

## 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.

Cambridge University Press  
978-0-521-87580-6 - Statistical Learning for Biomedical Data  
James D. Malley, Karen G. Malley and Sinisa Pajevic  
Frontmatter  
[More information](#)

---

## Practical Guides to Biostatistics and Epidemiology

### Series advisors

Susan Ellenberg, *University of Pennsylvania School of Medicine*  
Robert C. Elston, *Case Western Reserve University School of Medicine*  
Brian Everitt, *Institute for Psychiatry, King's College London*  
Frank Harrell, *Vanderbilt University Medical Center Tennessee*  
Jos W.R. Twisk, *Vrije Universiteit Medical Centre, Amsterdam*

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**

Cambridge University Press  
978-0-521-87580-6 - Statistical Learning for Biomedical Data  
James D. Malley, Karen G. Malley and Sinisa Pajevic  
Frontmatter  
[More information](#)

---

# 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  
978-0-521-87580-6 - Statistical Learning for Biomedical Data  
James D. Malley, Karen G. Malley and Sinisa Pajevic  
Frontmatter  
[More information](#)

---

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](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9780521875806](http://www.cambridge.org/9780521875806)

© James D. Malley, Karen G. Malley and Sinisa Pajevic 2011  
James D. Malley and Sinisa Pajevic's contributions are works of the  
United States Government and are not protected by copyright in the  
United States.

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

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

Cambridge University Press has no responsibility for the persistence or  
accuracy of URLs for external or third-party internet websites referred to  
in this publication, and does not guarantee that any content on such  
websites is, or will remain, accurate or appropriate.

## Contents

	<i>Preface</i>	<i>page xi</i>
	<i>Acknowledgments</i>	<i>xii</i>
<b>Part I</b>	<b>Introduction</b>	<b>1</b>
1	Prologue	3
	1.1 Machines that learn – some recent history	3
	1.2 Twenty canonical questions	7
	1.3 Outline of the book	9
	1.4 A comment about example datasets	11
	1.5 Software	12
	Note	13
2	The landscape of learning machines	14
	2.1 Introduction	14
	2.2 Types of data for learning machines	15
	2.3 Will that be supervised or unsupervised?	17
	2.4 An unsupervised example	18
	2.5 More lack of supervision – where are the parents?	20
	2.6 Engines, complex and primitive	20
	2.7 Model richness means what, exactly?	22
	2.8 Membership or probability of membership?	25
	2.9 A taxonomy of machines?	27
	2.10 A note of caution – one of many	30
	2.11 Highlights from the theory	30
	Notes	36
3	A mangle of machines	41
	3.1 Introduction	41

<b>vi</b>	<b>Contents</b>	
	3.2 Linear regression	41
	3.3 Logistic regression	42
	3.4 Linear discriminant	43
	3.5 Bayes classifiers – regular and naïve	45
	3.6 Logic regression	47
	3.7 $k$ -Nearest neighbors	48
	3.8 Support vector machines	50
	3.9 Neural networks	53
	3.10 Boosting	54
	3.11 Evolutionary and genetic algorithms	55
	Notes	56
4	Three examples and several machines	57
	4.1 Introduction	57
	4.2 Simulated cholesterol data	58
	4.3 Lupus data	61
	4.4 Stroke data	62
	4.5 Biomedical <i>means</i> unbalanced	63
	4.6 Measures of machine performance	64
	4.7 Linear analysis of cholesterol data	66
	4.8 Nonlinear analysis of cholesterol data	67
	4.9 Analysis of the lupus data	70
	4.10 Analysis of the stroke data	75
	4.11 Further analysis of the lupus and stroke data	79
	Notes	87
<b>Part II</b>	<b>A machine toolkit</b>	<b>89</b>
5	Logistic regression	91
	5.1 Introduction	91
	5.2 Inside and around the model	92
	5.3 Interpreting the coefficients	93
	5.4 Using logistic regression as a decision rule	94
	5.5 Logistic regression applied to the cholesterol data	94
	5.6 A cautionary note	98
	5.7 Another cautionary note	101
	5.8 Probability estimates and decision rules	102

