Ideas and Tools for a Data Mining Style of Analysis

John Maindonald

July 2, 2009

Contents

Purpose and Range of Content

Data mining gathers up some of the ways in which a section of the computer scientist community has sought to respond to the impact of computer technology on the collection and use of data. The effects have spread across commerce, government and society, as well as across science. The same or similar technology may be used across these different areas of application.

Other responses, with much in common with data mining, have the names "machine learning", and "analytics". Machine learning has grown out of an engineering context, while the name "analytics" is widely used in a business context. Other such names are in use also, most of them with a focus on a specific area of application.

Data mining is sometimes presented as a collection of algorithms. Validation issues are, largely, left to one side. My intention is to widen the scope, to describe an enterprise that builds on advances in data collection and data analysis technology in ways that pay serious regard to model validation issues. In this account, statistical considerations have a large role. Providing this is understood, the enterprise that is in mind can suitably be described as "data mining".

These notes elaborate to some limited extent on points that I plan to cover in my talk, and extend into areas that will not be covered. Sections 1– 4 are important for the talk. Remaining sections are for follow-up study.

1 Resampling and Other Computer Intensive Methods

Both for the talk and for the practical sessions, there are several methodologies that will be important. These are:

Simulation: A model is proposed that generated the data. What are its statistical properties? One answer is repeated simulations. Simple uses of this idea are:

- Simulate repeated sampling of values that follow a normal distribution.
- Simulate repeated sampling of values that follow a non-normal distribution, perhaps an exponential distribution.
- Marks are placed on the circumference of a roulette wheel that divide it into three perhaps unequal parts. Labeling the outcomes A, B and C, one can simulate, eg, a result from 1000 spins.

In statistical learning contexts, theoretical accuracy assessments have limited usefulness. It is frequently necessary to rely on empirical methods. This is especially true for classification models. Several important methods of this type will now be described.

Training/Test:

- Split data into training and test sets
- Train (NB: steps **1 & 2**); use test data for assessment.

This is the most widely applicable methodology. The test data can in principle be taken from the target population. If however data from the actual target is available when the model is fitted, one would want to use this for model fitting. Thus, in practice, testing on data from the actual target often has to be an "after-the-event" kind of check.

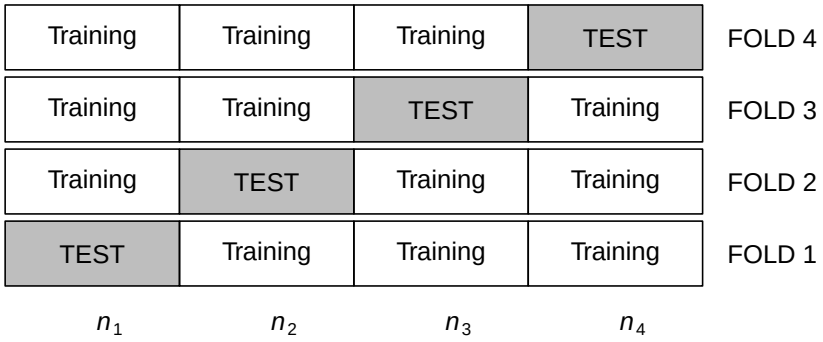

Figure 1: Schematic, designed to show how cross-validation works. For present illustrative purposes, data are split into four nearly equal parts, with n_1, n_2, n_3, and n_4 observations respectively. At the first iteration, the n_1 observations in the first part are set aside for testing, with remaining observations used for training, and so on. The split into a part that is used for testing, and the remaining observations that are used for testing, has the name *fold*. Once calculations for all 4 folds are complete, predicted values are available for all observations, in each case from a model that was trained independently of those observations.

When there is very adequate data, a training/test split of the data achieves all that is needed. It is not necessary to look for a method, such as will now be described, that makes better use of the data.

Cross-validation Simple version: Train on subset 1, test on subset 2
Then; Train on subset 2, test on subset 1

More generally, data are split into k parts (eg, $k = 10$). Use each part in turn for testing, with other data used to train.

Cross-Validation Steps are:

- Split data into k parts (in Figure 1, k=4)
- At the ith repeat or *fold* ($i = 1, \ldots k$) use:
  the ith part for *testing*, the other k-1 parts for *training*.
- Combine the performance estimates from the k folds.

Bootstrap Sampling Bootstrap samples are with replacement samples, of the same size as the initial sample.

Here are two bootstrap samples from the numbers 1 . . . 10

```
1 3 6 6 6 6 6 8 9 10 (5 6's; omits 2,4,5,7)
2 2 3 4 6 8 8 9 10 10 (2 2's, 2 8's, 2 10's; omits 1,5,7)
1 1 1 1 3 3 4 6 7 8 (4 1's, 2 3's; omits 2,5,9,10)
```

Here is how bootstrap sampling can be used in practice:

- Take repeated (with replacement) random samples of the observations, of the same size as the initial sample.

- Repeat analysis on each new sample (NB: In the example above, repeat **both** step 1 (selection) & 2 (analysis).
- Variability between samples indicates statistical variability.
- Combine separate analysis results into an an overall result.

