive introduction to machine learning techniques along with the basics of Python. The book leverages algorithms of machine learning in a unique way of describing real-life applications. Though not mandatory, some experience with subject knowledge will fasten the learning process.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The author declares no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENT

Declared none.

Deepti Chopra
Jagan Institute of Management Studies
Sector 5, Rohini, Delhi-110085
India

&
Roopal Khurana
Railtel Corporation of India Ltd
IT Park, Shastri Park
Delhi-110053
India

CHAPTER 1

Introduction to Python

Abstract: Python is considered one of the most simple and efficient programming languages. Its object-oriented programming approach and elegant syntax make it a powerful programming language. Python is an interpreted language. Its dynamic typing and high level data structures make it an ideal language for application development in various areas and on multiple platforms. Today, Python is widely used in the areas of machine learning and data science. The following chapter discusses Python, the utility of Python in machine learning and data science, ecosystem of Python in machine learning and various libraries in Python required for machine learning.

Keywords: Data science, Jupyter, Machine learning, Matplotlib, Numpy, Python, Scikit learn, SciPy.

INTRODUCTION

Python was developed by Guido van Rossum in 1990s. The name of the language ‘Python’ was taken from “Monty Python’s Flying Circus”, which was one of the favorite TV shows of Guido van Rossum. Python has a simple syntax that was designed as a language that could be used easily by beginners yet proven to be one of the most powerful languages for advanced developers. Python is an object-oriented programming language that can be used on various platforms. The syntax used in Python is very simple as compared to other programming languages [1]. Today, Python is considered a very popular programming language among students, researchers, developers, *etc.* Python is extensively used by tech giants such as Netflix, Facebook, Google, *etc.* Python offers numerous applications [2], [3]. These include the following:

Web Development

Nowadays, Python is used widely in web development. Some of the frameworks for web development in Python are: Django, Pyramid, Flask, *etc.* These frameworks are known to incorporate characteristics such as scalability, flexibility, security, *etc.*

Deepti Chopra & Roopal Khurana

Game Development

PySoy and PyGame are two python libraries that are used for game development.

Artificial Intelligence and Machine Learning

There are a large number of open-source libraries which can be used while developing AI/ML applications.

Desktop GUI

Desktop GUI offers many toolkits and frameworks using which we can build desktop applications. PyQt, PyGtk, PyGUI are some of the GUI frameworks.

Today, Python is used extensively for doing research especially in the areas of bioinformatics, mathematics, biology, *etc.* It is a part of Computer Science curriculum for many universities.

It is not just companies that seek through python. Python is used in various fields such as Artificial Intelligence, Astronomy, Internet of Things and Social Science.

In this chapter, we will discuss Python, set up Python environment and the importance of using Python in Data Science. We will also discuss tools and libraries used in Python Programming.

SETTING UP PYTHON ENVIRONMENT

Python is available on different platforms such as Windows, Linux and Mac OS X. We can open Window terminal and type "python" ; this will return the version of python if it is already installed.

Current documentation, source code, news and updated version of Python are available at: https://www.python.org/

We may download documentation of python in different formats such as PDF, HTML and PostScript format from https://www.python.org/doc/.

For installing Python, we need to download the binary code according to our platform. If binary code for our platform is not available, then we need to compile the code on c compiler manually.

Steps involved in installing Python on Unix/Linux include the following:

Check if python is already installed on machine by going to terminal using Ctrl+Alt+T. For Python2, type python —version and For Python3.x, type python3.x —version. In case, Python is already installed, then the version of Python installed is returned.

If Python is not installed then follow the following steps:

- Open the URL, https://www.python.org/downloads/.
- Download and extract files from zipped code available for Linux/Unix.
- Execute ./configure script
- Make, make install

The above steps install python libraries at /usr/local/lib/pythonYY. Here 'YY' represents the version of Python installed.

Steps Involved In Installing Python On Windows Include The Following:

- Open the URL, https://www.python.org/downloads/.
- Click on the link python-PQR.msi file and download it. Here, 'PQR' refers to the version of python we wish to install.
- Run the file and this installs python.

Steps involved in installing Python on Macintosh include the following

- Open the URL, https://www.python.org/downloads/.
- MacPython is used for older version of Mac; for Mac which are released before 2003.

Setting Up Path

The executable files and programs may be present in different directory locations. Path consists of a list of directories that comprise executable files that may be searched by the Operating System. Unix is case-sensitive and Windows is not case-sensitive. So, path is 'PATH' in Unix and 'path' in Windows.

Setting Up Path In The Unix/linux

Add python directory to the path using following ways:

- In csh shell, type set env PATH "$PATH:/usr/local/bin/python"
- In bash shell, type export PATH="$PATH:/usr/local/bin/python"
- In ksh shell, type PATH="$PATH:/usr/local/bin/python"

We can invoke python using different ways. One way to invoke python is by typing "python" at the shell command prompt. We may also type "help",

CHAPTER 2

Introduction To Machine Learning

Abstract: Machine Learning is referred to as the subset of Artificial Intelligence. It involves a machine being learned without programming it explicitly. In Machine Learning, machines try to improve their performance with the help of past experiences or by using training examples. The following chapter discusses Machine Learning, Selection of training set, Selection of target function, Selection of Function Approximation Algorithm, Perspectives and issues involved while building a machine learning system,In-sample and Out-of-sample error and Applications of Machine Learning.

Keywords: Data science, Jupyter, Machine learning, Matplotlib, Numpy, Python, Scikit learn, SciPy.

INTRODUCTION

Machine Learning (ML) is a subset of Artificial Intelligence. It is a field in computational intelligence that involves the analysis and interpretation of structures and patterns present in the data and the machine is able to do learning, decision making and reasoning with these patterns and be able to draw intelligent solutions without human intervention. Using ML, a user can input huge amount of data, perform training, give test data and get the automatic result. If there are any errors identified in the result, then machine learning algorithm can make changes, correct errors and use this information to do better decision making in future.

Machine learning comprises 3 components:

- The computational algorithm, expert in doing Machine Learning task.
- Variables and features based on which decisions may be taken.
- Actual or human expected output to determine the accuracy of the system.

Initially, machine learning model is provided with training data for which the result is known. Machine Learning algorithm is executed and adjustments are accomplished till the result of ML model is exactly similar to the actual model. Once, the results of ML model is as desired, then the ML model is said to be

Deepti Chopra & Roopal Khurana

trained and testing data is fed as input to this trained ML model to get human like intelligent results.

Machine Learning basically comprises pattern matching as well as data exploration in order to obtain automatic result with least human intervention.

For any business or corporate world, data is a lifeline. Data driven decisions made by any organisation decide its fate in terms of organisation's growth, progress, sale, *etc.* Machine learning driven decisions if applied correctly, can escalate organisation's growth and enable it to excel compared to its competitive counterparts.

Machine Learning applications are developed in various areas such as computer vision, natural language processing, pattern recognition and image processing. Machine Learning applications are being built in various industries and sectors such as healthcare, automotive, finance, innovation, travel, utilities, energy, hospitality, life science, feedstock *etc.*

Machine learning applications are advancing with time and they have become indispensable in organisations for taking quick and intelligent decisions with no or least human intervention. Machine learning systems are used in: Recommendation systems in e-commerce system, social networking sites or websites as chatbot, displaying advertisement on the basis of user-liked content, in healthcare, medical diagnosis, self-driving cars, *etc.*

DESIGN A LEARNING SYSTEM

Machine Learning is a process in which a machine gets trained from the training data fed into it and the machine is able to improve its performance from past experiences and can give human-like intelligent output [1]. When training data is fed to a Machine Learning model, the Machine learning algorithm will develop a mathematical model. Using this mathematical model, when test data is sent to the machine learning model, it provides efficient results. For example, in designing a driverless car, training of the car is done on how to drive in different road conditions, what should be the speed and how to stop at a signal or when there is an obstacle. Also, the more the training is performed on Machine learning systems, the more accurate and desirable the results are obtained.

