
Foreword

Machine learning is the latest in a long line of attempts to distill human knowledge and reasoning into a form that is suitable for constructing machines and engineering automated systems. As machine learning becomes more ubiquitous and its software packages become easier to use, it is natural and desirable that the low-level technical details are abstracted away and hidden from the practitioner. However, this brings with it the danger that a practitioner becomes unaware of the design decisions and, hence, the limits of machine learning algorithms.

The enthusiastic practitioner who is interested to learn more about the magic behind successful machine learning algorithms currently faces a daunting set of pre-requisite knowledge:

- Programming languages and data analysis tools
- Large-scale computation and the associated frameworks
- Mathematics and statistics and how machine learning builds on it

At universities, introductory courses on machine learning tend to spend early parts of the course covering some of these pre-requisites. For historical reasons, courses in machine learning tend to be taught in the computer science department, where students are often trained in the first two areas of knowledge, but not so much in mathematics and statistics.

Current machine learning textbooks primarily focus on machine learning algorithms and methodologies and assume that the reader is competent in mathematics and statistics. Therefore, these books only spend one or two chapters on background mathematics, either at the beginning of the book or as appendices. We have found many people who want to delve into the foundations of basic machine learning methods who struggle with the mathematical knowledge required to read a machine learning textbook. Having taught undergraduate and graduate courses at universities, we find that the gap between high school mathematics and the mathematics level required to read a standard machine learning textbook is too big for many people.

This book brings the mathematical foundations of basic machine learning concepts to the fore and collects the information in a single place so that this skills gap is narrowed or even closed.

Why Another Book on Machine Learning?

Machine learning builds upon the language of mathematics to express concepts that seem intuitively obvious but that are surprisingly difficult to formalize. Once formalized properly, we can gain insights into the task we want to solve. One common complaint of students of mathematics around the globe is that the topics covered seem to have little relevance to practical problems. We believe that machine learning is an obvious and direct motivation for people to learn mathematics.

“Math is linked in the popular mind with phobia and anxiety. You’d think we’re discussing spiders.” (Strogatz, 2014, page 281)

This book is intended to be a guidebook to the vast mathematical literature that forms the foundations of modern machine learning. We motivate the need for mathematical concepts by directly pointing out their usefulness in the context of fundamental machine learning problems. In the interest of keeping the book short, many details and more advanced concepts have been left out. Equipped with the basic concepts presented here, and how they fit into the larger context of machine learning, the reader can find numerous resources for further study, which we provide at the end of the respective chapters. For readers with a mathematical background, this book provides a brief but precisely stated glimpse of machine learning. In contrast to other books that focus on methods and models of machine learning (MacKay, 2003; Bishop, 2006; Alpaydin, 2010; Barber, 2012; Murphy, 2012; Shalev-Shwartz and Ben-David, 2014; Rogers and Girolami, 2016) or programmatic aspects of machine learning (Müller and Guido, 2016; Raschka and Mirjalili, 2017; Chollet and Allaire, 2018), we provide only four representative examples of machine learning algorithms. Instead, we focus on the mathematical concepts behind the models themselves. We hope that readers will be able to gain a deeper understanding of the basic questions in machine learning and connect practical questions arising from the use of machine learning with fundamental choices in the mathematical model.

We do not aim to write a classical machine learning book. Instead, our intention is to provide the mathematical background, applied to four central machine learning problems, to make it easier to read other machine learning textbooks.

Who Is the Target Audience?

As applications of machine learning become widespread in society, we believe that everybody should have some understanding of its underlying principles. This book is written in an academic mathematical style, which enables us to be precise about the concepts behind machine learning. We encourage readers unfamiliar with this seemingly terse style to persevere and to keep the goals of each topic in mind. We sprinkle comments and remarks throughout the text, in the hope that it provides useful guidance with respect to the big picture.

The book assumes the reader to have mathematical knowledge commonly

covered in high school mathematics and physics. For example, the reader should have seen derivatives and integrals before, and geometric vectors in two or three dimensions. Starting from there, we generalize these concepts. Therefore, the target audience of the book includes undergraduate university students, evening learners and learners participating in online machine learning courses.

In analogy to music, there are three types of interaction that people have with machine learning:

Astute Listener The democratization of machine learning by the provision of open-source software, online tutorials and cloud-based tools allows users to not worry about the specifics of pipelines. Users can focus on extracting insights from data using off-the-shelf tools. This enables non-tech-savvy domain experts to benefit from machine learning. This is similar to listening to music; the user is able to choose and discern between different types of machine learning, and benefits from it. More experienced users are like music critics, asking important questions about the application of machine learning in society such as ethics, fairness, and privacy of the individual. We hope that this book provides a foundation for thinking about the certification and risk management of machine learning systems, and allows them to use their domain expertise to build better machine learning systems.

