

Cambridge University Press

978-1-107-05713-5 - Understanding Machine Learning: From Theory to Algorithms

Shai Shalev-Shwartz and Shai Ben-David

Frontmatter

[More information](#)

## Understanding Machine Learning

Machine learning is one of the fastest growing areas of computer science, with far-reaching applications. The aim of this textbook is to introduce machine learning, and the algorithmic paradigms it offers, in a principled way. The book provides an extensive theoretical account of the fundamental ideas underlying machine learning and the mathematical derivations that transform these principles into practical algorithms. Following a presentation of the basics of the field, the book covers a wide array of central topics that have not been addressed by previous textbooks. These include a discussion of the computational complexity of learning and the concepts of convexity and stability; important algorithmic paradigms including stochastic gradient descent, neural networks, and structured output learning; and emerging theoretical concepts such as the PAC-Bayes approach and compression-based bounds. Designed for an advanced undergraduate or beginning graduate course, the text makes the fundamentals and algorithms of machine learning accessible to students and nonexpert readers in statistics, computer science, mathematics, and engineering.

Shai Shalev-Shwartz is an Associate Professor in the School of Computer Science and Engineering at The Hebrew University, Israel.

Shai Ben-David is a Professor in the School of Computer Science at the University of Waterloo, Canada.

Cambridge University Press  
978-1-107-05713-5 - Understanding Machine Learning: From Theory to Algorithms  
Shai Shalev-Shwartz and Shai Ben-David  
Frontmatter  
[More information](#)

Cambridge University Press  
978-1-107-05713-5 - Understanding Machine Learning: From Theory to Algorithms  
Shai Shalev-Shwartz and Shai Ben-David  
Frontmatter  
[More information](#)

# UNDERSTANDING MACHINE LEARNING

*From Theory to  
Algorithms*

**Shai Shalev-Shwartz**  
The Hebrew University, Jerusalem

**Shai Ben-David**  
University of Waterloo, Canada



Cambridge University Press  
978-1-107-05713-5 - Understanding Machine Learning: From Theory to Algorithms  
Shai Shalev-Shwartz and Shai Ben-David  
Frontmatter  
[More information](#)

CAMBRIDGE  
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

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](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9781107057135](http://www.cambridge.org/9781107057135)

© Shai Shalev-Shwartz and Shai Ben-David 2014

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

First published 2014  
Reprinted 2015

Printed in the United States of America

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

*Library of Congress Cataloging in Publication data*  
Shalev-Shwartz, Shai.  
Understanding machine learning : from theory to algorithms /  
Shai Shalev-Shwartz, The Hebrew University, Jerusalem,  
Shai Ben-David, University of Waterloo, Canada.  
pages cm  
Includes bibliographical references and index.  
ISBN 978-1-107-05713-5 (hardback)  
1. Machine learning. 2. Algorithms. I. Ben-David, Shai. II. Title.  
Q325.5.S475 2014  
006.3’1–dc23 2014001779

ISBN 978-1-107-05713-5 Hardback

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

Cambridge University Press  
978-1-107-05713-5 - Understanding Machine Learning: From Theory to Algorithms  
Shai Shalev-Shwartz and Shai Ben-David  
Frontmatter  
[More information](#)

Triple-S dedicates the book to triple-M

Cambridge University Press  
978-1-107-05713-5 - Understanding Machine Learning: From Theory to Algorithms  
Shai Shalev-Shwartz and Shai Ben-David  
Frontmatter  
[More information](#)

Contents

<i>Preface</i>	<i>page</i> xv
<b>1 Introduction</b>	1
1.1 What Is Learning?	1
1.2 When Do We Need Machine Learning?	3
1.3 Types of Learning	4
1.4 Relations to Other Fields	6
1.5 How to Read This Book	7
1.6 Notation	8
<b>Part 1 Foundations</b>	
<b>2 A Gentle Start</b>	13
2.1 A Formal Model – The Statistical Learning Framework	13
2.2 Empirical Risk Minimization	15
2.3 Empirical Risk Minimization with Inductive Bias	16
2.4 Exercises	20
<b>3 A Formal Learning Model</b>	22
3.1 PAC Learning	22
3.2 A More General Learning Model	23
3.3 Summary	28
3.4 Bibliographic Remarks	28
3.5 Exercises	28
<b>4 Learning via Uniform Convergence</b>	31
4.1 Uniform Convergence Is Sufficient for Learnability	31
4.2 Finite Classes Are Agnostic PAC Learnable	32
4.3 Summary	34
4.4 Bibliographic Remarks	35
4.5 Exercises	35