2 Key Motivations

2.1 Computing Technology is Transforming Science

Computing technology is clearly transforming science. The following is a summary of the changes. All except the final item feature strongly in the data mining literature:

- Huge data sets[1] have become common. Often these hold new types of data – web pages, medical images, expression arrays, genomic data, NMR spectroscopy, sky maps, ...
- Sheer size brings challenges. In the first place, these are challenges for data management. Challenges for data analysis are less simply described.
  - A key question is whether there are many observations, or many variables, or both.
  - Where the number of observations is large, there is often a structure (eg, changes in time) that requires attention.
- New algorithms; "algorithmic" models.
  - Examples are trees, random forests, Support Vector Machines, ...
  - They are algorithmic in the sense that the motivation was algorithmic.
- Automation, especially for data collection
  - There has been huge progress in automating many aspects of data analysis.
  - Except however in limited areas of application, complete automation remains a pipe dream. In those areas where it has proved effective, automation typically has a large setup and maintenance cost.
  - Automation is most feasible in applications where mistakes can be tolerated, where it is not necessary to be consistently correct.
- Synergy: computing power with new theory. Much of the new development in theoretical statistics is driven and facilitated by the demand to take advantage of new computational power.

All the above, except the final item, get strong emphasis in the data mining literature. The literature emphasizes the synergy with new algorithmic development, but pretty much ignores the synergy with new developments in statistical theory.

[1] Issues of data set size have generated some modest smount cf hype, as the frequent reference to *Big Data* in Weiss and Indurkha (1997)

2.2 Jargon – Mining, Learning and Training

Mining is used in two senses:

- Mining for data
- Mining to extract meaning is a scientific/statistical sense.

The pre-analysis data extraction & processing often relies heavily on computer technology. Additionally, there are design of data collection issues that demand more attention than has been common.

In mining to extract meaning, statistical considerations come to the fore. Computer power does however provide a major part of the "how".

Learning & Training

- The (computing) machine *learns* from data.
- Use *training* data to train the machine or software.

Modeling (or a machine?) that learns from the data

The reference is to modeling in which the data have a substantial role in determining the form of model. The demand for such models arises, in part, from the size of the data sets (many observations) that are now commonly available. Deviations from the strict form of model that are described by theory are more readily detectable. Statistical learning approaches are then used to accommodate deviations from the theoretical model. The plot of residuals in Figure 2 suggests that, for the data displayed there, a statistical learning approach may be appropriate.

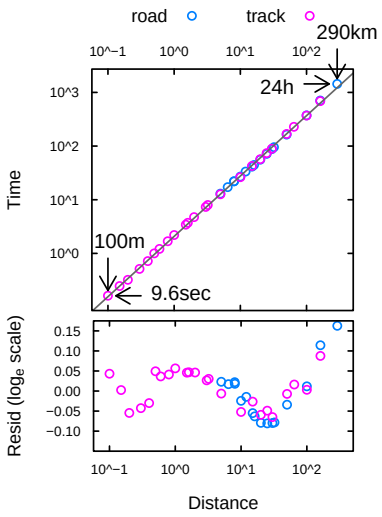

Figure 2: Record times versus distances, both on logarithmic scales, for track and road athletic races. With a ratio of largest to smallest time that is ~3000, differences from the line have to be large to be visually obvious. The plot of residuals shows that, for the longest race, the difference from the line is >15%. (Differences on a scale of natural logarithms, if small, are a little less than fractional differences.) There is a clear systematic pattern in the residuals.

For classification models, theory rarely gives much help in deciding on the form of model. A statistical learning approach is, to a greater or lesser extent, inherent in the nature of the modeling problem.

In the applications of statistical learning that are prominent in the data mining literature, the aim is usually prediction rather than the obtaining of interpretable model parameters. Hence the name "predictive modeling". Interpretation of model parameters raises additional issues that will be the subject of brief comment in a later section.

2.3 Statistical Learning Example – Continuous Outcome

The first example (Figure 2) is for a continuous outcome variable. Data are world record times for athletic track and road races, as at October 2006. The range of distances and times is huge, from 100m in 9.6sec to 292.2km in 24h.

We can fit a curve rather than a line, in what is a statistical learning approach. Here, a curve will be fitted to the residuals – this corrects for the biases in the line.

Questions are:

- Will the pattern be the same in 2030?
- Is it consistent across geographical regions?
- Does it partly reflect greater attention paid to some distances?
- So why/when the smooth, rather than the line?

Clearly the smooth curve (line, with 'corrections' from the line) would be useful to race organizers who wished to estimate the time at which a race winner could be expected to appear.

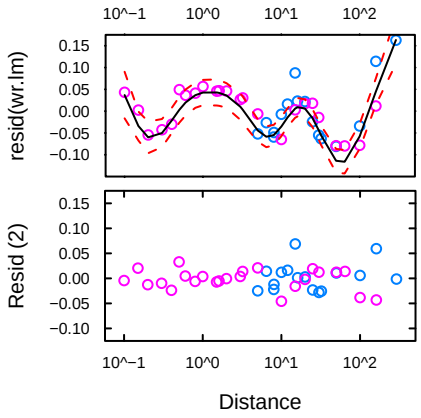

Figure 3: Here a smooth curve has been fitted to the residuals, using a routine that does the job pretty much automatically. It is assumed that residuals from the curve are independent. This assumption is especially crucial for the 95% pointwise confidence limits about the curve. The lower panel shows residuals from the smooth.

This is a very simple example of the use of the methodology. It generalizes, allowing the fitting of curves and surfaces, in principle in an arbitrary number of dimensions.[2] The ability to fit such curves and surfaces automatically is remarkable, relative to what was available a decade ago.

