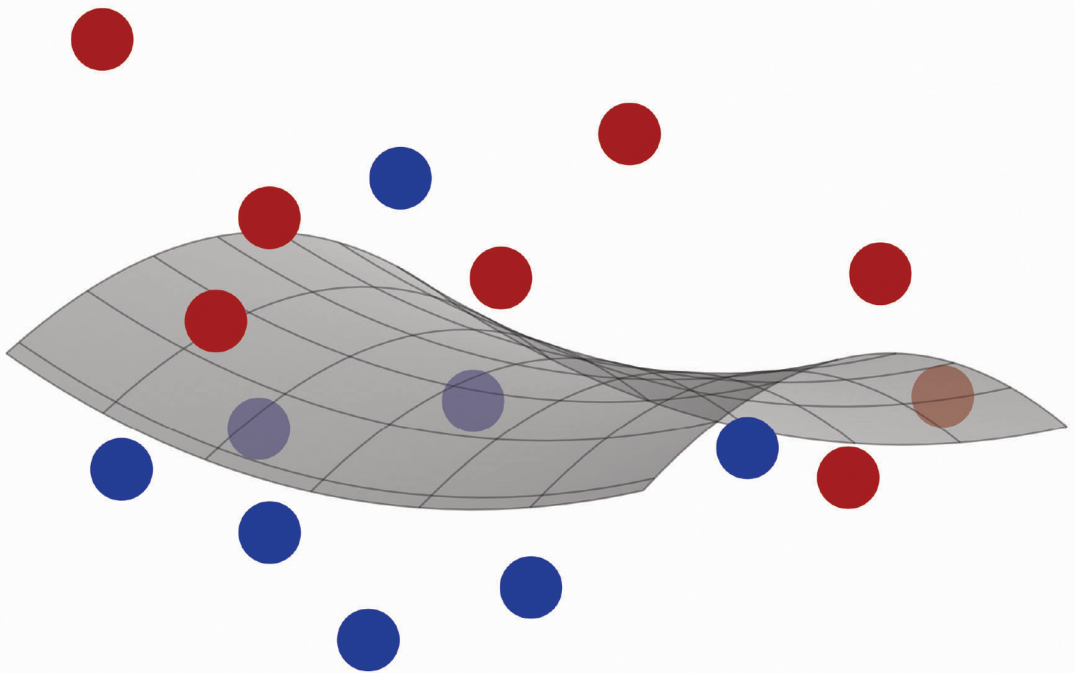


Foundations of Machine Learning

second edition



Mehryar Mohri,
Afshin Rostamizadeh,
and Ameet Talwalkar

Foundations of Machine Learning

second edition

Adaptive Computation and Machine Learning

Francis Bach, Editor

A complete list of books published in The Adaptive Computations and Machine Learning series appears at the back of this book.

Foundations of Machine Learning

second edition

Mehryar Mohri

Afshin Rostamizadeh

Ameet Talwalkar

The MIT Press
Cambridge, Massachusetts
London, England

© 2018 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in L^AT_EX by the authors. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Mohri, Mehryar, author. | Rostamizadeh, Afshin, author. | Talwalkar, Ameet, author.

Title: Foundations of machine learning / Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar.

Description: Second edition. | Cambridge, MA : The MIT Press, [2018] | Series: Adaptive computation and machine learning series | Includes bibliographical references and index.

Identifiers: LCCN 2018022812 | ISBN 9780262039406 (hardcover : alk. paper)

Subjects: LCSH: Machine learning. | Computer algorithms.

