Predictive Statistics

Analysis and Inference beyond Models

All scientific disciplines prize predictive success. Conventional statistical analyses, however, treat prediction as secondary, instead focusing on modeling and hence on estimation, testing, and detailed physical interpretation, tackling these tasks before the predictive adequacy of a model is established. This book outlines a fully predictive approach to statistical problems based on studying predictors; the approach does not require that predictors correspond to a model although this important special case is included in the general approach. Throughout, the point is to examine predictive performance before considering conventional inference. These ideas are traced through five traditional subfields of statistics, helping readers to refocus and adopt a directly predictive outlook. The book also considers prediction via contemporary 'blackbox' techniques and emerging data types and methodologies, where conventional modeling is so difficult that good prediction is the main criterion available for evaluating the performance of a statistical method. Well-documented open-source R code in a Github repository allows readers to replicate examples and apply techniques to other investigations.

BERTRAND S. CLARKE is Chair of the Department of Statistics at the University of Nebraska, Lincoln. His research focuses on predictive statistics and statistical methodology in genomic data. He is a fellow of the American Statistical Association, serves as editor or associate editor for three journals, and has published numerous papers in several statistical fields as well as a book on data mining and machine learning.

JENNIFER L. CLARKE is Professor of Food Science and Technology, Professor of Statistics, and Director of the Quantitative Life Sciences Initiative at the University of Nebraska, Lincoln. Her current interests include statistical methodology for metagenomics and also prediction, statistical computation, and multitype data analysis. She serves on the steering committee of the Midwest Big Data Hub and is Co-Principal Investigator on an award from the NSF focused on data challenges in digital agriculture.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics.
Recent titles include the following:

20. *Random Graph Dynamics*, by Rick Durrett
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort et al.
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mörters and Yuval Peres
31. *Probability (Fourth Edition)*, by Rick Durrett
33. *Stochastic Processes*, by Richard F. Bass
34. *Regression for Categorical Data*, by Gerhard Tutz
35. *Exercises in Probability (Second Edition)*, by Loïc Chaumont and Marc Yor
36. *Statistical Principles for the Design of Experiments*, by R. Mead, S. G. Gilmour and A. Mead
37. *Quantum Stochastics*, by Mou-Hsiung Chang
38. *Nonparametric Estimation under Shape Constraints*, by Piet Groeneboom and Geurt Jongbloed
39. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, by Jianfeng Yao, Shurong Zheng and Zhidong Bai
40. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, by Evarist Giné and Richard Nickl
41. *Confidence, Likelihood, Probability*, by Tore Schweder and Nils Lid Hjort
42. *Probability on Trees and Networks*, by Russell Lyons and Yuval Peres
43. *Random Graphs and Complex Networks (Volume 1)*, by Remco van der Hofstad
44. *Fundamentals of Nonparametric Bayesian Inference*, by Subhashis Ghosal and Aad van der Vaart
45. *Long-Range Dependence and Self-Similarity*, by Vladas Pipiras and Murad S. Taqqu
46. *Predictive Statistics*, by Bertrand S. Clarke and Jennifer L. Clarke

Predictive Statistics

Analysis and Inference beyond Models

Bertrand S. Clarke
University of Nebraska, Lincoln

Jennifer L. Clarke
University of Nebraska, Lincoln

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107028289
DOI: 10.1017/9781139236003

First published 2018

Printed in the United States of America by Sheridan Books, Inc.

A catalogue record for this publication is available from the British Library.

ISBN 978-1-107-02828-9 Hardback

Additional resources for this publication at www.cambridge.org/predictivestatistics

Contents

Expanded Contents

Preface

This book grew out of a nagging dissatisfaction with the various schools of thought in statistics and their increasing disjunction. Each one – frequentist, Bayes, survey sampling, information-theoretic, etc. – has its strengths and weaknesses, and comparisons amongst their different approaches to inference has energized statistical thinking. This dynamic has only grown stronger over the last decade as more challenging data types have become commonplace. Moreover, in contrasting the techniques advocated by the different schools of thought on harder problems, such as working with big data, high-dimensional data, or complex data, the nagging doubts have only become more insistent. Otherwise stated, the less data (or other information) relative to the believed complexity of the data generator that is available, the more the modeling contributes to an analysis and therefore the more the differences in schools of thought, which largely rest on modeling, become apparent.

Concisely, the era of big data, whether high-dimensional, streaming, multitype or otherwise 'big', is forcing us all to rethink statistics and its philosophy. Questions about how to measure the distance between points in high dimensions have to be addressed since that is one version of the curse of dimensionality, likewise, questions of sparsity – when it holds, when it fails, and how to deal with it in either case – and questions of data sets that have information which is not extractable within the traditional formulation of a 'random variable on a measure space' or a valid sample from a well-defined population. In these contexts, this book is a small first step to reorganize some of what we know in order to focus on predictive structure, which is one of the few properties that cuts across all the new, exciting, developments challenging us and our field.

