Machine Learning for Beginners

2nd Edition

Build and deploy Machine Learning systems using Python

Dr. Harsh Bhasin

www.bpbonline.com

First published: 2020

Second published: 2024

Published by BPB Online
WeWork
119 Marylebone Road
London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-5551-563-6

www.bpbonline.com

Dedicated to

My Mother

About the Author

Dr. Harsh Bhasin is a researcher and practitioner. Dr. Bhasin is currently associated with the Centre for Health Innovations, Manav Rachna International Institution of Research and Studies. Dr. Bhasin has completed his Ph. D. in ***Diagnosis and Conversion Prediction of Mild Cognitive Impairment Using Machine Learning*** from Jawaharlal Nehru University, New Delhi. He worked as a Deep Learning consultant for various firms and taught at various universities including Jamia Hamdard, MRU and DTU.

He has authored 11 books including *Programming in C#, Oxford University Press, 2014; Algorithms, Oxford University Press, 2015; Python for Beginners, New Age International, 2018; Python Basics, Mercury, 2019; Machine Learning, BPB Publications, 2020*, to name a few.

Dr. Bhasin has authored more than 40 papers published in conferences and renowned journals including Alzheimer's and Dementia, Soft Computing, Springer, BMC Medical Informatics & Decision Making, AI & Society, etc. He is the reviewer of a few renowned journals and has been the editor of a few special issues. He is also the recipient of a distinguished fellowship.

His areas of expertise includes Deep Learning, Algorithms, and Medical Imaging. Outside work, he is deeply interested in Hindi Poetry: the progressive era, and Hindustani Classical Music: percussion instruments.

About the Reviewers

- **Jerry Okafor** is an accomplished Senior Software Engineer with a deep passion for AI and ML. With over seven years in the software industry, Jerry has mastered the art of creating seamless user experiences and developing cutting-edge solutions.

  Driven by a curiosity for AI and ML, Jerry embarked on a journey into the realms of Natural Language Processing and Deep Learning, obtaining certifications from Udacity to solidify his expertise. His dedication to the field has led him to excel in various tech environments and work on groundbreaking projects.

  Notably, Jerry spearheaded the creation of an AI-powered chatbot that revolutionized customer support for a leading e-commerce platform. Additionally, he played a crucial role in designing a recommendation engine that personalized user experiences, resulting in a significant increase in user engagement.

  Outside of work, Jerry remains an active participant in AI and ML communities, sharing his knowledge through blog posts and tutorials. His commitment to democratizing AI drives him to inspire and empower others to explore this ever-evolving field. Jerry Okafor continues to push the boundaries of technology as a Senior Software Engineer and AI enthusiast, leaving an enduring impact on the industry and beyond.

- **Prashant Kikani** is an experienced software engineer with a demonstrated history of solving problems in the Natural Language Processing (NLP), Computer Vision, and Machine Learning (ML) fields. His significant achievements include the Master title in Kaggle competitions with a rank in the top 100 worldwide. He is also the author of the "Demystifying Artificial Intelligence" book focused on simplifying ML concepts for everyone. His other interests include gaming, finance and economy. His ultimate goal is to make a positive impact on millions of people's daily lives.

Acknowledgement

"Feeling gratitude and not expressing it is like wrapping a present and not giving it."

— William Arthur Ward

I am blessed to have met people who encouraged me to learn continuously. First of all, I would like to thank Professor Moin Uddin, former Pro-Vice-Chancellor, Delhi Technological University, for his unconditional support. He has deposed his faith in me when no one else did. Had it not been for his encouragement I would not have been able to achieve whatever I did. I would also like to express my sincere gratitude to the late Professor A. K. Sharma, former Dean, and Chairperson, the Department of Computer Science, YMCA, Faridabad, for his constant encouragement.

