

Springer Series in the Data Sciences

Mayer Alvo

Statistical Inference and Machine Learning for Big Data

 Springer

Springer Series in the Data Sciences

Series Editors

David Banks, Duke University, Durham, NC, USA

Jianqing Fan, Department of Financial Engineering, Princeton University, Princeton, NJ, USA

Michael Jordan, University of California, Berkeley, CA, USA

Ravi Kannan, Microsoft Research Labs, Bangalore, India

Yurii Nesterov, CORE, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium

Christopher Re, Department of Computer Science, Stanford University, Stanford, USA

Ryan J. Tibshirani, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

Larry Wasserman, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

Springer Series in the Data Sciences focuses primarily on monographs and graduate level textbooks. The target audience includes students and researchers working in and across the fields of mathematics, theoretical computer science, and statistics. Data Analysis and Interpretation is a broad field encompassing some of the fastest-growing subjects in interdisciplinary statistics, mathematics and computer science. It encompasses a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, including diverse techniques under a variety of names, in different business, science, and social science domains. Springer Series in the Data Sciences addresses the needs of a broad spectrum of scientists and students who are utilizing quantitative methods in their daily research. The series is broad but structured, including topics within all core areas of the data sciences. The breadth of the series reflects the variation of scholarly projects currently underway in the field of machine learning.

Mayer Alvo

Statistical Inference and Machine Learning for Big Data

 Springer

Mayer Alvo
Department of Mathematics and Statistics
University of Ottawa
Ottawa, ON, Canada

ISSN 2365-5674 ISSN 2365-5682 (electronic)
Springer Series in the Data Sciences
ISBN 978-3-031-06783-9 ISBN 978-3-031-06784-6 (eBook)
<https://doi.org/10.1007/978-3-031-06784-6>

Mathematics Subject Classification (2022): 62-00, 62Fxx, 62Gxx, 62Hxx, 62N99, 62M99

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book grew from a desire to provide both students and practitioners a reference text, which describes several modern and useful methods in statistics. In recent years, the subject of statistics has expanded at an exponential rate. The volume of data presently available has imposed new demands on both theory and implementation, stretching the capacity of sophisticated high-speed computers. This book represents an attempt to exhibit some of the mainstream topics in the literature. In Part I, we display various types of data most often encountered in applications and cross-reference some of the tools most often used for the analysis. Readers may use this part to direct their attention to the corresponding tools later described in this book. In Part II, we provide a brief summary of the basics in probability and statistical inference. Readers may wish to refresh their knowledge by quickly perusing the relevant chapters. The important Pearson system of distributions is discussed in preparation for its application to microarray data and to approximations to optimal nonparametric rank tests. Estimation and hypothesis testing are described followed by an exposition of classical Bayesian methods. The subject of multivariate analysis appears in Chap. 4. It includes references and applications to principal component analysis, factor analysis, canonical correlation, linear discriminant analysis, and multidimensional scaling. Nonparametric statistics forms the basis of Chap. 5. Much is included in addition to the usual one and two sample tests of location and scale. Specifically, the unified theory of hypothesis testing presented is then applied to the problems of testing for location and the detection of umbrella alternatives. Also included in this chapter are methods for spatial analysis which are then applied on COVID-19 data. In Chap. 6, we introduce the topic of exponential tilting and its implications to Tweedie's formula. Here, we see the importance of the Pearson system of distributions. In Chap. 7, we discuss the subject of contingency tables. Classical time series methods form the subject of Chap. 8 where we also include methods of state space modeling. Estimating equations are briefly described in Chap. 9 followed by a detailed focus on symbolic data analysis in Chap. 10. The latter topic is a modern addition to the toolbox for statisticians who deal with *big data*.