viii      **Contents**

<b>5</b>	<b>The Bias-Complexity Trade-off</b>	36
5.1	The No-Free-Lunch Theorem	37
5.2	Error Decomposition	40
5.3	Summary	41
5.4	Bibliographic Remarks	41
5.5	Exercises	41
<b>6</b>	<b>The VC-Dimension</b>	43
6.1	Infinite-Size Classes Can Be Learnable	43
6.2	The VC-Dimension	44
6.3	Examples	46
6.4	The Fundamental Theorem of PAC Learning	48
6.5	Proof of Theorem 6.7	49
6.6	Summary	53
6.7	Bibliographic Remarks	53
6.8	Exercises	54
<b>7</b>	<b>Nonuniform Learnability</b>	58
7.1	Nonuniform Learnability	58
7.2	Structural Risk Minimization	60
7.3	Minimum Description Length and Occam’s Razor	63
7.4	Other Notions of Learnability – Consistency	66
7.5	Discussing the Different Notions of Learnability	67
7.6	Summary	70
7.7	Bibliographic Remarks	70
7.8	Exercises	71
<b>8</b>	<b>The Runtime of Learning</b>	73
8.1	Computational Complexity of Learning	74
8.2	Implementing the ERM Rule	76
8.3	Efficiently Learnable, but Not by a Proper ERM	80
8.4	Hardness of Learning*	81
8.5	Summary	82
8.6	Bibliographic Remarks	82
8.7	Exercises	83
<b>Part 2 From Theory to Algorithms</b>		
<b>9</b>	<b>Linear Predictors</b>	89
9.1	Halfspaces	90
9.2	Linear Regression	94
9.3	Logistic Regression	97
9.4	Summary	99
9.5	Bibliographic Remarks	99
9.6	Exercises	99



	Contents	ix
<b>10 Boosting</b>	101	
10.1 Weak Learnability	102	
10.2 AdaBoost	105	
10.3 Linear Combinations of Base Hypotheses	108	
10.4 AdaBoost for Face Recognition	110	
10.5 Summary	111	
10.6 Bibliographic Remarks	111	
10.7 Exercises	112	
<b>11 Model Selection and Validation</b>	114	
11.1 Model Selection Using SRM	115	
11.2 Validation	116	
11.3 What to Do If Learning Fails	120	
11.4 Summary	123	
11.5 Exercises	123	
<b>12 Convex Learning Problems</b>	124	
12.1 Convexity, Lipschitzness, and Smoothness	124	
12.2 Convex Learning Problems	130	
12.3 Surrogate Loss Functions	134	
12.4 Summary	135	
12.5 Bibliographic Remarks	136	
12.6 Exercises	136	
<b>13 Regularization and Stability</b>	137	
13.1 Regularized Loss Minimization	137	
13.2 Stable Rules Do Not Overfit	139	
13.3 Tikhonov Regularization as a Stabilizer	140	
13.4 Controlling the Fitting-Stability Trade-off	144	
13.5 Summary	146	
13.6 Bibliographic Remarks	146	
13.7 Exercises	147	
<b>14 Stochastic Gradient Descent</b>	150	
14.1 Gradient Descent	151	
14.2 Subgradients	154	
14.3 Stochastic Gradient Descent (SGD)	156	
14.4 Variants	159	
14.5 Learning with SGD	162	
14.6 Summary	165	
14.7 Bibliographic Remarks	166	
14.8 Exercises	166	
<b>15 Support Vector Machines</b>	167	
15.1 Margin and Hard-SVM	167	
15.2 Soft-SVM and Norm Regularization	171	
15.3 Optimality Conditions and “Support Vectors”*	175	

x      **Contents**

15.4	Duality*	175
15.5	Implementing Soft-SVM Using SGD	176
15.6	Summary	177
15.7	Bibliographic Remarks	177
15.8	Exercises	178
<b>16</b>	<b>Kernel Methods</b>	179
16.1	Embeddings into Feature Spaces	179
16.2	The Kernel Trick	181
16.3	Implementing Soft-SVM with Kernels	186
16.4	Summary	187
16.5	Bibliographic Remarks	188
16.6	Exercises	188
<b>17</b>	<b>Multiclass, Ranking, and Complex Prediction Problems</b>	190
17.1	One-versus-All and All-Pairs	190
17.2	Linear Multiclass Predictors	193
17.3	Structured Output Prediction	198
17.4	Ranking	201
17.5	Bipartite Ranking and Multivariate Performance Measures	206
17.6	Summary	209
17.7	Bibliographic Remarks	210
17.8	Exercises	210
<b>18</b>	<b>Decision Trees</b>	212
18.1	Sample Complexity	213
18.2	Decision Tree Algorithms	214
18.3	Random Forests	217
18.4	Summary	217
18.5	Bibliographic Remarks	218
18.6	Exercises	218
<b>19</b>	<b>Nearest Neighbor</b>	219
19.1	$k$ Nearest Neighbors	219
19.2	Analysis	220
19.3	Efficient Implementation*	225
19.4	Summary	225
19.5	Bibliographic Remarks	225
19.6	Exercises	225
<b>20</b>	<b>Neural Networks</b>	228
20.1	Feedforward Neural Networks	229
20.2	Learning Neural Networks	230
20.3	The Expressive Power of Neural Networks	231
20.4	The Sample Complexity of Neural Networks	234
20.5	The Runtime of Learning Neural Networks	235
20.6	SGD and Backpropagation	236

