

## REFERENCES

- [1] J. A. GALLIAN, *Contemporary Abstract Algebra*, 5th ed., Houghton Mifflin College, 2001.
- [2] G. L. MULLEN, <http://www.math.psu.edu/mullen/>, February 20, 2002.

ROBERT A. BEEZER  
*University of Puget Sound*

**The Elements of Statistical Learning: Data Mining, Inference and Prediction.** By *T. Hastie, R. Tibshirani, and J. Friedman*. Springer-Verlag, New York, 2001. \$74.95. xvi+533 pp., hardcover. ISBN 0-387-95284-5.

Data mining is a field developed by computer scientists, but many of its crucial elements are imbedded in very important and subtle statistical concepts. Statisticians can play a very important role in the development of this field, but as was the case with artificial intelligence, expert systems, fuzzy set theory, genetic algorithms, and neural networks, the statistical research community has been slow to respond and by and large has stayed on the sidelines as spectators. I believe that this important book by Hastie, Tibshirani, and Friedman will change that.

Pattern recognition and cluster analysis are traditionally major subject areas in engineering and multivariate statistics. The terminology that statisticians use is generally different from that of the engineer, and the computer scientist often has a third set of terms. It is pattern recognition under the name discriminant analysis or classification that was at the heart of machine learning, artificial intelligence, and expert systems that were hot topics in the 1980s and 1990s, but often the user of these algorithms never knew the statistical basis for the techniques. Advances in neural networks and the so-called boosting methods also have statistical ideas behind their success.

The great thing about this text is that it provides this basis from the viewpoint of three statisticians (actually Jerry Friedman was a physicist first but got a lot of good sta-

tistical training from his collaboration with John Tukey over the years). Friedman's genius, exposure to Tukey, and years of studying high-dimensional data from linear accelerators have made him a first-rate statistician, and he has developed many innovations that are covered in this text including classification and regression trees, multivariate adaptive regression splines, and regularized discriminant analysis. Although he did not invent boosting, much of his recent research puts the method in such a context that its success is almost intuitive. His co-authors, though younger, have also made their mark in the field and each has written books before. Friedman wrote [1] with Breiman, Olshen, and Stone; Hastie and Tibshirani wrote [9] just after graduating from Stanford, and Tibshirani wrote [6] with Efron. All are considered major texts on important statistical techniques that are very useful in real world applications.

The tools in these three books just cited, tree classifiers, generalized additive models, and bootstrap methods, are also useful in data mining problems. Hence it is natural for these authors to contribute to advances in data mining from the perspective of statisticians using the statistical concepts of learning or gaining information from data. In addition, Tibshirani invented the lasso method and Hastie invented principal curves and surfaces. All these tools are covered in the book.

For computer scientists interested in data mining the authors could be criticized for concentrating on statistical ideas and specialized techniques from their own research while ignoring much of the rest of the literature. But this is partly the intent of the authors. They want to exhibit and emphasize the statistical aspects of data mining to encourage statisticians to move in and use these valuable techniques in a new arena. A good, balanced account of data mining providing both the statistical and computer science perspectives (both important to the practical implementation of data mining techniques) can be found in [8].

The book is very well written and color is used throughout. Color adds a dimension that can be used to help the reader visualize high-dimensional data, and it is also very useful to help the eye see patterns

and clusters more easily. This makes color effective in the book and not just a pleasing gimmick.

This is the first book of its kind to treat data mining from a statistical perspective that is comprehensive and up-to-date on the statistical methods. Among the important statistical methods that are covered in the book are under the category of supervised learning: regression, discriminant analysis, kernel methods, model assessment and selection, bootstrapping, maximum likelihood and Bayesian inference, additive models, classification and regression trees, multivariate adaptive regression splines, boosting, regularization methods, nearest neighbor classification, and neural networks. Similarly, under the category of unsupervised learning, clustering and association are covered. They also cover the latest developments in principal components, principal curves, multidimensional scaling, factor analysis, and projection pursuit. Most of the book is devoted to supervised learning, with one chapter at the end (Chapter 14) on unsupervised learning. In addition to their own work the authors cover the important new developments on support vector machines as developed by Vapnik in [11]. A more detailed treatment of support vector machines and kernel methods can be found in [10].

The authors wrote this book at an intermediate level that can be understood by nonstatisticians who have a good introduction to probability and statistics. Many of the important ideas are explained in an intuitive manner rather than through a very formal set of theorems and proofs. As an author of a book on the bootstrap [3] who has done research in classification error rate estimation (e.g., [4]), I was both pleased and disappointed with their treatment of that topic.

On the one hand, the treatment of the 632 estimator and the 632+ enhancement is clear and intuitive. The presentation is better and clearer than the original research articles. They have done a great job of distilling the important results and ideas. On the other hand, I am a bit disappointed because the historical perspective is lost and a number of simulation studies that came after Efron [5] and before Efron and Tibshirani [7] are overlooked.

