Contents

Preface *page* vii

1 Introduction 19
1.1 What Is Learning? 19
1.2 When Do We Need Machine Learning? 21
1.3 Types of Learning 22
1.4 Relations to Other Fields 24
1.5 How to Read This Book 25
1.5.1 Possible Course Plans Based on This Book 26
1.6 Notation 27

Part I Foundations 31

2 A Gentle Start 33
2.1 A Formal Model – The Statistical Learning Framework 33
2.2 Empirical Risk Minimization 35
2.2.1 Something May Go Wrong – Overfitting 35
2.3 Empirical Risk Minimization with Inductive Bias 36
2.3.1 Finite Hypothesis Classes 37
2.4 Exercises 41

3 A Formal Learning Model 43
3.1 PAC Learning 43
3.2 A More General Learning Model 44
3.2.1 Releasing the Realizability Assumption – Agnostic PAC Learning 45
3.2.2 The Scope of Learning Problems Modeled 47
3.3 Summary 49
3.4 Bibliographic Remarks 50
3.5 Exercises 50

4 Learning via Uniform Convergence 54
4.1 Uniform Convergence Is Sufficient for Learnability 54
4.2 Finite Classes Are Agnostic PAC Learnable 55

Understanding Machine Learning, © 2014 by Shai Shalev-Shwartz and Shai Ben-David
Published 2014 by Cambridge University Press.
Personal use only. Not for distribution. Do not post.
Please link to http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

4.3 Summary 58
4.4 Bibliographic Remarks 58
4.5 Exercises 58

5 The Bias-Complexity Tradeoff 60
5.1 The No-Free-Lunch Theorem 61
5.1.1 No-Free-Lunch and Prior Knowledge 63
5.2 Error Decomposition 64
5.3 Summary 65
5.4 Bibliographic Remarks 66
5.5 Exercises 66

6 The VC-Dimension 67
6.1 Infinite-Size Classes Can Be Learnable 67
6.2 The VC-Dimension 68
6.3 Examples 70
6.3.1 Threshold Functions 70
6.3.2 Intervals 71
6.3.3 Axis Aligned Rectangles 71
6.3.4 Finite Classes 72
6.3.5 VC-Dimension and the Number of Parameters 72
6.4 The Fundamental Theorem of PAC learning 72
6.5 Proof of Theorem 6.7 73
6.5.1 Sauer's Lemma and the Growth Function 73
6.5.2 Uniform Convergence for Classes of Small Effective Size 75
6.6 Summary 78
6.7 Bibliographic remarks 78
6.8 Exercises 78

7 Nonuniform Learnability 83
7.1 Nonuniform Learnability 83
7.1.1 Characterizing Nonuniform Learnability 84
7.2 Structural Risk Minimization 85
7.3 Minimum Description Length and Occam's Razor 89
7.3.1 Occam's Razor 91
7.4 Other Notions of Learnability – Consistency 92
7.5 Discussing the Different Notions of Learnability 93
7.5.1 The No-Free-Lunch Theorem Revisited 95
7.6 Summary 96
7.7 Bibliographic Remarks 97
7.8 Exercises 97

8 The Runtime of Learning 100
8.1 Computational Complexity of Learning 101

8.1.1 Formal Definition* 102
8.2 Implementing the ERM Rule 103
8.2.1 Finite Classes 104
8.2.2 Axis Aligned Rectangles 105
8.2.3 Boolean Conjunctions 106
8.2.4 Learning 3-Term DNF 107
8.3 Efficiently Learnable, but Not by a Proper ERM 107
8.4 Hardness of Learning* 108
8.5 Summary 110
8.6 Bibliographic Remarks 110
8.7 Exercises 110

Part II From Theory to Algorithms 115

9 Linear Predictors 117
9.1 Halfspaces 118
9.1.1 Linear Programming for the Class of Halfspaces 119
9.1.2 Perceptron for Halfspaces 120
9.1.3 The VC Dimension of Halfspaces 122
9.2 Linear Regression 123
9.2.1 Least Squares 124
9.2.2 Linear Regression for Polynomial Regression Tasks 125
9.3 Logistic Regression 126
9.4 Summary 128
9.5 Bibliographic Remarks 128
9.6 Exercises 128

10 Boosting 130
10.1 Weak Learnability 131
10.1.1 Efficient Implementation of ERM for Decision Stumps 133
10.2 AdaBoost 134
10.3 Linear Combinations of Base Hypotheses 137
10.3.1 The VC-Dimension of $L(B, T)$ 139
10.4 AdaBoost for Face Recognition 140
10.5 Summary 141
10.6 Bibliographic Remarks 141
10.7 Exercises 142

11 Model Selection and Validation 144
11.1 Model Selection Using SRM 145
11.2 Validation 146
11.2.1 Hold Out Set 146
11.2.2 Validation for Model Selection 147
11.2.3 The Model-Selection Curve 148

11.2.4 *k*-Fold Cross Validation 149
11.2.5 Train-Validation-Test Split 150
11.3 What to Do If Learning Fails 151
11.4 Summary 154
11.5 Exercises 154

