

Contents

	<i>Preface</i>	<i>page xi</i>
	<i>Acknowledgments</i>	<i>xii</i>
Part I	Introduction	1
1	Prologue	3
	1.1 Machines that learn – some recent history	3
	1.2 Twenty canonical questions	7
	1.3 Outline of the book	9
	1.4 A comment about example datasets	11
	1.5 Software	12
	Note	13
2	The landscape of learning machines	14
	2.1 Introduction	14
	2.2 Types of data for learning machines	15
	2.3 Will that be supervised or unsupervised?	17
	2.4 An unsupervised example	18
	2.5 More lack of supervision – where are the parents?	20
	2.6 Engines, complex and primitive	20
	2.7 Model richness means what, exactly?	22
	2.8 Membership or probability of membership?	25
	2.9 A taxonomy of machines?	27
	2.10 A note of caution – one of many	30
	2.11 Highlights from the theory	30
	Notes	36
3	A mangle of machines	41
	3.1 Introduction	41

vi	Contents	
	3.2 Linear regression	41
	3.3 Logistic regression	42
	3.4 Linear discriminant	43
	3.5 Bayes classifiers – regular and naïve	45
	3.6 Logic regression	47
	3.7 k -Nearest neighbors	48
	3.8 Support vector machines	50
	3.9 Neural networks	53
	3.10 Boosting	54
	3.11 Evolutionary and genetic algorithms	55
	Notes	56
4	Three examples and several machines	57
	4.1 Introduction	57
	4.2 Simulated cholesterol data	58
	4.3 Lupus data	61
	4.4 Stroke data	62
	4.5 Biomedical <i>means</i> unbalanced	63
	4.6 Measures of machine performance	64
	4.7 Linear analysis of cholesterol data	66
	4.8 Nonlinear analysis of cholesterol data	67
	4.9 Analysis of the lupus data	70
	4.10 Analysis of the stroke data	75
	4.11 Further analysis of the lupus and stroke data	79
	Notes	87
Part II	A machine toolkit	89
5	Logistic regression	91
	5.1 Introduction	91
	5.2 Inside and around the model	92
	5.3 Interpreting the coefficients	93
	5.4 Using logistic regression as a decision rule	94
	5.5 Logistic regression applied to the cholesterol data	94
	5.6 A cautionary note	98
	5.7 Another cautionary note	101
	5.8 Probability estimates and decision rules	102

vii	Contents	
	5.9 Evaluating the goodness-of-fit of a logistic regression model	103
	5.10 Calibrating a logistic regression	106
	5.11 Beyond calibration	111
	5.12 Logistic regression and reference models	113
	Notes	115
6	A single decision tree	118
	6.1 Introduction	118
	6.2 Dropping down trees	118
	6.3 Growing a tree	120
	6.4 Selecting features, making splits	120
	6.5 Good split, bad split	121
	6.6 Finding good features for making splits	124
	6.7 Misreading trees	125
	6.8 Stopping and pruning rules	127
	6.9 Using functions of the features	128
	6.10 Unstable trees?	129
	6.11 Variable importance – growing on trees?	132
	6.12 Permuting for importance	134
	6.13 The continuing mystery of trees	135
7	Random Forests – trees everywhere	137
	7.1 Random Forests in less than five minutes	137
	7.2 Random treks through the data	138
	7.3 Random treks through the features	139
	7.4 Walking through the forest	140
	7.5 Weighted and unweighted voting	140
	7.6 Finding subsets in the data using proximities	142
	7.7 Applying Random Forests to the Stroke data	144
	7.8 Random Forests in the universe of machines	151
	Notes	153
Part III	Analysis fundamentals	155
8	Merely two variables	157
	8.1 Introduction	157
	8.2 Understanding correlations	158
	8.3 Hazards of correlations	159

viii	Contents	
	8.4 Correlations big and small	163
	Notes	168
9	More than two variables	171
	9.1 Introduction	171
	9.2 Tiny problems, large consequences	172
	9.3 Mathematics to the rescue?	174
	9.4 Good models need not be unique	176
	9.5 Contexts and coefficients	179
	9.6 Interpreting and testing coefficients in models	181
	9.7 Merging models, pooling lists, ranking features	186
	Notes	190
10	Resampling methods	198
	10.1 Introduction	198
	10.2 The bootstrap	198
	10.3 When the bootstrap works	201
	10.4 When the bootstrap doesn't work	202
	10.5 Resampling from a single group in different ways	203
	10.6 Resampling from groups with unequal sizes	204
	10.7 Resampling from small datasets	206
	10.8 Permutation methods	207
	10.9 Still more on permutation methods	210
	Note	214
11	Error analysis and model validation	215
	11.1 Introduction	215
	11.2 Errors? What errors?	217
	11.3 Unbalanced data, unbalanced errors	218
	11.4 Error analysis for a single machine	219
	11.5 Cross-validation error estimation	222
	11.6 Cross-validation or cross-training?	224
	11.7 The leave-one-out method	226
	11.8 The out-of-bag method	227
	11.9 Intervals for error estimates for a single machine	228
	11.10 Tossing random coins into the abyss	230
	11.11 Error estimates for unbalanced data	232
	11.12 Confidence intervals for comparing error values	233

ix	Contents	
	11.13 Other measures of machine accuracy	236
	11.14 Benchmarking and winning the lottery	238
	11.15 Error analysis for predicting continuous outcomes	239
	Notes	240
Part IV	Machine strategies	245
12	Ensemble methods – let’s take a vote	247
	12.1 Pools of machines	247
	12.2 Weak correlation with outcome can be good enough	247
	12.3 Model averaging	250
	Notes	254
13	Summary and conclusions	255
	13.1 Where have we been?	255
	13.2 So many machines	257
	13.3 Binary decision or probability estimate?	259
	13.4 Survival machines? Risk machines?	259
	13.5 And where are we going?	260
	<i>Appendix</i>	263
	<i>References</i>	271
	<i>Index</i>	281

The color plate is situated between pages 244 and 245.