Steps involved in designing a learning system include the following:

- Selection of Training Set
- Selection of Target Function
- Selection of Function Approximation Algorithm.

We will discuss all the above-mentioned steps in detail in sub-sections.

Selection Of Training Set

The first step in designing a learning system is to design a training set. This training set will be used for training the system. The success of a machine learning model in terms of its accuracy depends on the quality and quantity of training set that is sent as an input to this model. Apart from training data, a machine learning model which is an artificial intelligence system must also learn from its experiences and use this knowledge to improve the quality of output in the future. For example, in a chess game, if a particular move leads to the losing situation, then a machine learning system must learn from this experience and suggest an alternative move that would lead to the winning situation. When a machine learning system has many training examples as well as feedback or result of each game; then a machine learning system will be able to train itself better and be able to depict good performance in the consecutive games. Thus, the performance of a machine learning system will escalate only when we input numerous examples considering all situations or cases and experiences on a machine learning system.

Selection Of Target Function

The next step in designing a machine learning system is to select a target function. The selection of a target function means that using a machine learning algorithm, a machine learning system is well equipped to take the next course of action to be performed. For example, in a game of chess, a machine learning system would be able to take decision to make the next intelligent move that would lead to a winning situation out of numerous available moves.

Selection Of A Function Approximation Algorithm

Choosing a training data will not decide the optimised moves taken by the machine learning systems. There needs to be the presence of numerous examples while training any system, of the outcome of this training whether success or failure is fed to the machine learning system as feedback. The success rate of the move helps the machine learning algorithm in deciding next optimised move to be taken and improve its performance with the least human intervention. For example. Deep Blue is the first supercomputer made by IBM that is based on Machine Learning concept and it won the chess game played against world chess champion Garry Kasparov in 1997.

CHAPTER 3

Linear Regression and Logistic Regression

Abstract: Supervised learning is a machine learning task of mapping the input to the output on the basis of labeled input-output example pairs. Supervised learning may be of two types: classification and regression. In this chapter, we will discuss linear regression in one variable, linear regression in multiple variables, gradient descent, and polynomial regression.

Keywords: Data science, Jupyter, Machine learning, Matplotlib, Numpy, Python, Scikit learn, SciPy.

INTRODUCTION

Supervised Learning involves training a machine using a labelled dataset so that the machine exhibits human-like intelligent behaviour. The training data in supervised learning comprises input data and corresponding output data. A Supervised Learning Algorithm is capable of performing mapping of a given input variable with the corresponding or related output variable. It involves two approaches: regression and classification. Regression is a supervised learning approach in which output in the form of real values is predicted. Classification is a supervised learning approach in which output in the form of discrete values is predicted.

LINEAR REGRESSION

Linear regression is a technique to depict the relationship between an independent variable x and a dependent variable y. Linear regression states that the relationship that exists between one or more input features and the relative output or target vector is approximately linear in nature. Linear regression finds the weighted sum of the input features along with the constant referred to as bias term or intercept [1]. Linear regression has numerous real-life applications. These applications fall into two categories:

If the application comprises forecasting, prediction, or error reduction, then linear regression may be applied to the data set values and make predictions in response.

Deepti Chopra & Roopal Khurana

When there is a need for variations in the response, then it may be attributed to the presence of other explanatory variables. Linear regression is used to find possible relationships between the variables in the field of behavioral, biological, and social sciences. In linear regression with one variable, the hypothesis is defined as:

$$h_\theta (x) = \theta_0+\theta_1*x$$

Here, *x* is referred to as an independent variable on which our hypothesis depends. For example, '*Rainfall*' measured in mm could be *x* and '*The Number of Umbrellas sold*' could be the hypothesis that we are trying to predict. θ_0 and θ_1 are referred to as the bias variable and weight variable, respectively and they together constitute the weight matrix.

Cost function is an equation that gives an estimate of how close we are to the hypothesis. The smaller the value of cost function, the closer we are from the required curve. So, we try to minimize the cost function in order to reduce errors. The mean squared error cost function is defined as follows:

$$J(\theta_0, \theta_1) \; = \; 1/2m \sum i = 1 \; to \; \; m(h_\theta(x^i) - y^i)^2$$

Here, *J* is a cost function, *m* refers to the number of data points in our data set, and *y* refers to the actual values that we will like to predict.

Linear regression is a part of predictive modelling in which the value of output corresponding to a given input depends on the previous values of data.

Our goal is to estimate the fit of the line. Best fit means that the accuracy in prediction is more and chances of error are least.

Types of Linear Regression include the following:

- Simple Linear Regression or Linear Regression in one variable- It comprises one independent variable. *E.g.* price of a new house to be purchased depends on its size.
- Multiple Linear Regression or Linear Regression in multiple variable- It comprises multiple independent variables. *E.g.* price of a new house to be purchased depends on multiple factors such as its size, economy, area or society *etc.*

Applications of Linear Regression:

1. It may be used for forecasting business trends and analysis.

2. It may be used for analyzing sales, marketing promotions, pricing, forecasts, *etc.*

3. It is used predominantly in the field of economics for predicting and analyzing the amount of export, expenditure, labour demand and supply, inventory investments, *etc.*

4. In a given biological system, a Linear Regression may be used for modelling causal relationships among various parameters.

Linear Regression In One Variable

Linear Regression is a simple linear regression that comprises one independent variable. Hypothesis function hθ helps in mapping input data with output data. The cost function is a square error function used in regression that helps in determining the difference accurately.

Gradient Descent algorithm helps in reducing the cost function value. In the regression problem, we try to map input values with the corresponding continuous output function. Linear Regression in one variable is also referred to as univariate linear regression. In Linear Regression using one variable, for a given one input x, there is a prediction of only one output y. In Linear Regression, the best fit line or regression line can be found by computing delta which refers to the difference between the actual data points and the line. These errors are squared and then summed. The line with least value of sum of squared errors is chosen as the regression line. In a Linear Regression graph, the independent variable is plotted on x-axis and the dependent variable is plotted on y-axis.

Linear Regression In Multiple Variables

Linear regression in multiple variables explains the relationship between a dependent variable y and many independent variables [2, 3, 4].

Instead of a 'Yes' or 'No' reply, the value of Y will be a number. It is a continuous dependent variable. Here, the theta values are referred to as regression weights and computed in such a way as to minimize the sum of the squared deviations.

Assumptions that need to be kept in mind while performing Linear Regression include the following:

CHAPTER 4

Support Vector Machine

Abstract: Support Vector Machine (SVM) may be defined as a machine learning algorithm that can be used for regression and classification. It is generally used for classification purposes. In this chapter, we will discuss Margin and Large Margin Methods and Kernel Methods.

Keywords: Data science, Jupyter, Machine learning, Matplotlib, Numpy, Python, Scikit learn, SciPy.

