Machine Learning Fundamentals

This lucid, accessible introduction to supervised machine learning presents core concepts in a focused and logical way that is easy for beginners to follow. The author assumes basic calculus, linear algebra, probability and statistics but no prior exposure to machine learning. Coverage includes widely used traditional methods such as SVMs, boosted trees, HMMs, and LDAs, plus popular deep learning methods such as convolution neural nets, attention, transformers, and GANs. Organized in a coherent presentation framework that emphasizes the big picture, the text introduces each method clearly and concisely "from scratch" based on the fundamentals. All methods and algorithms are described by a clean and consistent style, with a minimum of unnecessary detail. Numerous case studies and concrete examples demonstrate how the methods can be applied in a variety of contexts.

Hui Jiang is a Professor of Electrical Engineering and Computer Science at York University, where he has been since 2002. His main research interests include machine learning, particularly deep learning, and its applications to speech and audio processing, natural language processing, and computer vision. Over the past 30 years, he has worked on a wide range of research problems from these areas and published hundreds of technical articles and papers in the mainstream journals and top-tier conferences. His works have won the prestigious IEEE Best Paper Award and the ACL Outstanding Paper honor.

Simplicity is the ultimate sophistication.

—Leonardo da Vinci

Machine Learning Fundamentals

A Concise Introduction

Hui Jiang

York University, Toronto

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/9781108837040
DOI: 10.1017/9781108938051

First published 2021

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

ISBN 978-1-108-83704-0 Hardback
ISBN 978-1-108-94002-3 Paperback

Contents

Preface

Machine learning used to be a niche area originating out of pattern classification in electrical engineering and artificial intelligence in computer science. Today, machine learning has grown into a very diverse discipline spanning a variety of topics in mathematics, science, and engineering. Because of the widespread use and increased power of computers, machine learning has found a plethora of relevant applications in almost all engineering domains and has made a huge impact on our society. In particular, with the boom of deep learning in recent years, thousands of new researchers and practitioners across academia and industry join forces every year to tackle machine learning and its applications. In many universities, machine learning has become one of the most popular advanced elective courses, highly demanded by senior undergraduates and graduates in almost all computer science and electrical engineering programs. The number of industrial job positions in machine learning, deep learning, and data science has dramatically increased in recent years, and this trend is expected to continue for at least the next 10 years due to the availability of a huge amount of data over the internet and personal devices.

Why This Book?

There are already plenty of well-written textbooks for machine learning, most of which exhaustively cover a wide range of topics in machine learning. In teaching my machine learning courses, I found that they are too challenging for beginners because of the vast range of presented topics and the overwhelming technical details associated with them. Many beginners have trouble with the heavy mathematical notation and equations, whereas others drown in all the technical details and fail to grasp the essence of these machine learning methods.

In contrast, this book is intended to present the fundamental machine learning concepts, algorithms, and principles in a concise and lucid manner, without heavy mathematical machinery and excess detail. I have been selective in terms of the topics so that it can all be covered in an introductory course, rather than making it comprehensive enough to cover all machine learning topics. I chose to cover only relatively mature topics primarily related to supervised learning, which I believe are not only fundamental to the field of machine learning but also significant enough to have made an impact in both academia and industry. In other words, some satisfactory and feasible solutions have already been developed for these topics so that they are able to address not just toy problems

but many interesting problems arising in the real world. At the same time, I have tried to omit many minor issues surrounding the central topics so that beginners will not be distracted by these purely technical details.

Instead of covering the selected topics separately, one after another, I have tried to organize all machine learning topics into a coherent structure to give readers a big picture of the entire field. All topics are arranged into coherent groups, and the individual chapters are dedicated to covering all logically relevant methods in each group. After reading each chapter, readers can immediately understand the differences between them, grasp their relevance, and also know how these methods fit into the big picture of machine learning.

This book also aims to reflect the latest advancements in the field. I have included significant coverage on several important recent techniques, such as *transformers*, which have come to dominate many natural-language-processing tasks; *batch norm* and *ADAM optimization*, which are popular in learning large and deep neural networks; and recently popular deep generative models such as *variational autoencoders (VAEs)* and *generative adversarial nets (GANs)*.

For all topics in this book, I provide enough technical depth to explain the motivation, principles, and methodology in a professional manner. As much as possible, I derive the machine learning methods from scratch using rigorous mathematics to highlight the core ideas behind them. For critical theoretical results, I have included many important theorems and some light proofs. The important mathematical topics and methods that modern machine learning methods are built on are thoroughly reviewed in Chapter 2. However, readers do need a good background in *calculus*, *linear algebra*, and *probability and statistics* to be able to follow the descriptions and discussions in this book. Throughout the book, I have also done my best to present all technical content using clean and consistent mathematical notations and represent all algorithms in this book as concise linear algebra formulas, which can be translated almost line by line into efficient code using a programming language supporting vectorization, such as MATLAB or Python.