I am also thankful to the following academicians and professionals for encouragement and providing unconditional support to me:

- Prof. Sanjay Shrivastava, Vice Chancellor, MRIIRS, India
- Prof. I. K. Bhat, Vice Chancellor, MRU, India
- Prof. Prashant Jha, King's College London
- Prof. Sarita Sachdeva, Dean Research & AD, MRIIRS, India
- Prof. Tapas Kumar, Associate Dean, SET, MRIIRS, India.
- Prof. Maneesha Singh, ESIC, Faridabad
- Prof. Naresh Chauhan, former Head and Chairperson, Department of Computer Science, YMCA University of Science and Technology
- Dr. S K. Pal, Scientist, Defense Research and Development Organization
- Prof. Ranjeet Biswas, Former Professor, IIT Kharagpur

I am thankful to Mr. Nishant Kumar, The NorthCap University, India, for his contribution to editing, formatting, and developing some programs for this book. I would also like to thank my students and colleagues, for their critical reviews. I am also very thankful to the editorial team of BPB publications for providing valuable assistance.

I would like to express my sincere gratitude to my Mother: Ms. Vanita Bhasin and Sister: Ms. Swati Bhasin and the rest of the family, including my pets: Zoe & late Xena, and friends for their unconditional support to me.

Preface

Data is being collected by websites, mobile applications, dispensations (on various pretexts), and even by devices. This data must be analyzed to become useful. The patterns extracted by this data can be used for targeted marketing, national security, propagating believes and myths, and many other tasks. Machine Learning helps us in explaining the data by a simple model. It is currently being used in various disciplines ranging from Biology to Finance and hence has become one of the most important subjects.

There is an immediate need for a book that not only explains the basics but also includes implementations. The analysis of the models using various datasets needs to be explained, to find out which model can be used to explain a given data. Despite the presence of excellent books on the subject, none of the existing books covers all the above points.

This book covers major topics in Machine Learning. It begins with data cleansing and presents a brief overview of visualization.

Organization of the book

This book contains three sections namely: Fundamentals, Supervised Learning Techniques and Unsupervised Learning Techniques along with an introduction to Deep Learning. The names of the fifteen chapters and their brief overview are as follows:

Section I: Fundamentals

Chapter 1: An Introduction to Machine Learning - Machine Learning helps to analyze huge data, predict trends, find patterns, and so on. It is currently being used to diagnose diseases, for surveillance, for developing automated vehicles, etc. This chapter introduces Machine Learning, discusses its types, and how it is different from conventional algorithms. This chapter also presents an overview of the history of Machine Learning and its applications. Some exciting tools that use AI for writing your mails, drawing images, creating music etc have also been discussed in this chapter.

Chapter 2: The Beginning: Data Pre-Processing - Machine Learning pipeline includes four major steps: cleaning of data, extracting features, selecting relevant features, and applying learning algorithms. The first and most important task is to clean the data. The data may contain missing values due to the reasons discussed in the chapter. Since missing data will hamper the learning process or, worse, will make the model learn incorrectly, dealing with such values is essential. This chapter gives an overview of how to deal with such values. Cleaning of data will enhance the performance of our Machine Learning model and make the results more meaningful. The chapter also deals with data integration and normalization.

Chapter 3: Feature Selection - The pre-processing of data follows feature extraction. These features will be used to create a feature set, which will help in learning. At times, the features so obtained are huge in number. However, not all the features so obtained are equally important. Some of them are redundant, and some are noisy. The redundant features do not enhance the performance of a model, and the noisy features may degrade the performance of a model.

Therefore, a smaller, more relevant subset of features needs to be selected to carry out the required learning task efficiently and effectively.

Feature selection aims at better performance and reduced learning time. This chapter introduces some of the most important feature selection methods. The reader will be able to implement these methods and make the ML model effective and efficient.

Chapter 4: Feature Extraction - This chapter introduces feature extraction and discusses its importance. It also introduces a feature transformation method called Principal Component Analysis. Topics including extracting relevant features from this image, finding the distribution of microstructures etc. have been covered in the chapter. Feature extraction from audio data has also been discussed in this chapter. This includes finding frequencies of the audio input, finding the multiple frequencies of small samples, discrete wavelet transforms and so on.