2.4 Statistical Learning Example – Categorical Outcome

The example relates to glass fragments that were collected in the course of forensic work. Glass was of the following types. Numbers of pieces of glass of each of the

[2]Note in particular the abilities in the R package *mgcv*, documented in detail in Wood (2006).

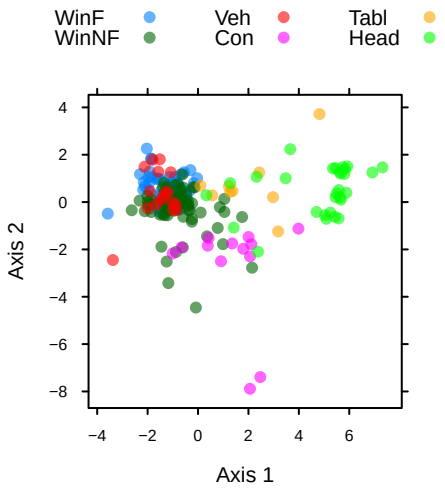

Figure 4: Visual representation of a classification rule, derived using *linear discriminant analysis*, for the forensic glass data. A six-dimensional pattern of separation between the categories has been collapsed down to two dimensions. Some categories may therefore be better distinguished than is evident from this figure.

different types are given:

Window float (70)	Window non-float (76)	Vehicle window (17)
Containers (13)	Tableware (9)	Headlamps (29)

Variables are %'s of `Na`, `Mg`, ..., plus refractive index. In all there are 214 rows of data (observations) × 10 columns (variables).

The aim is to find a rule that predicts the type of any new piece of glass. Figure 4 is a visual summary of the result from the use of a simple form of classification methodology, with the name *linear discriminant analysis*.

Points to note are:

- The methodology is implemented within a Bayesian framework. By default, the prior probabilities for the various categories are taken to be the relative frequencies for those categories. The classification rule changes if the frequencies are changed from the default.
- For any given classification rule, the overall accuracy (proportion correctly classified) changes if the prior probabilities are changed.
- For estimating the accuracy for a given target population, the prior probabilities should be the proportions in that population, not the proportions in the sample.

Note that these data date from 1987. Changes since then in glass mamufacturing may well mean that they have limited relevance to current circumstances. It would be hazardous to use a prediction rule derived from these data for predictions for recently collected glass fragments.

3 Purpose and Source/Target Issues

A data mining style of analysis, if it is done properly, offers exactly the same range of challenges as data analysis more generally. As with any data analysis, key considerations are:

A1: What is (are) the intended purpose(s) of the analysis?

A2: Is predictive accracy the main consideration? Or is the aim scientific of other insight, perhaps based on values of parameter estimates.

A3: How widely is it hoped that results will apply?

The answers may have strong implications for the any decision on how to handle the data. Data mining exercises typically are mainly interested in predictive accuracy. Questions of what interpretation can be placed on model parameters do however often arise, often as an afterthought to the main analysis. It is therefore important to understand the potential for misinterpretation.

Questions on which the analysis may help shed light are:

B1: Is a repeat of the analysis with a new set of data likely to turn up comparable results?

B2: How widely applicable is whatever can be learned from the data? (Will it apply to a city that is different from those that generated the sample data? Will it apply to a different time – next year, or two years' time?)

Question B2 is the really crucial question. It asks those who analyze or use the data to think about how the target population may relate to the source population? At the extremes:

T1: Source and target may be quite different. The notion that the data can be used to make statements that apply to the target is, at the very least, hopeful!

T2: Source and target may be closely identical.

Experiments, if properly designed, should ensure that source is closely identical with a target. This is not a guarantee that source will be closely identical with the target that is of main interest. Experiments with one or other strain of mice may or may not give results that generalize to other strains. Results may or may not have relevance, or some limited relevance, to humans.

Source/target issues have particular pertinence for data mining type applications, where data are not commonly experimental.

4 Analysis Types and Tools

The favoured methodologies come, mostly, under the broad heading of "regression". There is some use of regression methods with a continuous outcome variable. The data mining literature is however strongly focused on methodologies where the outcome is classification into one of two or more categories.

4.1 Methodologies

For work with classification data, there is some use of classical statistical tools. Such tools include logistic and multinomial logistic regression, and the closely related linear and quadratic discriminant analysis approaches. Theoretical accuracy assesments, making the usual independence assumptions, are asymptotic, and may or may not be useful in practice. Commonly, accuracy is estimated using the same approaches as are used for the newer "data mining" methodologies.

Prominent among the newer "data mining" methodologies that have found favour in the data mining and machine learning communities are:

- Classification trees and various extensions of trees:
  - Random forests and related "bagging methods";
  - "Boosting" methods, notably Adaboost
- Support Vector Machines
- Neural networks (these turn out, however, to have close connections with logistic regression).

4.2 Accuracy assessment

Primarily, the accuracy assessment methods that are discussed here assume that the target population is essentially the same as the source population from which the data have been obtained. Even for this limited purpose, there is serious scope for getting answers that can be grossly optimistic.

Accuracy assessment is important for its own sake. It is helpful to know what the finally fitted model has been able to achieve. Unless however there is effective accuracy assessment, it will not be possible to fit a good model:

- Many methods work by starting with an initial model, which is successively refinement. Too much refinement (over-fitting) will lead to a model with reduced predictive power. It is necessary to know when to stop.
- Good accuracy assessments are required so that a model fitted using one methodology can be compared with a model fitted using another methodology.

