Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, **Springer Series in Statistics, 2001, pp. xvi+533.**

This is a great book. All three authors have track records for clear exposition and are famously gifted for finding intuitive explanations that illuminate technical results. In this case, their work exceeds the high expectations that had been built up by their previous successes.

We have taught a large graduate course (for statisticians and computer scientists) in data mining from this book. In developing this course we spoke to many other faculty members at a range of institutions, and we found no one who did not enjoy reading and teaching from this text. Basically, there is no other book worth considering for such a course.

Nonetheless, there are a few small concerns. The book has beautiful graphics that brilliantly illustrate the points, provided that one is not red-green colorblind (about 12% of males, and a very small percentage of females). This oversight will plague many readers, and it is very unfortunate that a different choice of color schemes was not made early in the process of writing. We hope the second edition will correct this problem.

Second, the book does not directly address the problem of large dimension and small sample size (also known as "large p, small n data" or "short, fat data" from the shape of the $n \times p$ data matrix). Such problems are becoming common, in microarray analysis and many other applications, and although the authors cannot cover every facet of the field, this one deserves a chapter.

There are other omissions, which are much less critical but which it may be useful to point out to students, teachers, and researchers who are poised to buy this book:

- The book discusses reproducing kernel Hilbert spaces in Chapter 5.8, but it does not mention the connection between these and smoothing splines. That topic becomes pertinent in Chapter 12, which treats support vector machines.

- The treatment of Vapnik-Chernovennkis dimension in Chapter 7.9 is cursory, and the wording of a key point (that the dimension d requires the existence of some set of size

d that is shattered, not that any set of size d can be shattered) could be clarified. It would be good to talk more about how this concept is used, and to emphasize that the corresponding performance results are worst-case rather than average-case bounds.

- There are a number of tricks for extending the search that is done in complex model fitting. One strategy uses smart combinatorial techniques and experimental design (cf. Clyde, 1999) to find regions of the model space in which the fit is good. Another strategy uses "racing," a technique invented by Maron and Moore (1997) in which models are compared on a random small fraction of the data, and if a clear winner is found, the inferior model is dropped and a new comparison is made, but if no winner is found, a larger fraction is used. Neither of these methods scales well when p is large, but both are useful and can typically increase search by a factor of about 100, so it would be nice to mention these.

- The random forest approach, developed in technical reports by Leo Breiman and implemented by Adele Cutler, is a very hot area and it would tie nicely into the discussion of multidimensional scaling in Chapter 14.7, which is presently almost a dead-end in the exposition. Perhaps random forests were not obviously important at the time the book was written, but modern researchers will want to bring this material into any course on data mining.

- The success of overcomplete basis sets is a surprise to statisticians. For reasons related to properties of estimators, we have been trained to fit models in which the terms are selected from a basis (e.g., the Fourier series or decimated Haar wavelets). But computer scientists have obtained excellent performance by choosing terms from the union of the Fourier basis and Haar wavelets, and are now boldly exploring uncountable sets of basis functions. A good discussion of this counterintuitive success would be valuable.

- Another surprise for conventional statisticians is that when building classifiers, one can make use of unlableled sample points. This arises in cases where it is expensive to classify an object but cheap to measure the covariates (e.g., in document retrieval

problems it is expensive to determine if a document is relevant to a search query since it must be read by a human, but it is cheap to measure covariates such as counts of keywords). Blum and Mitchell (1998) lay out some of the issues in this area, and it would be good to include some discussion of this in, say, Chapter 4.

- There is significant and relevant material on consistency and the use of the Vapnik-Chernovenkis inequality in Devroye, Györfi, and Lugosi's *A Probabilistic Theory of Pattern Recognition* (another excellent book, though more technical and less intuitive). If a bit of this were inserted, perhaps after Chapter 7, it would add value.

This is a long list, but we must emphasize that these are relatively small points in comparison with the magisterial treatment the book affords to the really big ideas.

In particular, we admire the book for its:

- outstanding use of real data examples to motivate problems and methods;

- unified treatment of flexible inferential procedures in terms of maximization of an objective function subject to a complexity penalty;

- lucid explanation of the amazing performance of the AdaBoost algorithm in improving classification accuracy for almost any rule;

- clear account of support vector machines in terms of traditional statistical paradigms;

- regular introduction of some new insight, such as describing self-organizing maps as constrained k-means clustering.

These treatments are enormously valuable, and the writing never sacrifices understanding for the pursuit of a theoretical technicality.

No modern statistician or computer scientist should be without this book. It is a new high watermark in statistical writing, and the price ($70) is a bargain.

References

Blum, A., and Mitchell, T. (1998). "Combining Labeled and Unlabeled Data with Co-Training," in *Proceedings of the Conference on Computational Learning Theory*, 92–100.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

Clyde, M. (1999). "Bayesian Model Averaging and Model Search Strategies (with discussion)," in *Bayesian Statistics 6*, eds. J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith, Oxford University Press, pp. 157–185.

Maron, O., and Moore, A. (1997). "The Racing Algorithm: Model Selection for Lazy Learners," *Artificial Intelligence Review*, **11**, 193–225.

Institute of Statistics and Decision Sciences   David Banks

Duke University   Feng Liang