The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Trevor HASTIE, Robert TIBSHIRANI, and Jerome FRIEDMAN. New York: Springer-Verlag, 2001. ISBN 0-387-95284-5. viii + 533 pp. $74.95 (H).

In the words of the authors, the goal of this book was to "bring together many of the important new ideas in learning, and explain them in a statistical framework." The authors have been quite successful in achieving this objective, and their work is a welcome addition to the statistics and learning literatures. Statistics has always been interdisciplinary, borrowing ideas from diverse fields and repaying the debt with contributions, both theoretical and practical, to the other intellectual disciplines. For statistical learning, this cross-fertilization is especially noticeable. This book is a valuable resource, both for the statistician needing an introduction to machine learning and related fields and for the computer scientist wishing to learn more about statistics. Statisticians will especially appreciate that it is written in their own language.

The level of the book is roughly that of a second-year doctoral student in statistics, and it will be useful as a textbook for such students. In a stimulating article, Breiman (2001) argued that statistics has been focused too much on a "data modeling culture," where the model is paramount. Breiman argued instead for an "algorithmic modeling culture," with emphasis on black-box types of prediction. Breiman's article is controversial, and in his discussion, Efron objects that "prediction is certainly an interesting subject, but Leo's paper overstates both its role and our profession's lack of interest in it." Although I mostly agree with Efron, I worry that the courses offered by most statistics departments include little, if any, treatment of statistical learning and prediction. (Stanford, where Efron and the authors of this book teach, is an exception.) Graduate students in statistics certainly need to know more than they do now about prediction, machine learning, statistical learning, and data mining (not disjoint subjects). I hope that graduate courses covering the topics of this book will become more common in statistics curricula.

Most of the book is focused on supervised learning, where one has inputs and outputs from some system and wishes to predict unknown outputs corresponding to known inputs. The methods discussed for supervised learning include linear and logistic regression; basis expansion, such as splines and wavelets; kernel techniques, such as local regression, local likelihood, and radial basis functions; neural networks; additive models; decision trees based on recursive partitioning, such as CART; and support vector machines.

There is a final chapter on unsupervised learning, including association rules, cluster analysis, self-organizing maps, principal components and curves, and independent component analysis. Many statisticians will be unfamiliar with at least some of these algorithms. Association rules are popular for mining commercial data in what is called "market basket analysis." The aim is to discover types of products often purchased together. Such knowledge can be used to develop marketing strategies, such as store or catalog layouts. Self-organizing maps (SOMs) involve essentially constrained k-means clustering, where prototypes are mapped to a two-dimensional curved coordinate system. Independent components analysis is similar to principal components analysis and factor analysis, but it uses higher-order moments to achieve independence, not merely zero correlation between components.

A strength of the book is the attempt to organize a plethora of methods into a coherent whole. The relationships among the methods are emphasized. I know of no other book that covers so much ground. Of course, with such broad coverage, it is not possible to cover any single topic in great depth, so this book will encourage further reading. Fortunately, each chapter includes bibliographic notes surveying the recent literature. These notes and the extensive references provide a good introduction to the learning literature, including much outside of statistics. The book might be more suitable as a textbook if less material were covered in greater depth; however, such a change would compromise the book's usefulness as a reference, and so I am happier with the book as it was written.

David RUPPERT
Cornell University

REFERENCE

Breiman, L. (2001), "Statistical Modeling: The Two Cultures" (with discussion), *Statistical Science*, 16, 199–231.

Statistical Process Adjustment Methods for Quality Control.

Enrique DEL CASTILLO. New York: Wiley, 2002. ISBN 0-471-43574-0. xviii + 357 pp. $99.95 (H).

This book addresses the core issues of integration between statistical process control (SPC) and engineering process control (EPC). Traditionally, SPC techniques have been developed to monitor variables and, through a (usually off-line) cycle of diagnosing and correcting special causes, to reduce process variability. In contrast, EPC techniques have been developed to directly reduce process variability by adjusting or controlling input variables based on each (usually real-time) observation of the output variables. The area of SPC–EPC integration certainly owes its development to George E. P. Box and his collaborators. Box and Luceño (1997) concluded that:

To augment the monitoring aspects of statistical process control with appropriate techniques for process adjustment has long been an evident need. Some 35 years ago, in response to a paper that attempted such enhancement a discussant [Prof. J. H. Westcott in the discussion of "Some Statistical Aspects of Adaptive Optimization and Control" by Box and Jenkins, 1962] remarked, "I welcome this flirtation between control engineering and statistics. I doubt, however, whether they can yet be said to be going steady."

Box and Luceño went on to suggest that their book brings about the desired marriage, but I do not think that is completely true. With all due respect to their contribution, I think that perhaps we can celebrate an engagement, but there is a lot of room for improving the bridge between control and monitoring. I think that the best approach is to focus on the industrial statistics audience to foster an appreciation of control engineering (as opposed to focusing on control engineers to develop their statistical appreciation). With that in mind, I believe that this book goes a long way toward achieving this end. He states that the objective of his book is to "present process adjustment techniques based on EPC methods and to discuss them from the point of view of controlling the quality of a product." This product quality focus is a good point of connection. The book goes on to truly synthesize several sources across time series, statistics, and control theory, with a clear focus on quality control outcomes.

The book's organization makes a natural progression from process monitoring basics (Chap. 1), to stochastic-dynamic process modeling (Chaps. 2–4), to process control techniques (Chaps. 5–9). In the first chapter, Figure 1.23 provides a very nice flowchart guide to the use of the EPC and SPC techniques discussed in the book. SAS procedures that support aspects of the modeling and analysis are discussed in sufficient detail within the text. There are several examples of SAS code and output. The graphical user interface of MATLAB's system ID toolbox is presented and discussed briefly. Minitab's STAT functions are also frequently used for data analysis and plotting.

Although the author suggests that the book could be used in an undergraduate course, I think its level demands a certain amount of statistical and mathematical sophistication that would be beyond all but the very top undergraduate students. However, first-year or second-year graduate students would be very well prepared in the area by using this text. As a text for course instruction, this book certainly excels on the basis of exercises and real datasets. There are about 15 problems at the end of each chapter (a solutions manual was prepared by Rong Pan) and 18 data files and spreadsheets that serve to illuminate topics in each chapter. (In comparison, the Box and Luceño book has only one or two problems in most chapters and only three datasets.) The author's website, *www.ie.psu.edu/faculty/castillo/castillo.htm*, contains the electronic files, solutions manual, and errata in the first printing.

One of the past criticisms of the quality area is the perspective that quality is free, or that quality objectives should be pursued for purely intrinsic reasons. Six Sigma, of course, has sought to work against this misconception with a focus on bottom-line profitability of quality improvement. This book, with its strong focus on controlling the quality of products and processes, underscores the high relevance of quality control to industrial practice. To further this idea, I would have liked to see some strategic-level consideration of how statistical process adjustment may factor into a company's financial strength by creating opportunities that it may not have otherwise had.

Overall, I think that this is a great book that is well worth its price. Most of the text focuses on univariate system analysis, and it will help the reader appreciate the fundamentals. The final chapter gives a brief introduction to multivariate system analysis and suggests other avenues of future research in the SPC–EPC area. For those working in the manufacturing area—from either an academic or an industry point of view—this book is a valuable resource.

Harriet Black NEMBHARD
University of Wisconsin, Madison