Experienced Artist Skilled practitioners of machine learning can plug and play different tools and libraries into an analysis pipeline. The stereotypical practitioner would be a data scientist or engineer who understands machine learning interfaces and their use cases, and is able to perform wonderful feats of prediction from data. This is similar to a virtuoso playing music, where highly skilled practitioners can bring existing instruments to life and bring enjoyment to their audience. Using the mathematics presented here as a primer, practitioners would be able to understand the benefits and limits of their favorite method, and to extend and generalize existing machine learning algorithms. We hope that this book provides the impetus for more rigorous and principled development of machine learning methods.

Fledgling Composer As machine learning is applied to new domains, developers of machine learning need to develop new methods and extend existing algorithms. They are often researchers who need to understand the mathematical basis of machine learning and uncover relationships between different tasks. This is similar to composers of music who, within the rules and structure of musical theory, create new and amazing pieces. We hope this book provides a high-level overview of other technical books for people who want to become composers of machine learning. There is a great need in society for new researchers who are able to propose and explore novel approaches for attacking the many challenges of learning from data.

Acknowledgments

We are grateful to many people who looked at early drafts of the book and suffered through painful expositions of concepts. We tried to implement their ideas that we did not vehemently disagree with. We would like to especially acknowledge Christfried Webers for his careful reading of many parts of the book, and his detailed suggestions on structure and presentation. Many friends and colleagues have also been kind enough to provide their time and energy on different versions of each chapter. We have been lucky to benefit from the generosity of the online community, who have suggested improvements via <https://github.com>, which greatly improved the book.

The following people have found bugs, proposed clarifications and suggested relevant literature, either via <https://github.com> or personal communication. Their names are sorted alphabetically.

Abdul-Ganiy Usman	Ellen Broad
Adam Gaier	Fengkuangtian Zhu
Adele Jackson	Fiona Condon
Aditya Menon	Georgios Theodorou
Alasdair Tran	He Xin
Aleksandar Krnjaic	Irene Raissa Kameni
Alexander Makrigiorgos	Jakub Nabaglo
Alfredo Canziani	James Hensman
Ali Shafti	Jamie Liu
Amr Khalifa	Jean Kaddour
Andrew Tanggara	Jean-Paul Ebejer
Angus Gruen	Jerry Qiang
Antal A. Buss	Jitesh Sindhare
Antoine Toisoul Le Cann	John Lloyd
Areg Sarvazyan	Jonas Ngnawe
Artem Artemev	Jon Martin
Artyom Stepanov	Justin Hsi
Bill Kromydas	Kai Arulkumaran
Bob Williamson	Kamil Dreczkowski
Boon Ping Lim	Lily Wang
Chao Qu	Lionel Tondji Ngoupeyou
Cheng Li	Lydia Knüfing
Chris Sherlock	Mahmoud Aslan
Christopher Gray	Mark Hartenstein
Daniel McNamara	Mark van der Wilk
Daniel Wood	Markus Hegland
Darren Siegel	Martin Hewing
David Johnston	Matthew Alger
Dawei Chen	Matthew Lee

Maximus McCann	Shakir Mohamed
Mengyan Zhang	Shawn Berry
Michael Bennett	Sheikh Abdul Raheem Ali
Michael Pedersen	Sheng Xue
Minjeong Shin	Sridhar Thiagarajan
Mohammad Malekzadeh	Syed Nouman Hasany
Naveen Kumar	Szymon Brych
Nico Montali	Thomas Bühler
Oscar Armas	Timur Sharapov
Patrick Henriksen	Tom Melamed
Patrick Wieschollek	Vincent Adam
Pattarawat Chormai	Vincent Dutordoir
Paul Kelly	Vu Minh
Petros Christodoulou	Wasim Aftab
Piotr Januszewski	Wen Zhi
Pranav Subramani	Wojciech Stokowiec
Quyu Kong	Xiaonan Chong
Ragib Zaman	Xiaowei Zhang
Rui Zhang	Yazhou Hao
Ryan-Rhys Griffiths	Yicheng Luo
Salomon Kabongo	Young Lee
Samuel Ogunmola	Yu Lu
Sandeep Mavadia	Yun Cheng
Sarvesh Nikumbh	Yuxiao Huang
Sebastian Raschka	Zac Cranko
Senanayak Sesh Kumar Karri	Zijian Cao
Seung-Heon Baek	Zoe Nolan
Shahbaz Chaudhary	

Contributors through GitHub, whose real names were not listed on their GitHub profile, are:

SamDataMad	insad	empet
bumptiousmonkey	HorizonP	victorBigand
idoamihai	cs-maillist	17SKYE
deepakiim	kudo23	jessjing1995

We are also very grateful to Parameswaran Raman and the many anonymous reviewers, organized by Cambridge University Press, who read one or more chapters of earlier versions of the manuscript, and provided constructive criticism that led to considerable improvements. A special mention goes to Dinesh Singh Negi, our \LaTeX support, for detailed and prompt advice about \LaTeX -related issues. Last but not least, we are very grateful to our editor Lauren Cowles, who has been patiently guiding us through the gestation process of this book.