Chapter 5: Model Development - We need to gather appropriate data for the project, carry out the preprocessing, and then move forward. This chapter revisits the Machine Learning Pipeline, and then moves to the data splitting techniques. This is followed an informed discussion on the concepts of underfitting, overfitting, bias and variance. This chapter also discusses the methods to reduce bias and variance. Furthermore, the relation between underfitting and bias is also discussed in the chapter.

The reader can select the method for splitting the data, find if the model suffers from underfitting and overfitting, understand the concept of bias and variance, and methods to handle them. The chapter is immensely important to implement models that are efficient, and effective.

Section II: Supervised Learning

Chapter 6: Regression - This chapter introduces classification. It assigns one of the designated labels to a test sample and comes under supervised learning. Logistic Regression and Naïve Bayes are discussed and implemented in the chapter. The techniques discussed in this chapter use the concepts of Probability.

The chapter presents some basic experiments and expects the reader to understand the importance of empirical analysis in Machine Learning. This chapter will form the basis of complex ML-based projects like face recognition etc.

Chapter 7: K-Nearest Neighbors - This chapter discusses a non-parametric method, called K Nearest Neighbors , that does not require learning weights or biases. This method can generate non-linear decision boundaries, and it is used for classification and regression.

The chapter discusses the concept of Nearest Neighbor, the implementation of the algorithm from Scratch, and using SKLearn. The chapter also discuses Regression Using K- Nearest Neighbors, and a method for selecting the value of K.

Chapter 8: Classification: Logistic Regression and Naïve Bayes Classifier - This chapter introduces the reader to Regression. This chapter introduces gradient descent, which will not only help in implementing regression but also in the classification algorithms discussed in the following chapters. The algorithm assumes that the dependent variables depend linearly on the independent variables, which may not always the case. The regression technique based on the values of the nearest neighbors will overcome this limitation. This chapter also presents the results of the application of the above algorithms on different datasets, hence uncovering the applicability of an algorithm on diverse datasets and hence its robustness.

Chapter 9: Neural Network I: The Perceptron - This chapter describes neural networks, which are inspired by the neurons in the brain. This chapter starts with a brief description of the brain and the structure of neurons. The models, learning algorithms, and limitations of neural networks have been divided into two chapters. This chapter deals with the single-layer perceptron, and the next chapter discusses the multi-layer perceptron. This chapter also presents the Delta Learning Rule and discusses the applicability of Perceptron in the classification of two different datasets.

Chapter 10: Neural Network II: The Multi-Layer Perceptron - This chapter briefly explores the fascinating world of Multi-Layer Perceptron (MLP) and presents the feed-forward model and the back-propagation algorithm for learning. MLP's are capable of handling data that is not linearly separable. This chapter also presents the implementation of the multi-layer perceptron and its applicability to some of the non-linearly separable datasets.

Chapter 11: Support Vector Machines - Support Vector Machines are perhaps one of the best machine learning algorithms. They are elegant, effective, and even work for data having very large dimensions. These machines, handle the curse of dimensionality gracefully. These machines do not use the whole data to craft the separating hyperplane, but only a small subset of the training data called the support vectors. It makes these machines' memory efficient. Though the algorithm is based on the creation of hyperplane for linearly separable data, the model can be extended to non-linearly separable data using the kernel trick. Also, the concept of cost has been explained in the chapter, which allows the misclassification of the train data to achieve better performance on the test data.

The chapter explains the implementation of SVM using **sklearn.svm**. The reader will be able to appreciate the mathematical basis of SVM and use SVM for classifying the numeric data and the images using the experiments explained in this chapter.

Chapter 12: Decision Trees - This chapter discusses the importance of decision trees. After reading this chapter, the reader will be able to understand the concept of information gain and the formation of a tree using the concept of information gain. Trees can also be created using the Gini index. This chapter explains the application of Gini Index in selecting an attribute, at a particular level. The implement of decision trees using SKLearn has also been included in the chapter. Finally, the reader will be able to understand the procedures to curtail the depth of a tree.

