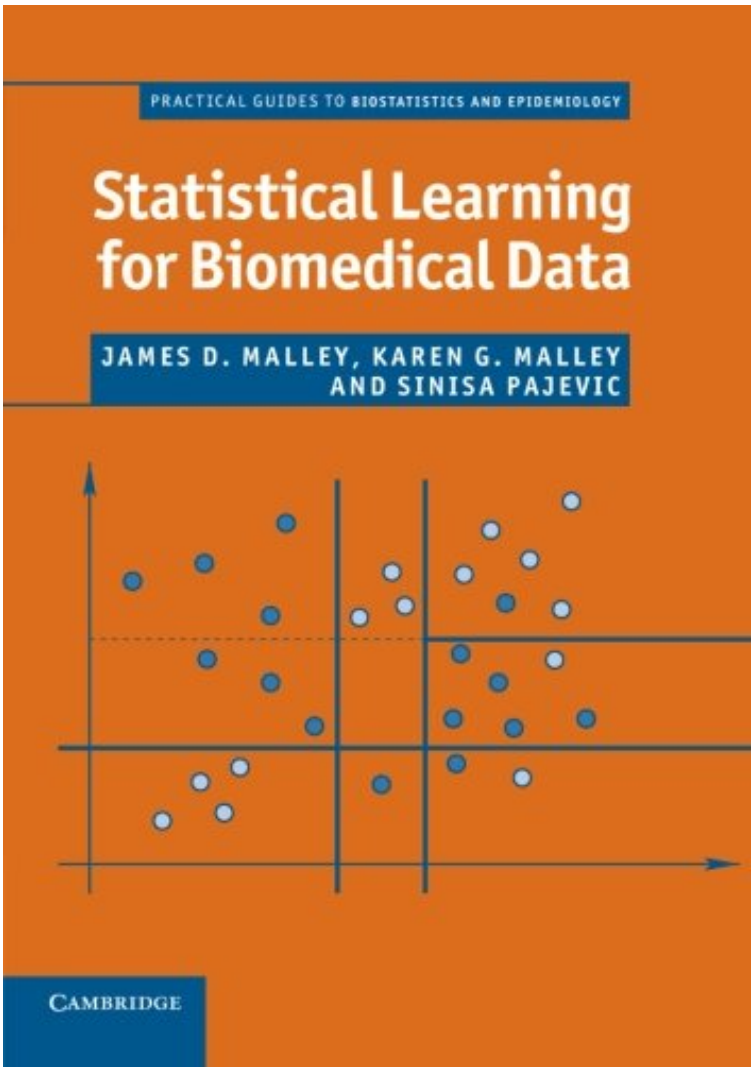
Statistical Learning for Biomedical Data (Practical Guides to Biostatistics and Epidemiology)

James D. Malley, Karen G. Malley, Sinisa Pajevic
*audiobook | *ebooks | Download PDF | ePub | DOC*



#1443515 in Books Cambridge University Press 2011-03-28Original language:EnglishPDF # 1 9.72 x .43 x 6.85l, 1.32 #File Name: 0521699096298 pages | File size: 15.Mb

James D. Malley, Karen G. Malley, Sinisa Pajevic : Statistical Learning for Biomedical Data (Practical Guides to Biostatistics and Epidemiology) before purchasing it in order to gage whether or not it would be worth my time, and all praised Statistical Learning for Biomedical Data (Practical Guides to Biostatistics and Epidemiology):

0 of 0 people found the following review helpful. Nice introductory book ont he topic.By W. YIPNice introductory book.7 of 7 people found the following review helpful. similar to Hastie-Tibshirani-FriedmanBy Michael R. ChernickThis is a book on statistical methods for pattern recognition/machine learning and data mining. The content is very similar to "The Elements of Statistical Learning: Data Mining Inference, and Prediction Second Edition" by