In our contemporary digitized society, *big data* is highly prized for its commercial and scientific value. One can argue that there are three general stages for processing *big data*: the preparation stage, the analysis stage, and the presentation stage. Statisticians are especially interested in the analysis stage, which has proven to be very challenging to handle using traditional methods that require human involvement. Consider for example the online resource ranking system. With millions of new entries appearing every day, no research group would be able to analyze all these entries, assign ranks, and maintain the online search and recommended system in a timely manner. Heterogeneous *big data* sets collected from multiple sources also present a challenge given the need to determine

Preface

correlation between the elements. The addition of interaction terms to statistical models can improve performance and prediction accuracy. It is however difficult to choose appropriate interaction terms in a multidimensional data set.

The human brain is an excellent data processor. However, it can be enormously challenged for processing *big data*. Artificial neural networks constitute a mathematical model inspired by the general structure of biological neural networks. They have proven to be useful for modeling complex relationships between inputs and outputs or to find patterns in data. They far surpass human or traditional model capabilities in applications such as audio *big data* analysis, visual *big data* analysis, and natural language processing.

Big data and neural networks are natural companions. Neural networks are very good at uncovering complex relations and abstract features from large data sets by combining multiple information sources, processing heterogeneous data, and effectively capturing dynamic change in data. In turn, the availability of extremely large volumes of data provides the necessary training examples to enable an effective training of deep neural networks. Part III focuses on the tools for machine learning. Chapter 11 deals with regression methods at length seen through the lens of machine learning. Neural networks are presented in Chap. 12. We conclude in Part IV with a look at Bayesian computational methods, which emphasize Bayesian Markov Chain Monte Carlo, and at Bayesian nonparametric statistics. The latter has been a popular topic since the 1970s when it was introduced by Ferguson.

This book provides both graduate and advanced undergraduate students with a reasonably detailed presentation of a variety of current subjects in statistics. It may also serve as a reference text for practitioners. Recommended prerequisites are two one-semester courses, one in probability and one in statistical inference at the third-year undergraduate level. In writing this book, I have benefited substantially from the assistance provided by my present postdoctoral fellow and former graduate student, Dr. Oleksii Volkov. He was the driving force behind the presentation on machine learning, and I am indebted to him for allowing me to share his knowledge of the subject and for his attention to detail in the presentation. I am grateful as well to my graduate MSc students Xiuwen Duan and Jingrui Mu. Xiuwen exploited the use of the Pearson system in connection with Tweedie's formula whereas Jingrui modeled COVID-19 data using Bayesian spatial-temporal methods. My undergraduate students Keyi Liu applied symbolic analysis to analyze data on Parkinson's and heart diseases whereas Josh Larock provided the computing analyses for the applications of multivariate analyses and non-parametric tests. To all four students many thanks for your insights and very hard work. It was a real pleasure sharing these times together.

Ottawa, ON, Canada

Mayer Alvo

Acknowledgments

This book was funded by the Natural Sciences and Engineering Research Council of Canada (grant OGP0009068), for which I am grateful.

I would also like to thank my wife Helen for her patience and support throughout the writing of this book. My children Anita and Daniel (Jennifer) as well as my grandchildren, Alyssa, Ryan, Maya, and Jacob often inquired about my progress and provided the additional motivation to complete the work. All in all, it was a labor of love rooted in a continuing admiration for the subjects of mathematics and statistics.

Contents

I. Introduction to Big Data	1
1. Examples of <i>Big Data</i>	5
1.1. Multivariate Data	5
1.2. Categorical Data	8
1.3. Environmental Data	10
1.4. Genetic Data	10
1.5. Time Series Data	11
1.6. Ranking Data	11
1.7. Social Network Data	12
1.8. Symbolic Data	13
1.9. Image Data	13
II. Statistical Inference for Big Data	15
2. Basic Concepts in Probability	17
2.1. Pearson System of Distributions	21
2.2. Modes of Convergence	27
2.3. Multivariate Central Limit Theorem	33
2.4. Markov Chains	34
3. Basic Concepts in Statistics	37
3.1. Parametric Estimation	37
3.2. Hypothesis Testing	46
3.3. Classical Bayesian Statistics	57
4. Multivariate Methods	63
4.1. Matrix Algebra	63
4.2. Multivariate Analysis as a Generalization of Univariate Analysis	64
4.2.1. The General Linear Model	67
4.2.2. One Sample Problem	68
4.2.3. Two-Sample Problem	69
4.3. Structure in Multivariate Data Analysis	71
4.3.1. Principal Component Analysis	71