	Contents	xi
20.7 Summary	240	
20.8 Bibliographic Remarks	240	
20.9 Exercises	240	
<b>Part 3 Additional Learning Models</b>		
<b>21 Online Learning</b>	245	
21.1 Online Classification in the Realizable Case	246	
21.2 Online Classification in the Unrealizable Case	251	
21.3 Online Convex Optimization	257	
21.4 The Online Perceptron Algorithm	258	
21.5 Summary	261	
21.6 Bibliographic Remarks	261	
21.7 Exercises	262	
<b>22 Clustering</b>	264	
22.1 Linkage-Based Clustering Algorithms	266	
22.2 $k$ -Means and Other Cost Minimization Clusterings	268	
22.3 Spectral Clustering	271	
22.4 Information Bottleneck*	273	
22.5 A High-Level View of Clustering	274	
22.6 Summary	276	
22.7 Bibliographic Remarks	276	
22.8 Exercises	276	
<b>23 Dimensionality Reduction</b>	278	
23.1 Principal Component Analysis (PCA)	279	
23.2 Random Projections	283	
23.3 Compressed Sensing	285	
23.4 PCA or Compressed Sensing?	292	
23.5 Summary	292	
23.6 Bibliographic Remarks	292	
23.7 Exercises	293	
<b>24 Generative Models</b>	295	
24.1 Maximum Likelihood Estimator	295	
24.2 Naive Bayes	299	
24.3 Linear Discriminant Analysis	300	
24.4 Latent Variables and the EM Algorithm	301	
24.5 Bayesian Reasoning	305	
24.6 Summary	307	
24.7 Bibliographic Remarks	307	
24.8 Exercises	308	
<b>25 Feature Selection and Generation</b>	309	
25.1 Feature Selection	310	
25.2 Feature Manipulation and Normalization	316	
25.3 Feature Learning	319	

xii      **Contents**

25.4	Summary	321
25.5	Bibliographic Remarks	321
25.6	Exercises	322
<b>Part 4    Advanced Theory</b>		
<b>26</b>	<b>Rademacher Complexities</b>	325
26.1	The Rademacher Complexity	325
26.2	Rademacher Complexity of Linear Classes	332
26.3	Generalization Bounds for SVM	333
26.4	Generalization Bounds for Predictors with Low $\ell_1$ Norm	335
26.5	Bibliographic Remarks	336
<b>27</b>	<b>Covering Numbers</b>	337
27.1	Covering	337
27.2	From Covering to Rademacher Complexity via Chaining	338
27.3	Bibliographic Remarks	340
<b>28</b>	<b>Proof of the Fundamental Theorem of Learning Theory</b>	341
28.1	The Upper Bound for the Agnostic Case	341
28.2	The Lower Bound for the Agnostic Case	342
28.3	The Upper Bound for the Realizable Case	347
<b>29</b>	<b>Multiclass Learnability</b>	351
29.1	The Natarajan Dimension	351
29.2	The Multiclass Fundamental Theorem	352
29.3	Calculating the Natarajan Dimension	353
29.4	On Good and Bad ERMs	355
29.5	Bibliographic Remarks	357
29.6	Exercises	357
<b>30</b>	<b>Compression Bounds</b>	359
30.1	Compression Bounds	359
30.2	Examples	361
30.3	Bibliographic Remarks	363
<b>31</b>	<b>PAC-Bayes</b>	364
31.1	PAC-Bayes Bounds	364
31.2	Bibliographic Remarks	366
31.3	Exercises	366
<b>Appendix A Technical Lemmas</b>		369
<b>Appendix B Measure Concentration</b>		372
B.1	Markov’s Inequality	372
B.2	Chebyshev’s Inequality	373
B.3	Chernoff’s Bounds	373
B.4	Hoeffding’s Inequality	375

	Contents	xiii
B.5	Bennet’s and Bernstein’s Inequalities	376
B.6	Slud’s Inequality	378
B.7	Concentration of $\chi^2$ Variables	378
<b>Appendix C</b>	<b>Linear Algebra</b>	380
C.1	Basic Definitions	380
C.2	Eigenvalues and Eigenvectors	381
C.3	Positive Definite Matrices	381
C.4	Singular Value Decomposition (SVD)	381
<i>References</i>		385
<i>Index</i>		395