<b>vii</b>	<b>Contents</b>	
	5.9 Evaluating the goodness-of-fit of a logistic regression model	103
	5.10 Calibrating a logistic regression	106
	5.11 Beyond calibration	111
	5.12 Logistic regression and reference models	113
	Notes	115
<b>6</b>	<b>A single decision tree</b>	<b>118</b>
	6.1 Introduction	118
	6.2 Dropping down trees	118
	6.3 Growing a tree	120
	6.4 Selecting features, making splits	120
	6.5 Good split, bad split	121
	6.6 Finding good features for making splits	124
	6.7 Misreading trees	125
	6.8 Stopping and pruning rules	127
	6.9 Using functions of the features	128
	6.10 Unstable trees?	129
	6.11 Variable importance – growing on trees?	132
	6.12 Permuting for importance	134
	6.13 The continuing mystery of trees	135
<b>7</b>	<b>Random Forests – trees everywhere</b>	<b>137</b>
	7.1 Random Forests in less than five minutes	137
	7.2 Random treks through the data	138
	7.3 Random treks through the features	139
	7.4 Walking through the forest	140
	7.5 Weighted and unweighted voting	140
	7.6 Finding subsets in the data using proximities	142
	7.7 Applying Random Forests to the Stroke data	144
	7.8 Random Forests in the universe of machines	151
	Notes	153
<b>Part III</b>	<b>Analysis fundamentals</b>	<b>155</b>
<b>8</b>	<b>Merely two variables</b>	<b>157</b>
	8.1 Introduction	157
	8.2 Understanding correlations	158
	8.3 Hazards of correlations	159

<b>viii</b>	<b>Contents</b>	
	8.4 Correlations big and small	163
	Notes	168
9	More than two variables	<b>171</b>
	9.1 Introduction	171
	9.2 Tiny problems, large consequences	172
	9.3 Mathematics to the rescue?	174
	9.4 Good models need not be unique	176
	9.5 Contexts and coefficients	179
	9.6 Interpreting and testing coefficients in models	181
	9.7 Merging models, pooling lists, ranking features	186
	Notes	190
10	Resampling methods	<b>198</b>
	10.1 Introduction	198
	10.2 The bootstrap	198
	10.3 When the bootstrap works	201
	10.4 When the bootstrap doesn't work	202
	10.5 Resampling from a single group in different ways	203
	10.6 Resampling from groups with unequal sizes	204
	10.7 Resampling from small datasets	206
	10.8 Permutation methods	207
	10.9 Still more on permutation methods	210
	Note	214
11	Error analysis and model validation	<b>215</b>
	11.1 Introduction	215
	11.2 Errors? What errors?	217
	11.3 Unbalanced data, unbalanced errors	218
	11.4 Error analysis for a single machine	219
	11.5 Cross-validation error estimation	222
	11.6 Cross-validation or cross-training?	224
	11.7 The leave-one-out method	226
	11.8 The out-of-bag method	227
	11.9 Intervals for error estimates for a single machine	228
	11.10 Tossing random coins into the abyss	230
	11.11 Error estimates for unbalanced data	232
	11.12 Confidence intervals for comparing error values	233



<b>ix</b>	<b>Contents</b>	
	11.13 Other measures of machine accuracy	236
	11.14 Benchmarking and winning the lottery	238
	11.15 Error analysis for predicting continuous outcomes	239
	Notes	240
<b>Part IV</b>	<b>Machine strategies</b>	<b>245</b>
12	Ensemble methods – let’s take a vote	247
	12.1 Pools of machines	247
	12.2 Weak correlation with outcome can be good enough	247
	12.3 Model averaging	250
	Notes	254
13	Summary and conclusions	255
	13.1 Where have we been?	255
	13.2 So many machines	257
	13.3 Binary decision or probability estimate?	259
	13.4 Survival machines? Risk machines?	259
	13.5 And where are we going?	260
	<i>Appendix</i>	263
	<i>References</i>	271
	<i>Index</i>	281

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).