Table of Symbols

Symbol	Typical meaning
$a, b, c, \alpha, \beta, \gamma$	Scalars are lowercase
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	Vectors are bold lowercase
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	Matrices are bold uppercase
$\mathbf{x}^\top, \mathbf{A}^\top$	Transpose of a vector or matrix
\mathbf{A}^{-1}	Inverse of a matrix
$\langle \mathbf{x}, \mathbf{y} \rangle$	Inner product of \mathbf{x} and \mathbf{y}
$\mathbf{x}^\top \mathbf{y}$	Dot product of \mathbf{x} and \mathbf{y}
$B = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$	(Ordered) tuple
$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]$	Matrix of column vectors stacked horizontally
$\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$	Set of vectors (unordered)
\mathbb{Z}, \mathbb{N}	Integers and natural numbers, respectively
\mathbb{R}, \mathbb{C}	Real and complex numbers, respectively
\mathbb{R}^n	n -dimensional vector space of real numbers
$\forall x$	Universal quantifier: for all x
$\exists x$	Existential quantifier: there exists x
$a := b$	a is defined as b
$a =: b$	b is defined as a
$a \propto b$	a is proportional to b , i.e., $a = \text{constant} \cdot b$
$g \circ f$	Function composition: “ g after f ”
\Longleftrightarrow	If and only if
\implies	Implies
\mathcal{A}, \mathcal{C}	Sets
$a \in \mathcal{A}$	a is an element of set \mathcal{A}
\emptyset	Empty set
$\mathcal{A} \setminus \mathcal{B}$	\mathcal{A} without \mathcal{B} : the set of elements in \mathcal{A} but not in \mathcal{B}
D	Number of dimensions; indexed by $d = 1, \dots, D$
N	Number of data points; indexed by $n = 1, \dots, N$
\mathbf{I}_m	Identity matrix of size $m \times m$
$\mathbf{0}_{m,n}$	Matrix of zeros of size $m \times n$
$\mathbf{1}_{m,n}$	Matrix of ones of size $m \times n$
\mathbf{e}_i	Standard/canonical vector (where i is the component that is 1)
\dim	Dimensionality of vector space
$\text{rk}(\mathbf{A})$	Rank of matrix \mathbf{A}
$\text{Im}(\Phi)$	Image of linear mapping Φ
$\ker(\Phi)$	Kernel (null space) of a linear mapping Φ
$\text{span}[\mathbf{b}_1]$	Span (generating set) of \mathbf{b}_1
$\text{tr}(\mathbf{A})$	Trace of \mathbf{A}
$\det(\mathbf{A})$	Determinant of \mathbf{A}
$ \cdot $	Absolute value or determinant (depending on context)
$\ \cdot\ $	Norm; Euclidean, unless specified
λ	Eigenvalue or Lagrange multiplier
E_λ	Eigenspace corresponding to eigenvalue λ

Symbol	Typical meaning
$\mathbf{x} \perp \mathbf{y}$	Vectors \mathbf{x} and \mathbf{y} are orthogonal
V	Vector space
V^\perp	Orthogonal complement of vector space V
$\sum_{n=1}^N x_n$	Sum of the x_n : $x_1 + \dots + x_N$
$\prod_{n=1}^N x_n$	Product of the x_n : $x_1 \cdot \dots \cdot x_N$
$\boldsymbol{\theta}$	Parameter vector
$\frac{\partial f}{\partial x}$	Partial derivative of f with respect to x
$\frac{df}{dx}$	Total derivative of f with respect to x
∇	Gradient
$f_* = \min_x f(x)$	The smallest function value of f
$x_* \in \arg \min_x f(x)$	The value x_* that minimizes f (note: $\arg \min$ returns a set of values)
\mathcal{L}	Lagrangian
\mathcal{L}	Negative log-likelihood
$\binom{n}{k}$	Binomial coefficient, n choose k
$\mathbb{V}_X[\mathbf{x}]$	Variance of \mathbf{x} with respect to the random variable X
$\mathbb{E}_X[\mathbf{x}]$	Expectation of \mathbf{x} with respect to the random variable X
$\text{Cov}_{X,Y}[\mathbf{x}, \mathbf{y}]$	Covariance between \mathbf{x} and \mathbf{y} .
$X \perp\!\!\!\perp Y \mid Z$	X is conditionally independent of Y given Z
$X \sim p$	Random variable X is distributed according to p
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\text{Ber}(\mu)$	Bernoulli distribution with parameter μ
$\text{Bin}(N, \mu)$	Binomial distribution with parameters N, μ
$\text{Beta}(\alpha, \beta)$	Beta distribution with parameters α, β

Table of Abbreviations and Acronyms

Acronym	Meaning
e.g.	Exempli gratia (Latin: for example)
GMM	Gaussian mixture model
i.e.	Id est (Latin: this means)
i.i.d.	Independent, identically distributed
MAP	Maximum a posteriori
MLE	Maximum likelihood estimation/estimator
ONB	Orthonormal basis
PCA	Principal component analysis
PPCA	Probabilistic principal component analysis
REF	Row-echelon form
SPD	Symmetric, positive definite
SVM	Support vector machine

