The Role of Models in Predictive Validation

John Maindonald
Centre for Bioinformation Science, Mathematical Sciences Institute
Australian National University, Canberra ACT 0200, Australia
john.maindonald@anu.edu.au

Introduction

Model choice and validation have a central role in data analysis, including predictive modeling. While standard diagnostics can help identify model inadequacies, it is natural to use predictive accuracy as the decisive criterion in the final choice of predictive model. A key point is that any assessment of predictive accuracy, theoretical or empirical, inevitably assumes a "data mechanism", i.e., a sampling or other stochastic model that relates model predictions to the population that is the target for predictions.

In a controversial paper Breiman [3] presents predictive accuracy as an obvious and natural criterion for model assessment. He criticises a statistical culture that has almost exclusively used models that assume a stochastic "data mechanism" that is thought to describe underlying scientific processes. Breiman argues for the wider use of algorithmic models, such as tree-based regression and neural nets, that "treat the data mechanism as unknown".

Disregard of data mechanisms has limits. Simple approaches to predictive validation assume, in effect, that the data are a random sample from the population to which predictions will be applied. This is often inappropriate. Cox [3, following comment] notes that predictions are often applied "under quite different conditions from the data". Indeed, the conditions may be so different that a realistic assessment of predictive accuracy is impossible!

In what follows I will (1) comment briefly on algorithmic models; (2) note that the interest is, in some contexts, in model parameters; (3) comment in more detail on predictive accuracy; (4) discuss implications for data mining and for the use of data bases.

How do Algorithmic Models Arise?

Possible motivations for algorithmic models include: (1) Normal theory i.i.d. (independent & identical distributions) and other simple models may be extended into contexts where the i.i.d. normal assumptions are clearly wrong; (2) Neural nets were offered as models of the brain's learning processes; (3) Tree-based methods mimic processes used in medical diagnosis, and in botanical or other taxonomic keys. Models that try to identify DNA sequences that code for genes are algorithmic in style; see Durbin et al [6]. They make unrealistic assumptions, e.g., that bases are strung together independently. Nevertheless the best of them can, with suitable tuning and validation, work well enough to be useful.

Models that are algorithmic in style have been popular in the data mining and machine learning communities, which have developed traditions of data analysis that have been largely separate from statistics. Some data miners promote their special skills in working with "large and complex data sets" where, often, major problems in pre-processing, data organization and data manipulation must be solved before any data analysis is possible. Once the "analyst" with the needed computing skills has a foot in the door, what reason is there to withdraw in favour of a statistician who will have to repeat the process of gaining familiarity with the data, and whose ability to adapt to the demands of the problem has still to be tested?

Interaction between different traditions who have different data handling and analysis skills is essential in getting maximum benefit from an evolving computer technology. This applies both in the handling of subtantial data analysis tasks and in training.

Models

Breiman seems uninterested in model parameters. Thus models have perhaps become, in Efron's [3, following comment] words, "black boxes with knobs to twiddle". I agree with Efron that "The whole point of science is to get inside black boxes and see what makes them work as they do". At the same time, the task of attaching meaning to model parameters, though not always impossible, is more difficult and beset with more traps than texts on regression methods commonly acknowledge. Rosenbaum [17] has a balanced and careful discussion.

In an interesting set of data [14, 5] where it has been possible to compare regression estimates for the effect of a labor training program with experimental estimates, a naive analysis gives regression estimates that go in the wrong direction. These data are further analyzed in Maindonald and Braun [16]. Most often, without experimental data to challenge the regression results, the biases in naïve regression parameter estimates will not come to attention.

Predictive accuracy

Where the output from the black box is the focus of interest, predictive accuracy may be the primary consideration. Here, I will draw attention to simple cases where care is required to determinine a sampling model (a "data mechanism") that relates model predictions to the relevant target population, as a basis for assessing predictive accuracy.

For example, a model may use floor area and perhaps socioeconomic status of the suburb to predict house price. The assessing of predictive accuracy for a single suburb requires a sample of houses within the one suburb. For assessing the accuracy of prediction for a new suburb, there must be a sample of suburbs. In principle, for that purpose, one house per suburb is enough! It is the number of suburbs, rather than the number of houses in each, that is likely to limit the accuracy of the assessment. A related and important point, which however I will not further discuss, is that a model that works well for prediction within suburbs may not work well for prediction across suburbs; see for example Cox and Wermuth [4].