4.3.2.	Factor Analysis	74
4.3.3.	Canonical Correlation	76
4.3.4.	Linear Discriminant Analysis	79
4.3.5.	Multidimensional Scaling	80
4.3.6.	Copula Methods	87
5.	Nonparametric Statistics	95
5.1.	Goodness-of-Fit Tests	96
5.2.	Linear Rank Statistics	98
5.3.	U Statistics	112
5.4.	Hoeffding's Combinatorial Central Limit Theorem	114
5.5.	Nonparametric Tests	116
5.5.1.	One-Sample Tests of Location	116
5.5.2.	Confidence Interval for the Median	119
5.5.3.	Wilcoxon Signed Rank Test	120
5.6.	Multi-Sample Tests	123
5.6.1.	Two-Sample Tests for Location	124
5.6.2.	Multi-Sample Test for Location	125
5.6.3.	Tests for Dispersion	126
5.7.	Compatibility	127
5.8.	Tests for Ordered Alternatives	128
5.9.	A Unified Theory of Hypothesis Testing	132
5.9.1.	Umbrella Alternatives	132
5.9.2.	Tests for Trend in Proportions	136
5.10.	Randomized Block Designs	142
5.11.	Density Estimation	144
5.11.1.	Univariate Kernel Density Estimation	145
5.11.2.	The Rank Transform	149
5.11.3.	Multivariate Kernel Density Estimation	149
5.12.	Spatial Data Analysis	154
5.12.1.	Spatial Prediction	156
5.12.2.	Point Poisson Kriging of Areal Data	160
5.13.	Efficiency	162
5.13.1.	Pitman Efficiency	162
5.13.2.	Application of Le Cam's Lemmas	168
5.14.	Permutation Methods	169
6.	Exponential Tilting and Its Applications	171
6.1.	Neyman Smooth Tests	171
6.2.	Smooth Models for Discrete Distributions	175
6.3.	Rejection Sampling	179
6.4.	Tweedie's Formula: Univariate Case	184
6.5.	Tweedie's Formula: Multivariate Case	188
6.6.	The Saddlepoint Approximation and Notions of Information	189

7.	Counting Data Analysis	195
7.1.	Inference for Generalized Linear Models	198
7.2.	Inference for Contingency Tables	200
7.3.	Two-Way Ordered Classifications	204
7.4.	Survival Analysis	209
7.4.1.	Kaplan-Meier Estimator	211
7.4.2.	Modeling Survival Data	214
8.	Time Series Methods	215
8.1.	Classical Methods of Analysis	215
8.2.	State Space Modeling	224
9.	Estimating Equations	229
9.1.	Composite Likelihood	234
9.2.	Empirical Likelihood	236
9.2.1.	Application to One-Sample Ranking Problems	239
9.2.2.	Application to Two-Sample Ranking Problems	243
10.	Symbolic Data Analysis	247
10.1.	Introduction	247
10.2.	Some Examples	247
10.3.	Interval Data	248
10.3.1.	Frequency	248
10.3.2.	Sample Mean and Sample Variance	251
10.3.3.	Realization In SODAS	253
10.4.	Multi-nominal Data	253
10.4.1.	Frequency	253
10.5.	Symbolic Regression	256
10.5.1.	Symbolic Regression for Interval Data	256
10.5.2.	Symbolic Regression for Modal Data	257
10.5.3.	Symbolic Regression in SODAS	257
10.6.	Cluster Analysis	258
10.7.	Factor Analysis	259
10.8.	Factorial Discriminant Analysis	260
10.9.	Application to Parkinson’s Disease	260
10.9.1.	Data Processing	261
10.9.2.	Result Analysis	262
10.9.2.1.	Viewer	262
10.9.2.2.	Descriptive Statistics	262
10.9.2.3.	Symbolic Regression Analysis	263
10.9.2.4.	Symbolic Clustering	263
10.9.2.5.	Principal Component Analysis	264
10.9.3.	Comparison with Classical Method	267

