

INTRODUCTION TO
Machine Learning

ETHEM ALPAYDIN

Contents

Series Foreword **xiii**

Figures **xv**

Tables **xxiii**

Preface **xxv**

Acknowledgments **xxvii**

Notations **xxix**

1 Introduction **1**

- 1.1 What Is Machine Learning? 1
- 1.2 Examples of Machine Learning Applications 3
 - 1.2.1 Learning Associations 3
 - 1.2.2 Classification 4
 - 1.2.3 Regression 8
 - 1.2.4 Unsupervised Learning 10
 - 1.2.5 Reinforcement Learning 11
- 1.3 Notes 12
- 1.4 Relevant Resources 14
- 1.5 Exercises 15
- 1.6 References 16

2 Supervised Learning **17**

- 2.1 Learning a Class from Examples 17
- 2.2 Vapnik-Chervonenkis (VC) Dimension 22
- 2.3 Probably Approximately Correct (PAC) Learning 24

2.4	Noise	25
2.5	Learning Multiple Classes	27
2.6	Regression	29
2.7	Model Selection and Generalization	32
2.8	Dimensions of a Supervised Machine Learning Algorithm	35
2.9	Notes	36
2.10	Exercises	37
2.11	References	38
3	<i>Bayesian Decision Theory</i>	39
3.1	Introduction	39
3.2	Classification	41
3.3	Losses and Risks	43
3.4	Discriminant Functions	45
3.5	Utility Theory	46
3.6	Value of Information	47
3.7	Bayesian Networks	48
3.8	Influence Diagrams	55
3.9	Association Rules	56
3.10	Notes	57
3.11	Exercises	57
3.12	References	58
4	<i>Parametric Methods</i>	61
4.1	Introduction	61
4.2	Maximum Likelihood Estimation	62
4.2.1	Bernoulli Density	62
4.2.2	Multinomial Density	63
4.2.3	Gaussian (Normal) Density	64
4.3	Evaluating an Estimator: Bias and Variance	64
4.4	The Bayes' Estimator	67
4.5	Parametric Classification	69
4.6	Regression	73
4.7	Tuning Model Complexity: Bias/Variance Dilemma	76
4.8	Model Selection Procedures	79
4.9	Notes	82
4.10	Exercises	82
4.11	References	83

5	<i>Multivariate Methods</i>	85
5.1	Multivariate Data	85
5.2	Parameter Estimation	86
5.3	Estimation of Missing Values	87
5.4	Multivariate Normal Distribution	88
5.5	Multivariate Classification	92
5.6	Tuning Complexity	98
5.7	Discrete Features	99
5.8	Multivariate Regression	100
5.9	Notes	102
5.10	Exercises	102
5.11	References	103
6	<i>Dimensionality Reduction</i>	105
6.1	Introduction	105
6.2	Subset Selection	106
6.3	Principal Components Analysis	108
6.4	Factor Analysis	116
6.5	Multidimensional Scaling	121
6.6	Linear Discriminant Analysis	124
6.7	Notes	127
6.8	Exercises	130
6.9	References	130
7	<i>Clustering</i>	133
7.1	Introduction	133
7.2	Mixture Densities	134
7.3	<i>k</i> -Means Clustering	135
7.4	Expectation-Maximization Algorithm	139
7.5	Mixtures of Latent Variable Models	144
7.6	Supervised Learning after Clustering	145
7.7	Hierarchical Clustering	146
7.8	Choosing the Number of Clusters	149
7.9	Notes	149
7.10	Exercises	150
7.11	References	150
8	<i>Nonparametric Methods</i>	153
8.1	Introduction	153

8.2	Nonparametric Density Estimation	154
8.2.1	Histogram Estimator	155
8.2.2	Kernel Estimator	157
8.2.3	k-Nearest Neighbor Estimator	158
8.3	Generalization to Multivariate Data	159
8.4	Nonparametric Classification	161
8.5	Condensed Nearest Neighbor	162
8.6	Nonparametric Regression: Smoothing Models	164
8.6.1	Running Mean Smoother	165
8.6.2	Kernel Smoother	166
8.6.3	Running Line Smoother	167
8.7	How to Choose the Smoothing Parameter	168
8.8	Notes	169
8.9	Exercises	170
8.10	References	170
9	Decision Trees	173
9.1	Introduction	173
9.2	Univariate Trees	175
9.2.1	Classification Trees	176
9.2.2	Regression Trees	180
9.3	Pruning	182
9.4	Rule Extraction from Trees	185
9.5	Learning Rules from Data	186
9.6	Multivariate Trees	190
9.7	Notes	192
9.8	Exercises	195
9.9	References	195
10	Linear Discrimination	197
10.1	Introduction	197
10.2	Generalizing the Linear Model	199
10.3	Geometry of the Linear Discriminant	200
10.3.1	Two Classes	200
10.3.2	Multiple Classes	202
10.4	Pairwise Separation	204
10.5	Parametric Discrimination Revisited	205
10.6	Gradient Descent	206
10.7	Logistic Discrimination	208

