

**Statistics, Data Mining and Machine Learning in Astronomy:
A Practical Python Guide for the Analysis of Survey Data**

ŽELJKO IVEZIĆ, ANDREW J. CONNOLLY, JACOB T. VANDERPLAS
UNIVERSITY OF WASHINGTON

AND ALEX GRAY
GEORGIA INSTITUTE OF TECHNOLOGY

Contents

Contents	5
Preface	7
I Introduction	9
1 About the Book and Supporting Material	11
1.1 What do data mining, machine learning and knowledge discovery mean?	11
1.2 What is this book about?	13
1.3 An incomplete survey of the relevant literature	16
1.4 Introduction to the Python language and the Git code management tool	20
1.5 Description of surveys and data sets used in examples	21
1.6 Plotting and visualizing the data in this book	38
1.7 How to efficiently use this book	48
References	49
2 Fast Computation on Massive Data Sets	51
2.1 Data types and data management systems	51
2.2 Analysis of algorithmic efficiency	52
2.3 Seven types of computational problems	54
2.4 Seven strategies for speeding things up	55
2.5 Case Studies: Speedup Strategies in Practice	58
References	72
II Statistical Frameworks and Exploratory Data Analysis	75
3 Probability and Statistical Distributions	77
3.1 Brief overview of probability and random variables	77
3.2 Descriptive statistics	85
3.3 Common univariate distribution functions	91
3.4 The central limit theorem	108
3.5 Bivariate and multivariate distribution functions	111
3.6 Correlation coefficients	120
3.7 Random number generation for arbitrary distributions	124
References	126
4 Classical Statistical Inference	129

4.1	Classical versus Bayesian statistical inference	129
4.2	Maximum likelihood estimation (MLE)	130
4.3	The goodness-of-fit and model selection	137
4.4	ML applied to Gaussian mixtures: the Expectation Maximization algorithm	140
4.5	Confidence estimates: the bootstrap and jackknife	146
4.6	Hypothesis testing	150
4.7	Comparison of distributions	156
4.8	Non-parametric modeling and histograms	168
4.9	Selection effects and luminosity function estimation	171
4.10	Summary	177
	References	177
5	Bayesian Statistical Inference	179
5.1	Introduction to the Bayesian Method	179
5.2	Bayesian priors	184
5.3	Bayesian parameter uncertainty quantification	188
5.4	Bayesian model selection	189
5.5	Non-uniform priors: Eddington, Malmquist and Lutz-Kelker biases	194
5.6	Simple examples of Bayesian analysis: parameter estimation	198
5.7	Simple examples of Bayesian analysis: model selection	229
5.8	Numerical methods for complex problems (MCMC)	236
5.9	Summary of pros and cons for classical and Bayesian methods	246
	References	250
III	Data Mining and Machine Learning	253
6	Searching for Structure in Point Data	255
6.1	Non-parametric density estimation	256
6.2	Nearest neighbor density estimation	263
6.3	Parametric density estimation	265
6.4	Finding clusters in data	274
6.5	Correlation functions	283
6.6	Which density estimation and clustering algorithms should I use?	287
	References	291
7	Dimensionality and its Reduction	293
7.1	The curse of dimensionality	293
7.2	The data sets used in this chapter	295
7.3	Principal Component Analysis	296
7.4	Non-negative Matrix Factorization	308
7.5	Manifold Learning	311
7.6	Independent Component Analysis and Projection Pursuit	318
7.7	Which dimensionality reduction technique should I use?	320
	References	322

8	Regression and Model Fitting	325
8.1	Formulation of the regression problem	325
8.2	Regression for linear models	329
8.3	Regularization and penalizing the likelihood	335
8.4	Principal component regression	341
8.5	Kernel regression	342
8.6	Locally linear regression.	343
8.7	Non-Linear Regression	344
8.8	Uncertainties in the data	346
8.9	Regression that is robust to outliers	348
8.10	Gaussian Process Regression	353
8.11	Overfitting, underfitting, and cross-validation	357
8.12	Which regression method should I use?	366
	References	368
9	Classification	371
9.1	Data Sets used in this Chapter	371
9.2	Assigning Categories: Classification	372
9.3	Generative Classification	374
9.4	K -Nearest-Neighbor Classifier	384
9.5	Discriminative Classification	386
9.6	Support Vector Machines	387
9.7	Decision Trees	391
9.8	Evaluating Classifiers: ROC Curves	400
9.9	Which classifier should I use?	403
	References	406
10	Time Series Analysis	407
10.1	Main concepts for time series analysis	408
10.2	Modeling toolkit for time series analysis	409
10.3	Analysis of periodic time series	430
10.4	Temporally localized signals	456
10.5	Analysis of stochastic processes	462
10.6	Which method should I use for time series analysis?	469
	References	470
IV	Appendices	473
A	An Introduction to Scientific Computing with Python	475
A.1	A Brief History of Python	475
A.2	The Scipy Universe	476
A.3	Getting Started with Python	478
A.4	IPython: basics of interactive computing	489
A.5	Introduction to Numpy	491
A.6	Visualization with Matplotlib	496