A suitable model might, after adjusting for floor area and socioeconomic status, have two components of variance: σ_h^2 associated with variation between houses in a single suburb, and σ_s^2 associated with variation between suburbs. Assessment of predictive accuracy for another house in one of the sampled suburbs, requires a good estimate of σ_h^2, while for a new house in a new suburb the pertinent variance is $\sigma_h^2 + \sigma_s^2$. Estimates of both these variances are available, given suitable data, by the use of the standard forms of empirical methodology: (1) the training/test set methodology, or (2) cross-validation, or (3) bootstrapping. A simple form of either of these empirical methods, which does not allow for a within and a between suburb component of variability, will be wrong.

Of course, the model I have suggested may be wrong, but that is only to emphasize that any assessment of predictive accuracy assumes a stochastic model or data mechanism. This is true both for empirically based and for theoretically based assessments of predictive accuracy. This is an issue even if analyses are regarded as "descriptive". In the search for "interesting" patterns in data, what part of the variation can be treated as "noise" and therefore smoothed away? Is the interest in house price patterns that are local to particular suburbs, at one time? Or is the interest in patterns that persist over time, or over different suburbs?

One of the data mining texts that I note [19] discusses implications of a simple form of grouping structure for cross-validation. The advice is that "each of the classes should be represented in about the right proportion in the training and test sets". This is fine if a pooled within group estimate of predictive accuracy is used and the generalization is to those same groups, but wrong if the aim is to generalize to new groups. I do not know of any statistics text that discusses such issues, from an empirical predictive validation point of view. Note however Hand et al's [11, p.227] insistence that estimates of score functions have a randomness that

comes both from the data used to train the model and from the data used to validate it.

In a study on student attitudes to science in schools in Australian Capital Territory and neighbouring regions of New South Wales, a multi-level model gave a between pupil component of variance equal to 3.05, a between class component of variance of 0.32, and a between school component of variance that was close to zero.[1] Thus, for prediction to a new class of size n, the standard error of the mean is $\sqrt{(}0.32 + 3.05/n)$. Subtlety, and suitable assumptions about the data mechanism, are required to get this result from a cross-validation. Notice that the two components make equal contributions to the variability of a class mean when the class size is about 10. If we had sampled from the original classes, and were predicting to a different sample from one of those same classes, the SEM would be $3.05/n$, i.e., cross-validation would need to be tuned to estimate a pooled within class estimate of standard error.

The Collection of Data into Databases

One reason for interest in algorithmic modeling has been the hope that it can help automate the gleaning of information from large databases. The collation of data that are scattered and fragmented, perhaps across the literature or across different institutions, can be a first step to its use for practical advice and for research, but does not guarantee that the database will yield reliable information! Important issues are access to crucial contextual information, the user choice of levels of detail, and the intended use.

There was a debate, in the early 1980s, between clinical medical researchers who took the view that databases holding largely observational data had an important role in the evaluation of new therapies, and those who wished to place the main reliance on randomized controlled trials. It is now clear that while observational databases can be useful, e.g., in drawing attention to side effects of medical treatment, they are an unreliable and potentially misleading source of evidence for deciding between alternative therapies. [13, 7, 9, 15].

As databases increase in size and complexity, it becomes more important to make data available in various summary forms as an alternative to full detail. In climate data that are recorded at one minute intervals, the detail can easily overwhelm a user for whom coarser information would be adequate. The forms of summary that are immediately available, and the accompanying background information, require tuning to the inferences that will be required. For some uses of the science attitudes data, information at the level of classes will be adequate. Other uses will require individual pupil results, identified by classes.

None of the data mining books that I reference handle this interaction between data structure and inference adequately. Data summary may be seen, as in Han and Kamber [10] as an important issue. Han and Kamber talk of such operations as *roll-up*, *drill-down*, *slice*, *dice* and *pivot* that may be performed on *data cubes*. Other terminology includes *data reduction*, *data cube aggregation*, *dimension reduction*, *data compression*, *numerosity reduction*, and *discretization and concept hierarchy generation*. These, once understood, make good sense in their proper place. The data structures are however too limited; why not the hierarchy of structure implied by the term "pyramid", which might be a natural framework for discussing the hierarchies of variation that are implicit in the house price and science attitudes examples? A further general comment is that it is inherently undesirable, and a barrier to effective communication with other traditions of data analysis, to use new terminology for well-worn ideas.