INTRODUCTION

Support Vector Machine (SVM) is one of the most popular supervised machine learning algorithms. It is used for performing classification and clustering tasks [1]. Usually, SVM is used in classification problems. Support Vector Machine creates a decision boundary referred to as a hyperplane that divides the n dimensional space into different classes or categories. After training an SVM model, the new data may be assigned the correct class to which it belongs.

SUPPORT VECTOR CLASSIFICATION

In SVM, the data item is plotted in an n-dimensional space, where n represents the number of features [1, 2]. A classification is performed by finding the hyperplane that can differentiate the two classes. Consider Fig. (**1**); Here, the classification of two different shapes is performed by finding the hyperplane:

Selecting the right hyperplane for a given problem can be done in the following ways:

1. Choose the hyperplane that classifies the data points in a better way. Consider Fig. (**2**). Here, we have three hyperplanes, namely A, B and C. We need to choose one hyperplane out of these. We choose hyperplane A as it classifies the data points efficiently as compared to other hyperplanes:

Deepti Chopra & Roopal Khurana

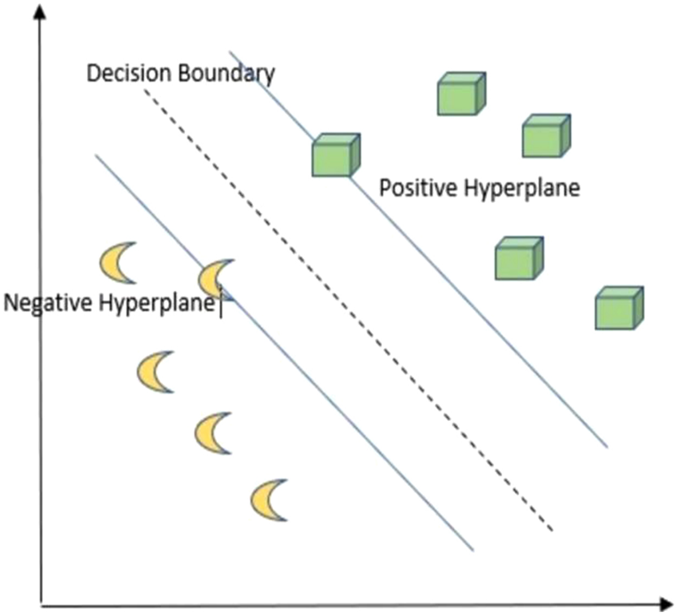

Fig. (1). The classification of shapes using hyperplane in SVM.

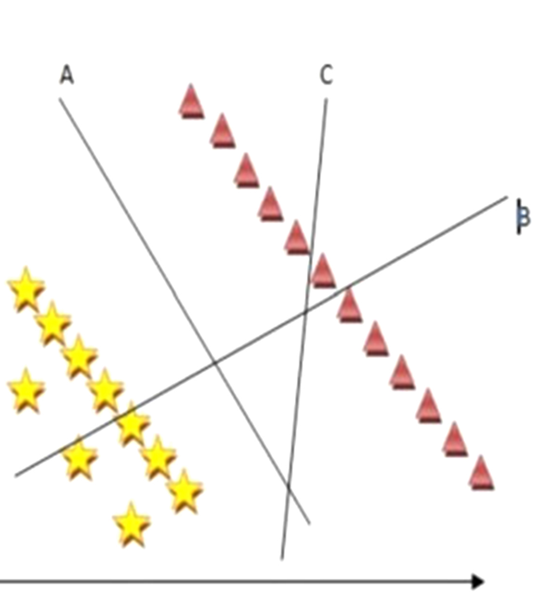

Fig. (2). Hyperplane A classifies the data points efficiently.

2. Use Margin and Large Margin Methods to find the appropriate hyperplane. The distance between the nearest data point and the hyperplane is referred to as the

margin. We must select the hyperplane that has a larger margin; this prevents a chance of miss classification. Consider Fig. (**3**). Here, the margin of hyperplane B is the largest as compared to the margin of hyperplanes A and C. So, we choose hyperplane B for classifying our data points.

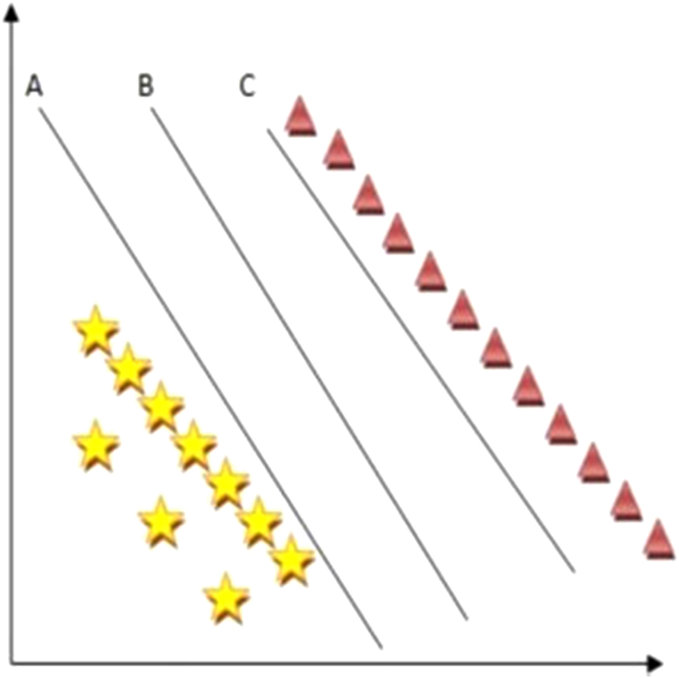

Fig. (3). Margin and Large Margin Methods.

3. Identify the correct hyperplane that can classify all the data points correctly without any error. Consider Fig. (**4**). Here, if we choose hyperplane B instead of hyperplane A, since hyperplane B has a larger margin as compared to hyperplane A, then hyperplane B doesn't classify all the data points correctly resulting in an error. Hyperplane A has a smaller margin but it classifies all the data points correctly. So, SVM will choose hyperplane A over hyperplane B.

4. SVM has the characteristic of ignoring the outliers or the noise. Consider Fig. (**5**). Here, the hyperplane chosen has a large margin and at the same time, the SVM ignores the star that lies in the other boundary and this star is treated as a noise or an outlier.

CHAPTER 5

Decision Trees

Abstract: Neural networks are a way to mimic the working of a human brain. Decision trees refer to the decision support structure that uses a tree to make decisions and draw all possible consequences. Decision trees are a way to display conditional control statements. In the following chapter, we will discuss decision trees, regression trees, stopping criterion and pruning loss functions in a decision tree, categorical attributes, multiway splits and missing values in decision trees, and instability in decision trees.

Keywords: Decision trees, Instability in decision trees, Multiway splits, Regression trees.

INTRODUCTION

Decision trees refer to the non-parametric method of supervised learning. Decision trees are used for the purpose of regression and classification.

In a Decision Tree algorithm, a class of a given dataset is predicted. We begin from the root node and follow the branch according to the condition or the property satisfied by the corresponding dataset.

Pruning in a Decision Tree involves the elimination of unwanted branches of a tree which will never contribute to the resultant path.

Features of Decision Trees include the following:

- It is referred to as a Supervised Machine Learning Algorithm which is non-parametric in nature.
- It is capable of representing both regression and classification tasks.
- A decision tree is represented by a hierarchical tree-like structure having components such as root node, internal node and leaf nodes.
- Decision tree employs greedy search and divide and conquer approach to solve the problem.
- For building a given tree, the CART algorithm is used. CART refers to the Classification and Regression Tree algorithm.

