A Hands-On Introduction to Machine Learning

Packed with real-world examples, industry insights, and practical activities, this textbook is designed to teach machine learning in a way that is easy to understand and apply. It assumes only a basic knowledge of technology, making it an ideal resource for students and professionals, including those who are new to computer science. All the necessary topics are covered, including supervised and unsupervised learning, neural networks, reinforcement learning, cloud-based services, and the ethical issues still posing problems within the industry. While Python is used as the primary language, many exercises will also have the solutions provided in R for greater versatility. A suite of online resources is available to support teaching across a range of different courses, including example syllabi, a solutions manual, and lecture slides. Datasets and code are also available online for students, giving them everything they need to practice the examples and problems in the book.

Dr. Chirag Shah is Professor at the University of Washington (UW) in Seattle, USA. Before UW, he was at Rutgers University. His research focuses on intelligent information access systems that are also fair, transparent, and trustworthy. Dr. Shah received his M.S. in Computer Science from the University of Massachusetts Amherst, and his Ph.D. in Information Science from the University of North Carolina Chapel Hill. He directs the InfoSeeking Lab and co-directs the Center for Responsibility in AI Systems & Experiences (RAISE). His research is supported by awards from the National Science Foundation (NSF), the National Institute of Health (NIH), the Institute of Museum and Library Services (IMLS), as well as Amazon, Google, and Yahoo!. Dr. Shah teaches undergraduate, masters, and Ph.D. programs at UW, focusing on data science and machine learning. He has designed MOOCs and has taught several tutorials and short courses at international venues. Dr. Shah has written several books, including his textbook *A Hands-On Introduction to Data Science*. He has visited and worked with many tech companies, including Amazon, Brainly, Getty Images, Microsoft Research, and Spotify.

A Hands-On Introduction to Machine Learning

CHIRAG SHAH
University of Washington

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009123303
DOI: 10.1017/9781009122092

First published 2023

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-009-12330-3 Hardback

Additional resources for this publication at www.cambridge.org/shah-ML

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To my students who helped me, challenged me, and made me learn in ways that went far beyond my own education and research.

Contents

Part II Supervised Learning

CAMBRIDGE
Cambridge University Press & Assessment
978-1-009-12330-3 — A Hands-On Introduction to Machine Learning
Chirag Shah
Frontmatter
More Information

Part IV Neural Networks

Part V Further Explorations

Preface

Machine learning (ML) is everywhere. From the news that gets recommended in our feeds to diagnosing cancer, from which image to display on a movie poster for a user to forecasting storms. Because of this, there is high demand for people to fill ML-related jobs. This is not something new; ML has been in play for decades. So why is it that suddenly there is a lot more chatter about ML, coming from all sorts of people and places? I believe there are two primary reasons. First is technical advancement. There is a new availability of large-scale data, and of tools that can work on such large-scale data with striking effectiveness. Most ML algorithms, as you will see in this book, require data – often, a lot of it. Many of these algorithms do better with more data and more diverse data. But that's just one part of the equation. The others are computational power, and the tools and the techniques that can process that data to create insights. Since the beginning of the twenty-first century we have made amazing advancements in all of these areas. Not all techniques are new, but years ago they were limited in their applicability due to sparsity of data or lack of computational power. We have also developed new hardware components, programming tools, and rich software libraries that allow us to more effectively and efficiently implement ML solutions, making them more viable in a number of areas that were ignored before.

The second reason is more subtle, but much more potent. It is about the social acceptability of ML. When I was in grad school we worked on ML (and artificial intelligence in general) with a few specific applications in mind – speech, vision, robotics. We were fascinated by neural networks primarily as a way to try to hypothesize how a human brain worked. But now ML has made its way into every aspect of our lives, and, more importantly, we have all accepted that new reality. Why is this so important? Because that acceptance fuels more development and usage of ML, not just in traditional computer science fields, but in finance, healthcare, travel, business, education, and almost any field you can imagine. More acceptance and usage mean more data generated, and more data means better tuning of those ML algorithms. It is this cycle that has triggered unprecedented amounts of funding and efforts in this field. It is for this reason that ML is now being taught not only in computer science and engineering disciplines, but in all kinds of majors in colleges and graduate schools. And it is with this as a core principle that this book has emerged.