Classification: LCC Q325.5 .M64 2018 | DDC 006.3/1--dc23 LC record available at <https://lccn.loc.gov/2018022812>

10 9 8 7 6 5 4 3 2 1

Contents

Preface	xiii
1 Introduction	1
1.1 What is machine learning?	1
1.2 What kind of problems can be tackled using machine learning?	2
1.3 Some standard learning tasks	3
1.4 Learning stages	4
1.5 Learning scenarios	6
1.6 Generalization	7
2 The PAC Learning Framework	9
2.1 The PAC learning model	9
2.2 Guarantees for finite hypothesis sets — consistent case	15
2.3 Guarantees for finite hypothesis sets — inconsistent case	19
2.4 Generalities	21
2.4.1 Deterministic versus stochastic scenarios	21
2.4.2 Bayes error and noise	22
2.5 Chapter notes	23
2.6 Exercises	23
3 Rademacher Complexity and VC-Dimension	29
3.1 Rademacher complexity	30
3.2 Growth function	34
3.3 VC-dimension	36
3.4 Lower bounds	43
3.5 Chapter notes	48
3.6 Exercises	50
4 Model Selection	61
4.1 Estimation and approximation errors	61
4.2 Empirical risk minimization (ERM)	62
4.3 Structural risk minimization (SRM)	64

4.4	Cross-validation	68
4.5	n -Fold cross-validation	71
4.6	Regularization-based algorithms	72
4.7	Convex surrogate losses	73
4.8	Chapter notes	77
4.9	Exercises	78
5	Support Vector Machines	79
5.1	Linear classification	79
5.2	Separable case	80
5.2.1	Primal optimization problem	81
5.2.2	Support vectors	83
5.2.3	Dual optimization problem	83
5.2.4	Leave-one-out analysis	85
5.3	Non-separable case	87
5.3.1	Primal optimization problem	88
5.3.2	Support vectors	89
5.3.3	Dual optimization problem	90
5.4	Margin theory	91
5.5	Chapter notes	100
5.6	Exercises	100
6	Kernel Methods	105
6.1	Introduction	105
6.2	Positive definite symmetric kernels	108
6.2.1	Definitions	108
6.2.2	Reproducing kernel Hilbert space	110
6.2.3	Properties	112
6.3	Kernel-based algorithms	116
6.3.1	SVMs with PDS kernels	116
6.3.2	Representer theorem	117
6.3.3	Learning guarantees	117
6.4	Negative definite symmetric kernels	119
6.5	Sequence kernels	121
6.5.1	Weighted transducers	122
6.5.2	Rational kernels	126
6.6	Approximate kernel feature maps	130
6.7	Chapter notes	135
6.8	Exercises	137
7	Boosting	145
7.1	Introduction	145
7.2	AdaBoost	146
7.2.1	Bound on the empirical error	149
7.2.2	Relationship with coordinate descent	150
7.2.3	Practical use	154

7.3	Theoretical results	154
7.3.1	VC-dimension-based analysis	154
7.3.2	L_1 -geometric margin	155
7.3.3	Margin-based analysis	157
7.3.4	Margin maximization	161
7.3.5	Game-theoretic interpretation	162
7.4	L_1 -regularization	165
7.5	Discussion	167
7.6	Chapter notes	168
7.7	Exercises	170
8	On-Line Learning	177
8.1	Introduction	178
8.2	Prediction with expert advice	178
8.2.1	Mistake bounds and Halving algorithm	179
8.2.2	Weighted majority algorithm	181
8.2.3	Randomized weighted majority algorithm	183
8.2.4	Exponential weighted average algorithm	186
8.3	Linear classification	190
8.3.1	Perceptron algorithm	190
8.3.2	Winnow algorithm	198
8.4	On-line to batch conversion	201
8.5	Game-theoretic connection	204
8.6	Chapter notes	205
8.7	Exercises	206
9	Multi-Class Classification	213
9.1	Multi-class classification problem	213
9.2	Generalization bounds	215
9.3	Uncombined multi-class algorithms	221
9.3.1	Multi-class SVMs	221
9.3.2	Multi-class boosting algorithms	222
9.3.3	Decision trees	224
9.4	Aggregated multi-class algorithms	228
9.4.1	One-versus-all	229
9.4.2	One-versus-one	229
9.4.3	Error-correcting output codes	231
9.5	Structured prediction algorithms	233
9.6	Chapter notes	235
9.7	Exercises	237
10	Ranking	239
10.1	The problem of ranking	240
10.2	Generalization bound	241
10.3	Ranking with SVMs	243

