The Role of Models in Predictive Validation
(Statistics for Budding Data Miners)

John Maindonald

Centre for Bioinformation Science
http://cbis.anu.edu.au

Australian National University

Motivation

- What is data mining? (see, e.g., Ripley & Fei Chen, ISI 2003).
- What should we make of explicit or implicit claims that:
  - "Algorithmic models, plus predictive validation, can be an alternative to more classical data analysis approaches." (c.f. Breiman 2001).
  - "Large & complex data sets need a different theory."
  - "Predictive validation is (always?) the key criterion for model choice." (c.f., again, Breiman 2001).
- Challenges to conventional approaches and training can be useful. Questions that then warrant attention include:
  - Is a statistics major needed/sufficient for data analysts?
  - What are other good routes for training data analysts, e.g., from computing/mathematics/physics/application area?
  - Is there an indispensable core of statistical ideas?

References

Breiman, L. 2001. Statistical modeling: the two cultures. (With invited comments). Statistical Science 16: 199-215.

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

Ripley, B.D. and Fei Chen, R. 2003. Data mining by scaling up open source software. Invited paper, Proceedings, 2003 session of the ISI.

- 3rd year Mathematics Dept course in “data mining” at ANU. [See http://datamining.anu.edu/]
  - Most (all?) students have done a course in probability.
  - The primary interest of tutors has been numerical analysis.
  - Models have Normal i.i.d. errors.
  - The style (EDA?) has similarities to Hastie et al [2000]?
  - Which aspects of data structure can be smoothed away, and which should be retained?
  - Are students of this course missing important ideas?
- Collection of data into data bases can be mindless, unless potential use is considered and demonstrated, and implications are considered for the form (level of aggregation, etc.) in which data will be stored. [Don’t collect unusable junk!]

Is Algorithmic Data Analysis Possible?

Breiman (2001) argues for wider acceptance of “algorithmic” models (including neural nets, decision trees and Support Vector Machines.) Breiman claims that these, and a reliance on predictive validation (PV) can avoid “data mechanism” assumptions.

PV: Fit model to a training set; validate on a test set. Thus:

- ○ Fit a sequence of models to the training set
- ○ Choose the model that best predicts for the test set

CV: In some applications, “cross-validation” (CV) allows improved estimates of “error”. For this split the data into, e.g., ten parts. Then each tenth becomes in turn the test set, with the remaining nine tenths the training set.

[But i) should the test set be houses, or suburbs, or . . .?
ii) what should split into 10 parts? Houses, or Suburbs, or . . .?]

Predictive Validation (PV)

- Most (all) data analyses are in some sense predictive. Unless results have a relevance beyond the specific set of data, what is the point of any analysis?
- There can be different generalizations from the same data – e.g., generalizations in space, in time, to other business units. **Different generalizations require different sampling mechanisms, with different predictive accuracies.**
- Predictive accuracy assessment can impose useful "reality" checks, forcing careful thinking about how the model will be used.
- There are other considerations than predictive accuracy per se.

c.f. "Predictive/Descriptive Accuracy . . . is the dominant measure for most data mining models since it measures how well the data is summarized and generalized into the model."
[Kantardzic, M. 2002. Data Mining: Concepts, Models, Methods and Algorithms. Wiley.]

Different Generalizations – Different Accuracies

Example: Predict house prices, based on floor area, no. of bedrooms, etc., in different suburbs of Canberra (or Berlin!).

- A prediction from data for one suburb may not be useful for other suburbs.
- Even if a model can be found that is useful for different predictive generalizations (same/different suburbs), the predictive accuracy estimates will change, depending on whether predictions are:
  - for the same suburbs as in the data;
  - for a new set of suburbs;
  - for the same suburbs, for 2004.
  - for a new set of suburbs, for 2004.
- In general, the statistics texts, if they address these issues at all, address them in the context of "advanced" theoretical treatments.
- The data mining texts (also Breiman 2001) ignore such issues.

Sometimes, it is necessary to say the obvious!

Commentary

- The training/test set methodology is too limited for many important practical contexts.
- Models have deterministic and stochastic parts. The implicit model

  $y_i = f(x_i) + \square_i$, (with □ ~ i.i.d. normal)

  is often a poor model for the real world!
- There can be algorithmic analyses, i.e., analyses that follow steps that are specified by an algorithm.