This book is not just for computer science majors, but also for those who want to understand ML and even build ML solutions for whatever they care about. It is organized in a way that provides a very easy entry point for almost anyone to become introduced to ML, but it also has enough fuel to take one from that beginning stage to a place where they feel comfortable applying learning algorithms to solve real-life problems. In addition to providing the basics of ML and intelligent systems, the book teaches standard tools and techniques for using them. It also examines the implications of the use of ML in areas such as privacy, ethics, and fairness. Finally, as the name suggests, this text is meant to provide a hands-on introduction to these topics. Almost everything presented in the book is accompanied by examples and exercises – sometimes done by hand and other times using the tools taught here. In teaching these topics myself, I have found this to be a very effective method.

The rest of this preface explains how the book is organized, how it could be used for fulfilling various teaching needs, and what specific requirements a student needs to meet to make the most out of it.

Requirements and Expectations

This book is intended for advanced undergraduates or graduate students in information science, computer science, business, education, psychology, sociology, and related fields who are interested in ML. It is not meant to provide an in-depth treatment of any programming language, tool, or platform. One of the strengths of this book is the very low barrier it provides for anyone to enter the realm of ML.

The book assumes no prior exposure to programming or technology. It does, however, expect the reader to be comfortable with computational thinking (see Chapter 1) and the basics of statistics (covered in Appendix A). The reader should also have general computer literacy, including the ability to download, install, and configure software, do file operations, and use online resources. Each chapter lists specific requirements and expectations, many of which can be met by going over some other parts of the book (usually an earlier chapter or an appendix).

Almost all the tools and software used in this book are free. There is no requirement for a specific operating system or computer architecture, but it is assumed that the reader has a relatively modern computer with reasonable storage, memory, and processing power. In addition, a reliable and preferably high-speed internet connection is required for several parts of this book.

Structure of the Book

The book is organized in five parts. Part I provides enough background and basics for anyone coming from non-technical fields so that they can get the most out of the rest of the book. It starts by providing a clear introduction to ML and various related concepts, along with computational thinking. Immediately after that, a chapter on Python is provided. Remember, this is not a programming book, so while this part introduces Python, it is not meant to provide a comprehensive treatment to any programming platforms. Instead, the focus here is on giving a quick introduction with lots of examples and hands-on practice, and then showing how various ML functions can be used with appropriate packages. It is useful to know that the rest of the book uses Python as the standard language for most of the examples and problem-solving, but many of those exercises will also have solutions in R in case an instructor wants that option while teaching ML. In addition to Python, this part also has a chapter on using cloud-based services for doing ML. This is becoming increasingly popular as it circumvents the need to buy one's own high-end hardware or worry about scalability of one's ML apps. The three popular platforms – Amazon Web Services (AWS), Google Cloud, and Microsoft Azure – will be introduced here.

Parts II and III are for two main branches of ML: supervised learning and unsupervised learning, respectively. Each of these parts could be its own course, as they cover many techniques for solving a large variety of problems. For each technique or algorithm, an intuitive explanation is given first, sometimes with a helpful math derivation, followed by a couple of hands-on exercises to see how it allows us to solve real-life problems. There are also plenty of exercises provided at the end of each of the chapters that could be useful for homework assignments or hands-on projects.

Part IV is about neural networks. My students often think this subject is something very new or advanced in ML/AI. And I often tell them that it was already pretty old and well established when I was a student. Here, we assume no background, and start from the beginning – all the way back in the 1950s with basic neural architectures. We talk about learning laws and discuss how they are appropriate for different kind of problems – supervised or unsupervised learning. Then we continue moving through the landscape of various models, all the way to the new ones in deep learning that keep making the headlines, including CNN, RNN, LSTM, and BERT. Once again, the idea here is not to worry about the theoretical depths of these models, but to focus on understanding the intuitions behind them and learn when and how to apply them, with lots of hands-on exercises.