4.2.1 Key Methodologies for Empirical Accuracy Assessment

In statistical learning contexts, theoretical accuracy assessments have limited usefulness. It is frequently necessary to rely on empirical methods. This is especially true for classification models. Section 1 described the methods that will be important here.

Continuous outcome data: For regression with a continuous outcome, normal theory accuracy estimates can, if the independent normal error assumptions are not too badly wrong, work quite well. Note however that:

- If the model is selected from a wide class of models, or if there is extensive variable selection (e.g., select the best 3 explanatory variables out of 10), the accuracy estimates may be grossly optimistic.
- If observations are not independent, accuracy estimates may again be wrong, usually optimistic. Note however that that the situation can in special cases be rescued by choosing a more realistic model for the "error". Some of the possibilities are:
  - For data that are collected over time, models are available that can account for the likely sequential correlation.

- Variation is often multi-layered – variation between different countries, variation between humans in an individual country, variation between clinical assessment made on the same human, and so on. Again modeling approaches are available that can account for such different sources of variation.
- Spatial models are another possibility.

Empirical methods for accuracy assessment can in principal be adapted for use where there is a complex error structure. This does however require a clear understanding of the theoretical issues, and is not straightforward.

Categorical outcome data: Here, the theoretical accuracy estimates that are available for certain of the methods rely on asymptotic approximations. For 'algorithmic' methods, including tree-based methods and random forests, theoretical results have limited relevance. Accuracy assessment almost inevitably relies on empirical methods.

The empirical methods do if used correctly cope with the effects of model and variable selection.

4.2.2 Keeping Score

As has been pointed out, the source population from which data have been obtained will often not be exactly the same as the target population. There are two important issues here:

- Where predictive modeling of a comparable type is being carried out repeatedly, the analyst should keep a record of the comparison between after-the-event model performance and predicted performance, e.g., from cross-validation on the original data.
- Often, predictions are successively made ahead in time. If a long enough data series is available, time series methods may be apprpropriate. In effect, past changes from one time to the next are used as a guide to likely future changes.

The modeling of changes over time is in principle a good idea. Do not however put too much faith in the model. A lesson from the recent financial crisis is, surely, that it is unwise to put much faith in any financial model that does not allow for occasional "shocks". The warnings in Taleb (2004) merit attention.

5 Challenges from Size of Dataset

Sheer size brings challenges. In the first place, these are challenges for data management. The analysis challenges that result from data set size are not, however, primarily those of scaling up regression and related models for use with large datasets.

The analysis challenges are most important for models with a complex error structure – repeated measures and times series, spatial models, and so on. They are much less important for the use of regression models.

Some of the issues are:

- Additional structure often comes with increased size – data may be less homogeneous, span larger regions of time or space, be from more countries

- Or there may extensive information about not much!
  - e.g., temperatures, at second intervals, for a day.
  - SEs from modeling that ignores this structure may be misleadingly small.
- In large homogeneous datasets, spurious effects are a risk
  - Small SEs increase the risk of detecting spurious effects that arise, e.g., from sampling bias (likely in observational data) and/or from measurement error.

Where a straightforward use of a regression model really does seem appropriate, there can be advantages in structuring the analysis as a series of separate regressions:

- As an example, consider data that have been collected over a non-trivial interval of time. It is then sensible, as a check, to do separate analyses for separate times.
- Where there is no time or other such structure in the data, the analysis can usefully be repeated for separate random samples of the data. Variation in model parameters and predictions under such repeated sampling provides a check on theoretically based estimates of standard errors. If the empirical standard errors are larger, it is those that should be believed.

6 The Interpretation of Model Parameters

Consider data[3] that gives record times for Northern Ireland mountain races.

The "obvious" simple model has log(*time*) as a linear function of log(*dist*) and log(*climb*).

$$\widehat{\log(\text{time})} = -5.0 + 0.68\ \log(\text{dist}) + 0.47\ \log(\text{climb})$$

Note the coefficients 0.68 (for log(*dist*)) and 0.47 (for log(*climb*))! Do they make sense? Thus, the coefficient 0.68 for log(*dist*) implies that the relative rate of increase of time with distance is, if climb is held constant, 68% of the relative rate of increase of distance. If a second kilometer is added to a 1 kilometer race, the time per unit distance will be better than for the 1 kilometer race.

The clue is that the coefficient predicts what will happen if climb is held constant. The one kilometer race then involves much steeper climbing (and decent) than the two kilometer race.

More interpretable coefficients can be obtained by regressing on log(*dist*) and log(*climb*/*dist*). The comparison between different distances is then fair.

For a meaningful interpretation of model parameters, it is necessary to be sure that:

- All major variables or factors that affect the outcome have been accounted for.
- Those variables and factors operate, at least to a first order of approximation, independently.

Rosenbaum (2002) suggests approaches that are often useful in the attempt to give meaningful interpretations to coefficients that are derived from observational data.

[3] from `http://www.nimra.org.uk/calendar.asp`

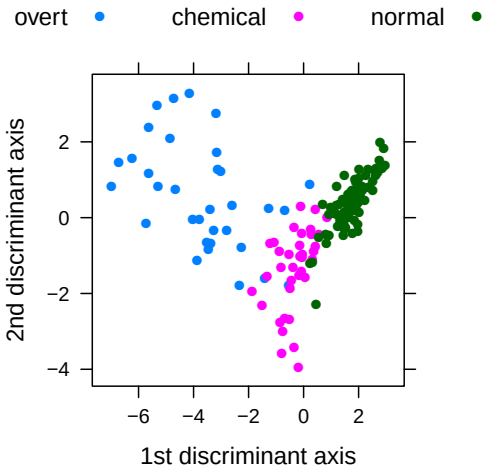

Figure 5: This is a visual summary of results from the use of linear discriminant analysis. Classification accuracy, estimated by leave-one-out cross-validation, was 89%.