Deepti Chopra & Roopal Khurana

The advantages of decision trees include the following:

- It is simple and easy to understand.
- It is helpful in solving decision related problems efficiently.
- It provides all possible different kinds of results from a given problem.
- Compared to other Machine Learning algorithms; it requires less data cleaning.

REGRESSION TREES

Decision tree algorithms may be used for the process of the prediction of results based on the given data. Decision tree algorithms are of two types, namely: classification tree and regression tree algorithm. **The classification and Regression Tree** (**CART**) methodology came into existence in 1984. It was introduced by *Leo Breiman*, *Jerome Friedman*, *Richard Olshen*, and *Charles Stone*. In a classification tree algorithm, the outcome variable is categorical or fixed. For example, using the classification tree algorithm, we may decide what type of car a customer will purchase. In the regression tree algorithm, the target outcome value is a real number. For example, the selling price of residential places may be predicted using the regression tree algorithm. In the classification tree algorithm, the data set is split into classes such as Yes or No. In the regression tree algorithm, the target variable is continuous, for example, temperature, price, *etc.*

STOPPING CRITERION AND PRUNING LOSS FUNCTIONS IN DECISION TREE

The pruning technique is associated with the decision trees that can perform a reduction in the size of the decision trees by eliminating the parts of the tree that do not classify instances. Overfitting can be prevented by including pruning along with the decision trees. Overfitting occurs when the training is done so thoroughly that it also learns noise along with the pattern. Under-fitting occurs when the amount of training is so insufficient that all the patterns cannot be identified. Pruning means that the tree is cut back. *Quinlan* in 1987 suggested a simple method for pruning decision trees, referred to as reduced error pruning. In reduced error pruning, internal nodes are traversed from the bottom to the top and pruned only if it doesn't reduce the tree's accuracy. *Olaru* and *Wehenkel* in 2003 suggested the use of minimum error pruning. In minimum error pruning, for every node we perform a comparison of 1-probability error rate estimation without and with pruning. Pessimistic pruning is a fast method of pruning in which the nodes are traversed in a top to down manner. If a given internal node is pruned then all descendants of this internal node are not sent for the pruning process. Optimal pruning is used to guarantee optimality and is based on the concept of dynamic programming. In optimal pruning, the tree obtained after pruning is much smaller as compared to the original tree and the number of internal nodes is much smaller as compared to the number of leaves.

CATEGORICAL ATTRIBUTES, MULTIWAY SPLITS AND MISSING VALUES IN DECISION TREES

CART refers to the decision tree algorithm that either generates binary regression or classification trees based on whether the target variable is numeric or categorical. Optimal partitioning must be followed while performing the partitioning of decision trees. In CART, the same variables may be reused in the different parts of decision trees. Splitting may be binary or multiway.

In a binary splitting, each node is further divided into at most two subgroups, whereas in the case of multiway splitting, each node is further split into multiple subgroups. Decision trees are very easy to comprehend if we follow multiway splitting as a particular attribute rarely reappears while traversing a path from the root to the leaf.

There are several methods used for dealing with missing values in decision trees. Missing values may be ignored or may be assigned some other category.

Missing values instances may be distributed among the child nodes

as follows:

1. Everything goes to a node having the largest number of instances.

2. Distribution is done among all child nodes but with minimum weights, which is proportional to the number of instances from every child node.

3. Distribution is done randomly according to the categorical distribution of a single child node.

4. Sort, build, and use input features that decide how the distribution of instances is done in a child node.

ISSUES IN DECISION TREE LEARNING

Preventing Overfitting of Data

Decision Trees is a supervised machine-learning approach used for performing regression and classification [1]. Overfitting occurs when the model trains data along with noise and fails to collect the important patterns. A decision tree in an

CHAPTER 6

Neural Network

Abstract: A Support Vector Machine is used for the purpose of classification and regression. In the previous chapter, we discussed Support Vector Machine, margin and large margin methods, and kernel methods. A neural network refers to a parallel computing device that attempts to mimic the model of the brain. In the following chapter, we will discuss the early models of Neural Networks as well as the Perceptron learning model, back propagation, and Stochastic Gradient Descent.

Keywords: Artificial neural network, Backpropagation, Data science, Machine learning,, Neural network, Perceptron learning, Python.

INTRODUCTION

Neural network is a means of performing a machine learning task, in which a computer learns by the analysis of training examples. Following are the objectives covered in this chapter:

- To know early models and perceptron learning.
- Understanding back propagation.
- Understanding stochastic gradient descent.

EARLY MODELS

Neural network is one of the subfields of machine learning. Neural network accepts the input data, performs training on the data, and produces the output based on the training performed [1]. In 1943, *Warren Mc Culloch* and *Walter Pits* described the working of neurons. They modeled neural networks with the help of electrical circuits in order to explain the working of neurons in the brain. In 1949, *Donald Hebb* wrote *The Organization of Behavior*, which pointed that the connection between the two neurons is enhanced if they are fired together. In 1959, *Bernard Widrow* and *Marcian Hoff* at Stanford developed neural network based models called **'ADALINE'** and **'MADALINE'. ADALINE** stands for **Adaptive Linear Elements** and **MADALINE** stands for **Multiple Adaptive Linear Elements**. ADALINE was used for the recognition of binary patterns. For a given stream of bits, it can predict the occurrence of the next bit. MADALINE

Deepti Chopra & Roopal Khurana

stands for Multiple Adaptive Linear Elements. It is the first neural network that is applied to real-world problems and is still in use for commercial purposes.

PERCEPTRON LEARNING

Perceptron is based on a neutron, which is the basic processing unit of the brain. A neuron comprises dendrites, a cell body, and an axon. Signal flows from the axon to the dendrites. An action signal is fired by a neuron when a particular threshold is met by a cell. This action either takes place or it does not. There is no concept of partial firing by a neuron. Consider Fig. (**1**) depicting the Perceptron model:

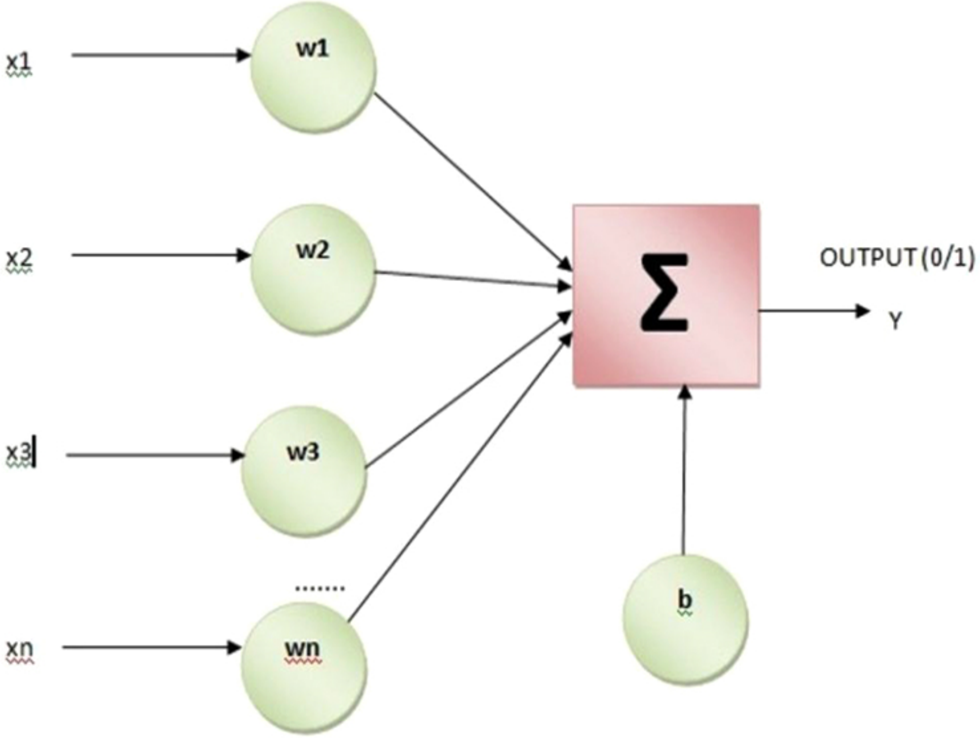

Fig. (1). Perceptron model.