  But what is an algorithmic model? A model with no error term?
- Algorithmic analyses are fine, providing they do the required job.

  But what is the job that is required? [There can be > 1 jobs.]
- Different uses ("jobs") for the model will lead to different predictive accuracy assessments.

  [Target population &/or sampling mechanism change with the "job".]

The Predictive Accuracy Assessment Process

- Samples are from the target population, a/c a sampling mechanism.
- Predictions are made for the sample data.
- Predictions are compared with observed sample values, and an accuracy estimate calculated.

The sampling mechanism reflects the intended generalization.

- For prediction to new houses in the same suburbs: test each prediction on a house in the same suburb.
- New suburbs: test on new houses in new suburbs.
- Predict mean for a new suburb: a stochastic model is necessary. ((Number of houses varies between suburbs.)
- Prediction to 2004, use pre-2000 data to predict for 2000, pre-2001 data to predict for 2001, etc.

Predictive Accuracy – Four Circumstances

A. Data are from the target population, with a sampling mechanism that accurately reflects the intended use of the model.
[Use the cross-validation estimate of predictive accuracy.]

B. Test data are from the target population, with a sampling mechanism that reflects the intended use of the model.
[Use the test data to derive an estimate of predictive accuracy.]

C. The sampling mechanism for the target data differs from the mechanism that yielded the data in A, or yielded the test data in B. However, there is a model that predicts how predictive accuracy will change with the change in sampling mechanism.
[Thus, in the "attitudes to science" data, the predictive accuracy for the mean of a new class depends on the number in the class.]

D. A realistic test set and associated sampling mechanism may not be available, making a soundly based accuracy estimate unavailable. An informed guess may be necessary.

Borderline Cases

There may be a test set that is a plausible proxy for the target population. E.g.,

- The evaluation of algorithms for gene prediction.
  [Ideally, they should be effective in finding new genes!]
- Algorithms for finding protein homologies.

The Informed Guess

Commonly, predictions are required for a future time, but the only accuracy estimates are for prediction for another observation at the same time.

- Is there any relevant past experience with comparable predictions?
- Is the accuracy for prediction in time likely to be similar to that for prediction in space, or smaller, or larger?
- Such ballpark estimates can be badly astray.

“Data Mining” Texts – What do they say?

- Witten and Frank: These authors note that grouping in the data has implications for generalization, but get the details wrong.
- Hastie et al.: Discuss only simple forms of the use of test sets & cross-validation, i.e., as in A & B above.
- Hand et al.: “The score function (e.g., estimate of PA) has a randomness [that] arises both from the data used to train the model and from the data used to validate it.” Does not discuss the need to get the right contribution from the validation (test) data.
- Kantardzic: “Predictive/Descriptive Accuracy: This is the dominant measure for most data mining models since it measures how well the data is summarized and generalized into the model.”
  (But means for assessing predictive accuracy are not discussed.)

Key Ideas (Once stated; they seem trite!)

The first level of exposition can (should?) be intuitive, i.e., ask

- What generalization is required?
- What generalizations do these data allow?
  (To the same suburb? To different suburbs?)
- Emphasize:
  - Data from a single suburb (or hospital, or . . .) allows (at best) generalization within that suburb (or hospital, …)
  - For generalization to multiple suburbs, the sample size *n* is the number of suburbs.
    [More data within a suburb improves the accuracy of the sample summary for that suburb.]

Multi-level models seem the simplest context in which to get across key theoretical ideas. Start with balanced multi-level models.

Introduce sequential correlation structures early on.

Issues/Questions

- Are predictive validation (PV) ideas accessible, at least intuitively, without a strong grounding in statistical theory?
  [The challenge is to make them thus accessible!]
- Should PV get more attention in specialist statistics courses?
  [A side benefit is that accessible and practically oriented statistical expositions might quickly diffuse through to the data mining and machine learning communities.]
- A cynical view has been that
  "Data Mining = Massive Data Bases + Bad Statistics"
  The need is to ensure that
  "Data Mining = Well-designed Data Bases + Insightful Statistics"
- Data base design is about more than making data accessible.
  Think about the generalizations that:
  i) are of interest to users; ii) can be answered by available data.
  and think about consequent implications for design.