Hastie, Tibshirani and Friedman. It differs by being (1) less technical, (2) less comprehensive and (3) attentive exclusively to biomedical applications. The book is well-written and provides nice graphics and numerous applications. The main problem I find with the book is that in an effort to be less technical, the informal description can at times be incorrect or misleading. As examples consider Chapter 10 on resampling in the sections on the bootstrap. On pages 198-199 they write "The bootstrap method is an instance of data sampling. When applied to error analysis it is a generalization of the cross-validation and the leave-one-out schemes: these schemes are discussed here." But in fact the bootstrap is a different type of resampling method and cross-validation is not a special case of bootstrapping. They go on to say "We propose to reuse the data to effectively enlarge the apparent size of the dataset." This sentence is misleading. While bootstrap sampling takes many samples of size n, the bootstrap samples, that are similar to the original sample and thus it mimics independently taking samples of size n from the population, it does nothing to enlarge the size of the dataset. Although they are careful to say "apparent size", I think it conveys an incorrect message suggesting that the bootstrap adds something to the original data that provides new information. The bootstrap does not do that. Such discourse bothers me because adding information by resampling is one of the common misconceptions that people have about the bootstrap. I prefer to say that resampling attempts to milk out all the information that is avialble from the original sample. I have been faced with clients wanting me to bootstrap because their sample size is small and they believe the bootstrap can turn their small sample into a much larger one. It does not do that and if n is small, say between 2 and 8 the bootstrap doesn't work very well either.Aside from these slips, the book is good for its intended audience, the users of biomedical data. It covers the topics in the Hastie, Tibshirani and Friedman book in a way that is easier for their intended audience to understand while Hastie, Tibshirani and Friedman's book is geared to statisticians or graduate students in statistics. Topics include linear regression, logistic regression, linear dicriminant analysis, decision trees and ensembles of trees (e.g. Brieman's random forests) and learning machines including, logic regression,kth nearest neighbors, support vector machines, neural networks, and genetic algorithms.Some important concepts are explained well, including the difference between supervied and unsupervised learning,and when specific methods work well and when they don't. They also describe twenty canonical questions and point the reader to the sections where these questions are answered. They provide many important examples from biomedical research and illustrate the methods to solve these problems along with the pitfalls of some methods.In discussing resampling methods for classification they point out that error rate estimates such as the bootstrap 632+ method have been shown to work well in small samples where many alternative methods suffer from large biases or high variability.In discussing when the bootstrap works they provide a general description of why and how it works but not specific reasons why it works well in certain instances. On the other hand when describing when it doesn't work they provide a nice example with some intuition as to why estimating extremes of a distribution is difficult when bootstrapping. But they avoid the mathematical explanation which is inconsistency of the estimate which is the essential problem. Although the mathematical details of consistency as a form of convergence as the sample size increases is not necessary, identifying the problem of inconsistency is. Also often when the ordinary bootstrap approach fails, modifications work. In particular, the authors do not mention m-out-of-n bootstrap which is a remedy to inconsistency for extremes as well as other cases where the bootstrap fails.2 of 4 people found the following review helpful. clear, elegant, eminently readableBy Roy RubinfeldEminently readable, elegant, interesting and informative, the style of the book is reminiscent of Lewis Thomas's bestsellers like "Lives of a Cell."This book describes the "tipping point" of where biostatistical analyses are headed using excellent analogies that made clear sense to me (a physician involved in clinical research). Some of the analogies demonstrate a palpable reverence for the way Nature operates and can be studied in creative, innovative ways that will displace some older legacy statistical methods. Buy this book.

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(TM), neural nets, support vector machines, nearest neighbors and boosting.

"Some important concepts are explained well including the difference between supervised and unsupervised learning, and describing when specific methods work well and when they don't. They also list twenty canonical questions and point the reader to the sections in the book where these questions are answered. They provide many important examples from biomedical research and illustrate the methods to solve these problems along with the pitfalls of some

of them. ... Overall, I think this is a good reference source for biomedical researchers involved in data mining or classification, but the reader should beware of the arguments that are loosely explained." Michael R. Chernick, Significance"The book is well written and provides nice graphics and numerous applications... the book is good for its intended audience, the users of biomedical data." Michael R. Chernick, Technometrics"While biomedical applications of the statistical learning machines described in this book are becoming more apparent, they are not widely practiced. This book provides an excellent overview for the neophyte as to the nuts and bolts of certain statistical learning machines and the major issues involved in development and evaluation of specific machines. The authors display a nice dance between exuberance and caution; they do not attempt to advance any particular machine learning approach, instead emphasizing the processes and the need to use several approaches depending on data context." Wendy J. Mack, University of Southern California, Los Angeles for American Journal of EpidemiologyAbout the AuthorJames D. Malley is a Research Mathematical Statistician in the Mathematical and Statistical Computing Laboratory, Division of Computational Bioscience, Center for Information Technology, at the National Institutes of Health.Karen G. Malley is president of Malley Research Programming, Inc. in Rockville, Maryland, providing statistical programming services to the pharmaceutical industry and the National Institutes of Health. She also serves on the global council of the Clinical Data Interchange Standards Consortium (CDISC) user network, and the steering committee of the Washington, DC area CDISC user network.Sinisa Pajevic is a Staff Scientist in the Mathematical and Statistical Computing Laboratory, Division of Computational Bioscience, Center for Information Technology, at the National Institutes of Health.