7 Notes on Methods for Use with Classification Data

Many different methods are available. The data mining literature encourages wide experimentation. The stance taken here is that it is better to start with several well-understood methodologies, and get to understand those, before branching out more widely. Three methodologies will be described here – linear discriminant analysis (LDA), trees and random forests. Trees will be of interest mainly for understanding the random forest methodology.

Linear Discriminant Analysis: The implementation that will be used here is `lda()` in R's *MASS* package. This methodology has the following attractions:

- It is simple, in the style of classical regression.
- It leads naturally to a 2-D plot.
- The plot may suggest trying methods that build in weaker assumptions.

Details are in Section 8.

Random forests: The implementation that will be used here is `randomForest()` in R's *randomForest* package. This methodology has the following attractions:

- It is clear why it works and does not overfit.
- No other method consistently outperforms it.
- It is simple, and highly automatic to use.

7.1 Example used to illustrate the methodologies

The three classes are from a clinical classification of Diabetes – `overt` (overt diabetic), `chemical` (chemical diabetic), and `normal` (normal). Can these be derived directly from five available clinical measures?

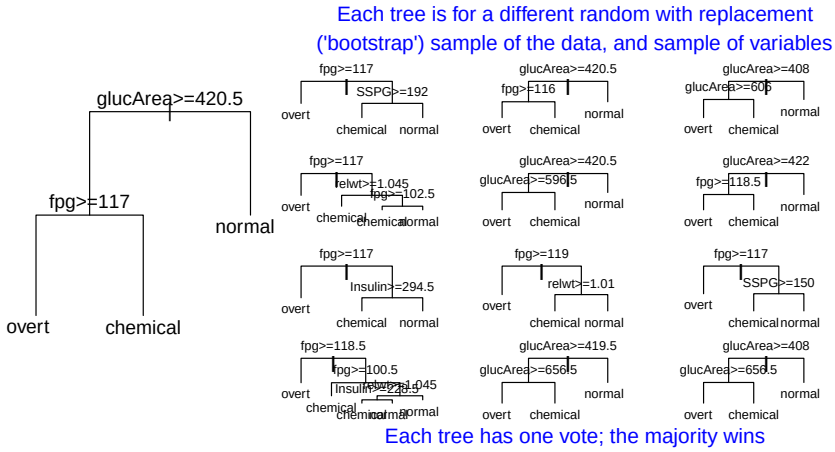

Figure 6: The left panel is a classification tree that was derived from tree-based classification. Each tree in the right panel is for a different bootstrap sample of the diabetes data. Additionally, a different random sample of variables is used for each different tree. The final classification is determined by a random vote over all trees.

The clinical measures (explanatory variables) are `relwt` (relative weight), `fpg` (fasting plasma glucose), `glucArea` (glucose area), `Insulin` (insulin area), and `SSPG` (steady state plasma glucose).

Figure 5 is a visual summary of results from the use of linear discriminant analysis. The right panel shows the classification tree obtained from tree-based regression.

Tree-based classification proceeds by constructing a sequence of decision steps. At each node, the split is used that best separates the data into two groups. Here tree-based regression does unusually well (CV accuracy = 97.2%), perhaps because it is well designed to reproduce a simple form of sequential decision rule that has been used by the clinicians.

How is 'best' defined? Splits are chosen so that the Gini index of "impurity" is minimized. Other criteria are possible, but this is how `randomForest()` constructs its trees.

The random forest methodology will usually improve (but not here), sometimes quite dramatically, on tree-based classification. Figure 6 shows trees that have been fitted to different bootstrap samples of the diabetes data. Typically 500 or more trees are fitted, without a stopping rule. Individual trees are likely to overfit. As each tree is for a different random sample of the data, there is no overfitting overall.

Figure 7 is a visual summary of the random forest classification result.

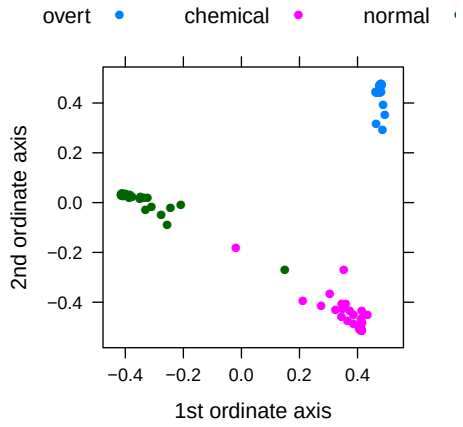

Figure 7: The plot is an attempt to represent, in two dimensions, the random forest result. This plot tries hard to reflect probabilities of group membership assigned by the analysis. It does not result from a 'scaling' of the feature space.

8 Linear Methods for Classification – Theory

See Ripley (1996); Venables and Ripley (2002); Maindonald & Braun (2007, Section 12.2).

8.1 Linear methods, vs strongly non-parametric methods

Linear methods may be contrasted with the strongly non-parametric random forest method that uses an ensemble of trees. See Maindonald & Braun (2007, Section 11.7). A good strategy for getting started on an analysis where predictive accuracy is of primary importance is to fit a linear discriminant model with main effects only, comparing the accuracy from a random forest analysis. If the random forest analysis gives little or no improvement, the linear discriminant model may be hard to better. There is much more that can be said, but this may be a good starting strategy.