10.4	RankBoost	244
10.4.1	Bound on the empirical error	246
10.4.2	Relationship with coordinate descent	248
10.4.3	Margin bound for ensemble methods in ranking	250
10.5	Bipartite ranking	251
10.5.1	Boosting in bipartite ranking	252
10.5.2	Area under the ROC curve	255
10.6	Preference-based setting	257
10.6.1	Second-stage ranking problem	257
10.6.2	Deterministic algorithm	259
10.6.3	Randomized algorithm	260
10.6.4	Extension to other loss functions	262
10.7	Other ranking criteria	262
10.8	Chapter notes	263
10.9	Exercises	264
11	Regression	267
11.1	The problem of regression	267
11.2	Generalization bounds	268
11.2.1	Finite hypothesis sets	268
11.2.2	Rademacher complexity bounds	269
11.2.3	Pseudo-dimension bounds	271
11.3	Regression algorithms	275
11.3.1	Linear regression	275
11.3.2	Kernel ridge regression	276
11.3.3	Support vector regression	281
11.3.4	Lasso	285
11.3.5	Group norm regression algorithms	289
11.3.6	On-line regression algorithms	289
11.4	Chapter notes	290
11.5	Exercises	292
12	Maximum Entropy Models	295
12.1	Density estimation problem	295
12.1.1	Maximum Likelihood (ML) solution	296
12.1.2	Maximum a Posteriori (MAP) solution	297
12.2	Density estimation problem augmented with features	297
12.3	Maxent principle	298
12.4	Maxent models	299
12.5	Dual problem	299
12.6	Generalization bound	303
12.7	Coordinate descent algorithm	304
12.8	Extensions	306
12.9	L_2 -regularization	308

12.10	Chapter notes	312
12.11	Exercises	313
13	Conditional Maximum Entropy Models	315
13.1	Learning problem	315
13.2	Conditional Maxent principle	316
13.3	Conditional Maxent models	316
13.4	Dual problem	317
13.5	Properties	319
13.5.1	Optimization problem	320
13.5.2	Feature vectors	320
13.5.3	Prediction	321
13.6	Generalization bounds	321
13.7	Logistic regression	325
13.7.1	Optimization problem	325
13.7.2	Logistic model	325
13.8	L_2 -regularization	326
13.9	Proof of the duality theorem	328
13.10	Chapter notes	330
13.11	Exercises	331
14	Algorithmic Stability	333
14.1	Definitions	333
14.2	Stability-based generalization guarantee	334
14.3	Stability of kernel-based regularization algorithms	336
14.3.1	Application to regression algorithms: SVR and KRR	339
14.3.2	Application to classification algorithms: SVMs	341
14.3.3	Discussion	342
14.4	Chapter notes	342
14.5	Exercises	343
15	Dimensionality Reduction	347
15.1	Principal component analysis	348
15.2	Kernel principal component analysis (KPCA)	349
15.3	KPCA and manifold learning	351
15.3.1	Isomap	351
15.3.2	Laplacian eigenmaps	352
15.3.3	Locally linear embedding (LLE)	353
15.4	Johnson-Lindenstrauss lemma	354
15.5	Chapter notes	356
15.6	Exercises	356
16	Learning Automata and Languages	359
16.1	Introduction	359