Whom Is This Book For?

This book is primarily written as a textbook for an introductory course on machine learning for senior undergraduate students in computer science and computer/ software/electrical engineering programs or first-year graduate students in many science, engineering, and applied mathematics programs who are interested in basic machine learning methods for their own research problems. I also hope it will be useful as a self-study or reference book for researchers who wish to apply machine learning methods to solve their own problems, as well

as industrial practitioners who want to understand the concepts and principles behind the popular machine learning methods they implement. Given the large number of machine learning software programs and toolkits freely available today, it is often not hard to write code to run fairly complicated machine learning algorithms. However, in many cases, knowledge of the principles and mathematics behind these algorithms is required to tune these algorithms in order to deliver optimal results for the task at hand.

Online Resources

This book is accompanied by the following GitHub repository:

`https://github.com/iNCML/MachineLearningBook`

This website provides a variety of supplementary materials to support this book, including the following:

- Lecture slides per chapter
- Code samples for some lab projects (MATLAB or Python)

Meanwhile, readers and instructors can also provide their feedback, suggestions, and comments on this book as *issues* through the GitHub repository. I will reply to these requests as much as possible.

How to Use This Book

I have made much effort to keep this book succinct and only cover the most important issues for each selected topic. I encourage readers to read all chapters in order because I have tried my best to arrange a wide range of machine learning topics in a coherent structure. For each machine learning method, I have thoroughly covered the motivation, main ideas, concepts, methodology, and algorithms in the main text and sometimes have left extensive issues and extra technical details or extensions as chapter-end exercises. Readers may optionally follow these links to work on these exercises and practice the main ideas discussed in the text.

- **For a Semester-Long Course**
  Instructors may use this book as the primary or alternate textbook for a standard semester-long introductory course (about 10–12 weeks) on machine learning in the fourth year of a computer science, engineering, or applied mathematics program. I suggest covering the following topics in order:

- Chapter 1: Introduction (0.5 week)
- Chapter 2: Mathematical Foundation (1.5 weeks)
- Chapter 4: Feature Extraction (1 week)
- Chapter 5: Statistical Learning Theory (0.5 week)
  - §5.1 Formulation of Discriminative Models
  - §5.2 Learnability
- Chapter 6: Linear Models (1.5 weeks)
- Chapter 7: Learning Discriminative Models (1 week)
  - §7.1 General Framework
  - §7.2 Ridge and LASSO
  - §7.3 Matrix Factorization
- Chapter 8: Neural Networks (2 weeks)
- Chapter 9: Ensemble Learning (1 week)
- Chapter 10: Overview of Generative Models (1 week)
- Chapter 11: Unimodal Models (1 week)
  - §11.1 Gaussian Models
  - §11.2 Multinomial Models
  - §11.3 Markov Chain Models
- Chapter 12 Mixture Models (1 week)
  - §12.1 Formulation
  - §12.2 EM Method
  - §12.3 Gaussian Mixture Models

► **For a Year-Long Full Course**

Instructors may also use this book as the primary or alternate textbook for a year-long full course on machine learning (20–24 weeks) to give balanced coverage of both discriminative and generative models. The first half focuses on the mathematical preparation and discriminative models, whereas the second half gives full exposure to a variety of topics in generative models, including Chapter 13: Entangled Models, Chapter 14: Bayesian Learning, and Chapter 15: Graphical Models.

If time is tight, instructors may skip some optional topics, such as §4.3 Manifold Learning, §7.4 Dictionary Learning, §11.4 Generalized Linear Models, §12.4 Hidden Markov Models, or §14.4 Gaussian Processes.

► **For Self-Study**

All self-study readers are strongly recommended to go through the book in order. This will give a smooth transition from one topic to another, generally progressing gradually from easy topics to hard ones. Depending on one's own interests, readers may choose to skip any of the following advanced topics without affecting the understanding of other parts:

- §4.3 Manifold Learning
- §7.4 Dictionary Learning
- §11.4 Generalized Linear Models
- §12.4 Hidden Markov Models
- §14.4 Gaussian Processes

Acknowledgments

Writing a textbook is a very challenging task. This book would not have been possible without help and supports from a large number of people.

Most content in this book evolved from the lecture notes I have used for many years to teach a machine learning course in the Department of Electrical Engineering and Computer Science at York University in Toronto, Canada. I am grateful to York University for the long-standing support of my teaching and research there.

