

MATHEMATICS FOR MACHINE LEARNING

Marc Peter Deisenroth
A. Aldo Faisal
Cheng Soon Ong

Contents

<i>Foreword</i>	1
Part I Mathematical Foundations	9
1 Introduction and Motivation	11
1.1 Finding Words for Intuitions	12
1.2 Two Ways to Read This Book	13
1.3 Exercises and Feedback	16
2 Linear Algebra	17
2.1 Systems of Linear Equations	19
2.2 Matrices	22
2.3 Solving Systems of Linear Equations	27
2.4 Vector Spaces	35
2.5 Linear Independence	40
2.6 Basis and Rank	44
2.7 Linear Mappings	48
2.8 Affine Spaces	61
2.9 Further Reading	63
Exercises	64
3 Analytic Geometry	70
3.1 Norms	71
3.2 Inner Products	72
3.3 Lengths and Distances	75
3.4 Angles and Orthogonality	76
3.5 Orthonormal Basis	78
3.6 Orthogonal Complement	79
3.7 Inner Product of Functions	80
3.8 Orthogonal Projections	81
3.9 Rotations	91
3.10 Further Reading	94
Exercises	96
4 Matrix Decompositions	98
4.1 Determinant and Trace	99

4.2	Eigenvalues and Eigenvectors	105
4.3	Cholesky Decomposition	114
4.4	Eigendecomposition and Diagonalization	115
4.5	Singular Value Decomposition	119
4.6	Matrix Approximation	129
4.7	Matrix Phylogeny	134
4.8	Further Reading	135
	Exercises	137
5	Vector Calculus	139
5.1	Differentiation of Univariate Functions	141
5.2	Partial Differentiation and Gradients	146
5.3	Gradients of Vector-Valued Functions	149
5.4	Gradients of Matrices	155
5.5	Useful Identities for Computing Gradients	158
5.6	Backpropagation and Automatic Differentiation	159
5.7	Higher-Order Derivatives	164
5.8	Linearization and Multivariate Taylor Series	165
5.9	Further Reading	170
	Exercises	170
6	Probability and Distributions	172
6.1	Construction of a Probability Space	172
6.2	Discrete and Continuous Probabilities	178
6.3	Sum Rule, Product Rule, and Bayes' Theorem	183
6.4	Summary Statistics and Independence	186
6.5	Gaussian Distribution	197
6.6	Conjugacy and the Exponential Family	205
6.7	Change of Variables/Inverse Transform	214
6.8	Further Reading	221
	Exercises	222
7	Continuous Optimization	225
7.1	Optimization Using Gradient Descent	227
7.2	Constrained Optimization and Lagrange Multipliers	233
7.3	Convex Optimization	236
7.4	Further Reading	246
	Exercises	247
Part II Central Machine Learning Problems		249
8	When Models Meet Data	251
8.1	Data, Models, and Learning	251
8.2	Empirical Risk Minimization	258
8.3	Parameter Estimation	265
8.4	Probabilistic Modeling and Inference	272
8.5	Directed Graphical Models	278

<i>Contents</i>	iii
8.6 Model Selection	283
9 Linear Regression	289
9.1 Problem Formulation	291
9.2 Parameter Estimation	292
9.3 Bayesian Linear Regression	303
9.4 Maximum Likelihood as Orthogonal Projection	313
9.5 Further Reading	315
10 Dimensionality Reduction with Principal Component Analysis	317
10.1 Problem Setting	318
10.2 Maximum Variance Perspective	320
10.3 Projection Perspective	325
10.4 Eigenvector Computation and Low-Rank Approximations	333
10.5 PCA in High Dimensions	335
10.6 Key Steps of PCA in Practice	336
10.7 Latent Variable Perspective	339
10.8 Further Reading	343
11 Density Estimation with Gaussian Mixture Models	348
11.1 Gaussian Mixture Model	349
11.2 Parameter Learning via Maximum Likelihood	350
11.3 EM Algorithm	360
11.4 Latent-Variable Perspective	363
11.5 Further Reading	368
12 Classification with Support Vector Machines	370
12.1 Separating Hyperplanes	372
12.2 Primal Support Vector Machine	374
12.3 Dual Support Vector Machine	383
12.4 Kernels	388
12.5 Numerical Solution	390
12.6 Further Reading	392
<i>References</i>	395
<i>Index</i>	407

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 of 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 github.com, which greatly improved the book.

The following people have found bugs, proposed clarifications and suggested relevant literature, either via 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 Ngawane
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