Perceptron can be used for solving binary classification problems where the sample that needs to be identified belongs to two classes. Many features or inputs are sent to the linear unit of a Perceptron and it generates one binary output.

So, a single neuron neural network is referred to as a Perceptron. It accepts the input and the weight, performs the weighted sum of inputs, and applies an activation function over it. It accepts and generates only binary values. One of the limitations of the Perceptron learning model is that it can solve only linearly separable problems.

BACKPROPAGATION

Back propagation is also referred to as Gradient Computation. Back propagation learning algorithm comprises two phases, namely: Gradient Computation Phase and Weight Updation Phase. The first phase is the Propagation phase. It involves the following steps:

Forward Propagation - Here, the training input pattern is sent to the neural network and the propagation's output activation is generated.

Backward Propagation – Here, the input is the propagation's output activation that is sent to the neural network, and it generates the deltas of all the output and hidden neurons.

The second phase is the Weight Updation phase. It involves the following steps:

1. The gradient of the weight is calculated by multiplying the output delta and the input activation.

2. A ratio or percentage of the gradient is subtracted from the weight. This ratio or percentage affects the quality of learning and speed. It is referred to as the learning rate. A neuron is able to train faster if the learning rate is higher. If the learning rate is lower, then the training is considered accurate.

AN ILLUSTRATIVE EXAMPLE: FACE RECOGNITION

Facial Recognition is an area of research in the field of pattern recognition and image processing. Face recognition systems are very important and are in great demand in commercial as well as law enforcement applications. A lot of work has already been done in the field of face recognition systems but still research is going on in this field to overcome the challenges that are faced in the face recognition system. One of the challenges faced in face recognition systems is an increase in the database size increases the time of recognition. Other issues in recognition include: changes in scene illumination, pose changes, updations in expressions, orientation, *etc.*

Face Recognition application is important in the field of Computer Vision, Image Processing, Psychology, Security, multimedia, *etc.* In Face Recognition System, a person’s face is captured using a camera. Facial recognition analyzes the characteristics of a person's face images input through a digital video camera or online face capturing. The first step in Face Recognition System is to collect the

CHAPTER 7

Supervised Learning

Abstract: Supervised learning involves training using well “labelled” training data, and on the basis of the training data, machines are able to predict the output. The labelled data means that the correct output is attached along with the corresponding input data. In supervised learning, as the name suggests, the training data acts as the supervisor and provides training to the machine to predict the correct output. This chapter discusses Statistical Decision Theory, Gaussian & Normal Distribution, Conditionally Independent Binary Components, Learning Beliefs Network and Nearest-Neighbour Methods.

Keywords: Belief networks, Normal distribution, Statistical decision theory, Supervised learning.

INTRODUCTION

Supervised Learning is a machine learning technique that involves mapping of input data with the corresponding output [1, 2]. Supervised learning is performed using supervised learning algorithm that provides a mapping function which can, input data with the output data. Real-life application of supervised learning is: Fraud Detection, Spam Filtering, Risk Assessment, Image Classification *etc* [3, 4].

USING STATISTICAL DECISION THEORY

Decision Theory is defined as the study of choices being made by the customers, professionals, voters, *etc*. Decision Theory is of two types: Normative Decision Theory and Optimal Decision Theory. In the analysis of decision theory, analysis is carried out regarding how an optimal decision is made, what are the characteristics of an optimal decision maker and how an optical decision maker can arrive at an outcome.

Decision Theory may be defined as the study of choices of agents or persons. It assists in understanding various choices that are made at the time of decision by customers, professionals, voters, *etc*.

Deepti Chopra & Roopal Khurana

The three kinds of uncertainty found in decision theory are: Actions, States and Consequences. Marketers use decision theory as an excellent tool for understanding consumer behaviour.

Normative Decision performs analysis of the result of decisions. Normative Decision Theory is responsible for taking optimal decisions on the basis of the result obtained. It is responsible for obtaining results for a specific situation. In Optimal Decision Theory, a detailed analysis and an investigation is carried out on how and why the choices are made by the individuals and agents after decision. Decision Theory is helpful in predicting consumer behaviour. It is helpful in knowing whether the product is helpful in understanding why customer chooses a particular product, and which transportation is more suitable for carrying out finished products? It is helpful in getting information that the product will be successful or not. The three different uncertainties found in decision making include- state, actions and consequences. States represent the existing list of facts that may affect the decision. Consequences represent the characteristics of decisions that are taken by the decision maker. Action is the bond between the state and consequences.

Gaussian or Normal Distribution

Gaussian distribution is also referred to as Normal distribution or Gauss or Laplace - Gauss distribution. For a given real valued random variable, gauss distribution represents the continuous probability distribution. The probability density function of the gaussian or normal distribution is represented as follows:

$$f(x)=(1/\sigma 21/2\pi)e-1/2[(x-\mu)/\sigma]2$$

Here, μ represents the expectation or mean. It also represents median or mode. The parameter σ represents the standard deviation. $\sigma 2$ represents the variance. A variable having Gaussian distribution is said to exhibit normal distribution or normal deviation. Gaussian distribution is used for the representation of real-value random variables whose distribution is unknown. Gaussian distribution makes use of the central limit theorem. According to the central limit theorem, for a given random variable, having a finite value of variance and mean, as the number of observations of these random variables increases, its distribution converges to normal. One of the characteristics of a Gaussian distribution is that a linear combination of various normal deviates is a normal deviation. A Gaussian distribution is also sometimes referred to as a bell curve. But, there are other distributions as well that may have a bell curve. A special case of a Gaussian distribution is referred to as a Unit Normal distribution or Standard Normal distribution when $\mu=0$ and $\sigma=1$ and the density or probability density function is represented as follows:

$$\varphi(z)=[\{(e-(z2)/2\}/21/2\pi]$$

Here, the mean and variance of z is 0 and the standard deviation of z is 1.

Conditionally Independent Binary Components

The term conditional independence for a given set of multiple variables was given by Dawid in 1980.

An important concept for probability distributions over multiple variables is that of conditional independence (Dawid, 1980).

Consider three variables p, q, and r, and imagine that the conditional distribution of p, given q and r, is such that it is not dependent on the value of q, then:s

$$p(p|q, r) = p(p|r).$$

We can say that p is conditionally independent of q given r. The above equation can be rewritten in a different way if we consider joint distribution of p and q conditioned on r, which we can write in the following form:

$$p(p, q|r) = p(p|q, r)p(q|r)$$

$$= p(p|r)p(q|r)$$

In the above equation, product rule has been used. So, it is contained on q. The joint distribution of p and q factories into a product of marginal distribution of p as well as marginal distribution of q. It means that p and q are statistically independent given r.

