

Contents

Preface

ix

I Introduction

1 About the Book and Supporting Material 3

- 1.1 What Do Data Mining, Machine Learning, and Knowledge Discovery Mean? 3
- 1.2 What Is This Book About? 5
- 1.3 An Incomplete Survey of the Relevant Literature 8
- 1.4 Introduction to the Python Language and the Git Code Management Tool 12
- 1.5 Description of Surveys and Data Sets Used in Examples 13
- 1.6 Plotting and Visualizing the Data in This Book 29
- 1.7 How to Efficiently Use This Book 35
- References 37


2 Fast Computation on Massive Data Sets 41

- 2.1 Data Types and Data Management Systems 41
- 2.2 Analysis of Algorithmic Efficiency 42
- 2.3 Seven Types of Computational Problem 44
- 2.4 Eight Strategies for Speeding Things Up 45
- 2.5 Case Studies: Speedup Strategies in Practice 48
- References 60

II Statistical Frameworks and Exploratory Data Analysis

3 Probability and Statistical Distributions 65

- 3.1 Brief Overview of Probability and Random Variables 66
- 3.2 Descriptive Statistics 73
- 3.3 Common Univariate Distribution Functions 80
- 3.4 The Central Limit Theorem 98
- 3.5 Bivariate and Multivariate Distribution Functions 100
- 3.6 Correlation Coefficients 108
- 3.7 Random Number Generation for Arbitrary Distributions 111
- References 114

4	Classical Statistical Inference	115
4.1	Classical vs. Bayesian Statistical Inference	115
4.2	Maximum Likelihood Estimation (MLE)	116
4.3	The Goodness of Fit and Model Selection	123
4.4	ML Applied to Gaussian Mixtures: The Expectation Maximization Algorithm	126
4.5	Confidence Estimates: The Bootstrap and the Jackknife	132
4.6	Hypothesis Testing	135
4.7	Comparison of Distributions	141
4.8	Nonparametric Modeling and Histograms	153
4.9	Selection Effects and Luminosity Function Estimation	157
4.10	Summary	162
	References	162
5	Bayesian Statistical Inference	165
5.1	Introduction to the Bayesian Method	166
5.2	Bayesian Priors	170
5.3	Bayesian Parameter Uncertainty Quantification	174
5.4	Bayesian Model Selection	175
5.5	Nonuniform Priors: Eddington, Malmquist, and Lutz–Kelker Biases	180
5.6	Simple Examples of Bayesian Analysis: Parameter Estimation	185
5.7	Simple Examples of Bayesian Analysis: Model Selection	211
5.8	Numerical Methods for Complex Problems (MCMC)	217
5.9	Hierarchical Bayesian Modeling	228
5.10	Approximate Bayesian Computation	232
5.11	Summary of Pros and Cons for Classical and Bayesian Methods	234
	References	237
	Data Mining and Machine Learning	
6	Searching for Structure in Point Data	243
6.1	Nonparametric Density Estimation	244
6.2	Nearest-Neighbor Density Estimation	251
6.3	Parametric Density Estimation	253
6.4	Finding Clusters in Data	263
6.5	Correlation Functions	269
6.6	Which Density Estimation and Clustering Algorithms Should I Use?	273
	References	277
7	Dimensionality and Its Reduction	281
7.1	The Curse of Dimensionality	281
7.2	The Data Sets Used in This Chapter	283

7.3	Principal Component Analysis	283
7.4	Nonnegative Matrix Factorization	295
7.5	Manifold Learning	297
7.6	Independent Component Analysis and Projection Pursuit	304
7.7	Which Dimensionality Reduction Technique Should I Use?	306
	References	309
8	Regression and Model Fitting	311
8.1	Formulation of the Regression Problem	311
8.2	Regression for Linear Models	315
8.3	Regularization and Penalizing the Likelihood	321
8.4	Principal Component Regression	326
8.5	Kernel Regression	327
8.6	Locally Linear Regression	328
8.7	Nonlinear Regression	329
8.8	Uncertainties in the Data	331
8.9	Regression That Is Robust to Outliers	332
8.10	Gaussian Process Regression	337
8.11	Overfitting, Underfitting, and Cross-Validation	341
8.12	Which Regression Method Should I Use?	349
	References	351
9	Classification	353
9.1	Data Sets Used in This Chapter	353
9.2	Assigning Categories: Classification	354
9.3	Generative Classification	356
9.4	K -Nearest-Neighbor Classifier	366
9.5	Discriminative Classification	367
9.6	Support Vector Machines	370
9.7	Decision Trees	373
9.8	Deep Learning and Neural Networks	381
9.9	Evaluating Classifiers: ROC Curves	391
9.10	Which Classifier Should I Use?	393
	References	397
10	Time Series Analysis	399
10.1	Main Concepts for Time Series Analysis	400
10.2	Modeling Toolkit for Time Series Analysis	401
10.3	Analysis of Periodic Time Series	420
10.4	Temporally Localized Signals	447
10.5	Analysis of Stochastic Processes	449
10.6	Which Method Should I Use for Time Series Analysis?	459
	References	460

IV Appendices

A	An Introduction to Scientific Computing with Python	467
A.1	A Brief History of Python	467
A.2	The SciPy Universe	468
A.3	Getting Started with Python	470
A.4	IPython: The Basics of Interactive Computing	482
A.5	Introduction to NumPy	484
A.6	Visualization with Matplotlib	489
A.7	Overview of Useful NumPy/SciPy Modules	492
A.8	Efficient Coding with Python and NumPy	497
A.9	Wrapping Existing code in Python	501
A.10	Other Resources	502
B	AstroML: Machine Learning for Astronomy	505
B.1	Introduction	505
B.2	Dependencies	505
B.3	Tools Included in AstroML v1.0	506
B.4	Open Source Deep Learning Libraries	507
C	Astronomical Flux Measurements and Magnitudes	509
C.1	The Definition of the Specific Flux	509
C.2	Wavelength Window Function for Astronomical Measurements	509
C.3	The Astronomical Magnitude Systems	510
D	SQL Query for Downloading SDSS Data	513
E	Approximating the Fourier Transform with the FFT	515
	References	518
	<i>Visual Figure Index</i>	521
	<i>Index</i>	529