Statistical Learning for Biomedical Data

This book is for anyone who has biomedical data and needs to identify variables that predict an outcome, for two-group outcomes such as tumor/not tumor, survival/death, or response from treatment. Statistical learning machines are ideally suited to these types of prediction problems, especially if the variables being studied may not meet the assumptions of traditional techniques.

Learning machines come from the world of probability and computer science, but are not yet widely used in biomedical research. This introduction brings learning machine techniques to the biomedical world in an accessible way, explaining the underlying principles in nontechnical language and using extensive examples and figures. The authors connect these new methods to familiar techniques by showing how to use the learning machine models to generate smaller, more easily interpretable, traditional models.

Coverage includes single decision trees, multiple-tree techniques such as Random Forests™, neural nets, support vector machines, nearest neighbors, and boosting.

Practical Guides to Biostatistics and Epidemiology

This series of short and practical but authoritative books is for biomedical researchers, clinical investigators, public health researchers, epidemiologists, and non-academic and consulting biostatisticians who work with data from biomedical and epidemiological and genetic studies. Some books explore a modern statistical method and its applications, others may focus on a particular disease or condition and the statistical techniques most commonly used in studying it.

The series is for people who use statistics to answer specific research questions. Books will explain the application of techniques, specifically the use of computational tools, and emphasize the interpretation of results, not the underlying mathematical and statistical theory.

Published in the series
Applied Multilevel Analysis, by **Jos W.R. Twisk**
Secondary Data Sources for Public Health, by **Sarah Boslaugh**
Survival Analysis for Epidemiologic and Medical Research, by **Steve Selvin**

Statistical Learning for Biomedical Data

James D. Malley
National Institutes of Health, Bethesda, MD

Karen G. Malley
Malley Research Programming, Inc., Rockville, MD

Sinisa Pajevic
National Institutes of Health, Bethesda, MD

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521875806

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

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

ISBN 978-0-521-87580-6 Hardback
ISBN 978-0-521-69909-9 Paperback

Contents

The color plate is situated between pages 244 and 245.

Preface

Statistical learning machines live at the triple-point of statistical data analysis, pure mathematics, and computer science. Learning machines form a still rapidly expanding family of technologies and strategies for analyzing an astonishing variety of data. Methods include pattern recognition, classification, and prediction, and the discovery of networks, hidden structure, or buried relationships. This book focuses on the problem of using biomedical data to classify subjects into just two groups. Connections are drawn to other topics that arise naturally in this setting, including how to find the most important predictors in the data, how to validate the results, how to compare different prediction models ("engines"), and how to combine models for better performance than any one model can give. While emphasis is placed on the core ideas and strategies, keeping mathematical gadgets in the background, we provide extensive plain-text translations of recent important mathematical and statistical results. Important learning machine topics that we don't discuss, but which are being studied actively in the research literature, are described in Chapter 13: Summary and conclusions.

Acknowledgments

This book evolved from a short lecture series at the invitation of Andreas Ziegler (University of Lübeck, Germany). Both he and Inke König (also at the University of Lübeck) live at the intersection of loyal friend, insightful analyst, and enabler of sound technical advances. We owe them our profound gratitude. Thank you both, Andreas and Inke.

Many others have also contributed significantly to this project. Collaborators at the National Institutes of Health have been numerous and helpful, over a long period of time. We especially single out: Joan Bailey-Wilson and Larry Brody (both at the National Human Genome Research Institute); Deanna Greenstein and Jens Wendland (both at the National Institute of Mental Health); Jack Yanovski, MD (National Institute for Child Health and Human Development); Michael Ward, MD, and Abhijit Dasgupta (both at the National Institute of Arthritis and Musculoskeletal and Skin Diseases); and John Tisdale (National Institute of Diabetes and Digestive and Kidney Diseases).

Other researchers, whose insight and support we also greatly appreciate, include: Luc Devroye (McGill University, Montreal); Gérard Biau (Université Pierre et Marie Curie, Paris); Carolin Stobl (Ludwig-Maximilians-Universität, München); Adele Cutler (Utah State University, Logan); Nathalie Japkowicz (Université d'Ottawa); Anna Jerebko (Siemens Healthcare, Germany); Jørgen Hilden (University of Copenhagen); and Paul Glasziou (University of Oxford).

We dedicate this book to Leo Breiman (1928–2005).