Finally, we come to Part V, which introduces reinforcement learning. This branch of ML has always been very important, but in the past it was primarily used within robotics. In recent years, as we have encountered problems involving large amounts of data and not enough labels or annotations, reinforcement learning has found its way into many interesting and important problems. In this part we will also see how ML is used in research and development – both in academia and industry. We will also look at some of the ways ML systems are evaluated, including A/B testing. The last chapter is one of the most important ones – it shows where ML fails, not in the technical realm, but at a societal level, where it introduces and propagates biases and a lack of fairness and equity. These issues have become very important in recent years, and I strongly recommend anyone studying or teaching ML not to ignore them.

The book is full of extra material that either adds more value and knowledge to the ML theories and practices being covered, or provides a broader and deeper treatment of some of the topics. Throughout the book there are several FYI boxes that provide important and relevant information without interrupting the flow of the text, allowing the student to be aware of various issues related to ML and AI without being overwhelmed by them. There are also "ML in Practice" boxes at several places that provide insights from how ML techniques covered in the book are used in industry. These insights stem primarily from my own experiences working at and with various tech companies.

The appendices of this book provide quick reference to various formulas from differential calculus and probability, as well as helpful pointers and instructions for installing and configuring various tools used in the book. Another appendix provides a listing of various sources for obtaining datasets to further practice ML, or even to participate in ML challenges to win some cool prizes and recognition. There is also an appendix that provides helpful information related to ML jobs in various fields and what skills one should have to apply for those jobs.

The book also has an online appendix (OA), accessible through the book's website at www.cambridge.org/shah-ML, which is regularly updated to reflect any changes in data and other resources. The primary purpose for this OA is to provide you with the most current and up-to-date datasets or links to datasets that you can download and use in the dozens of examples and try-it-yourself exercises in the chapters, as well as data problems at the ends of the chapters. Look for the OA icon which will inform you that you can find the needed resource in the OA. In the description of the exercise you will see the specific number (e.g., OA 4.7) that tells you where exactly you should go in the OA. In addition to the OA document, at this website you will also find code and datasets used in Hands-On Exercises, as well as an errata document that will list any corrections in the book after its printing.

Using This Book in Teaching

The book is deliberately organized around teaching ML to beginner computer science (CS) students or intermediate to advanced non-CS students. The book is modular, making it easier for students and teachers to cover topics at the desired depth. This makes the book suitable for use as a main reference book or textbook for an ML curriculum. The following is a suggested curriculum path in ML using this book. It contains six courses, each lasting a semester or a quarter:

- preparation for ML (beginner): Chapters 2 and 3, appropriate appendices;
- introduction to ML (beginner): Chapters 1 and 2, with some elements from Parts II and III as needed;
- ethical issues in ML (beginner): Chapters 1 and 13;
- neural networks (intermediate): Chapters 9 and 10; and
- ML at scale (advanced): Chapters 3 and 12.

This book's website contains a Resources tab with a section labeled "For Instructors." This section contains sample syllabi for various courses that could be taught using this book, PowerPoint slides for each chapter, and other useful resources such as sample midterms and final exams. These resources make it easier for someone teaching this course for the first time to adapt the text as needed for his or her own ML curriculum.

Each chapter also has several conceptual questions and hands-on problems. The conceptual questions could be used for in-class discussions, homework, or quizzes. For each new technique or problem covered in this book, there are at least two hands-on problems. One of these could be used in the class and the other could be given for homework or as an exam. Most hands-on exercises in chapters are also immediately followed by hands-on homework exercises that a student could try for further practice, or that an instructor could assign as homework or as an in-class practice assignment.

