Journal of Statistical Software

January 2007, Volume 17, Book Review 5. http://www.jstatsoft.org/

Reviewer: John Maindonald
Australian National University

Pattern Recognition and Machine Learning

Christopher M. Bishop
Springer-Verlag, New York, 2006.
ISBN 0-387-31073-8. xx + 738 pp. USD 74.95.
http://research.microsoft.com/~cmbishop/PRML/

This beautifully produced book is intended for advanced undergraduates, PhD students, and researchers and practitioners, primarily in machine learning or allied areas. The theoretical framework is, as far as possible, that of Bayesian decision theory, taking advantage of the computational tools now available for practical implementation of such methods. Readers should have a good grasp of calculus and linear algebra and, preferably, some prior familiarity with probability theory.

Two examples are used in the first chapter for motivation – recognition of handwritten digits, and polynomial curve fitting. These typify two classes of problem which are the subject of this book. These are: regression with a categorical outcome variable, otherwise known as discriminant analysis or supervised classification; and regression with a continuous or perhaps ordinal outcome variable.

A strong feature is the use of geometric illustration and intuition, noting however that 2- or 3-dimensional analogues are not always effective for higher numbers of dimensions. There is helpful commentary that explains why, e.g., linear models might be useful in one context and neural networks in another. The discussion of Support Vector Machines notes, among other limitations, that the generation of decision values rather than probabilities prevents use of a Bayesion decision theoretic framework.

Chapters 1 and 2 develop Bayesian decision theory, and introduce commonly used families of probability distributions. Chapters 3 and 4 cover linear models for regression and classification. Chapters that follow treat neural networks, kernel methods, sparse kernel machines (including support vector machines), graphical models, mixture models and EM, approximate inference (variational Bayes and expectation propagation), sampling methods (leading into MCMC and Gibbs sampling), latent variables (PCA, factor analysis and extensions), hidden Markov models and linear dynamical systems, and combining models (Bayesian model averaging, committees, boosting and conditional mixture models). The discussion of MCMC and Gibbs sampling makes no direct mention of the practical issues of checking mixing and stopping rules. These have perhaps been left over for the upcoming companion volume, due in 2008, that will address practical issues in the implementation of machine learning methods.

The chapter on sequential data introduces Markov and hidden Markov models (HMMs) as models for dependence such as is commonly found in time series, noting however that "these models are equally applicable to all forms of sequential data, not just time series". Surely, HMMs have proved more applicable to these other forms of sequential data than to standard time series. As the author notes, applications have been in modeling biological sequence data, natural language modeling and handwriting recognition. There is an insightful discussion of the limitations of HMMs, and of extensions that may overcome some of these limitations. The final section of the chapter discusses linear dynamical systems and approaches to state space models that have been explored in the machine learning literature. I did not find any mention of papers from the statistical time series literature, whether on state space or other models for sequential data. There are other dependence structures, including multi-level models, that are subsumed under the discussion (in the chapter on graphical models) of Bayesian hierarchical models.

Statisticians will recognise most of the topics covered, even if not entirely comfortable with models that feature more strongly in the machine learning literature than in the statistical literature. Among the seven datasets used for illustration, only the MNIST handwritten digits dataset seems to me strongly evocative of machine learning rather than statistics.

The preface notes that "no attempt has been made to provide accurate historical attribution of ideas". Nevertheless the book conveys an impression of the history that statisticians will find odd. "Pattern recognition has its origin in engineering", whereas machine learning (including its statistical methodology?) "grew out of computer science". Sidebars describe major historical figures, from Jacob Bernoulli in the 17th century through to Frank Rosenblatt (of perceptron fame) and Claude Shannon (information theorist) in the 20th century. Thomas Bayes finds a place in the pantheon, as does the Joseph Willard Gibbs of Gibbs sampling. Ronald Fisher might seem a notable omission. A large proportion of the references are to the modern statistical literature (the chapter on sequential data is an exception).

A large and legitimate difference from most roughly comparable advanced statistics texts is that there is no discussion of hypothesis testing and confidence intervals. There is much helpful comment on model choice, but little on model diagnostics. Neither "residual(s)" nor "diagnostics" appears in the index. Perhaps model diagnostics will receive attention in the upcoming companion volume.

This is an impressive and interesting book that might form the basis of several advanced statistics courses. It would be a good choice for a reading group. It might with profit be set alongside a text such as Wasserman (2003), exploring the differences in purpose, style, detailed content and emphasis. For style and emphasis however, but not content, Ripley (1996) would be a better match.

Color graphs and diagrams are used liberally to aid insight and understanding. The figures are all available from the book's web page. Each chapter has a large number of carefully graded exercises. most of them testing or extending theory. The book will be supported by an extensive collection of additional material, including lecture slides, solutions to selected exercises (with other exercise solutions available from Springer), datasets, and MATLAB software that implements most of the algorithms that are described. A companion volume that is scheduled for publication in 2008 will deal with practical aspects.

For effective student use, the MATLAB software and the promised companion volume will be crucial. These two volumes, and the MATLAB software, should comprise a powerful set

of learning resources, for statistics as well as for machine learning. I hope that someone will quickly adapt functions from existing R packages, and/or provide whatever additional functions are needed, for use of the R system as an alternative.

References

Ripley, BD (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Wasserman, L (2003). *All of Statistics. A Concise Course in Statistical Inference*. Springer-Verlag.

Reviewer:

John Maindonald
Australian National University
Centre for Mathematics and Its Applications
Canberra, ACT 0200
E-mail: john.maindonald@anu.edu.au
URL: http://www.maths.anu.edu.au/~johnm/