

UNDERSTANDING MACHINE LEARNING

*From Theory to
Algorithms*

Shai Shalev-Shwartz

The Hebrew University, Jerusalem

Shai Ben-David

University of Waterloo, Canada



CAMBRIDGE
UNIVERSITY PRESS

Contents

<i>Preface</i>	<i>page xv</i>
1 Introduction	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
Part 1 Foundations	
2 A Gentle Start	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
3 A Formal Learning Model	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
4 Learning via Uniform Convergence	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

5	The Bias-Complexity Trade-off	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
6	The VC-Dimension	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
7	Nonuniform Learnability	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
8	The Runtime of Learning	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
Part 2 From Theory to Algorithms		
9	Linear Predictors	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

10	Boosting	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
11	Model Selection and Validation	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
12	Convex Learning Problems	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
13	Regularization and Stability	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
14	Stochastic Gradient Descent	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
15	Support Vector Machines	167
15.1	Margin and Hard-SVM	167
15.2	Soft-SVM and Norm Regularization	171
15.3	Optimality Conditions and “Support Vectors”*	175

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
16	Kernel Methods	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
17	Multiclass, Ranking, and Complex Prediction Problems	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
18	Decision Trees	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
19	Nearest Neighbor	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
20	Neural Networks	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

20.7	Summary	240
20.8	Bibliographic Remarks	240
20.9	Exercises	240
Part 3 Additional Learning Models		
21	Online Learning	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
22	Clustering	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
23	Dimensionality Reduction	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
24	Generative Models	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
25	Feature Selection and Generation	309
25.1	Feature Selection	310
25.2	Feature Manipulation and Normalization	316
25.3	Feature Learning	319

25.4	Summary	321
25.5	Bibliographic Remarks	321
25.6	Exercises	322
Part 4 Advanced Theory		
26	Rademacher Complexities	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 ℓ_1 Norm	335
26.5	Bibliographic Remarks	336
27	Covering Numbers	337
27.1	Covering	337
27.2	From Covering to Rademacher Complexity via Chaining	338
27.3	Bibliographic Remarks	340
28	Proof of the Fundamental Theorem of Learning Theory	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
29	Multiclass Learnability	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 ERM's	355
29.5	Bibliographic Remarks	357
29.6	Exercises	357
30	Compression Bounds	359
30.1	Compression Bounds	359
30.2	Examples	361
30.3	Bibliographic Remarks	363
31	PAC-Bayes	364
31.1	PAC-Bayes Bounds	364
31.2	Bibliographic Remarks	366
31.3	Exercises	366
Appendix A Technical Lemmas		369
Appendix B Measure Concentration		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

B.5	Bennet's and Bernstein's Inequalities	376
B.6	Slud's Inequality	378
B.7	Concentration of χ^2 Variables	378
Appendix C Linear Algebra		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