Journal of Statistical Software

February 2009, Volume 29, Book Review 12. http://www.jstatsoft.org/

Reviewer: John Maindonald
Australian National University

Statistical Learning from a Regression Perspective

Richard A. Berk
Springer-Verlag, New York, 2008.
ISBN 978-0-387-77500-5. 358+xviii pp. USD 69.95.
http://www.crim.upenn.edu/faculty/profiles/statistical_learning.html

Statistical learning is a name for methods that experiment with many different models for the data, and which have been a strong focus of data mining and machine learning traditions of data analysis. Here, regression encompasses classification as well as regression with a continuous or ordinal outcome variable. Regression splines and regression smoothers are the point of departure, before moving on to other methodologies – classification and regression trees, bagging, random forests, boosting and support vector machines.

Berk suggests that statistical learning may be the "muscle car" version of exploratory data analysis. Presumably the "muscle" is computing muscle. I would prefer to say that statistical learning is a "muscle car" version of regression methodology that emphasizes exploratory uses.

Equally appropriate titles would be "Regression from a Statistical Learning Perspective", or "Regression as Statistical Learning". Berk comments:

> The stance taken is within the spirit of procedures such as stepwise regression, but beyond allowing the data to determine which predictors are useful, the data are allowed to help determine which predictor functions are most appropriate. (p. 43)

What are the new possibilities? What are the traps? Hype has been a huge problem. New procedures have sometimes been introduced "with the hype of a big-budget movie". As Berk goes on to say:

> . . . none of the techniques has ever lived up to their most optimistic billing. Widespread misuse that has further increased the gap between promised performance and actual performance. (p. xi)

The tone is therefore cautious and critical, "some might even say dark". Would-be prospectors in this territory are warned not to believe every claim that is made, and to evaluate all claims critically. More positively, there are new ideas and insights, and interesting new perspectives on more traditional methods. The issues cut right across all uses of regression methodology, from whatever perspective.

How should performance be measured? Berk comments that:

> In much of the statistical learning literature on which we rely, the true test of a procedure is not how well it fits the data on hand, but how well it forecasts. (p. 41)

This important point is not elaborated. Berk's discussion stays pretty much within a framework that assumes the use of a training and a test sample that are from the "same population". Cross-validation and the bootstrap may be viewed as different forms of elaboration of a split into training and test data. Often, however, it is optimistic to assume that this provides a good measure of performance when results are finally put to use. Almost always, the true target population will be forward in time from the population that generated the sample data, and may be different in space. This difference may or may not matter. (Note that "out of sample" prediction is sometimes used to describe prediction for a target is different to an extent from the population from which the data have been sampled.)

Data miners, and statisticians who have ventured into statistical learning, have generally relied on a theory that assumes independent observations. This is true, for example of Hastie, Tibshirani, and Friedman (2001). Machine learning texts (such as Bishop 2006) have shown some willingness to consider alternatives to independence. Berk has little to say about error distributions *per se*, but does at several points note the importance of the *data generation process*. He notes that for the observational data that are the focus of attention, good approximations to simple random sampling from the relevant population are unusual. Probabilistic data generation processes that might in important special cases improve on the independence assumption, for example hierarchical and/or time series error processes, are not contemplated. This limits the range of problems to which the discussion has direct relevance.

The strength of this book is its extensive discussion of practical issues. Algorithmic details are a starting point for discussing why and how methods work, comparison with other methodologies, limitations and strengths, and so on. Throughout the book, examples are worked through in detail. Each chapter except the first and the last end with a section headed "Software Considerations", followed by "Summary and Conclusions" and data analysis exercises.

The computing for this book was done in R. Detailed code and solutions to exercises are available from the book's web page.

Chapter 1 introduces key ideas. Berk suggests that a data analysis may tell any one of four kinds of story – a causal story, a conditional ($Y \mid X$) story without causal interpretation, a data summary story, or a forecasting story. The data summary can proceed irrespective of the data generation process. But what then? The summary may give a highly biased view of a population that is of interest. Or the issue may be biases in a stochastic process that has generated the data. Either way, it is necessary to ask: "What has been summarized? Is it relevant to questions that are of interest?"

The discussion then moves on to loss functions, linear estimators, degrees of freedom, and model evaluation and selection (R^2, AIC, BIC, cross-validation and generalized cross-validation). There are, finally, brief notes on, among other matters, the bias-variance tradeoff, overfitting and interpretability. Note in passing that the *bias* in "bias-variance tradeoff" arises from modeling the data, not as in the previous paragraph from the sampling or other process that generated the data.

Chapter 2 discusses regression splines and regression smoothers. There is extensive discussion of various alternative shrinkage methods, none without problems, In Berk's account, these are important as examples of the use of penalty methods to smooth the fitted values, an idea that is much more widely applicable. Penalization methods, and a section on inference, occupy most of the rest of the chapter.

Chapter 3 is on classification and regression trees. Chapter 4 discusses bagging, or bootstrap aggregation. Trees are fitted to multiple bootstrap samples, with the classification determined by a majority vote over all the trees. Random forests, the subject of Chapter 5, can often improve on bagging by deriving each of the multiple trees from a bootstrap sample of observations and a random sample of variables. As Berk comments, no other method consistently shows, for classification, better predictive accuracy than random forests. Random forests has the virtue that there is very limited opportunity for tuning, and does not overfit. (It does not overfit, on the assumption that the data can be treated as a random sample from the relevant population.)

Boosting (Chapter 6) makes an interesting contrast with random forests. Boosting works with all the data at once. At each of successive iterations, it seeks to improve prediction for observations that were misclassified at the previous iteration. It derives the final prediction as a suitably weighted average over all iterations. There are many variants, including some that inject a random element, with more appearing all the time. It is not totally clear why boosting commonly performs as well as it does, and does not usually overfit.

Chapter 7, on support vector machines (SVMs), concludes the discussion of different methodologies. SVMs have the property that the only observations that figure directly in the fitting process are those that are near the classification boundary. To get good results, extensive tuning may be necessary, which brings the risk of over-fitting.

A final chapter is devoted to broader implications and craft lore. I especially like the subsection "Matching Your Goals to What You Can Credibly Do".

Regression methods, both the theory and the practice, remain a work in progress, for applications where the form of the model cannot be prescribed *a priori*. Berk has made a valiant attempt at telling a coherent story, albeit one in which any theory assumes independent observations, and in which Bayesian approaches get only passing mention. While there is a great deal more that might be said about the current state of the art, Berk has made a good start in pulling together commentary on issues of major importance.

Parting words sum up a theme that recurs throughout:

> At the very least, demand that all results to be taken seriously rest on test data or their equivalent. And if the results do not make subject matter sense, skepticism is a sensible stance. Ask that each step in the data analysis, including how the data were obtained, be reviewed. If anomalies persist, consider getting an independent third party involved. (p. 341)

Amen and Amen!

References

Bishop CM (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, New York.

Hastie T, Tibshirani R, Friedman J (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York.

Reviewer:

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

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 29, Book Review 12 *Published:* 2009-02-09
February 2009