

Contents

<i>Series Foreword</i>	xvii
<i>Figures</i>	xix
<i>Tables</i>	xxix
<i>Preface</i>	xxxix
<i>Acknowledgments</i>	xxxiii
<i>Notes for the Second Edition</i>	xxxv
<i>Notations</i>	xxxix
1 Introduction	1
1.1 What Is Machine Learning?	1
1.2 Examples of Machine Learning Applications	4
1.2.1 Learning Associations	4
1.2.2 Classification	5
1.2.3 Regression	9
1.2.4 Unsupervised Learning	11
1.2.5 Reinforcement Learning	13
1.3 Notes	14
1.4 Relevant Resources	16
1.5 Exercises	18
1.6 References	19
2 Supervised Learning	21
2.1 Learning a Class from Examples	21

2.2	Vapnik-Chervonenkis (VC) Dimension	27
2.3	Probably Approximately Correct (PAC) Learning	29
2.4	Noise	30
2.5	Learning Multiple Classes	32
2.6	Regression	34
2.7	Model Selection and Generalization	37
2.8	Dimensions of a Supervised Machine Learning Algorithm	41
2.9	Notes	42
2.10	Exercises	43
2.11	References	44
3	<i>Bayesian Decision Theory</i>	47
3.1	Introduction	47
3.2	Classification	49
3.3	Losses and Risks	51
3.4	Discriminant Functions	53
3.5	Utility Theory	54
3.6	Association Rules	55
3.7	Notes	58
3.8	Exercises	58
3.9	References	59
4	<i>Parametric Methods</i>	61
4.1	Introduction	61
4.2	Maximum Likelihood Estimation	62
4.2.1	Bernoulli Density	63
4.2.2	Multinomial Density	64
4.2.3	Gaussian (Normal) Density	64
4.3	Evaluating an Estimator: Bias and Variance	65
4.4	The Bayes' Estimator	66
4.5	Parametric Classification	69
4.6	Regression	73
4.7	Tuning Model Complexity: Bias/Variance Dilemma	76
4.8	Model Selection Procedures	80
4.9	Notes	84
4.10	Exercises	84
4.11	References	85
5	<i>Multivariate Methods</i>	87
5.1	Multivariate Data	87

5.2	Parameter Estimation	88
5.3	Estimation of Missing Values	89
5.4	Multivariate Normal Distribution	90
5.5	Multivariate Classification	94
5.6	Tuning Complexity	99
5.7	Discrete Features	102
5.8	Multivariate Regression	103
5.9	Notes	105
5.10	Exercises	106
5.11	References	107
6	<i>Dimensionality Reduction</i>	109
6.1	Introduction	109
6.2	Subset Selection	110
6.3	Principal Components Analysis	113
6.4	Factor Analysis	120
6.5	Multidimensional Scaling	125
6.6	Linear Discriminant Analysis	128
6.7	Isomap	133
6.8	Locally Linear Embedding	135
6.9	Notes	138
6.10	Exercises	139
6.11	References	140
7	<i>Clustering</i>	143
7.1	Introduction	143
7.2	Mixture Densities	144
7.3	<i>k</i> -Means Clustering	145
7.4	Expectation-Maximization Algorithm	149
7.5	Mixtures of Latent Variable Models	154
7.6	Supervised Learning after Clustering	155
7.7	Hierarchical Clustering	157
7.8	Choosing the Number of Clusters	158
7.9	Notes	160
7.10	Exercises	160
7.11	References	161
8	<i>Nonparametric Methods</i>	163
8.1	Introduction	163
8.2	Nonparametric Density Estimation	165

8.2.1	Histogram Estimator	165
8.2.2	Kernel Estimator	167
8.2.3	k -Nearest Neighbor Estimator	168
8.3	Generalization to Multivariate Data	170
8.4	Nonparametric Classification	171
8.5	Condensed Nearest Neighbor	172
8.6	Nonparametric Regression: Smoothing Models	174
8.6.1	Running Mean Smoother	175
8.6.2	Kernel Smoother	176
8.6.3	Running Line Smoother	177
8.7	How to Choose the Smoothing Parameter	178
8.8	Notes	180
8.9	Exercises	181
8.10	References	182
9	<i>Decision Trees</i>	185
9.1	Introduction	185
9.2	Univariate Trees	187
9.2.1	Classification Trees	188
9.2.2	Regression Trees	192
9.3	Pruning	194
9.4	Rule Extraction from Trees	197
9.5	Learning Rules from Data	198
9.6	Multivariate Trees	202
9.7	Notes	204
9.8	Exercises	207
9.9	References	207
10	<i>Linear Discrimination</i>	209
10.1	Introduction	209
10.2	Generalizing the Linear Model	211
10.3	Geometry of the Linear Discriminant	212
10.3.1	Two Classes	212
10.3.2	Multiple Classes	214
10.4	Pairwise Separation	216
10.5	Parametric Discrimination Revisited	217
10.6	Gradient Descent	218
10.7	Logistic Discrimination	220
10.7.1	Two Classes	220