Berry and Linoff [1] include good practical advice, with minimum jargon, that makes good sense to anyone who works with data. There is a nice discussion of the different types of data (*a hierarchy of data*, p. 363) that enter into the construction of a database – there are *rules* (what has been learned from the data), *metadata*, *database schema*, *summary data*, and *operational data*. Structural hierachy, as implied by different levels of tabular summary, is

[1]The author of the study was Francine Adams, who was an ANU honours student.

another type of hierachy that fits readily within this framework.

Weiss and Indurkhya [18] use jargon cavalierly, to the detriment of effective communication. For these authors (p.2), "A defining characteristic of data mining is 'big data'."

Finally, note three books that have a statistical perspective. Hastie et al [12] and Hand et al [11] are important contributions to dialogue between statisticians, data miners and machine learners. Neither book directly addresses the role of "data mechanism" in predictive validation, though Hand et al give attention to a range of related statistical issues. Bock and Diday [2] emphasizes description more than analysis. It describes frameworks for data description and database storage that could be starting points for addressing predictive validation issues.

Acknowledgement: Discussion with Markus Hegland (MSI, ANU) has helped hone my ideas. His collection of data mining texts was a crucial resource for the writing of this paper.

REFERENCES

1. Berry, M.J. and Linoff, G. 1997. *Data Mining Techniques: for Marketing, Sales and Customer Support.* Wiley.
2. Bock, H.-H. and Diday, E., eds. 2000. *Analysis of Symbolic Data.* Springer-Verlag.
3. Breiman, L. 2001. *Statistical Science* 16:199- 231.
4. Cox, D.R. and Wermuth, N. 1996. *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman and Hall.
5. Dehejia, R.H. and Wahba, S. 1999. *Journal of the American Statistical Association* 94:1053-1062.
6. Durbin, R.S., Eddy, A., Krogh, A. and Mitchison, G. 1998. *Biological Sequence Analysis.*
7. Feinstein, A. 1984. *Statistics in Medicine* 3:341-345.
8. Freedman, D. 1999. *Statistical Science* 14:243-258.
9. Green, S.B. and Byar, D.P. 1984. *Statistics in Medicine* 3:361-370.
10. Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques.* Morgan Kaufmann.
11. Hand, D., Mannila, H. and Smyth, P. 2001. *Principles of Data Mining.* MIT Press.
12. Hastie, T., Tibshirani, R. and Friedman, J. 2001. *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer-Verlag.
13. Jorgensen, M. and Gentleman, R. 1998. *Chance* 11:34-39 & 42
14. Lalonde, R. 1986. *American Economic Review* 76:604-620.
15. Maindonald, J.H. 1998. New approaches to using scientific data – statistics, data mining & related technologies in research & research training. ANU Graduate School Occasional Paper 98/2. Available from http://www.anu.edu.au/graduate/pubs/occasional-papers/GS98_2.pdf
16. Maindonald, J.H. and Braun, W.J., in press. *Data Analysis and Graphics Using R: An Example-Based Approach.* Cambridge University Press.
17. Rosenbaum, P.R. 1999. *Statistical Science* 14:259-278, with following discussion, pp. 279-304.
18. Weiss, M.W. and Indurkhya, N. 1998. *Predictive Data mining.* Morgan Kaufmann.
19. Witten, I.H. and Frank, E. 2000. *Data Mining. Practical machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann.

RÉSUMÉ

Toute évaluation de la précision , théorique ou empirique, d'une prévision nécessite la définition d'un modèle qui relie les données utilisées pour déterminer le modèle de prévision à la population cible sur laquelle on calculera les prévisions. Du fait que l'information contenue dans la base de données est utilisée pour construire et valider le modèle de prévision, ce point concerne aussi les concepteurs de base de données. Il y a des liens très importants entre les concepts de bases de données et les résultats de statistique inférentielle qui apparaissent lorsqu'on utilise des données et qu'il faut enseigner aux gestionnaires de données. Les concepts de modélisation statistique pertinents incluent les effets fixes et aléatoires, la hiérarchie et la structure de corrélation séquentielle.