Also the  $e_0$  bootstrap estimate studied in detail by Chatterjee and Chatterjee [2] is called the leave-one-out bootstrap. Although this is a more appropriate name for it, it is a different name than what appears in the literature, and they change the notation for it as it appears in the formula for the 632 estimate. It would have been better to at least refer to the old notation for the benefit of readers who might take a look at the related literature.

Aside from this minor complaint I found the book to be both innovative and fresh. It provides an important contribution to data mining and statistical pattern recognition. It should become a classic. The level is between intermediate and advanced. It is especially good for statisticians interested in high-dimensional and high-volume data such as can be found in telephone records, satellite images, and genetic microarrays. It can be used for an advanced special topics course in statistics for graduate students. Hand, Mannila, and Smyth [8] is the only comparable text. That text also gives the reader an appreciation for issues of selecting scalable methodologies. Not all good statistical techniques are scalable and so some may be too slow to be practical even with today's amazingly fast machines.

#### REFERENCES

- [1] L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.
- [2] S. CHATTERJEE AND S. CHATTERJEE, *Estimation of misclassification probabilities by bootstrap methods*, Commun. Statist. Simul. Comput., 12 (1983), pp. 645–656.
- [3] M. R. CHERNICK, *Bootstrap Methods: A Practitioner's Guide*, Wiley, New York, 1999.
- [4] M. R. CHERNICK, V. K. MURTHY, AND C. D. NEALY, *Application of bootstrap and other resampling techniques: evaluation of classifier performance*, Pattern Recogn. Lett., 4 (1985), pp. 167–178.
- [5] B. EFRON, *Estimation of the error rate of a prediction rule: Improvements on cross-validation*, J. Amer. Statist. Assoc., 79 (1983), pp. 316–331.

- [6] B. EFRON AND R. TIBSHIRANI, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
- [7] B. EFRON AND R. TIBSHIRANI, *Improvements on cross-validation: The 632+ bootstrap method*, J. Amer. Statist. Assoc., 92 (1997), pp. 548–560.
- [8] D. HAND, H. MANNILA, AND P. SMYTH, *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001.
- [9] T. J. HASTIE AND R. J. TIBSHIRANI, *Generalized Additive Models*, Chapman and Hall, London, 1990.
- [10] B. SCHOLKOPF AND A. J. SMOLA, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2002.
- [11] V. VAPNIK, *Statistical Learning Theory*, Wiley, New York, 1998.

MICHAEL R. CHERNICK  
*Novo Nordisk Pharmaceuticals Inc.*

**Practical Methods for Optimal Control Using Nonlinear Programming.** By John T. Betts. SIAM, Philadelphia, 2001. \$51.00. x+190 pp., hardcover. ISBN 0-89871-488-5.

Solving optimal control problems via direct transcription into a nonlinear programming (NLP) problem has become an accepted method, perhaps the preferred method, in the last 15 years. The method became known, particularly in this country, primarily through a 1987 paper by Charles Hargraves and Stephen Paris, colleagues of John Betts at Boeing; this paper is frequently cited by users of the method. Hargraves and Paris introduced a simplification of an existing transcription by collocation, making it truly direct, i.e., shed of the use of any Lagrange multiplier variables. Their method, representing system states by cubic polynomials and using implicit integration (by Simpson's rule) of the system equations, proved to be straightforward to implement and robust and has been used to solve many useful problems.

Until this book by John Betts, there hasn't been a reference on the use of this method for solving optimal control problems; the method has still been accessible through a large body of papers by various authors describing their use of related meth-

ods and, in many cases, their innovations and improvements to the method. If one knows where to look and which researchers to follow, this is still a reasonable way to become familiar with the method, but it requires a knowledge of the literature that not everyone will have, particularly if they are not working in an aerospace-related field, since the great majority of these papers are found in aerospace journals. There is thus a real need for such a book.

The book is quite brief, five chapters comprising 176 pages exclusive of the appendix and index. It is devoid of proofs and can be said to be oriented to the practitioner. The first two chapters are an introduction to nonlinear programming and its application to large, sparse systems, which is what result from the use of collocation. This may be more than many users need, as many will employ an NLP solver, such as NPSOL or SNOPT, or E04VCF from the NAG library as a black box. However, I think these two chapters are a valuable introduction to the sequential quadratic programming method used by these routines, and this basic knowledge is particularly useful to have "when things go wrong."

The third chapter discusses several transcriptions that can be used, both collocation and shooting methods. Chapter 4 begins by describing, very briefly, the analytic necessary conditions of the problem, i.e., the Euler-Lagrange equations. However, its purpose in the book is almost exclusively to show why indirect methods are to be avoided, in favor of direct transcription, whenever possible. I think the brevity of this section is unfortunate. In particular, Betts shows that the NLP Lagrange multipliers can be interpreted as discrete approximations to the continuous adjoint variables on the optimal trajectory. But then no significance of this correspondence is described. In fact, it can be very useful; with this information the Euler-Lagrange equations can be integrated backward, determining the optimality of the solution and even ways to improve a suboptimal trajectory, as others have shown.

The remainder of Chapter 4 deals with the subject of mesh refinement, i.e., where mesh points are to be added to improve the accuracy of the solution. I think this