Conditional independence has been used for creating probabilistic models for pattern recognition by performing simplification of complex structures as well as computations required in applying inference under various conditions.

If the product of conditional distributions is given in form of an expression for a particular directed graph, then we can check if any conditional independence property holds or not. Looking at the graphical models; we can derive the conditional independence property of joint distribution. This graphical framework

CHAPTER 8

Unsupervised Learning

Abstract: Unsupervised learning is a complex processing task involving the identification of patterns in datasets having data points that are neither labeled nor classified. Unsupervised learning is a kind of machine learning algorithm that can be used to draw useful conclusions without the presence of labeled responses in the input data. In the following chapter, we will discuss Clustering (K-means Clustering, Hierarchical Clustering), and Principal Component Analysis.

Keywords: Clustering, Principal component analysis, Unsupervised learning.

INTRODUCTION

In unsupervised learning, an uncategorized and unlabeled data is sent to the AI system and the algorithms act on this data without any prior training. In unsupervised learning, it is important to know the following outcomes:

- To know about unsupervised learning
- Understanding Clustering and Clustering Algorithms
- (K-means Clustering, Hierarchical Clustering)
- Understanding Principal Component Analysis

CLUSTERING

The term Clustering was first used in 1932 in anthropology by *Driver* and *Kroeber*. It was used in 1938 in psychology by *Joseph Zubin* and in 1939 by *Robert Tryon*. It was used for trait theory classification in personality psychology in 1943 by *Cattell*. Clustering also referred to as cluster analysis is a process of grouping together similar objects into the same group called cluster in such a way that objects in one cluster are not similar to the objects in another cluster. Cluster analysis is used in many fields such as data compression, pattern recognition and image processing, machine learning, computer graphics, information retrieval, and bioinformatics [1]. In the following chapter, we will discuss clustering algorithms such as k-means clustering algorithm and hierarchical clustering algorithm.

Deepti Chopra & Roopal Khurana

K-means Clustering

K-means clustering involves the partitioning of observations into k clusters in which a given observation belongs to the cluster having the closest mean [2, 3].

K-means Clustering for solving general problems involves the following steps:

- Cleaning and Transforming data
- Choosing the value of K and K-means algorithm is made to run.
- Reviewing the result obtained
- Iterating over different values of K

Hierarchical Clustering

Hierarchical clustering is also referred to as hierarchical cluster analysis. This method involves building a hierarchy of clusters [1]. Two approaches used in hierarchical clustering analysis include the following:

- Agglomerative clustering
- Divisive clustering

Agglomerative clustering is a ‘bottom-up’ approach that involves each individual cluster, and the pairs of clusters merge and move higher in the hierarchy. Divisive clustering is a ‘top-down’ approach in which a single cluster is further split into multiple clusters as we proceed down in the hierarchy.

Cluster dissimilarity is a metric that measures the distance between the pair of observations and the linkage criterion that specifies a dissimilarity between the pair of observations. Cluster dissimilarity is used in agglomerative clustering to decide which cluster would join together to form the bigger cluster. It is also used in divisive clustering to decide which cluster would finally break.

Principal Component Analysis (PCA)

PCA refers to the dimensionality reduction methodology that can be used for reducing the dimensions of large data sets into smaller ones, while still preserving as much information as possible [4].

Steps performed in PCA include the following:

1. Standardization

2. Computation of Covariance Matrix

3. Identification of Principal Components by computing the Eigen Values and Eigen Vectors of Covariance Matrix

4. Creating a Feature Vector

5. Recasting the data

In the first step, standardization is important because if there are some values that are large and some that are small, then the larger ranges would dominate over the smaller ranges. So, standardization is performed to solve this problem. Standardization can be done by subtracting the mean from the value and dividing it by the standard deviation. This is represented as follows:

Z=(value-mean)/standard deviation

In the second step, the covariance matrix is computed to find whether there exists redundant information in the input data. In the third step, the principle components are generated, which are nothing but the variables that are formed by the linear combination of the initial variables. In the fourth step, a feature vector is created by considering those principle components that are of a higher significance over the ones with a lower significance.

PYTHON CODE

1. Consider the following code in python using k-means clustering:

```
print(  doc  ) import numpy as np

import matplotlib.pyplot as plt

from mpl_toolkits.mplot3d import Axes3D

from sklearn.cluster import KMeans from sklearn import datasets

np.random.seed(5)

irisdata = datasets.load_iris()

P = irisdata.data

q = irisdata.target
```

In the following code, n_init is set to 1 instead of 10 which is a default value. So, this bad initialization has an impact on the classification process as it reduces the number of times an algorithm runs with different centroid seeds.

CHAPTER 9

Theory of Generalisation

Abstract: Unsupervised learning is a kind of machine learning algorithm that can be used to draw useful conclusions without the presence of labeled responses in the input data. In the previous chapter, we discussed clustering (k-means clustering, hierarchical clustering) and Principal Component Analysis. In this chapter, we will discuss training versus testing, bounding the testing error, and the VC dimension.

Keywords: Testing, Training, VC dimension.

INTRODUCTION

Training data helps the algorithm to learn from experience. In supervised learning, each observation comprises an input variable and the corresponding target variable. For building a model, a training set is implemented and for validating a test set, a testing set is required. The data set is divided into the training set and the test set [1, 2].

In machine learning, a model is created in order to perform testing on the test data. To fit the model, training data is used and to perform testing, test data is used. It is not necessary to use 70% of the data set for developing the training set and the rest for the purpose of testing. It depends on the data set that is being used and the task that needs to be accomplished.

BOUNDING THE TESTING ERROR

Principal Component Analysis refers to the dimensionality reduction methodology that can be used for reducing the dimensions of large data sets into smaller ones, while still preserving as much amount of information as possible.

Steps performed in Principal Component Analysis include the following:

1. Standardization

2. Computation of Covariance Matrix

Deepti Chopra & Roopal Khurana

3. Identification of Principal Components by computing the Eigen Values and the Eigen Vectors of the Covariance Matrix

4. Creating a Feature Vector

5. Recasting the data

The first step, standardization, is important because if there are some values that are large and some small, then the larger ranges would dominate over the smaller ranges. So, standardization is performed to solve this problem. Standardization can be done by subtracting the mean from the value and dividing it by the standard deviation. This is represented as follows:

Z=(value-mean)/standard deviation

In the second step, the covariance matrix is computed to find whether there exists redundant i.

VAPNIK CHERVONENKIS INEQUALITY

VC dimension was originally given by *Vladimir Vapnik* and *Alexey Chervonenkis*. VC dimension may be defined as a measure of some of the features in terms of complexity, flexibility, richness or the expressive power of the set of functions that may be learned using a statistical binary classification algorithm.

Uses of VC dimension include the following:

- VC dimension is used in the statistical learning theory for the prediction of the probabilistic upper bound of the test error of a classification model [3, 4].
- VC dimension is also used in sample complexity bounds. Sample complexity may be defined as the linear function of the VC dimension of the hypothesis space.
- VC dimension is used in computational geometry for the prediction of the complexity of approximation algorithms.

PROOF OF VC INEQUALITY

The main outcome in Statistical learning is VC inequality which is a difference in generalisation and empirical risk.

The VC inequality is represented for binary hypothesis classes.

$$\{h{:}X \rightarrow \{0,1\}\}$$

It represents the upper bound on the absolute difference between the given empirical and the true risk, when they are defined with 0-1 loss .The VC dimension can be defined for binary classifiers only.