Cambridge University Press  
978-1-107-05713-5 - Understanding Machine Learning: From Theory to Algorithms  
Shai Shalev-Shwartz and Shai Ben-David  
Frontmatter  
[More information](#)

## Preface

The term *machine learning* refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets. We are surrounded by a machine learning-based technology: Search engines learn how to bring us the best results (while placing profitable ads), antispam software learns to filter our e-mail messages, and credit card transactions are secured by a software that learns how to detect frauds. Digital cameras learn to detect faces and intelligent personal assistance applications on smart-phones learn to recognize voice commands. Cars are equipped with accident-prevention systems that are built using machine learning algorithms. Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy.

One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide an explicit, fine-detailed specification of how such tasks should be executed. Taking examples from intelligent beings, many of our skills are acquired or refined through *learning* from our experience (rather than following explicit instructions given to us). Machine learning tools are concerned with endowing programs with the ability to “learn” and adapt.

The first goal of this book is to provide a rigorous, yet easy-to-follow, introduction to the main concepts underlying machine learning: What is learning? How can a machine learn? How do we quantify the resources needed to learn a given concept? Is learning always possible? Can we know whether the learning process succeeded or failed?

The second goal of this book is to present several key machine learning algorithms. We chose to present algorithms that on one hand are successfully used in practice and on the other hand give a wide spectrum of different learning techniques. Additionally, we pay specific attention to algorithms appropriate for large-scale learning (a.k.a. “Big Data”), since in recent years, our world has become increasingly “digitized” and the amount of data available for learning is dramatically increasing. As a result, in many applications data is plentiful and computation

time is the main bottleneck. We therefore explicitly quantify both the amount of data and the amount of computation time needed to learn a given concept.

The book is divided into four parts. The first part aims at giving an initial rigorous answer to the fundamental questions of learning. We describe a generalization of Valiant’s Probably Approximately Correct (PAC) learning model, which is a first solid answer to the question “What is learning?” We describe the Empirical Risk Minimization (ERM), Structural Risk Minimization (SRM), and Minimum Description Length (MDL) learning rules, which show “how a machine can learn.” We quantify the amount of data needed for learning using the ERM, SRM, and MDL rules and show how learning might fail by deriving a “no-free-lunch” theorem. We also discuss how much computation time is required for learning. In the second part of the book we describe various learning algorithms. For some of the algorithms, we first present a more general learning principle and then show how the algorithm follows the principle. While the first two parts of the book focus on the PAC model, the third part extends the scope by presenting a wider variety of learning models. Finally, the last part of the book is devoted to advanced theory.

We made an attempt to keep the book as self-contained as possible. However, the reader is assumed to be comfortable with basic notions of probability, linear algebra, analysis, and algorithms. The first three parts of the book are intended for first-year graduate students in computer science, engineering, mathematics, or statistics. It can also be accessible to undergraduate students with the adequate background. The more advanced chapters can be used by researchers intending to gather a deeper theoretical understanding.

**ACKNOWLEDGMENTS**

The book is based on Introduction to Machine Learning courses taught by Shai Shalev-Shwartz at Hebrew University and by Shai Ben-David at the University of Waterloo. The first draft of the book grew out of the lecture notes for the course that was taught at Hebrew University by Shai Shalev-Shwartz during 2010–2013. We greatly appreciate the help of Ohad Shamir, who served as a teaching assistant for the course in 2010, and of Alon Gonen, who served as TA for the course in 2011–2013. Ohad and Alon prepared a few lecture notes and many of the exercises. Alon, to whom we are indebted for his help throughout the entire making of the book, has also prepared a solution manual.

We are deeply grateful for the most valuable work of Dana Rubinstein. Dana has scientifically proofread and edited the manuscript, transforming it from lecture-based chapters into fluent and coherent text.

Special thanks to Amit Daniely, who helped us with a careful read of the advanced part of the book and wrote the advanced chapter on multiclass learnability. We are also grateful for the members of a book reading club in Jerusalem who have carefully read and constructively criticized every line of the manuscript. The members of the reading club are Maya Alroy, Yossi Arjevani, Aharon Birnbaum, Alon Cohen, Alon Gonen, Roi Livni, Ofer Meshi, Dan Rosenbaum, Dana Rubinstein, Shahar Somin, Alon Vinnikov, and Yoav Wald. We would also like to thank Gal Elidan, Amir Globerson, Nika Haghtalab, Shie Mannor, Amnon Shashua, Nati Srebro, and Ruth Urner for helpful discussions.