10.10.	Application to Cardiovascular Disease Analysis	267
10.10.1.	Results of the Analysis	269
10.10.2.	Comparison with the Classical Method	273
III. Machine Learning for Big Data		275
11.	Tools for Machine Learning	277
11.1.	Regression Models	277
11.2.	Simple Linear Regression	279
11.2.1.	Least Squares Method	280
11.2.2.	Statistical Inference on Regression Coefficients	282
11.2.3.	Verifying the Assumptions on the Error Terms	284
11.3.	Multiple Linear Regression	289
11.3.1.	Multiple Linear Regression Model	289
11.3.2.	Normal Equations	290
11.3.3.	Statistical Inference on Regression Coefficients	291
11.3.4.	Model Fit Evaluation	292
11.4.	Regression in Machine Learning	296
11.4.1.	Optimization for Linear Regression in Machine Learning	298
11.4.1.1.	Gradient Descent	300
11.4.1.2.	Feature Standardization	301
11.4.1.3.	Computing Cost on a Test Set	303
11.5.	Classification Models	306
11.5.1.	Logistic Regression	307
11.5.1.1.	Optimization with Maximal Likelihood for Logistic Regression	308
11.5.1.2.	Statistical Inference	310
11.5.2.	Logistic Regression for Binary Classification	311
11.5.2.1.	Kullback-Leibler Divergence	312
11.5.3.	Logistic Regression with Multiple Response Classes	316
11.5.4.	Regularization for Regression Models in Machine Learning	317
11.5.4.1.	Ridge Regression	319
11.5.4.2.	Lasso Regression	320
11.5.4.3.	The Choice of Regularization Method	321
11.5.5.	Support Vector Machines (SVM)	321
11.5.5.1.	Introduction	321
11.5.5.2.	Finding the Optimal Hyperplane	322
11.5.5.3.	SVM for Nonlinearly Separable Data Sets	325
11.5.5.4.	Illustrating SVM	325
12.	Neural Networks	329
12.1.	Feed-Forward Networks	329
12.1.1.	Motivation	330
12.1.2.	Introduction to Neural Networks	333

12.1.3.	Building a Deep Feed-Forward Network	334
12.1.4.	Learning in Deep Networks	340
12.1.4.1.	Quantitative Model	341
12.1.4.2.	Binary Classification Model	342
12.1.5.	Generalization	342
12.1.5.1.	A Machine Learning Approach to Generalization	345
12.2.	Recurrent Neural Networks	350
12.2.1.	Building a Recurrent Neural Network	350
12.2.2.	Learning in Recurrent Networks	352
12.2.3.	Most Common Design Structures of RNNs	354
12.2.4.	Deep RNN	357
12.2.5.	Bidirectional RNN	359
12.2.6.	Long-Term Dependencies and LSTM RNN	361
12.2.7.	Reduction for Exploding Gradients	364
12.3.	Convolution Neural Networks	366
12.3.1.	Convolution Operator for Arrays	368
12.3.1.1.	Properties of the Convolution Operator	369
12.3.2.	Convolution Layers	372
12.3.3.	Pooling Layers	375
12.4.	Text Analytics	376
12.4.1.	Introduction	376
12.4.2.	General Architecture	378

IV. Computational Methods for Statistical Inference 383

13. Bayesian Computation Methods 385

13.1.	Data Augmentation Methods	385
13.2.	Metropolis-Hastings Algorithm	387
13.3.	Gibbs Sampling	389
13.4.	EM Algorithm	390
13.4.1.	Application to Ranking	391
13.4.2.	Extension to Several Populations	398
13.5.	Variational Bayesian Methods	400
13.5.1.	Optimization of the Variational Distribution	402
13.6.	Bayesian Nonparametric Methods	404
13.6.1.	Dirichlet Prior	404
13.6.2.	The Poisson-Dirichlet Prior	408
13.6.3.	Simulation of Bayesian Posterior Distributions	408
13.6.4.	Other Applications	410

Index 427