Chapter 13: An Introduction to Ensemble Learning - Decision Trees often are not considered good predictors or classifiers. This is because even a small change in training data can drastically change the performance of the decision tree. That is, they have high variance. Boosting and Bagging help us to handle this problem.

Boosting is an ensemble method that uses many predictors and returns the majority vote as a result of classification. In the case of regression, the average of these predictors is returned as the result. This chapter introduces Boosting and discusses the types of Boosting.

Bootstrap aggregating (Bagging) is used to decrease the variance of such predictors in which we generally average the performance of many such predictors. In the bootstrap, we create resampled data of size 'n' by sampling from our observed data with replacement. The empirical distribution of data is used to estimate the true unknown data-generating distribution. In resampling (with replacement) from the observed data, not all the training samples will appear in each sample and each bootstrap sample contains around two third of the data points.

After reading the chapter, the reader will be able to develop efficient and effective models to classify data and carry out regression. The models discussed in this chapter gracefully handle the problems of Decision trees and generate better results than the classifiers discussed till now.

Section III: Unsupervised Learning and Deep Learning

Chapter 14: Clustering - This chapter introduces an unsupervised learning technique called clustering. The creation of groups from unorganized data is referred to as clustering. Ideally, the items in a cluster should be as similar as possible and distinct from items of other groups. This similarity can be found by any standard similarity measure like Euclidian distance, Manhattan distance, and so on. To carry out clustering, one needs to decide the similarity measure, figure out how to evaluate a cluster, and an algorithm for clustering. The evaluation of a cluster requires finding inter-cluster separation and intra-cluster cohesion. This chapter discusses the above issues. This chapter also addresses the question of finding the number of clusters.

Chapter 15: Deep Learning - This chapter gives a brief overview of DL, explains how it is different from the conventional Machine Learning pipeline, the factors responsible for the growth of DL, the DL architectures and finally gives an overview of the applications of DL.

The implementations have been given in Python, therefore cheat sheets of NumPy, Pandas, and Matplotlib have been included in the appendix.

Code Bundle and Coloured Images

Please follow the link to download the
Code Bundle and the ***Coloured Images*** of the book:

https://rebrand.ly/znye9xp

The code bundle for the book is also hosted on GitHub at **https://github.com/bpbpublications/Machine-Learning-for-Beginners-2nd-Edition**. In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

Section I: Fundamentals

"By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."

—Eliezer Yudkowsky

CHAPTER 1

An Introduction to Machine Learning

Introduction

When you turn on the location on your mobile, upload your pictures on Facebook or Instagram, fill out online forms, browse websites, or even order items from apps, your data is collected. What do companies do with this huge data? They analyze it and find your preferences and this helps them in targeted marketing. The advertisements being shown to you generally depend on the data collected. This data helps marketing professionals lure you into buying something you need or are even remotely interested in. Likewise, the dispensation may keep track of suspicious activities using this data, may track the source of transactions, or gather other important information. However, this is easier said than done. The analysis of this huge data cannot be done using conventional methods.

To understand this, let us consider another example. Suppose Hari visits YouTube every day and watches videos related to Indian Classical Music, Hindi Poetry, and watch Lizzie McGuire. His friend Tarush goes to YouTube and watches Beer Biceps and other videos related to workouts. After some time, YouTube starts suggesting different videos to both of them. While Hari is shown a video related to Lizzie McGuire's reboot or Dinkar, in the recommended videos list, Tarush is shown the recommendations of videos related to workouts. This is done using a Machine Learning algorithm that uses the profile information, context, and the video being watched.

Machine Learning comes to the rescue of those wanting to analyze huge amounts of data, predict trends, find patterns, and so on. It is currently being used to diagnose diseases, for surveillance, for developing automated vehicles, etc. This chapter introduces Machine Learning, discusses the types of Machine Learning, and gives a brief overview of its history. This chapter also presents an overview of the history of Machine Learning and its applications.