16.2	Finite automata	360
16.3	Efficient exact learning	361
16.3.1	Passive learning	362
16.3.2	Learning with queries	363
16.3.3	Learning automata with queries	364
16.4	Identification in the limit	369
16.4.1	Learning reversible automata	370
16.5	Chapter notes	375
16.6	Exercises	376
17	Reinforcement Learning	379
17.1	Learning scenario	379
17.2	Markov decision process model	380
17.3	Policy	381
17.3.1	Definition	381
17.3.2	Policy value	382
17.3.3	Optimal policies	382
17.3.4	Policy evaluation	385
17.4	Planning algorithms	387
17.4.1	Value iteration	387
17.4.2	Policy iteration	390
17.4.3	Linear programming	392
17.5	Learning algorithms	393
17.5.1	Stochastic approximation	394
17.5.2	TD(0) algorithm	397
17.5.3	Q-learning algorithm	398
17.5.4	SARSA	402
17.5.5	TD(λ) algorithm	402
17.5.6	Large state space	403
17.6	Chapter notes	405
	Conclusion	407
A	Linear Algebra Review	409
A.1	Vectors and norms	409
A.1.1	Norms	409
A.1.2	Dual norms	410
A.1.3	Relationship between norms	411
A.2	Matrices	411
A.2.1	Matrix norms	411
A.2.2	Singular value decomposition	412
A.2.3	Symmetric positive semidefinite (SPSD) matrices	412

B	Convex Optimization	415
B.1	Differentiation and unconstrained optimization	415
B.2	Convexity	415
B.3	Constrained optimization	419
B.4	Fenchel duality	422
B.4.1	Subgradients	422
B.4.2	Core	423
B.4.3	Conjugate functions	423
B.5	Chapter notes	426
B.6	Exercises	427
C	Probability Review	429
C.1	Probability	429
C.2	Random variables	429
C.3	Conditional probability and independence	431
C.4	Expectation and Markov's inequality	431
C.5	Variance and Chebyshev's inequality	432
C.6	Moment-generating functions	434
C.7	Exercises	435
D	Concentration Inequalities	437
D.1	Hoeffding's inequality	437
D.2	Sanov's theorem	438
D.3	Multiplicative Chernoff bounds	439
D.4	Binomial distribution tails: Upper bounds	440
D.5	Binomial distribution tails: Lower bound	440
D.6	Azuma's inequality	441
D.7	McDiarmid's inequality	442
D.8	Normal distribution tails: Lower bound	443
D.9	Khintchine-Kahane inequality	443
D.10	Maximal inequality	444
D.11	Chapter notes	445
D.12	Exercises	445
E	Notions of Information Theory	449
E.1	Entropy	449
E.2	Relative entropy	450
E.3	Mutual information	453
E.4	Bregman divergences	453
E.5	Chapter notes	456
E.6	Exercises	457

F	Notation	459
	Bibliography	461
	Index	475

Preface

This book is a general introduction to machine learning that can serve as a reference book for researchers and a textbook for students. It covers fundamental modern topics in machine learning while providing the theoretical basis and conceptual tools needed for the discussion and justification of algorithms. It also describes several key aspects of the application of these algorithms.

We have aimed to present the most novel theoretical tools and concepts while giving concise proofs, even for relatively advanced results. In general, whenever possible, we have chosen to favor succinctness. Nevertheless, we discuss some crucial complex topics arising in machine learning and highlight several open research questions. Certain topics often merged with others or treated with insufficient attention are discussed separately here and with more emphasis: for example, a different chapter is reserved for multi-class classification, ranking, and regression.

Although we cover a very wide variety of important topics in machine learning, we have chosen to omit a few important ones, including graphical models and neural networks, both for the sake of brevity and because of the current lack of solid theoretical guarantees for some methods.

The book is intended for students and researchers in machine learning, statistics and other related areas. It can be used as a textbook for both graduate and advanced undergraduate classes in machine learning or as a reference text for a research seminar. The first three or four chapters of the book lay the theoretical foundation for the subsequent material. Other chapters are mostly self-contained, with the exception of chapter 6 which introduces some concepts that are extensively used in later ones and chapter 13, which is closely related to chapter 12. Each chapter concludes with a series of exercises, with full solutions presented separately.

The reader is assumed to be familiar with basic concepts in linear algebra, probability, and analysis of algorithms. However, to further help, we have included an extensive appendix presenting a concise review of linear algebra, an introduction to convex optimization, a brief probability review, a collection of concentration

inequalities useful to the analyses and discussions in this book, and a short introduction to information theory.

