

Contents

Preface	xv
Notation	xxiii
List of Figures	xxxi
List of Tables	xlvii
I INTRODUCTION TO MACHINE LEARNING AND DATA ANALYTICS	1
1 Machine Learning for Predictive Data Analytics	3
1.1 What Is Predictive Data Analytics?	3
1.2 What Is Machine Learning?	5
1.3 How Does Machine Learning Work?	7
1.4 Inductive Bias Versus Sample Bias	12
1.5 What Can Go Wrong with Machine Learning?	13
1.6 The Predictive Data Analytics Project Lifecycle: CRISP-DM	15
1.7 Predictive Data Analytics Tools	17
1.8 The Road Ahead	19
1.9 Exercises	21
2 Data to Insights to Decisions	23
2.1 Converting Business Problems into Analytics Solutions	23
2.1.1 Case Study: Motor Insurance Fraud	25
2.2 Assessing Feasibility	26
2.2.1 Case Study: Motor Insurance Fraud	27
2.3 Designing the Analytics Base Table	28
2.3.1 Case Study: Motor Insurance Fraud	31
2.4 Designing and Implementing Features	32
2.4.1 Different Types of Data	34

2.4.2	Different Types of Features	34
2.4.3	Handling Time	36
2.4.4	Legal Issues	39
2.4.5	Implementing Features	41
2.4.6	Case Study: Motor Insurance Fraud	42
2.5	Summary	44
2.6	Further Reading	47
2.7	Exercises	48
3	Data Exploration	53
3.1	The Data Quality Report	54
3.1.1	Case Study: Motor Insurance Fraud	55
3.2	Getting to Know the Data	55
3.2.1	The Normal Distribution	61
3.2.2	Case Study: Motor Insurance Fraud	62
3.3	Identifying Data Quality Issues	63
3.3.1	Missing Values	64
3.3.2	Irregular Cardinality	64
3.3.3	Outliers	65
3.3.4	Case Study: Motor Insurance Fraud	66
3.4	Handling Data Quality Issues	69
3.4.1	Handling Missing Values	69
3.4.2	Handling Outliers	70
3.4.3	Case Study: Motor Insurance Fraud	71
3.5	Advanced Data Exploration	72
3.5.1	Visualizing Relationships between Features	72
3.5.2	Measuring Covariance and Correlation	81
3.6	Data Preparation	87
3.6.1	Normalization	87
3.6.2	Binning	89
3.6.3	Sampling	91
3.7	Summary	94
3.8	Further Reading	95
3.9	Exercises	96
II	PREDICTIVE DATA ANALYTICS	115
4	Information-Based Learning	117
4.1	Big Idea	117
4.2	Fundamentals	120

<i>Contents</i>	ix
4.2.1 Decision Trees	121
4.2.2 Shannon’s Entropy Model	123
4.2.3 Information Gain	127
4.3 Standard Approach: The ID3 Algorithm	132
4.3.1 A Worked Example: Predicting Vegetation Distributions	135
4.4 Extensions and Variations	141
4.4.1 Alternative Feature Selection and Impurity Metrics	142
4.4.2 Handling Continuous Descriptive Features	146
4.4.3 Predicting Continuous Targets	149
4.4.4 Tree Pruning	153
4.4.5 Model Ensembles	158
4.5 Summary	169
4.6 Further Reading	170
4.7 Exercises	172
5 Similarity-Based Learning	181
5.1 Big Idea	181
5.2 Fundamentals	182
5.2.1 Feature Space	183
5.2.2 Measuring Similarity Using Distance Metrics	184
5.3 Standard Approach: The Nearest Neighbor Algorithm	187
5.3.1 A Worked Example	188
5.4 Extensions and Variations	191
5.4.1 Handling Noisy Data	191
5.4.2 Efficient Memory Search	196
5.4.3 Data Normalization	204
5.4.4 Predicting Continuous Targets	208
5.4.5 Other Measures of Similarity	211
5.4.6 Feature Selection	223
5.5 Summary	230
5.6 Further Reading	233
5.7 Epilogue	234
5.8 Exercises	236
6 Probability-Based Learning	243
6.1 Big Idea	243
6.2 Fundamentals	245
6.2.1 Bayes’ Theorem	248
6.2.2 Bayesian Prediction	251
6.2.3 Conditional Independence and Factorization	256

6.3	Standard Approach: The Naive Bayes Model	261
6.3.1	A Worked Example	262
6.4	Extensions and Variations	265
6.4.1	Smoothing	265
6.4.2	Continuous Features: Probability Density Functions	269
6.4.3	Continuous Features: Binning	280
6.4.4	Bayesian Networks	284
6.5	Summary	300
6.6	Further Reading	303
6.7	Exercises	305
7	Error-Based Learning	311
7.1	Big Idea	311
7.2	Fundamentals	312
7.2.1	Simple Linear Regression	312
7.2.2	Measuring Error	315
7.2.3	Error Surfaces	317
7.3	Standard Approach: Multivariable Linear Regression with Gradient Descent	319
7.3.1	Multivariable Linear Regression	319
7.3.2	Gradient Descent	321
7.3.3	Choosing Learning Rates and Initial Weights	328
7.3.4	A Worked Example	330
7.4	Extensions and Variations	332
7.4.1	Interpreting Multivariable Linear Regression Models	332
7.4.2	Setting the Learning Rate Using Weight Decay	334
7.4.3	Handling Categorical Descriptive Features	336
7.4.4	Handling Categorical Target Features: Logistic Regression	338
7.4.5	Modeling Non-Linear Relationships	351
7.4.6	Multinomial Logistic Regression	357
7.4.7	Support Vector Machines	361
7.5	Summary	367
7.6	Further Reading	370
7.7	Exercises	371
8	Deep Learning	381
8.1	Big Idea	382
8.2	Fundamentals	383
8.2.1	Artificial Neurons	384
8.2.2	Artificial Neural Networks	388