8.2 Notation and types of model

Observations are rows of a matrix $\mathbf{X}$ with p columns. The vector $\mathbf{x}$, is a row of $\mathbf{X}$, but in column vector form. The outcome is categorical, one of g classes.

Methods discussed here will all work with monotone functions of the columns of $\mathbf{X}$. By allowing columns that are non-linear monotone functions of the initial variables, additive non-linear effects can be accommodated.

As before, observations are rows of a matrix $\mathbf{X}$ with p columns. The vector $\mathbf{x}$, is a row of $\mathbf{X}$, but in column vector form.

The outcome is categorical, one of g classes, where now g may be greater than 2. The matrix $\mathbf{W}$ estimates the within class variance-covariance matrix, while $\mathbf{B}$ estimates the between class variance-covariance matrix. Details of the estimators used are not immediately important. Note however that they may differ somewhat between computer programs.

8.3 `lda()` and `qda()`

The functions that will be used here are `lda()` and `qda()`, from the *MASS* package for R. The function `lda()` implements linear discriminant analysis, while `qda()` imple-

ments quadratic discriminant analysis. Quadratic discriminant analysis is an adaptation of linear discriminant analysis to handle data where the variance-covariance matrices of the different classes are markedly different. For $g = 2$ the logistic regression model, fitted using R's `glm()` function, is closely analagous to the linear discriminant model, fitted using `lda()`.

An attractive feature of `lda()` is that it yields "scores" that can be plotted. Let $r = min(g - 1, p)$. Recall that p is the number of columns of a version of the model matrix that lacks an initial column of ones. Then assuming that $\mathbf{X}$ has no redundant columns, there will be r sets of scores. The r sets of scores can be examined using a pairs plot. Often, most of the information is in the fist two or three dimensions. Such plots may be insightful even for data where `lda()` is inadequate as a classification tool.

8.3.1 `lda()` and `qda()` – theory

The functions `lda()` and `qda()` in the *MASS* package implement a Bayesian decision theory approach.

- A prior probability π_c is assigned to the cth class ($i = 1, \ldots g$).
- The density $p(\mathbf{x}|c)$ of $\mathbf{x}$, conditional on the class c, is assumed multivariate normal, i.e., rows of $\mathbf{X}$ are sampled independently from a multivariate normal distribution.
- For linear discrimination, classes are assumed to have a common covariance matrix Σ, or more generally a common $p(\mathbf{x}|c)$. For quadratic discrimination, different $p(\mathbf{x}|c)$ are allowed for different classes.
- Use Bayes' formula to derive $p(c|\mathbf{x})$. The allocation rule that gives the largest expected accuracy chooses the class with maximal $p(c|\mathbf{x})$; this is the Bayes' rule.
- More generally, assign cost L_{ij} to allocating a case of class i to class j, and choose c to minimize $\sum_i L_{ic} p(i|\mathbf{x})$.

Note that `lda()` and `qda()` use the prior weights, if specified, as weights in combining the within class variance-covariance matrices.

Using Bayes' formula

$$\begin{aligned} p(c|\mathbf{x}) &= \frac{\pi_c p(\mathbf{x}|c)}{p(\mathbf{x})} \\ &\propto \pi_c p(\mathbf{x}|c) \end{aligned}$$

The Bayes' rule maximizes $p(c|\mathbf{x})$. For this it is sufficient, for any given $\mathbf{x}$, to maximize

$$\pi_c p(\mathbf{x}|c)$$

or, equivalently, to maximize

$$\log(\pi_c) + \log(p(\mathbf{x}|c))$$

Now assume $p(\mathbf{x}|c)$ is multivariate normal, i.e.,

$$p(\mathbf{x}|c) = (2\pi)^{\frac{p}{2}} \mid \Sigma_c \mid)^{-\frac{1}{2}} exp(-\frac{1}{2} Q_c)$$

where

$$Q_c = (\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c)$$

Then

$$\log(\pi_c) + \log(p(\mathbf{x}|c)) = \log(\pi_c) - \frac{1}{2} Q_c + \frac{p}{2} \log(2\pi) - \frac{1}{2} \log(\mid \Sigma_c \mid)$$

Leaving off the $\log(2\pi)$ and multiplying by -2, this is equivalent to minimization of

$$Q_c + \log(\mid \Sigma_c \mid) - 2\log(\pi_c) = (\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c) + \log(\mid \Sigma_c \mid) - 2\log(\pi_c)$$

The observation $\mathbf{x}$ is assigned to the group for which

$$(\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c) + \log(\mid \Sigma_c \mid) - 2\log(\pi_c)$$

is smallest.

Set $\mu_c = \bar{\mathbf{x}}_c$, and replace $\mid \Sigma_c \mid$ by an estimate $\mathbf{W}_c$.
[Note that the usual estimate of the variance-covariance matrix (or matrices) is positive definite, providing that the same observations are used in calculating all elements in the variance-covariance matrix and $\mathbf{X}$ has no redundant columns.]

Then $\mathbf{x}$ is assigned to the group to which, after adjustments for possible differences in π_c and $\mid \Sigma_c \mid$, the Mahalanobis distance

$$(\mathbf{x} - \bar{\mathbf{x}}_c)^T \mathbf{W}_c^{-1} (\mathbf{x} - \bar{\mathbf{x}}_c)$$

of $\mathbf{x}$ from $\mathbf{x}_c$ is smallest.