12 Convex Learning Problems 156
12.1 Convexity, Lipschitzness, and Smoothness 156
12.1.1 Convexity 156
12.1.2 Lipschitzness 160
12.1.3 Smoothness 162
12.2 Convex Learning Problems 163
12.2.1 Learnability of Convex Learning Problems 164
12.2.2 Convex-Lipschitz/Smooth-Bounded Learning Problems 166
12.3 Surrogate Loss Functions 167
12.4 Summary 168
12.5 Bibliographic Remarks 169
12.6 Exercises 169

13 Regularization and Stability 171
13.1 Regularized Loss Minimization 171
13.1.1 Ridge Regression 172
13.2 Stable Rules Do Not Overfit 173
13.3 Tikhonov Regularization as a Stabilizer 174
13.3.1 Lipschitz Loss 176
13.3.2 Smooth and Nonnegative Loss 177
13.4 Controlling the Fitting-Stability Tradeoff 178
13.5 Summary 180
13.6 Bibliographic Remarks 180
13.7 Exercises 181

14 Stochastic Gradient Descent 184
14.1 Gradient Descent 185
14.1.1 Analysis of GD for Convex-Lipschitz Functions 186
14.2 Subgradients 188
14.2.1 Calculating Subgradients 189
14.2.2 Subgradients of Lipschitz Functions 190
14.2.3 Subgradient Descent 190
14.3 Stochastic Gradient Descent (SGD) 191
14.3.1 Analysis of SGD for Convex-Lipschitz-Bounded Functions 191
14.4 Variants 193
14.4.1 Adding a Projection Step 193
14.4.2 Variable Step Size 194
14.4.3 Other Averaging Techniques 195

14.4.4 Strongly Convex Functions* 195
14.5 Learning with SGD 196
14.5.1 SGD for Risk Minimization 196
14.5.2 Analyzing SGD for Convex-Smooth Learning Problems 198
14.5.3 SGD for Regularized Loss Minimization 199
14.6 Summary 200
14.7 Bibliographic Remarks 200
14.8 Exercises 201

15 **Support Vector Machines** 202
15.1 Margin and Hard-SVM 202
15.1.1 The Homogenous Case 205
15.1.2 The Sample Complexity of Hard-SVM 205
15.2 Soft-SVM and Norm Regularization 206
15.2.1 The Sample Complexity of Soft-SVM 208
15.2.2 Margin and Norm-Based Bounds versus Dimension 208
15.2.3 The Ramp Loss* 209
15.3 Optimality Conditions and “Support Vectors”* 210
15.4 Duality* 211
15.5 Implementing Soft-SVM Using SGD 212
15.6 Summary 213
15.7 Bibliographic Remarks 213
15.8 Exercises 214

16 **Kernel Methods** 215
16.1 Embeddings into Feature Spaces 215
16.2 The Kernel Trick 217
16.2.1 Kernels as a Way to Express Prior Knowledge 221
16.2.2 Characterizing Kernel Functions* 222
16.3 Implementing Soft-SVM with Kernels 222
16.4 Summary 224
16.5 Bibliographic Remarks 225
16.6 Exercises 225

17 **Multiclass, Ranking, and Complex Prediction Problems** 227
17.1 One-versus-All and All-Pairs 227
17.2 Linear Multiclass Predictors 230
17.2.1 How to Construct Ψ 230
17.2.2 Cost-Sensitive Classification 232
17.2.3 ERM 232
17.2.4 Generalized Hinge Loss 233
17.2.5 Multiclass SVM and SGD 234
17.3 Structured Output Prediction 236
17.4 Ranking 238

17.4.1 Linear Predictors for Ranking 240
17.5 Bipartite Ranking and Multivariate Performance Measures 243
17.5.1 Linear Predictors for Bipartite Ranking 245
17.6 Summary 247
17.7 Bibliographic Remarks 247
17.8 Exercises 248

18 Decision Trees 250
18.1 Sample Complexity 251
18.2 Decision Tree Algorithms 252
18.2.1 Implementations of the Gain Measure 253
18.2.2 Pruning 254
18.2.3 Threshold-Based Splitting Rules for Real-Valued Features 255
18.3 Random Forests 255
18.4 Summary 256
18.5 Bibliographic Remarks 256
18.6 Exercises 256

19 Nearest Neighbor 258
19.1 k Nearest Neighbors 258
19.2 Analysis 259
19.2.1 A Generalization Bound for the 1-NN Rule 260
19.2.2 The "Curse of Dimensionality" 263
19.3 Efficient Implementation* 264
19.4 Summary 264
19.5 Bibliographic Remarks 264
19.6 Exercises 265

20 Neural Networks 268
20.1 Feedforward Neural Networks 269
20.2 Learning Neural Networks 270
20.3 The Expressive Power of Neural Networks 271
20.3.1 Geometric Intuition 273
20.4 The Sample Complexity of Neural Networks 274
20.5 The Runtime of Learning Neural Networks 276
20.6 SGD and Backpropagation 277
20.7 Summary 281
20.8 Bibliographic Remarks 281
20.9 Exercises 282