Structure

The main topics covered in this chapter are as follows:

- Conventional Algorithm and Machine Learning
- Types of Learning
- Working
- Applications of Machine Learning
- History of Machine Learning
- Some exciting tools that use AI

Objectives

After reading this chapter, the reader will be able to understand the definition and types of Machine Learning. The reader will be introduced to the ML pipeline and apprised of the applications of Machine Learning. The chapter also gives a brief overview of the history of Machine Learning. Finally, we will have a look at some of the exciting tools based on Machine Learning capable of creating images, generating music, and so on.

Conventional Algorithm and Machine Learning

The algorithmic solution of a problem requires the input data and a program to produce an output. Here, a program is a set of instructions and output is generated by applying those instructions to the input data. In a Machine Learning Algorithm, the system takes the Input Data along with the examples of Output (in the case of supervised learning) and creates a model, which establishes (or tries to establish) some relation between the input and the output. Learning, in general, is improving the outcome using experience (E). How do we know that we have improved? The performance measure tells the performance of our model. As per Tom Michel, Machine Learning can be defined as follows:

"If the performance measure (P) improves with experience on task (T), then the system is said to have learned."

Here, the Task (T) can be Classification, Regression, clustering, and so on. The data constitutes Experience (E). The Performance Measure (P) can be any accuracy, specificity, sensitivity, F measure, Sum of Squared errors, and so on. These terms will be defined as we proceed. To understand this, let us consider an example of disease classification, using Magnetic Resonance Imagining. If the number of patients correctly classified (accuracy) as diseased is considered a performance measure, then this problem can be defined as follows:

- **T**: Classify given patients as diseased or not diseased
- **P**: Accuracy
- **E**: The MRI images of a patient

The task will be accomplished by pre-processing the given data, extracting relevant features from the pre-processed data, selecting the most important features, and applying a classification algorithm followed by post-processing. In general, a Machine Learning pipeline constitutes the following steps, as shown in *Figure 1.1*:

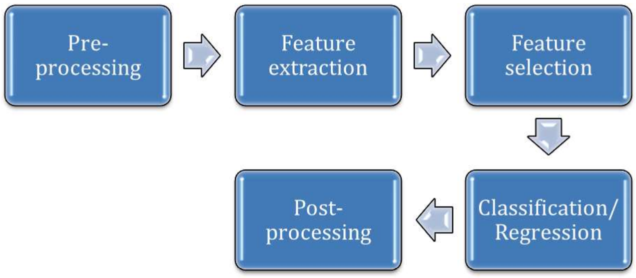

Figure 1.1: *Machine Learning Pipeline*

These terms will become clear in the following chapters. The following chapters will discuss pre-processing, Feature Selection, Feature Extraction, and supervised learning techniques. Having seen the definition of Machine Learning, let us now have a look at its types.

Types of Learning

Machine Learning can be classified as Supervised, Unsupervised, Semi-Supervised, and reinforcement. Let us understand these terms before proceeding any further.

Supervised Machine Learning

This type of learning uses the labels of the data in the training-set to predict the label of a sample in the test-set. The training set acts as a teacher in this type of algorithm, which supervises the training process. The data in these algorithms contain samples and their correct labels. The training process tries to uncover the pattern hidden in the data. That is, the learning aims to relate the label Y with the data X as $y = f(x)$, where x is a sample, and y is the label.

If this label is a discrete value, then the process is termed **Classification**. If y is a real value, then it is called **Regression**. Regression and classification algorithms have been discussed in the following chapters.

Examples of classification are face detection, voice detection, object detection, and so on. Classification essentially means placing the given sample into one of the predefined categories. Examples of regression include predicting the price of a commodity, predicting temperature, housing price, and so on.

Unsupervised Learning

This type of learning uses input *Data(X)* but no labels. The learning aims to learn about the data by grouping the like samples or by deducing the associations. Since there is no teacher