If a common variance-covariance matrix $\mathbf{W}_c = \mathbf{W}$ can be assumed, a linear transformation is available to a space in which the Mahalanobis distance becomes a Eucliean distance. Replace $\mathbf{x}$ by

$$\mathbf{z} = (U^T)^{-1} \mathbf{x}$$

and $\bar{\mathbf{x}}_c$ by $\bar{\mathbf{z}}_c = (U^T)^{-1} \bar{\mathbf{x}}_c$ where U is an upper triangular matrix such that $U^T U = \Sigma$. Then

$$(\mathbf{x} - \mu_c)^T \mathbf{W}^{-1} (\mathbf{x} - \mu_c) = (\mathbf{z} - \bar{\mathbf{z}}_c)^T (\mathbf{z} - \bar{\mathbf{z}}_c)$$

which in the new space is the squared Euclidean distance to from $\mathbf{z}$ to $\bar{\mathbf{z}}_c$.

Note: For estimation of the posterior probabilities, the simplest approach is that described above. Thus, replace $p(c|\mathbf{x}; \theta)$ by $p(c|\mathbf{x}; \hat{\theta})$ for calculation of posterior probabilities (the 'plug-in' rule). Here, θ is the vector of parameters that must be estimated. The R functions `predict.lda()` and `predict.qda()` offer the alternative estimate `method="predictive"`, which takes account of uncertainty in $p(c|\mathbf{x}; \hat{\theta})$. Note also `method="debiased"`, which may be a reasonable compromise between `method="plugin"` and `method="predictive"`

8.4 Canonical discriminant analysis

Here we assume a common variance-covariance matrix. As described above, replace $\mathbf{x}$ by

$$\mathbf{z} = \mathbf{U^{T}}^{-1} \mathbf{x}$$

where $\mathbf{U}$ is an upper triangular matrix such that $\mathbf{U^T U} = \mathbf{W}$.

The between classes variance-covariance matrix becomes

$$\tilde{\mathbf{B}} = \mathbf{U}^{\mathbf{T}^{-1}}\mathbf{B}\mathbf{U}^{-1}$$

The ratio of between to within class variance of the linear combination $\alpha^T\mathbf{z}$ is then

$$\frac{\alpha^T\tilde{\mathbf{B}}\alpha}{\tilde{\alpha}^T\tilde{\alpha}}$$

The matrix $\tilde{\mathbf{B}}$ admits the principal components decomposition

$$\tilde{\mathbf{B}} = \lambda_1\mathbf{u_1}\mathbf{u_1^T} + \lambda_2\mathbf{u_2}\mathbf{u_2^T} + \ldots + \lambda_r\mathbf{u_r}\mathbf{u_r^T}$$

The choice $\alpha = \mathbf{u_1}$ maximizes the ratio of the between to the within group variance, a fraction λ_1 of the total. The choice $\alpha = \mathbf{u_2}$ accounts for the next largest proportion λ_2, and so on.

The vectors $\mathbf{u}_1, \ldots \mathbf{u}_r$ are known as "linear discriminants" or "canonical variates". Scores, which are conveniently centered about the mean over the data as a whole, are available on each observation for each discriminant. These locate the observations in r-dimensional space, where r is at most $\min(g-1, p)$. A simple rule is to assign observations to the group to which they are nearest, i..e., the distance d_c is smallest in a Euclidean distance sense.

For plotting in two dimensions, one takes the first two sets of discriminant scores. A point $\mathbf{z}_i$ that is represented as

$$\zeta_{i1}\mathbf{u}_1 + \zeta_{i2}\mathbf{u}_2 + ... + \zeta_{ir}\mathbf{u}_r$$

is plotted in two dimensions as (ζ_{i1}, ζ_{i2}), or in three dimensions as $(\zeta_{i1}, \zeta_{i2}, \zeta_{i3})$. The amounts by which the original columns of $\mathbf{x}_i$ need to be multiplied to give ζ_{i1} are given by the first column of the list element `scaling` in the `lda` object. For ζ_{i2}, the elements are those in the second column, and so on. See the example below.

As variables have been scaled so that within group variance-covariance matrix is the identity, the variance in the transformed space is the same in every direction. An equal scaled plot should therefore be used to plot the scores.

8.4.1 Linear Discriminant Analysis – Fisherian and other

Fisher's linear disciminant analysis was a version of canonical discriminant analysis that used a single discriminant axis. The more general case, where there can be as many as $r = \min(g-1, p)$ discriminant functions, is described here.

The theory underlying `lda()` assigns $\mathbf{x}$ to the class that maximizes the likelihood. This is equivalent to choosing the class c that minimizes $d_c + \log(\pi_c)$, where if the same estimates are used for $\mathbf{W}$ and $\mathbf{B}$, d_c is the distance as defined for Fisherian linear discriminant analysis. Recall that π_c is the prior probability of class c.

The output from `lda()` includes the list element `scaling`, which is a matrix with one row for each column of $\mathbf{X}$ and one column for each discriminant function that is calculated. This gives the discriminant(s) as functions of the values in the matrix $\mathbf{X}$.

8.4.2 Example – analysis of the forensic glass data

The data frame `fgl` in the *MASS* gives 10 measured physical characteristics for each of 214 glass fragments that are classified into 6 different types.