<i>Contents</i>	xi
8.2.3 Neural Networks as Matrix Operations	390
8.2.4 Why Are Non-Linear Activation Functions Necessary?	394
8.2.5 Why Is Network Depth Important?	395
8.3 Standard Approach: Backpropagation and Gradient Descent	403
8.3.1 Backpropagation: The General Structure of the Algorithm	404
8.3.2 Backpropagation: Backpropagating the Error Gradients	407
8.3.3 Backpropagation: Updating the Weights in a Network	413
8.3.4 Backpropagation: The Algorithm	418
8.3.5 A Worked Example: Using Backpropagation to Train a Feedforward Network for a Regression Task	421
8.4 Extensions and Variations	434
8.4.1 Vanishing Gradients and ReLUs	434
8.4.2 Weight Initialization and Unstable Gradients	447
8.4.3 Handling Categorical Target Features: Softmax Output Layers and Cross-Entropy Loss Functions	463
8.4.4 Early Stopping and Dropout: Preventing Overfitting	472
8.4.5 Convolutional Neural Networks	477
8.4.6 Sequential Models: Recurrent Neural Networks and Long Short-Term Memory Networks	499
8.5 Summary	521
8.6 Further Reading	523
8.7 Exercises	524
9 Evaluation	533
9.1 Big Idea	533
9.2 Fundamentals	534
9.3 Standard Approach: Misclassification Rate on a Hold-Out Test Set	535
9.4 Extensions and Variations	540
9.4.1 Designing Evaluation Experiments	540
9.4.2 Performance Measures: Categorical Targets	547
9.4.3 Performance Measures: Prediction Scores	556
9.4.4 Performance Measures: Multinomial Targets	572
9.4.5 Performance Measures: Continuous Targets	574
9.4.6 Evaluating Models after Deployment	578
9.5 Summary	585
9.6 Further Reading	586
9.7 Exercises	588

III	BEYOND PREDICTION	595
10	Beyond Prediction: Unsupervised Learning	597
10.1	Big Idea	597
10.2	Fundamentals	598
10.3	Standard Approach: The k -Means Clustering Algorithm	600
10.3.1	A Worked Example	601
10.4	Extensions and Variations	605
10.4.1	Choosing Initial Cluster Centroids	605
10.4.2	Evaluating Clustering	607
10.4.3	Choosing the Number of Clusters	612
10.4.4	Understanding Clustering Results	613
10.4.5	Agglomerative Hierarchical Clustering	616
10.4.6	Representation Learning with Auto-Encoders	624
10.5	Summary	628
10.6	Further Reading	629
10.7	Exercises	631
11	Beyond Prediction: Reinforcement Learning	637
11.1	Big Idea	637
11.2	Fundamentals	638
11.2.1	Intelligent Agents	639
11.2.2	Fundamentals of Reinforcement Learning	640
11.2.3	Markov Decision Processes	643
11.2.4	The Bellman Equations	651
11.2.5	Temporal-Difference Learning	654
11.3	Standard Approach: Q-Learning, Off-Policy Temporal-Difference Learning	657
11.3.1	A Worked Example	659
11.4	Extensions and Variations	664
11.4.1	SARSA, On-Policy Temporal-Difference Learning	664
11.4.2	Deep Q Networks	668
11.5	Summary	674
11.6	Further Reading	677
11.7	Exercises	679

<i>Contents</i>	xiii
IV CASE STUDIES AND CONCLUSIONS	683
12 Case Study: Customer Churn	685
12.1 Business Understanding	685
12.2 Data Understanding	688
12.3 Data Preparation	691
12.4 Modeling	697
12.5 Evaluation	698
12.6 Deployment	702
13 Case Study: Galaxy Classification	703
13.1 Business Understanding	704
13.1.1 Situational Fluency	706
13.2 Data Understanding	707
13.3 Data Preparation	713
13.4 Modeling	719
13.4.1 Baseline Models	719
13.4.2 Feature Selection	722
13.4.3 The 5-Level Model	722
13.5 Evaluation	725
13.6 Deployment	727
14 The Art of Machine Learning for Predictive Data Analytics	729
14.1 Different Perspectives on Prediction Models	731
14.2 Choosing a Machine Learning Approach	735
14.2.1 Matching Machine Learning Approaches to Projects	738
14.2.2 Matching Machine Learning Approaches to Data	739
14.3 Beyond Prediction	740
14.4 Your Next Steps	741
V APPENDICES	743
A Descriptive Statistics and Data Visualization for Machine Learning	745
A.1 Descriptive Statistics for Continuous Features	745
A.1.1 Central Tendency	745
A.1.2 Variation	746
A.2 Descriptive Statistics for Categorical Features	749
A.3 Populations and Samples	750
A.4 Data Visualization	752
A.4.1 Bar Plots	752
A.4.2 Histograms	752

A.4.3	Box Plots	755
B	Introduction to Probability for Machine Learning	757
B.1	Probability Basics	757
B.2	Probability Distributions and Summing Out	761
B.3	Some Useful Probability Rules	762
B.4	Summary	763
C	Differentiation Techniques for Machine Learning	765
C.1	Derivatives of Continuous Functions	766
C.2	The Chain Rule	768
C.3	Partial Derivatives	768
D	Introduction to Linear Algebra	771
D.1	Basic Types	771
D.2	Transpose	772
D.3	Multiplication	772
D.4	Summary	774
	Bibliography	775
	Index	787