

Contents

Preface	xi
1 Introduction	1
1.1 Applications and problems	1
1.2 Definitions and terminology	3
1.3 Cross-validation	5
1.4 Learning scenarios	7
1.5 Outline	8
2 The PAC Learning Framework	11
2.1 The PAC learning model	11
2.2 Guarantees for finite hypothesis sets — consistent case	17
2.3 Guarantees for finite hypothesis sets — inconsistent case	21
2.4 Generalities	24
2.4.1 Deterministic versus stochastic scenarios	24
2.4.2 Bayes error and noise	25
2.4.3 Estimation and approximation errors	26
2.4.4 Model selection	27
2.5 Chapter notes	28
2.6 Exercises	29
3 Rademacher Complexity and VC-Dimension	33
3.1 Rademacher complexity	34
3.2 Growth function	38
3.3 VC-dimension	41
3.4 Lower bounds	48
3.5 Chapter notes	54
3.6 Exercises	55
4 Support Vector Machines	63
4.1 Linear classification	63
4.2 SVMs — separable case	64

4.2.1	Primal optimization problem	64
4.2.2	Support vectors	66
4.2.3	Dual optimization problem	67
4.2.4	Leave-one-out analysis	69
4.3	SVMs — non-separable case	71
4.3.1	Primal optimization problem	72
4.3.2	Support vectors	73
4.3.3	Dual optimization problem	74
4.4	Margin theory	75
4.5	Chapter notes	83
4.6	Exercises	84
5	Kernel Methods	89
5.1	Introduction	89
5.2	Positive definite symmetric kernels	92
5.2.1	Definitions	92
5.2.2	Reproducing kernel Hilbert space	94
5.2.3	Properties	96
5.3	Kernel-based algorithms	100
5.3.1	SVMs with PDS kernels	100
5.3.2	Representer theorem	101
5.3.3	Learning guarantees	102
5.4	Negative definite symmetric kernels	103
5.5	Sequence kernels	106
5.5.1	Weighted transducers	106
5.5.2	Rational kernels	111
5.6	Chapter notes	115
5.7	Exercises	116
6	Boosting	121
6.1	Introduction	121
6.2	AdaBoost	122
6.2.1	Bound on the empirical error	124
6.2.2	Relationship with coordinate descent	126
6.2.3	Relationship with logistic regression	129
6.2.4	Standard use in practice	129
6.3	Theoretical results	130
6.3.1	VC-dimension-based analysis	131
6.3.2	Margin-based analysis	131
6.3.3	Margin maximization	136
6.3.4	Game-theoretic interpretation	137

6.4	Discussion	140
6.5	Chapter notes	141
6.6	Exercises	142
7	On-Line Learning	147
7.1	Introduction	147
7.2	Prediction with expert advice	148
7.2.1	Mistake bounds and Halving algorithm	148
7.2.2	Weighted majority algorithm	150
7.2.3	Randomized weighted majority algorithm	152
7.2.4	Exponential weighted average algorithm	156
7.3	Linear classification	159
7.3.1	Perceptron algorithm	160
7.3.2	Winnow algorithm	168
7.4	On-line to batch conversion	171
7.5	Game-theoretic connection	174
7.6	Chapter notes	175
7.7	Exercises	176
8	Multi-Class Classification	183
8.1	Multi-class classification problem	183
8.2	Generalization bounds	185
8.3	Uncombined multi-class algorithms	191
8.3.1	Multi-class SVMs	191
8.3.2	Multi-class boosting algorithms	192
8.3.3	Decision trees	194
8.4	Aggregated multi-class algorithms	198
8.4.1	One-versus-all	198
8.4.2	One-versus-one	199
8.4.3	Error-correction codes	201
8.5	Structured prediction algorithms	203
8.6	Chapter notes	206
8.7	Exercises	207
9	Ranking	209
9.1	The problem of ranking	209
9.2	Generalization bound	211
9.3	Ranking with SVMs	213
9.4	RankBoost	214
9.4.1	Bound on the empirical error	216
9.4.2	Relationship with coordinate descent	218