The following may help make sense of the information in the list element `scaling`.

```
library(MASS}
fgl.lda <- lda(type ~ ., data=fgl)
scores <- predict(fgl.lda, dimen=5)$x  # Default is dimen=2
## Now calculate scores from other output information
checkscores <- model.matrix(fgl.lda)[, -1] %*% fgl.lda$scaling
## Center columns about mean
checkscores <- scale(checkscores, center=TRUE, scale=FALSE)
plot(scores[,1], checkscores[,1])  # Repeat for remaining columns
## Check other output information
fgl.lda
```

93% of the information, as measured by the trace, is in the first two discriminants.

8.5 Two groups – comparison with logistic regression

Logistic regression, which can be handled using R's function `glm()`, is a special case of a Generalized Linear Model (GLM). The approach is to model $p(c|\mathbf{x};\hat{\theta})$ using a parametric model that may be the same logistic model as for linear and quadratic discriminant analysis.

In this context it is convenient to change notation slightly, and give $\mathbf{X}$ an initial column of ones. In the linear model and generalized linear model contexts, $\mathbf{X}$ has the name "model matrix".

The vector $\mathbf{x}$ is a row of $\mathbf{X}$, but in column vector form. Then if π is the probability of membershipin the second group, the model assumes that

$$\log(\pi/(1-\pi) = \beta'\mathbf{x}$$

where β is a constant.

Compare logistic regression with linear discriminant analysis:

- Inference is conditional on the observed $\mathbf{x}$. A model for $p(\mathbf{x}|c)$ is not required. Results are therefore more robust against the distribution $p(\mathbf{x}|c)$.
- Parametric models with "links" other than the logit $f(\pi) = \log(\pi/(1-\pi)$ are available. Where there are sufficient data to check whether one of these other links may be more appropriate, this should be done. Or there may be previous experience with comparable data that suggests use of a link other than the logit.
- Observations can be given prior weights.
- There is no provision to adjust predictions to take account of prior probabilities, though this can be done as an add-on to the analysis.
- The fitting procedure minimizes the deviance, which is twice the difference between the loglikelihood for the model that is fitted and the loglikelihood for a 'saturated' model in which predicted values from the model equal observed values. This does not necessarily maximize predictive accuracy.
- Standard errors and Wald statistics (roughly comparable to t-statistics) are provided for parameter estimates. These are based on approximations that may fail if predicted proportions are close to 0 or 1 and/or the sample size is small.

8.6 Linear models vs highly non-parametric approaches

The linearity assumptions are restrictive, even allowing for the use of regression spline terms to model non-linear effects. It is not obvious how to choose the appropriate degree for each of a number of terms. The attempt to investigate and allow for interaction effects adds further complications. In order to make progress with the analysis, it may be expedient to rule out any but the most obvious interaction effects. These issues affect regression methods (including GLMs) as well as discriminant methods.

On a scale in which highly parametric methods lie at one end and highly non-parametric methods at the other, linear discriminant methods lie at the parametric end, and tree-based methods and random forests at the non-parametric extreme. An attraction of tree-based methods and random forests is that model choice can be pretty much automated.

8.7 Low-dimensional Graphical Representation

In linear discriminant analysis, discriminant scores in as many dimensions as seem necessary are used to classify the points. These scores can be plotted. Each pair of dimensions gives a two-dimensional projection of the data. If there are three groups and at least two explanatory variables, the two-dimensional plot is a complete summary of the analysis. Even where higher numbers of dimensions are required, two dimensions may capture most of the information. This can be checked.

With most other methods, a low-dimensional representation does not arise so directly from the analysis. The following approach, which can be used directly with random forests, can be adapted for use with other methods. The proportion of trees in which any pair of points appear together at the same node may be used as a measure of the "proximity" between that pair of points. Then, subtracting proximity from one to obtain a measure of distance, an ordination method can be used to find a representation of those points in a low-dimensional space.

References

Berk, R. 2008. *Statistical Learning from a Regression Perspective.*
[Berk's insightful commentary injects needed reality checks into the discussion of data mining and statistical learning.]

Maindonald, J.H. 2006. Data Mining Methodological Weaknesses and Suggested Fixes. Proceedings of Australasian Data Mining Conference (Aus06)'[4]

Maindonald, J. H. and Braun, W. J. 2007. *Data Analysis and Graphics Using R – An Example-Based Approach.* 2nd edn, Cambridge University Press.[5]
[Statistics, with hints of data mining!]

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Venables, W. N. and Ripley, B. D. 2002. *Modern Applied Statistics with* S. Springer-Verlag, 4edn.

[4] `http://www.maths.anu.edu.au/~johnm/dm/ausdm06/ausdm06-jm.pdf` and `http://wwwmaths.anu.edu.au/~johnm/dm/ausdm06/ohp-ausdm06.pdf`

[5] `http://www.maths.anu.edu.au/~johnm/r-book.html`

Rosenbaum, P. R., 2002. *Observational Studies*. Springer, 2ed.
[Observational data from an experimental perspective.]

Taleb, Naseem, 2004. *Fooled By Randomness: The Hidden Role Of Chance In Life And In The Markets.* Random House, 2ed.
[Has many insightful comments about the over-interpretation of phenomena in which randomness is likely to have a large role.]

Wood, S. N., 2006. *Generalized Additive Models*. An Introduction with R. Chapman & Hall/CRC.
[This has an elegant treatment of linear models and generalized linear models, as a lead-in to methods for fitting curves and surfaces.]