I also thank Zoubin Ghahramani, David Blei, and Huy Vu for granting permission to use their materials in this book.

Many people have helped to significantly improve this book by proofreading the early draft and providing valuable comments and suggestions, including Dong Yu, Kelvin Jiang, Behnam Asadi, Jia Pan, Yong Ge, William Fu, Xiaodan Zhu, Chao Wang, Jiebo Luo, Hanjia Lyu, Joyce Luo, Qiang Huo, Chunxiao Zhou, Wei Zhang, Maria Koshkina, Zhuoran Li, Junfei Wang, and Parham Eftekhar. My special thanks to all of them!

Finally, I would like to thank my family, Iris and Kelvin, and my parents for their endless support and love throughout the time of writing this book as well as my career and life.

Notation

This list describes some of the symbols that are used within this book.

$\boldsymbol{\mu}$	The mean vector of a multivariate Gaussian
$\boldsymbol{\Sigma}$	The covariance matrix of a multivariate Gaussian
$\mathbb{E}[\cdot]$	The expectation or the mean
$\mathbb{E}_X[\cdot]$	The expectation with respect to X
$\mathbb{H}$	Model space
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^n$	The set of n-dimensional real vectors
$\mathbb{R}^{m\times n}$	The set of $m \times n$ real matrices
$\mathbb{W}$	The set of all parameters in a neural network
$\mathbf{S}$	The sample covariance matrix
$\mathbf{w} * \mathbf{x}$	The convolution sum of $\mathbf{w}$ and $\mathbf{x}$
$\mathbf{w} \cdot \mathbf{x}$	The inner product of two vectors $\mathbf{w}$ and $\mathbf{x}$
$\mathbf{w} \odot \mathbf{x}$	The element-wise multiplication of $\mathbf{w}$ and $\mathbf{x}$
$\mathbf{W}$	A weight matrix
$\mathbf{w}$	A weight vector
$\mathbf{x}$	A feature vector
$\nabla f(\mathbf{x})$	The gradient of a function $f(\mathbf{x})$
$\Pr(A)$	The probability of an event A
$\|\mathbf{w}\|$	The norm (or L_2 norm) of a vector $\mathbf{w}$
$\|\mathbf{w}\|_p$	The L_p norm of a vector $\mathbf{w}$
$f(\mathbf{x};\boldsymbol{\theta})$	A function of $\mathbf{x}$ with the parameter $\boldsymbol{\theta}$
$f_{\boldsymbol{\theta}}(\mathbf{x})$	A function of $\mathbf{x}$ with the parameter $\boldsymbol{\theta}$
$l(\boldsymbol{\theta})$	A log-likelihood function of the model parameter $\boldsymbol{\theta}$
$m \ll n$	m is much less than n
$p(\mathbf{x},\mathbf{y})$	A joint distribution of $\mathbf{x}$ and $\mathbf{y}$
$p(\mathbf{y}\,\|\,\mathbf{x})$	A conditional distribution of $\mathbf{y}$ given $\mathbf{x}$
$p_{\boldsymbol{\theta}}(\mathbf{x})$	A probability distribution of $\mathbf{x}$ with the parameter $\boldsymbol{\theta}$
$Q(\mathbb{W};\mathbf{x})$	An objective function of the model parameters $\mathbb{W}$ given the data $\mathbf{x}$
$\boldsymbol{\theta}$	Model parameter

Summary of the General Notation Rules

Notation	Meaning	Examples
Lowercase letters	A scalar	x, y, n, m, x_i, x_{ij}
	A function	$f(\cdot)$, $p(\cdot)$, $g(\cdot)$, $h(\cdot)$
Lowercase letters in bold	A column vector	$\mathbf{w}$, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, $\mathbf{a}$, $\mathbf{b}$ $\boldsymbol{\mu}$, $\boldsymbol{\nu}$
Uppercase letters	A random variable	X, Y, X_i, X_j
	A function	$Q(\cdot)$, $\Phi(\cdot,\cdot)$
Uppercase letters in bold	A matrix	$\mathbf{A}$, $\mathbf{W}$, $\mathbf{S}$ $\boldsymbol{\Sigma}$, $\boldsymbol{\Phi}$
Uppercase letters in blackboard bold	A set of numbers	$\mathbb{N}$, $\mathbb{R}$
	A set of parameters	$\mathbb{B}$, $\mathbb{W}$, $\mathbb{V}$
Uppercase letters in calligraphy	A set of data	$\mathcal{D}$, $\mathcal{D}_N$