9.4.3	Margin bound for ensemble methods in ranking	220
9.5	Bipartite ranking	221
9.5.1	Boosting in bipartite ranking	222
9.5.2	Area under the ROC curve	224
9.6	Preference-based setting	226
9.6.1	Second-stage ranking problem	227
9.6.2	Deterministic algorithm	229
9.6.3	Randomized algorithm	230
9.6.4	Extension to other loss functions	231
9.7	Discussion	232
9.8	Chapter notes	233
9.9	Exercises	234
10	Regression	237
10.1	The problem of regression	237
10.2	Generalization bounds	238
10.2.1	Finite hypothesis sets	238
10.2.2	Rademacher complexity bounds	239
10.2.3	Pseudo-dimension bounds	241
10.3	Regression algorithms	245
10.3.1	Linear regression	245
10.3.2	Kernel ridge regression	247
10.3.3	Support vector regression	252
10.3.4	Lasso	257
10.3.5	Group norm regression algorithms	260
10.3.6	On-line regression algorithms	261
10.4	Chapter notes	262
10.5	Exercises	263
11	Algorithmic Stability	267
11.1	Definitions	267
11.2	Stability-based generalization guarantee	268
11.3	Stability of kernel-based regularization algorithms	270
11.3.1	Application to regression algorithms: SVR and KRR	274
11.3.2	Application to classification algorithms: SVMs	276
11.3.3	Discussion	276
11.4	Chapter notes	277
11.5	Exercises	277
12	Dimensionality Reduction	281
12.1	Principal Component Analysis	282

12.2	Kernel Principal Component Analysis (KPCA)	283
12.3	KPCA and manifold learning	285
12.3.1	Isomap	285
12.3.2	Laplacian eigenmaps	286
12.3.3	Locally linear embedding (LLE)	287
12.4	Johnson-Lindenstrauss lemma	288
12.5	Chapter notes	290
12.6	Exercises	290
13	Learning Automata and Languages	293
13.1	Introduction	293
13.2	Finite automata	294
13.3	Efficient exact learning	295
13.3.1	Passive learning	296
13.3.2	Learning with queries	297
13.3.3	Learning automata with queries	298
13.4	Identification in the limit	303
13.4.1	Learning reversible automata	304
13.5	Chapter notes	309
13.6	Exercises	310
14	Reinforcement Learning	313
14.1	Learning scenario	313
14.2	Markov decision process model	314
14.3	Policy	315
14.3.1	Definition	315
14.3.2	Policy value	316
14.3.3	Policy evaluation	316
14.3.4	Optimal policy	318
14.4	Planning algorithms	319
14.4.1	Value iteration	319
14.4.2	Policy iteration	322
14.4.3	Linear programming	324
14.5	Learning algorithms	325
14.5.1	Stochastic approximation	326
14.5.2	TD(0) algorithm	330
14.5.3	Q-learning algorithm	331
14.5.4	SARSA	334
14.5.5	TD(λ) algorithm	335
14.5.6	Large state space	336
14.6	Chapter notes	337

Conclusion	339
A Linear Algebra Review	341
A.1 Vectors and norms	341
A.1.1 Norms	341
A.1.2 Dual norms	342
A.2 Matrices	344
A.2.1 Matrix norms	344
A.2.2 Singular value decomposition	345
A.2.3 Symmetric positive semidefinite (SPSD) matrices	346
B Convex Optimization	349
B.1 Differentiation and unconstrained optimization	349
B.2 Convexity	350
B.3 Constrained optimization	353
B.4 Chapter notes	357
C Probability Review	359
C.1 Probability	359
C.2 Random variables	359
C.3 Conditional probability and independence	361
C.4 Expectation, Markov's inequality, and moment-generating function	363
C.5 Variance and Chebyshev's inequality	365
D Concentration inequalities	369
D.1 Hoeffding's inequality	369
D.2 McDiarmid's inequality	371
D.3 Other inequalities	373
D.3.1 Binomial distribution: Slud's inequality	374
D.3.2 Normal distribution: tail bound	374
D.3.3 Khintchine-Kahane inequality	374
D.4 Chapter notes	376
D.5 Exercises	377
E Notation	379
References	381
Index	397