The inequality can be generalized with the help of different complexity measures (like Rademacher complexity or pseudo-dimension).

CONCLUSION

In this chapter, we learned about training versus testing, bounding the testing error, and VC dimension. In the next chapter, we will discuss how to detect bias and how to fix bias or achieve fairness in ML.

EXERCISES

1. Explain the difference between training and testing in machine learning.

2. Explain bounding the testing error.

REFERENCES

[1] C.A. Corneanu, S. Escalera, and A.M. Martinez, "Computing the testing error without a testing set", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 2677-2685, 2020.
[http://dx.doi.org/10.1109/CVPR42600.2020.00275]

[2] Z. Yuan, Y. Yan, R. Jin, and T. Yang, "Stagewise training accelerates convergence of testing error over SGD", *Adv. Neural Inf. Process. Syst.,* p. 32, 2019.

[3] R.P. Lamberts, J. Swart, R.W. Woolrich, T.D. Noakes, and M.I. Lambert, "Measurement error associated with performance testing in well-trained cyclists: application to the precision of monitoring changes in training status", *International SportMed Journal,* vol. 10, no. 1, pp. 33-44, 2009.

[4] S.L. Wise, L.E. Lukin, and L.L. Roos, "Teacher beliefs about training in testing and measurement", *J. Teach. Educ.,* vol. 42, no. 1, pp. 37-42, 1991.
[http://dx.doi.org/10.1177/002248719104200106]

CHAPTER 10

Bias and Fairness in Ml

Abstract: In machine learning and AI, future predictions are based on past observations, and bias is based on prior information. Harmful biases occur because of human biases which are learned by an algorithm from the training data. In the previous chapter, we discussed training versus testing, bounding the testing error, and VC dimension. In this chapter, we will discuss bias and fairness.

Keywords: Bias, Confidence intervals, Fairness, Hypothesis testing.

INTRODUCTION

Bias is referred to as a disproportionate prejudice or inclination towards a particular thing or an idea. Bias may be found in the following different fields:

- Research
- Statistics
- Social sciences

In machine learning, algorithmic biases are referred to as unwarranted associations. Algorithmic biases are the bugs that can be harmful to the business and people. Following are the objectives of this chapter:

- Understanding how to identify bias.
- Understanding how to achieve fairness in ML.

HOW TO DETECT BIAS?

Bias is one of the popular topics that one encounters while building AI-based models. Many uncommon and common biases may be found in the following stages of AI model development:

- Data collection
- Data preprocessing

Deepti Chopra & Roopal Khurana

- Data analysis
- Modeling

During data collection, biases may take place. This happens due to the occurrence of outliers and errors that happen while collecting data.

Biases that are found during the data collection process include the following:

- **Selection bias**: While preparing the sample data, the selection of data must be done in a proper manner to avoid bias. For example, if the participants are students who are to undergo tests, then they may include the bias results.
- **The Framing Effect**: Survey questions are framed in such a manner that biases are avoided and displays positivity in sentences, or else biases crop up.
- **Systematic bias**: It occurs because of faulty equipment. It leads to repeatable and consistent errors.
- **Response bias**: It occurs due to questions that are answered incorrectly by the participants.

During data preprocessing, the following steps may be undertaken:

1. Outlier detection

2. Missing values

3. Filtering data

Outliers lead to a disproportionate effect on many of the analyses that are conducted.

While dealing with the missing values, if all the missing values are replaced by the mean values, then it would mean being biased towards a particular group that is closer to the mean.

Biases may be found during the process of data analysis. Biases may be found using the following approaches:

- **Missing graphs**: Incorrect conclusions may be drawn from a distorted graph that provides incorrect information.
- **Confirmation bias**: It involves the tendency to focus and confirm information that is related to someone's preconceptions.

When performing data modelling, it is very important to detect biases. For example, Amazon created a hiring algorithm that showed gender bias by favoring men as high-potential candidates. A model that has high variance focuses on training data and doesn't generalize well. Data always behaves in the same way in high bias. When we increase bias, variance decreases and *vice versa* . In supervised machine learning, training is performed on the input variables in such a manner that there is closeness between the predicted values and the actual values. The error refers to the difference between the actual and the predicted values. There are 3 types of errors in supervised machine learning:

- Bias error
- Variance error
- Noise

Bias and variance are reducible errors that we can minimize to a large extent. Noise is said to be an irreducible error that cannot be eliminated.

HOW TO FIX BIASES OR ACHIEVE FAIRNESS IN ML?

There are already many definitions of fairness as per literature and these cover the following elements:

- Equalized odds
- Unawareness
- Individual fairness
- Demographic parity
- Counterfactual fairness
- Predictive rate parity

We should avoid including a sensitive attribute as one of the features in training data. There are many ways to mitigate biases. Some of these techniques include:

- Preprocessing
- In-processing
- Post-processing

The pre-processing approach takes place before the development of a model. Its main intent is to eliminate the underlying bias from the data set before modelling. This is one of the basic approaches to removing biases from the data. In-processing is the process of removing biases during the training phase. In post-processing, the elimination of biases takes place after the training phase is over.

APPENDIX

Case Study- Use of Machine Learning in Harley Davidson

The owner of New York dealership Asaf Jacobi had to sell one or two Harley-Davidson motorcycles in a coming week. One day, during Winter, he was walking in Riverside Park and he met Or Shani, the CEO of Adgorithms. Asaf Jacobi discussed the low sales number with Shani. Shani then suggested him to use Albert, which is AI enabled and has been developed by Algorithms for marketing purposes. Albert can work on different platforms like Facebook and Google. When Asaf Jacobi used Albert; he was able to sell 15 motorcycles that week. The sales were increasing day by day. The sale just doubled of what he managed to sell during summer. In the first month, the new lead was 15%. In the third month, the lead grew by 29-30%. Jacobi now had to expand his current call center so as to handle the current growing business. Today AI and Machine Learning have been used by all leading firms such as Amazon, Google, Facebook, *etc.* AI and Machine Learning have helped these companies to reach the target customers using personalised marketing campaigns. Using Albert in case of Harley Davidson, store traffic increased as it generated leads which were target customers that had shown interest in the form that was filled. Albert is provided content from Harley Davidson in order to improve its sales. Albert could analyse the behaviour of target customers. Albert targeted people who have completed their purchases, have been adding items to the cart and are considered among the top 25% people to visit the Harley Davidson website. Albert ran tests in various groups and got data related to high-value customers. Albert tried to improvise digital marketing campaigns by providing successful conversations. Albert could gather and do the processing of millions of interactions in a minute. So, it saved a lot of time. For the same task, humans would have taken a lot of time.

CONCLUSION

Incorporating AI and Machine Learning specially in business such as in the case of Harley Davidson would give great success by increasing decision-making power, taking the right decisions, reducing marketing time, *etc.* It will definitely lead to an escalation in its growth in terms of sales in no time.