10.7.2	Multiple Classes	224
10.8	Discrimination by Regression	228
10.9	Notes	230
10.10	Exercises	230
10.11	References	231
11	<i>Multilayer Perceptrons</i>	233
11.1	Introduction	233
11.1.1	Understanding the Brain	234
11.1.2	Neural Networks as a Paradigm for Parallel Processing	235
11.2	The Perceptron	237
11.3	Training a Perceptron	240
11.4	Learning Boolean Functions	243
11.5	Multilayer Perceptrons	245
11.6	MLP as a Universal Approximator	248
11.7	Backpropagation Algorithm	249
11.7.1	Nonlinear Regression	250
11.7.2	Two-Class Discrimination	252
11.7.3	Multiclass Discrimination	254
11.7.4	Multiple Hidden Layers	256
11.8	Training Procedures	256
11.8.1	Improving Convergence	256
11.8.2	Overtraining	257
11.8.3	Structuring the Network	258
11.8.4	Hints	261
11.9	Tuning the Network Size	263
11.10	Bayesian View of Learning	266
11.11	Dimensionality Reduction	267
11.12	Learning Time	270
11.12.1	Time Delay Neural Networks	270
11.12.2	Recurrent Networks	271
11.13	Notes	272
11.14	Exercises	274
11.15	References	275
12	<i>Local Models</i>	279
12.1	Introduction	279
12.2	Competitive Learning	280

12.2.1	Online k -Means	280	
12.2.2	Adaptive Resonance Theory	285	
12.2.3	Self-Organizing Maps	286	
12.3	Radial Basis Functions	288	
12.4	Incorporating Rule-Based Knowledge	294	
12.5	Normalized Basis Functions	295	
12.6	Competitive Basis Functions	297	
12.7	Learning Vector Quantization	300	
12.8	Mixture of Experts	300	
12.8.1	Cooperative Experts	303	
12.8.2	Competitive Experts	304	
12.9	Hierarchical Mixture of Experts	304	
12.10	Notes	305	
12.11	Exercises	306	
12.12	References	307	
13	Kernel Machines	309	
13.1	Introduction	309	
13.2	Optimal Separating Hyperplane	311	
13.3	The Nonseparable Case: Soft Margin Hyperplane	315	
13.4	ν -SVM	318	
13.5	Kernel Trick	319	
13.6	Vectorial Kernels	321	
13.7	Defining Kernels	324	
13.8	Multiple Kernel Learning	325	
13.9	Multiclass Kernel Machines	327	
13.10	Kernel Machines for Regression	328	
13.11	One-Class Kernel Machines	333	
13.12	Kernel Dimensionality Reduction	335	
13.13	Notes	337	
13.14	Exercises	338	
13.15	References	339	
14	Bayesian Estimation	341	
14.1	Introduction	341	
14.2	Estimating the Parameter of a Distribution	343	
14.2.1	Discrete Variables	343	
14.2.2	Continuous Variables	345	
14.3	Bayesian Estimation of the Parameters of a Function	348	

14.3.1	Regression	348	
14.3.2	The Use of Basis/Kernel Functions		352
14.3.3	Bayesian Classification	353	
14.4	Gaussian Processes	356	
14.5	Notes	359	
14.6	Exercises	360	
14.7	References	361	
15	<i>Hidden Markov Models</i>	363	
15.1	Introduction	363	
15.2	Discrete Markov Processes	364	
15.3	Hidden Markov Models	367	
15.4	Three Basic Problems of HMMs	369	
15.5	Evaluation Problem	369	
15.6	Finding the State Sequence	373	
15.7	Learning Model Parameters	375	
15.8	Continuous Observations	378	
15.9	The HMM with Input	379	
15.10	Model Selection in HMM	380	
15.11	Notes	382	
15.12	Exercises	383	
15.13	References	384	
16	<i>Graphical Models</i>	387	
16.1	Introduction	387	
16.2	Canonical Cases for Conditional Independence		389
16.3	Example Graphical Models	396	
16.3.1	Naive Bayes' Classifier	396	
16.3.2	Hidden Markov Model	398	
16.3.3	Linear Regression	401	
16.4	d-Separation	402	
16.5	Belief Propagation	402	
16.5.1	Chains	403	
16.5.2	Trees	405	
16.5.3	Polytrees	407	
16.5.4	Junction Trees	409	
16.6	Undirected Graphs: Markov Random Fields		410
16.7	Learning the Structure of a Graphical Model		413
16.8	Influence Diagrams	414	