In our field, we have relied too much on our models by not assessing them as extensively as we should. We have not looked enough at their stability. We have not, as a rule, considered a sufficient number of alternative models to be sure that the model we used was reasonable. With few exceptions, we have not done sequential searches over modeling strategies to find a reasonable model, given a certain amount of data, and then modified it in view of getting more data. Also, we have not assessed the robustness of our inferences to our modeling strategies sufficiently. The present authors are as guilty of this as anyone else. In short, we have contented ourselves with the bromide that even if the model is wrong it may be useful, in the hope that if there is a true model (and here we argue that often there isn't) we have found at least a part of it. However, that ain't necessarily so.

This book is an attempt to focus more heavily on the data than the formalism and to focus more heavily on the performance of predictors rather than the fit or physical interpretation of a model or other construct. As a consequence, testing and estimation are given short

shrift. In fact, the general enterprise of inference by modeling, testing, and estimation seems premature until a lot more is known about a data generator than that it is described by a simple model that may be useful even when it's not true. In reality, the situation is usually worse than that for conventional analysis because the inferences are generated from one data set and a pile of assumptions, often dubious. Granted, in the hands of capable statisticians with enough persistence, most schools of thought will yield useful inferences. However, such success reflects the insight and doggedness of the statistician more than the efficacy of the methods. Consequently, it is hoped that one effect of focusing on the data, here via prediction, will be to energize the debate about what the central goals of statistics should be and how to go about achieving them.

An idea that recurs throughout this book is the concept of a problem class based on the relationship between the data generator and a class of predictors. The emphasis is on predictors that do not correspond to a perfect model for the data generator. In particular, cases in which the model used is only an approximation, or in which there is no true model, are frequently considered.

This book is in three parts. Part I outlines a general approach to statistics based on prediction. No claim is made that it is complete, merely that it is an alternative to various established schools of thought and deserves more attention than it has received. It is based on the prequential (predictive sequential) ideas that emerged in the early 1980s from A. P. Dawid and M. West, amongst others. There are four chapters, outlining the importance of prediction, defining a predictive paradigm, explaining why modeling, while sometimes useful, is not as good an approach as prediction, and finally looking at some familiar predictors. The view here is also more general: other schools of thought are incorporated into the predictive approach by regarding them as techniques for generating predictors that may be tested and studied. Thus, other schools of thought are not 'wrong' so much as incomplete: one school's techniques may not yield good predictors as readily as those of another.

Part II is a review of five major fields within statistics (time series, longitudinal data, survival analysis, nonparametrics, and model selection) from the predictive standpoint. The material is not new; the perspective on it is. The point of Part II is to demonstrate the feasibility of the predictive view: that it is a valid way to think about traditional branches of statistics and is computationally feasible. The five specific subfields were chosen because they are quite different from each other, suggesting the wide applicability of a general predictive view. They are also fields where the problems are so complicated that prediction is obviously important.

Part III is brings prediction up to the present. Starting with prediction in more contemporary model classes such as trees, neural nets, kernel methods, and penalized methods, it moves on to a chapter on ensemble methods, including Bayes model averaging, bagging, stacking, boosting and median methods. Even more than in previous chapters, computing is stressed to verify that the perspective advocated here is feasible. The final chapter is intended to bring predictive concepts to branches of statistics that have either recently emerged or recently changed character through, e.g., big data, changes in data collection, or new applications that have made prediction more important. Having dealt with terrestrial matters, the last chapter also indulges in some moon-gazing, speculating on which problems become more interesting when a predictive view is taken.

On the one hand this book does not require much mathematical background; a strong, determined MS student in statistics, mathematics, engineering, computer science, or other highly quantitative field should be able to follow the formal derivations. On the other hand, the book is primarily conceptual and so makes demands on the prospective reader that likely require more sophistication than a typical MS student, even a strong one, would have. Thus, our primary target audience is mid-career PhD students, practicing statisticians, and researchers in statistical fields. The authors sincerely hope that, whether or not these audiences agree with the perspective expressed in this book, they will find this perspective worth their time to understand.

For those interested in examining the R code or data used for the many examples in this text, please visit the catalog page on the Cambridge University Press website:

www.cambridge.org/predictivestatistics

This page includes a link to the github repository containing all relevant R code and data. The repository is structured so that each chapter has a branch. All code is provided under GNU Public License 3.0.

As with every book, there are people who should be thanked. First, all the people who supplied the data sets we used for examples. Second, all the people who, over the past four years put up with us obsessing over this book; we apologize for endlessly bending your ear. Third, Diana Gillooly of Cambridge University Press, with whom we had many conversations about the content, organization, and orientation of this book. Fourth, those colleagues who encouraged us in our folly. (You know who you are!) We forbear from mentioning names for fear they will regret encouraging us.

Finally, we have consistently tried to be engaging and sometimes provocative. Of course, some people will disagree with us and some errors may remain despite our best efforts. We are reminded of the (possibly mythical) story of a French physicist who, when asked about a colleague's work, pondered a few moments and finally responded: 'It's not even wrong.' In the spirit of that witticism, we apologize in advance for any errors that remain, whether technical or philosophical, hoping that they will at least be interesting.

Bertrand S. Clarke
Jennifer L. Clarke