Deepti Chopra & Roopal Khurana

SUBJECT INDEX

A

Acid, malic 54
Algorithm 19, 25, 26, 78, 79, 101, 103, 105, 113, 116, 120
Applications of linear regression 32

B

Bayesian networks 100

C

Cells 7, 37, 84
  diagonal 37
  off-diagonal 37
Chatbots 16, 25
Chi-squared automatic interaction detector (CHAID) 79
  Classification 26, 35, 64, 74, 75, 76, 78, 101, 105
  and regression tree (CART) 26, 74, 75, 76, 78
  probabilistic 64
  process 35, 101, 105
Classification problems 58, 84
  binary 84
Cluster analysis 103
Clustering 101, 103, 104, 111, 113
  agglomerative 104
  algorithms 103
Code cells 7
Computational 15, 114
  algorithm 15
  geometry 114
  intelligence 15
Computation(s) 5, 64, 86, 87, 88, 99, 101
  energy 88
  numerical 5
Condensed nearest neighbour (CNN) 101
Conditional independence 99, 100, 101, 102
Cyber 26
  fraud 26
  surveillance 26

D

Data 6, 20, 103
  analytics 6
  augmentation 20
  compression 103
Datasets, non-linear separable 86
Decision(s) 15, 16, 17, 58, 62, 63, 69, 70, 71, 74, 88, 97, 98
  algorithms 88
  boundary 58, 62, 63, 69
  function 69
  intelligent 16
Detection 36, 65, 86
  credit card fraudulent transaction 36
Devices, portable 23
Dynamically modifying network structures 88
Dynamic neural networks 88

E

Electrical circuits 83
Error 18, 22, 23, 31, 32, 33, 35, 45, 49, 50, 60, 87, 117, 118, 120, 121
  algorithm 87
  bayes 18
  minimization mechanism 87
  resubstitution 22

F

Facial recognition 85
Features 86
  facial 86
  global 86
Filtering data 117
Financial 26

data 26
frauds 26
Fourier transform 8
Function 35, 45, 64, 65, 69, 70, 71, 87, 92, 114
linear 114

G

Gaussian 64, 72, 98
Processes 64, 72
Generating split points 78
Google neural machine translation 27

H

Healthcare 16, 26, 28
medical 26
sector 28
Helicopter aerobatics 28
Heterogeneity 78
Heuristics technique 101
Hierarchical clustering 101, 103, 104, 111, 113
algorithm 103
analysis 104
Homogeneous data 8
Homoscedasticity 33
Human-like intelligence 26, 27
Human-like intelligent 16, 30
behaviour 30
output 16
Hyperplane 58, 59, 60, 61, 62, 64, 72, 73
circular 64, 72
linear 64, 72

I

Image 9, 16, 65, 72, 85, 89, 103
processing 9, 16, 65, 72, 85, 89, 103
recognition 65, 72
segmentation 89
Infinite impulse response network 88
Input activation 85, 94
Input data 30, 32, 83, 97, 103, 105, 109, 113, 120
mapping 32
Instability problem 80

K

Kernel ridge regression 64

L

Learning algorithms 58, 86
popular supervised machine 58
Learning belief networks 100
Learning system 16, 17, 18
effective machine 18
efficient machine 18
Linear relation 108
Logistic 35, 52
function 35
regression 52
Long short-term memory networks 88

M

Machine learning 5, 13, 14, 15, 16, 17, 19, 21, 25, 27, 28, 58, 97, 103, 113, 120, 121
algorithm 5, 15, 16, 17, 25, 58, 103, 113, 120, 121
building 14
methods 27
neural 27
process 13, 28
project 21
system 15, 16, 17, 19, 28
technique 97
Machine translation 65
Malignant tumors 53, 54
Memory 88, 120
gated 88
Method 34, 59, 60, 83, 86
global feature-based 86
margin and large margin 59, 60, 83
namely-local feature-based 86

nonlinear regression 34
Monty python's flying circus 1
Multiple 31, 83, 84
adaptive linear elements 83, 84
linear regression 31
Multiple variables 30, 31, 32, 34, 42, 43, 44, 99
linear regression in 30, 32, 42

N

Natural language processing tasks 5
Nearest 97, 100
neighbour methods 97, 100
neighbours values 100
Neural network(s) 5, 18, 19, 83, 84, 85, 86, 87, 88, 89, 91, 93, 95, 96, 121
and linear regression 121
artificial 83, 88
static 88
traditional 88
Neurons 83, 84, 85
Noise 33, 60, 71, 75, 76, 86, 101, 118
removal 86
Normative decision theory 97, 98
Numerical python 7

O

Objects 7, 24, 27, 89, 90, 100, 103
multidimensional array 7
Online 23, 26, 85, 120, 121
fraud detection 26
learning 120, 121
track 26
transportation networks 23
Output 17, 19, 28, 30, 31, 32, 35, 44, 50, 52, 54, 55, 84, 87, 88, 89, 94
binary 84
of linear regression 44

P

Parallel 83, 121
algorithm 121
computing device 83
Perceptron learning 83, 84, 89, 95
Post pruning technique 77
Pre-pruning technique 77
Principal component analysis (PCA) 101, 103, 104, 107, 108, 110, 111, 113
Probability density function 98
Problem 21, 22, 39, 64, 74, 75, 100, 104, 105, 114
non-linear separation 64
travelling salesman 100
Process 16, 21, 25, 75, 89, 100, 103, 117, 118, 119
data generation 119
Processing, natural language 4, 16
Programming languages 1, 4
object-oriented 1
popular 1
Propagation learning algorithm 85
Protein remote homology detection 65
Pruning 75, 77
process 75
techniques 75, 77
Python 2, 4, 5, 13
file 4
machine learning 5
programming 2, 13

R

Random variables 64, 98, 100
Recurrent neural network (RNN) 88
Regression 32, 63, 64
algorithm 63
problems 32, 64
Regression tree 74, 75, 78, 81, 95
algorithm 75
Relationship 30, 32, 33, 34, 35, 39, 41, 88
linearity 35
Reliability-machine learning systems 19

Risk 26, 114, 115
  empirical 114
  financial 26

S

Social 16, 24
  media services 24
  networking sites 16
Spam filtering techniques 25
Splitting 18, 22, 53, 55, 56, 76, 78
  binary 76
Square error function 32
Squared probabilities 78
Statistical learning theory 114
Stochastic gradient descent optimization 87
  algorithm 87
Supervised 101
  learning technique 101
  metric learning 101
Support vector regression 63
System 10, 15, 17, 19, 20, 25, 27, 33, 86, 103
  artificial intelligence 17
  automatic translation 27
  security programs 25

T

Text iris 80
Transactions 26
  illegal 26
Transportation 98

V

Validation data 18
Variables 30, 31, 32, 33, 35, 39, 43, 98, 99, 100, 105
  binary 35

W

Wearable sensors 26
Weight updation phase 85

Deepti Chopra

Dr. Deepti Chopra has done PhD in the area of Natural Language Processing from Banasthali Vidyapith. Currently, she is working as Associate Professor at JIMS Rohini, Sector 5. Dr. Chopra is an author of five books and two MOOCs. Two of her books have been translated into Chinese and one has been translated into Korean. She has 2 Australian Patents and 1 Indian Patent to her credit. Dr. Chopra has several publications in various International Conferences and journals of repute. Her areas of interest include Artificial Intelligence, Natural Language Processing and Computational Linguistics. Her primary research works involve machine translation, information retrieval, and cognitive computing.

Roopal Khurana

Mr. Roopal Khurana is working as Assistant General Manager at Railtel Corporation of India Ltd., IT Park, Shastri Park, Delhi. Currently, he is working in the field of Data Networking, MPLS Technology. He has done BTech in Computer Science and Engineering from GLA University, Mathura, India. He is a technology enthusiast. Previously, he has worked with companies, such as Orange and Bharti Airtel.