A.7	Overview of Useful NumPy/SciPy Modules	499
A.8	Efficient coding with Python and NumPy	505
A.9	Wrapping existing code in Python	508
A.10	Other Resources	509
B	AstroML: Machine Learning for Astronomy	513
B.1	Introduction	513
B.2	Dependencies	513
B.3	Tools Included in AstroML v0.1	514
C	Astronomical flux measurements and magnitudes	517
C.1	The definition of the specific flux	517
C.2	Wavelength window function for astronomical measurements	517
C.3	The astronomical magnitude systems	518
D	SQL query for downloading SDSS data	521
E	Approximating the Fourier Transform with the FFT	523
	References	526

Preface

Astronomy and astrophysics are witnessing dramatic increases in data volume as detectors, telescopes and computers become ever more powerful. During the last decade, sky surveys across the electromagnetic spectrum have collected hundreds of terabytes of astronomical data for hundreds of millions of sources. Over the next decade, the data volume will enter the petabyte domain, and provide accurate measurements for billions of sources. Astronomy and physics students are not traditionally trained to handle such voluminous and complex data sets. Furthermore, standard analysis methods employed in astronomy often lag far behind rapid progress in statistics and computer science. The main purpose of this book is to help minimize the time it takes a student to become an effective researcher.

This book provides the interface between astronomical data analysis problems and modern statistical methods. It is aimed at physical and data-centric scientists who have an understanding of the science drivers for analyzing large data sets but may not be aware of developments in statistical techniques over the last decade. The book targets researchers who want to use existing methods for analysis of large data sets, rather than those interested in the development of new methods. Theoretical discussions are limited to the minimum required to understand the algorithms. Nevertheless, extensive and detailed references to relevant specialist literature are interspersed throughout the book.

We present an example-driven compendium of modern statistical and data mining methods, together with carefully chosen examples based on real modern data sets, and of current astronomical applications that will illustrate each method introduced in the book. The book is loosely organized by practical analysis problems, and offers a comparative analysis of different techniques, including discussions of the advantages and shortcomings of each method, and their scaling with the sample size. The exposition of the material is supported by appropriate publicly available Python code (available from the book website, rather than fully printed here) and data to enable a reader to reproduce all the figures and examples, evaluate the techniques, and adapt them to their own field of interest. To some extent, this book is an analog of the well-known Numerical Recipes book, but aimed at the analysis of massive astronomical data sets, with more emphasis on modern tools for data mining and machine learning, and with freely available code.

From the start, we desired to create a book which, in the spirit of reproducible research, would allow readers to easily replicate the analysis behind every example and figure. We believe this feature will make the book uniquely valuable as a practical guide. We chose to implement this using Python, a powerful and flexible programming language that is quickly becoming a standard in astronomy (a number of next-generation large astronomical surveys and projects use Python, e.g., JVLA, ALMA, LSST). The Python code base associated with this book, called AstroML, is maintained as a live web repository (GitHub), and is intended to be a growing collection of well-documented and well-tested tools for astronomical research. Any astronomical researcher who

is currently developing software for analysis of massive survey data is invited and encouraged to contribute their own tools to the code.

The target audience for this text includes senior undergraduate and graduate students in physics and astronomy, as well as researchers using large data sets in a scientific context. Familiarity with calculus and other basic mathematical techniques is assumed, but no extensive prior knowledge in statistics is required (e.g., we assume that readers have heard before of the Gaussian distribution, but not necessarily of the Lorentzian distribution). Though the examples in this book are aimed at researchers in the fields of astronomy and astrophysics, the organization of the book allows for easy mapping of relevant algorithms to problems from other fields. After the first introductory Chapter, data organization and some aspects of fast computation are discussed in Chapter 2, statistical foundations are reviewed in Chapters 3–5 (statistical distributions, maximum likelihood and other classical statistics, and Bayesian methodology), exploratory data analysis is described in Chapters 6 and 7 (Searching for Structure in Point Data; Dimensionality and its Reduction), and data-based prediction methods are described in Chapters 8–10 (Regression and Model Fitting; Classification; Time Series Analysis).

Finally, we are indebted to a number of colleagues whose careful reading and resulting comments significantly improved this book. A summer study group consisting of Bob Abel, Yusra AlSayyad, Lauren Anderson, Vaishali Bhardwaj, James Davenport, Alexander Fry, Bryce Kalmbach, and David Westman identified many rough edges in the manuscript and tested the AstroML code. We thank Alan Weinstein for help and advice with LIGO data, and Carl Carter-Schwendler for motivational and expert discussions about Bayesian statistics. In addition, Tim Axelrod, Andy Becker, Joshua Bloom, Tamás Budavári, David Hogg, Robert Lupton, Chelsea MacLeod, Lovro Palaversa, Fernando Perez, Maria Süveges, Przemek Woźniak, and two anonymous reviewers provided extensive expert comments. Any remaining errors are entirely our own.

We dedicate this book to Cristin, Ian, Nancy, Pamela, Tom, and Vedrana for their support, encouragement, and understanding during the periods of intensive work and absent-mindedness along the way to this finished text.

Authors, Seattle and Atlanta, 2012