Springer Series in the Data Sciences

Mayer Alvo

Statistical Inference and Machine Learning for Big Data

Springer

Springer Series in the Data Sciences

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

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

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 nonparametric 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