Our goal has been to give a unified presentation of multiple topics and areas, as opposed to a more specialized presentation adopted by some books which favor a particular viewpoint, such as for example a Bayesian view, or a particular topic, such as for example kernel methods. The theoretical foundation of this book and its deliberate emphasis on proofs and analysis make it also very distinct from many other presentations.

In this second edition, we have updated the entire book. The changes include a different writing style in most chapters, new figures and illustrations, many simplifications, some additions to existing chapters, in particular chapter 6 and chapter 17, and several new chapters. We have added a full chapter on model selection (chapter 4), which is an important topic that was only briefly discussed in the previous edition. We have also added a new chapter on Maximum Entropy models (chapter 12) and a new chapter on Conditional Maximum Entropy models (chapter 13) which are both essential topics in machine learning. We have also significantly changed the appendix. In particular, we have added a full section on Fenchel duality to appendix B on convex optimization, made a number of changes and additions to appendix D dealing with concentration inequalities, added appendix E on information theory, and updated most of the material. Additionally, we have included a number of new exercises and their solutions for existing and new chapters.

Most of the material presented here takes its origins in a machine learning graduate course (*Foundations of Machine Learning*) taught by the first author at the Courant Institute of Mathematical Sciences in New York University over the last fourteen years. This book has considerably benefited from the comments and suggestions from students in these classes, along with those of many friends, colleagues and researchers to whom we are deeply indebted.

We are particularly grateful to Corinna Cortes and Yishay Mansour who made a number of key suggestions for the design and organization of the material presented in the first edition, with detailed comments that we have fully taken into account and that have greatly improved the presentation. We are also grateful to Yishay Mansour for using a preliminary version of the first edition of the book for teaching, and for reporting his feedback to us.

We also thank for discussions, suggested improvement, and contributions of many kinds the following colleagues and friends from academic and corporate research laboratories: Jacob Abernethy, Cyril Allauzen, Kareem Amin, Stephen Boyd, Aldo Corbisiero, Giulia DeSalvo, Claudio Gentile, Spencer Greenberg, Lisa Hellerstein, Sanjiv Kumar, Vitaly Kuznetsov, Ryan McDonald, Andr  s Mu  oz Medina, Tyler Neylon, Peter Norvig, Fernando Pereira, Maria Pershina, Borja de Balle Pigem,

Ashish Rastogi, Michael Riley, Dmitry Storcheus, Ananda Theertha Suresh, Umar Syed, Csaba Szepesvári, Toshiyuki Tanaka, Eugene Weinstein, Jason Weston, Scott Yang, and Ningshan Zhang.

Finally, we thank the MIT Press publication team for their help and support in the development of this text.

1

Introduction

This chapter presents a preliminary introduction to machine learning, including an overview of some key learning tasks and applications, basic definitions and terminology, and the discussion of some general scenarios.

1.1 What is machine learning?

Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions. Here, *experience* refers to the past information available to the learner, which typically takes the form of electronic data collected and made available for analysis. This data could be in the form of digitized human-labeled training sets, or other types of information obtained via interaction with the environment. In all cases, its quality and size are crucial to the success of the predictions made by the learner.

An example of a learning problem is how to use a finite sample of randomly selected documents, each labeled with a topic, to accurately predict the topic of unseen documents. Clearly, the larger is the sample, the easier is the task. But the difficulty of the task also depends on the quality of the labels assigned to the documents in the sample, since the labels may not be all correct, and on the number of possible topics.

Machine learning consists of designing efficient and accurate prediction *algorithms*. As in other areas of computer science, some critical measures of the quality of these algorithms are their time and space complexity. But, in machine learning, we will need additionally a notion of *sample complexity* to evaluate the sample size required for the algorithm to learn a family of concepts. More generally, theoretical learning guarantees for an algorithm depend on the complexity of the concept classes considered and the size of the training sample.

Since the success of a learning algorithm depends on the data used, machine learning is inherently related to data analysis and statistics. More generally, learning