Strengths and Unique Features of This Book

Machine learning has a very visible presence these days, and it is not surprising that there are currently several books and much material related to the field available. *A Hands-On Introduction to Machine Learning* is different from the other books in several ways:

- It is targeted at students with very basic experience with technology. Students who fit within that category are those majoring in information science, business, psychology, sociology, education, health, cognitive science, and indeed any area in which ML can be applied. The study of ML should not be limited to those studying CS or statistics. This book is intended for a broader audience.
- The book starts by introducing the field of ML without any expectation of prior knowledge on the part of the reader. It then introduces the reader to some foundational ideas and techniques that are independent of technology. This does two things: (1) it provides an easier access point for a student without a strong technical background; and (2) it presents material that will continue to be relevant even when tools and technologies have changed.
- Based on my own teaching and curriculum development experiences, I have found that most ML books on the market can be divided into two categories: they are either too technical, making them suitable only for a limited audience; or they are structured to be simply informative, making it hard for the reader to actually use and apply data science tools and techniques. *A Hands-On Introduction to Machine Learning* is aimed at a nice middle ground. On one hand, it does not simply describe ML, but also teaches real hands-on tools (Python, cloud computing) and techniques (from basic regression to neural networks and deep learning). On the other hand, it does not require students to have a strong technical background to be able to learn and practice ML.
- *A Hands-On Introduction to Machine Learning* also examines the implications of the use of data in areas such as privacy, ethics, and fairness. For instance, it discusses how unbalanced data used without enough care with an ML technique could lead to biased (and often unfair) predictions.
- The book provides many examples of real-life applications, as well as practices ranging from small to big data. For instance, Chapter 1 has an example of working with movie recommendations, and this one is done by hand, without using any tools or programming. In Chapter 5 we see how multiple linear regression can be easily implemented using Python to learn how advertising spending on various media could influence sales. Chapter 6 includes an example that uses Python to analyze data about wines to predict which ones are of high quality. Chapters 8–11 on supervised and unsupervised learning have many real-life and general interest problems from different fields. Many of the examples can be worked by hand or with everyday software, without requiring specialized tools. This makes it easier for a student to grasp a concept without having to worry about programming structures, which allows the book to be used for non-majors as well as professional certificate courses.

- Each chapter has plenty of in-chapter exercises where I walk the reader through solving a data problem using a new technique, homework exercises to do more practice, and more hands-on problems (often using real-life data) at the ends of the chapters. There are 51 hands-on solved exercises, 50 try-it-yourself exercises, and 81 end-of-chapter problems.
- The book is supplemented by a generous set of material for instructors. These instructor resources include curriculum suggestions (even full-length syllabi for some courses), slides for each chapter, datasets, program scripts, answers and solutions to each exercise, and sample mid-term exams and final projects.

Acknowledgments

While this book lists only one author, it was a result of many years of collaboration with many people – some of them gave inspiration, some provided raw material, and others made critical corrections. It is nearly impossible to list them all, but I would be remiss if I didn't at least attempt to thank a few.

I want to start by acknowledging the profound impact on my education and life that my late father, Rajendrakumar Shah, had. He encouraged and supported me to pursue my dreams even when he didn't understand them. He is still guiding me through my life.

Writing a book is a very arduous and often lonely task. But I could always count on my dear wife, Lori, to help me through it. She wrangled and even managed to hide many of life's complexities that allowed me to focus on this project. Her unconditional support is perhaps more than what I deserve but was necessary to make this book possible. My smart and sweet daughters – Sophie, Zoe, and Sarah – were just as much instrumental in this support. At the beginning of this summer, my oldest daughter, Sophie, who had just turned 10, asked me about AI and ML. As I proceeded to talk to her with excitement, I realized how widespread these technologies have become in our society today that elementary school kids are talking and learning about them. Thanks, sweetheart, for continuing to inspire and challenge me.

I started my journey on the path of AI at the Indian Institute of Technology (IIT) – first in Bombay (Mumbai) and then in Madras (Chennai). I'm grateful to Prof. Pushpak Bhattacharyya (IITB) and Prof. B Yegnanarayana (IITM) for introducing me to the wonderful worlds of natural language processing and neural networks. Thanks to their tremendous patience and support, I continued my appreciation and exploration of these topics even after I left India.