Part III Additional Learning Models 285

21 Online Learning 287
21.1 Online Classification in the Realizable Case 288

21.1.1 Online Learnability 290
21.2 Online Classification in the Unrealizable Case 294
21.2.1 Weighted-Majority 295
21.3 Online Convex Optimization 300
21.4 The Online Perceptron Algorithm 301
21.5 Summary 304
21.6 Bibliographic Remarks 305
21.7 Exercises 305

22 Clustering 307
22.1 Linkage-Based Clustering Algorithms 310
22.2 k-Means and Other Cost Minimization Clusterings 311
22.2.1 The k-Means Algorithm 313
22.3 Spectral Clustering 315
22.3.1 Graph Cut 315
22.3.2 Graph Laplacian and Relaxed Graph Cuts 315
22.3.3 Unnormalized Spectral Clustering 317
22.4 Information Bottleneck* 317
22.5 A High Level View of Clustering 318
22.6 Summary 320
22.7 Bibliographic Remarks 320
22.8 Exercises 320

23 Dimensionality Reduction 323
23.1 Principal Component Analysis (PCA) 324
23.1.1 A More Efficient Solution for the Case $d \gg m$ 326
23.1.2 Implementation and Demonstration 326
23.2 Random Projections 329
23.3 Compressed Sensing 330
23.3.1 Proofs* 333
23.4 PCA or Compressed Sensing? 338
23.5 Summary 338
23.6 Bibliographic Remarks 339
23.7 Exercises 339

24 Generative Models 342
24.1 Maximum Likelihood Estimator 343
24.1.1 Maximum Likelihood Estimation for Continuous Random Variables 344
24.1.2 Maximum Likelihood and Empirical Risk Minimization 345
24.1.3 Generalization Analysis 345
24.2 Naive Bayes 347
24.3 Linear Discriminant Analysis 347
24.4 Latent Variables and the EM Algorithm 348

24.4.1 EM as an Alternate Maximization Algorithm 350
24.4.2 EM for Mixture of Gaussians (Soft k-Means) 352
24.5 Bayesian Reasoning 353
24.6 Summary 355
24.7 Bibliographic Remarks 355
24.8 Exercises 356

25 Feature Selection and Generation 357
25.1 Feature Selection 358
25.1.1 Filters 359
25.1.2 Greedy Selection Approaches 360
25.1.3 Sparsity-Inducing Norms 363
25.2 Feature Manipulation and Normalization 365
25.2.1 Examples of Feature Transformations 367
25.3 Feature Learning 368
25.3.1 Dictionary Learning Using Auto-Encoders 368
25.4 Summary 370
25.5 Bibliographic Remarks 371
25.6 Exercises 371

Part IV Advanced Theory 373

26 Rademacher Complexities 375
26.1 The Rademacher Complexity 375
26.1.1 Rademacher Calculus 379
26.2 Rademacher Complexity of Linear Classes 382
26.3 Generalization Bounds for SVM 383
26.4 Generalization Bounds for Predictors with Low ℓ_1 Norm 386
26.5 Bibliographic Remarks 386

27 Covering Numbers 388
27.1 Covering 388
27.1.1 Properties 388
27.2 From Covering to Rademacher Complexity via Chaining 389
27.3 Bibliographic Remarks 391

28 Proof of the Fundamental Theorem of Learning Theory 392
28.1 The Upper Bound for the Agnostic Case 392
28.2 The Lower Bound for the Agnostic Case 393
28.2.1 Showing That $\boldsymbol{m(\epsilon,\delta) \geq 0.5 \log(1/(4\delta))/\epsilon^2}$ 393
28.2.2 Showing That $m(\epsilon, 1/8) \geq 8d/\epsilon^2$ 395
28.3 The Upper Bound for the Realizable Case 398
28.3.1 From ϵ-Nets to PAC Learnability 401

29 Multiclass Learnability 402

29.1 The Natarajan Dimension 402

29.2 The Multiclass Fundamental Theorem 403

29.2.1 On the Proof of Theorem 29.3 403

29.3 Calculating the Natarajan Dimension 404

29.3.1 One-versus-All Based Classes 404

29.3.2 General Multiclass-to-Binary Reductions 405

29.3.3 Linear Multiclass Predictors 405

29.4 On Good and Bad ERMs 406

29.5 Bibliographic Remarks 408

29.6 Exercises 409

30 Compression Bounds 410

30.1 Compression Bounds 410

30.2 Examples 412

30.2.1 Axis Aligned Rectangles 412

30.2.2 Halfspaces 412

30.2.3 Separating Polynomials 413

30.2.4 Separation with Margin 414

30.3 Bibliographic Remarks 414

31 PAC-Bayes 415

31.1 PAC-Bayes Bounds 415

31.2 Bibliographic Remarks 417

31.3 Exercises 417

Appendix A Technical Lemmas 419

Appendix B Measure Concentration 422

Appendix C Linear Algebra 430

Notes 435

References 437

Index 447