- 16.9 Notes 414
- 16.10 Exercises 417
- 16.11 References 417

17 Combining Multiple Learners 419

- 17.1 Rationale 419
- 17.2 Generating Diverse Learners 420
- 17.3 Model Combination Schemes 423
- 17.4 Voting 424
- 17.5 Error-Correcting Output Codes 427
- 17.6 Bagging 430
- 17.7 Boosting 431
- 17.8 Mixture of Experts Revisited 434
- 17.9 Stacked Generalization 435
- 17.10 Fine-Tuning an Ensemble 437
- 17.11 Cascading 438
- 17.12 Notes 440
- 17.13 Exercises 442
- 17.14 References 443

18 Reinforcement Learning 447

- 18.1 Introduction 447
- 18.2 Single State Case: K -Armed Bandit 449
- 18.3 Elements of Reinforcement Learning 450
- 18.4 Model-Based Learning 453
 - 18.4.1 Value Iteration 453
 - 18.4.2 Policy Iteration 454
- 18.5 Temporal Difference Learning 454
 - 18.5.1 Exploration Strategies 455
 - 18.5.2 Deterministic Rewards and Actions 456
 - 18.5.3 Nondeterministic Rewards and Actions 457
 - 18.5.4 Eligibility Traces 459
- 18.6 Generalization 461
- 18.7 Partially Observable States 464
 - 18.7.1 The Setting 464
 - 18.7.2 Example: The Tiger Problem 465
- 18.8 Notes 470
- 18.9 Exercises 472
- 18.10 References 473

19	<i>Design and Analysis of Machine Learning Experiments</i>	475
19.1	Introduction	475
19.2	Factors, Response, and Strategy of Experimentation	478
19.3	Response Surface Design	481
19.4	Randomization, Replication, and Blocking	482
19.5	Guidelines for Machine Learning Experiments	483
19.6	Cross-Validation and Resampling Methods	486
19.6.1	<i>K</i> -Fold Cross-Validation	487
19.6.2	5×2 Cross-Validation	488
19.6.3	Bootstrapping	489
19.7	Measuring Classifier Performance	489
19.8	Interval Estimation	493
19.9	Hypothesis Testing	496
19.10	Assessing a Classification Algorithm's Performance	498
19.10.1	Binomial Test	499
19.10.2	Approximate Normal Test	500
19.10.3	<i>t</i> Test	500
19.11	Comparing Two Classification Algorithms	501
19.11.1	McNemar's Test	501
19.11.2	<i>K</i> -Fold Cross-Validated Paired <i>t</i> Test	501
19.11.3	5 × 2 cv Paired <i>t</i> Test	502
19.11.4	5 × 2 cv Paired <i>F</i> Test	503
19.12	Comparing Multiple Algorithms: Analysis of Variance	504
19.13	Comparison over Multiple Datasets	508
19.13.1	Comparing Two Algorithms	509
19.13.2	Multiple Algorithms	511
19.14	Notes	512
19.15	Exercises	513
19.16	References	514
A	<i>Probability</i>	517
A.1	Elements of Probability	517
A.1.1	Axioms of Probability	518
A.1.2	Conditional Probability	518
A.2	Random Variables	519
A.2.1	Probability Distribution and Density Functions	519
A.2.2	Joint Distribution and Density Functions	520
A.2.3	Conditional Distributions	520
A.2.4	Bayes' Rule	521

	A.2.5	Expectation	521
	A.2.6	Variance	522
	A.2.7	Weak Law of Large Numbers	523
A.3		Special Random Variables	523
	A.3.1	Bernoulli Distribution	523
	A.3.2	Binomial Distribution	524
	A.3.3	Multinomial Distribution	524
	A.3.4	Uniform Distribution	524
	A.3.5	Normal (Gaussian) Distribution	525
	A.3.6	Chi-Square Distribution	526
	A.3.7	t Distribution	527
	A.3.8	F Distribution	527
A.4		References	527
<i>Index</i>			529