A big reason I could embark on this book journey is that I'm an educator myself. The first drafts of almost everything in this book have come from my own teaching and advising. I have been fortunate enough to have access to some of the brightest and most curious minds in the world at major US universities, such as the University of Washington and Rutgers University. My colleagues and students from these places over the last many years have contributed substantively to my understanding of the material presented here; I have learned a lot by teaching. I would specifically call out Andrea Berg, Smart Chang, Yuzhen Qu, Daniel Saelid, and Bingbing Wen for their contributions to some of the writeups, problems, and solutions.

I have also had the opportunity to spend time at tech industry places such as Amazon, Getty Images, Microsoft, and Spotify. Through my work in applied science at these places, I learned a lot about making machine learning work in real-life applications. Some of those lessons have been poured into this book, especially with the "ML in Practice" boxes. I'm

thankful for these opportunities and the wonderful collaborators I had at these places, who are too many to list.

I am very grateful to the wonderful staff of Cambridge University Press for guiding me through the development of this book from the beginning. I would first call out Lauren Cowles and Stefanie Seaton. Having worked with them before for my textbook on data science, I knew how wonderful and supportive they were and decided to work with them again. They certainly did not disappoint! Through many rounds of feedback and design decisions, they once again ensured that the book meets the highest standards of quality and accessibility that one would expect from the Press. I am also grateful to Madelyn Glymour, who painstakingly not only proofread the whole manuscript, but also corrected several technical errors.

Finally, I want to thank the University of Washington iSchool and the Whiteley Center at Friday Harbor for supporting two writing retreats in 2021–2022 that allowed me to do critical writing work on this book.

I am almost certain that I have forgotten many more people to thank here, but they should know that it was a result of my forgetfulness and not ungratefulness.

About the Author

Dr. Chirag Shah is Professor in the Information School (iSchool) at the University of Washington (UW) in Seattle. He is also Adjunct Professor with the Paul G. Allen School of Computer Science & Engineering as well as the Human Centered Design & Engineering (HCDE) department. Before UW, he was at Rutgers University. He received his Ph.D. in Information Science from the University of North Carolina (UNC) Chapel Hill, and MS in Computer Science from the University of Massachusetts (UMass) Amherst.

Dr. Shah's research involves creating smart systems for search and recommendation. Examples include intelligent assistants and task managers. He also focuses on making these systems bias-free. He applies techniques from data science and ML for most of this work. He has published several books and peer-reviewed articles in these areas. He developed the Coagmento system for collaborative and social searching, IRIS (Information Retrieval and Interaction System) for investigating and implementing interactive IR activities, as well as several systems for collecting and analyzing data from social media channels, including the award-winning ContextMiner, InfoExtractor, TubeKit, and SOCRATES systems. He is the Founding Co-Director of Responsibility in AI Systems & Experiences (RAISE) at UW, a center focused on research and education around issues of bias, fairness, accountability, transparency, and ethics in AI. He is also the Founder and Director of the InfoSeeking Lab, where he investigates issues related to information seeking, interactive information retrieval, and social media. These research projects are supported by grants from the National Science Foundation (NSF), National Institute of Health (NIH), Institute of Museum and Library Services (IMLS), Amazon, Google, and Yahoo!. He has also served as a consultant to the United Nations Data Analytics on various data science projects involving social and political issues, peacekeeping, climate change, and energy.

He spent his sabbatical in 2018 at Spotify, working on voice-based search and recommendation problems. In 2019, as an Amazon Scholar, he worked with Amazon's Personalization team on applications involving personalized and task-oriented recommendations.

In 2020 he was a Visiting Researcher at Microsoft Research (MSR) AI and worked on an intelligent task manager. In 2021 he visited Getty Images to work on improving their search platform with embedding-based deep semantic approaches.

Dr. Shah teaches undergraduate, masters, and Ph.D. programs at UW, focusing on data science and ML. He teaches both on campus and online. He has also created a course for Coursera (Social Media Data Analytics) and taught several tutorials and short courses at international venues. Dr. Shah has written several books, including the textbook *A Hands-On Introduction to Data Science*, published by Cambridge University Press.