10.7.1	Two Classes	208
10.7.2	Multiple Classes	211
10.8	Discrimination by Regression	216
10.9	Support Vector Machines	218
10.9.1	Optimal Separating Hyperplane	218
10.9.2	The Nonseparable Case: Soft Margin Hyperplane	221
10.9.3	Kernel Functions	223
10.9.4	Support Vector Machines for Regression	225
10.10	Notes	227
10.11	Exercises	227
10.12	References	228
11	Multilayer Perceptrons	229
11.1	Introduction	229
11.1.1	Understanding the Brain	230
11.1.2	Neural Networks as a Paradigm for Parallel Processing	231
11.2	The Perceptron	233
11.3	Training a Perceptron	236
11.4	Learning Boolean Functions	239
11.5	Multilayer Perceptrons	241
11.6	MLP as a Universal Approximator	244
11.7	Backpropagation Algorithm	245
11.7.1	Nonlinear Regression	246
11.7.2	Two-Class Discrimination	248
11.7.3	Multiclass Discrimination	250
11.7.4	Multiple Hidden Layers	252
11.8	Training Procedures	252
11.8.1	Improving Convergence	252
11.8.2	Overtraining	253
11.8.3	Structuring the Network	254
11.8.4	Hints	257
11.9	Tuning the Network Size	259
11.10	Bayesian View of Learning	262
11.11	Dimensionality Reduction	263
11.12	Learning Time	266
11.12.1	Time Delay Neural Networks	266
11.12.2	Recurrent Networks	267

11.13	Notes	268
11.14	Exercises	270
11.15	References	271

12 Local Models 275

12.1	Introduction	275
12.2	Competitive Learning	276
12.2.1	Online k -Means	276
12.2.2	Adaptive Resonance Theory	281
12.2.3	Self-Organizing Maps	282
12.3	Radial Basis Functions	284
12.4	Incorporating Rule-Based Knowledge	290
12.5	Normalized Basis Functions	291
12.6	Competitive Basis Functions	293
12.7	Learning Vector Quantization	296
12.8	Mixture of Experts	296
12.8.1	Cooperative Experts	299
12.8.2	Competitive Experts	300
12.9	Hierarchical Mixture of Experts	300
12.10	Notes	301
12.11	Exercises	302
12.12	References	302

13 Hidden Markov Models 305

13.1	Introduction	305
13.2	Discrete Markov Processes	306
13.3	Hidden Markov Models	309
13.4	Three Basic Problems of HMMs	311
13.5	Evaluation Problem	311
13.6	Finding the State Sequence	315
13.7	Learning Model Parameters	317
13.8	Continuous Observations	320
13.9	The HMM with Input	321
13.10	Model Selection in HMM	322
13.11	Notes	323
13.12	Exercises	325
13.13	References	325

14 Assessing and Comparing Classification Algorithms 327

14.1	Introduction	327
------	--------------	-----

14.2	Cross-Validation and Resampling Methods	330
14.2.1	K-Fold Cross-Validation	331
14.2.2	5×2 Cross-Validation	331
14.2.3	Bootstrapping	332
14.3	Measuring Error	333
14.4	Interval Estimation	334
14.5	Hypothesis Testing	338
14.6	Assessing a Classification Algorithm's Performance	339
14.6.1	Binomial Test	340
14.6.2	Approximate Normal Test	341
14.6.3	Paired t Test	341
14.7	Comparing Two Classification Algorithms	341
14.7.1	McNemar's Test	342
14.7.2	K-Fold Cross-Validated Paired t Test	342
14.7.3	5 × 2 cv Paired t Test	343
14.7.4	5 × 2 cv Paired F Test	344
14.8	Comparing Multiple Classification Algorithms: Analysis of Variance	345
14.9	Notes	348
14.10	Exercises	349
14.11	References	350
15	Combining Multiple Learners	351
15.1	Rationale	351
15.2	Voting	354
15.3	Error-Correcting Output Codes	357
15.4	Bagging	360
15.5	Boosting	360
15.6	Mixture of Experts Revisited	363
15.7	Stacked Generalization	364
15.8	Cascading	366
15.9	Notes	368
15.10	Exercises	369
15.11	References	370
16	Reinforcement Learning	373
16.1	Introduction	373
16.2	Single State Case: K -Armed Bandit	375
16.3	Elements of Reinforcement Learning	376

16.4	Model-Based Learning	379
16.4.1	Value Iteration	379
16.4.2	Policy Iteration	380
16.5	Temporal Difference Learning	380
16.5.1	Exploration Strategies	381
16.5.2	Deterministic Rewards and Actions	382
16.5.3	Nondeterministic Rewards and Actions	383
16.5.4	Eligibility Traces	385
16.6	Generalization	387
16.7	Partially Observable States	389
16.8	Notes	391
16.9	Exercises	393
16.10	References	394

A Probability 397

A.1	Elements of Probability	397
A.1.1	Axioms of Probability	398
A.1.2	Conditional Probability	398
A.2	Random Variables	399
A.2.1	Probability Distribution and Density Functions	399
A.2.2	Joint Distribution and Density Functions	400
A.2.3	Conditional Distributions	400
A.2.4	Bayes' Rule	401
A.2.5	Expectation	401
A.2.6	Variance	402
A.2.7	Weak Law of Large Numbers	403
A.3	Special Random Variables	403
A.3.1	Bernoulli Distribution	403
A.3.2	Binomial Distribution	404
A.3.3	Multinomial Distribution	404
A.3.4	Uniform Distribution	404
A.3.5	Normal (Gaussian) Distribution	405
A.3.6	Chi-Square Distribution	406
A.3.7	t Distribution	407
A.3.8	F Distribution	407
A.4	References	407