Data Mining Statistical Issues – Some Further Comments

John Maindonald (Centre for Mathematics and Its Applications, Australian National University)

June 30, 2009

The Tradition and Practice of Data Mining

- The tradition is recent, from computer science.
- The emphasis has been classification and clustering, perhaps also regression with a continuous outcome.
- 'Algorithmic' methods are preferred (trees, ...).
- Independence between cases is assumed.
  - Attention to temporal or spatial dependence is unusual
  - This assumption is taken for granted, also, in "data mining" books and papers written by statisticians
- Prediction is usually the aim, not interpretation of model coefficients or other parameters
  - Where the task merges over into interpretation of regression or other coefficients, there are many traps.
- The source/target distinction may be ignored.

New Names for Old Ideas – Exploratory Data Analysis, etc.

- Data Mining has the character of "muscular" EDA (Berk , 2006, who uses the term "musclecar")
- Statistical Learning and Machine Learning likewise connect strongly with computer science, but pay more attention to theory (some mix of algorithmics, mathematics, probability, statistics).
- *Analytics* has become popular as a name for applications in business and commerce.
- Several recent books have catchy invented titles (a bit in the style of catchy phrases such as "Data Mining" and "Machine Learning"!)
  - Super Crunchers (Ayres, 2006); The Numerati (Baker, 2008) Way to be Smart.

  There'll be more such titles. There'll also be more new data analysis streams, with their own new catchy names.

Examples – Learning in a Regression Context

Example 1: Northern Irish Hill Races (classical regression)

Data[1] are record times for Northern Ireland mountain races.

The "obvious" simple model has log(*time*) as a function of log(*dist*) and log(*climb*). But does the result make sense?, i.e.

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

Note the coefficients 0.68 and the 0.47! Do they make sense?

The only "learning" was the calculation of coefficients.

Example 2 – Lean Bodyfat Data (modern regression?)

Body `density` (252 men) is from underwater weighing.

Is a good estimate possible without the water, from weight scales and measuring tape: `age`, `weight`,`height`,`neck`, `chest`, `abdomen`, `hip`, `thigh`, `knee`, `ankle`, `biceps`, `forearm`, `wrist`

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

From Experiment to Observation to Statistical Learning

- **Experiment:** Variables/factors are manipulated under tight control.
- Data are **observational**, but theory suggests the form of model, or at least plausible models cf, the data on record times for hill races.
- **Statistical Learning**: Scientific understanding gives limited or no help
  - x, x^2 or $\log(x)$, or splines, maybe for many variables.
    The saving grace is that few may have much effect.
  - Broad strategy: Use the model that works!
    Why does it work? Good question!

Take home message: Apply proper checks that the model really does work!!

The glorious endeavour that we know today as science ...

. . . has grown out of the murk of sorcery, religious ritual, and cooking. But while witches, priests and chefs were developing taller and taller hats, scientists worked out a method for determining the validity of their results: they learned to ask "Are they reproducible?"

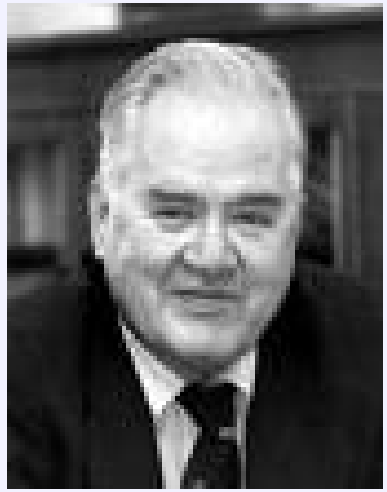
Scherr, G. H. 1983. Irreproducible Science: Editor's Introduction. In *The Best of the Journal of Irreproducible Results*, Workman Publishing, New York.

What does it mean for a result to be reproducible? Would that the answer was obvious!

. . . no isolated experiment . . . can suffice

We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. [R A Fisher (1935)]

Long ago Fisher . . . recognised that . . . solid knowledge came from a demonstrated ability to repeat experiments . . . This is unhappy for the investigator who would like to settle things once and for all, but consistent with the best accounts . . . of the scientific method . . . [Tukey (1991)]



Reproducibility is even harder to demonstrate when data are observational!

Common Methods for Assessing Accuracy

- Training/Test, with a random split of the avaialble data
  - Do not use the test data for tuning or variable selection (this is a form of cheating!)
- Cross-validation – a clever use of the training/test idea
  - Repeat any tuning and/or variable selection at each cross-validation "fold" (anything else is cheating, unless it can be shown that it does not induce bias!)
- Bootstrap approaches (built into random forests)
- Theoretical error estimates
  - Error estimates are not commonly available that account for tuning and/or variable selection effects.
  - Theoretical understanding helps structure thinking.

The Source/Target Issue

All assume the identity of target and source populations.

Accuracy Assessment, with respect to the target

- Accuracy measures mostly assume that results will be applied to the population that is the source of the data.
  - Academic papers rarely hint that accuracy will be reduced when results are used, e.g., to predict default rates for next years' applicants for credit.
- There is a dearth of data in the public domain that can be used to test how methods perform on realistic target populations that differ somewhat from the source.
  - Data sets that mix email spam with genuine email can however be readily collected and used for such an exercise.

Upscaling of Methodology – When is it an Issue?

- It is a large issue for methods that are computer-intensive and/or assume a complex error structure (time series, multi-level models, some types of image processing, . . .)
- For regression with iid errors, repeated samples typically give better insight than one large regression. (Why?)
- Summarization or other manipulation of data into a suitable form for analysis can be a huge challenge.

Methodologies for Low-Dimensional Representations

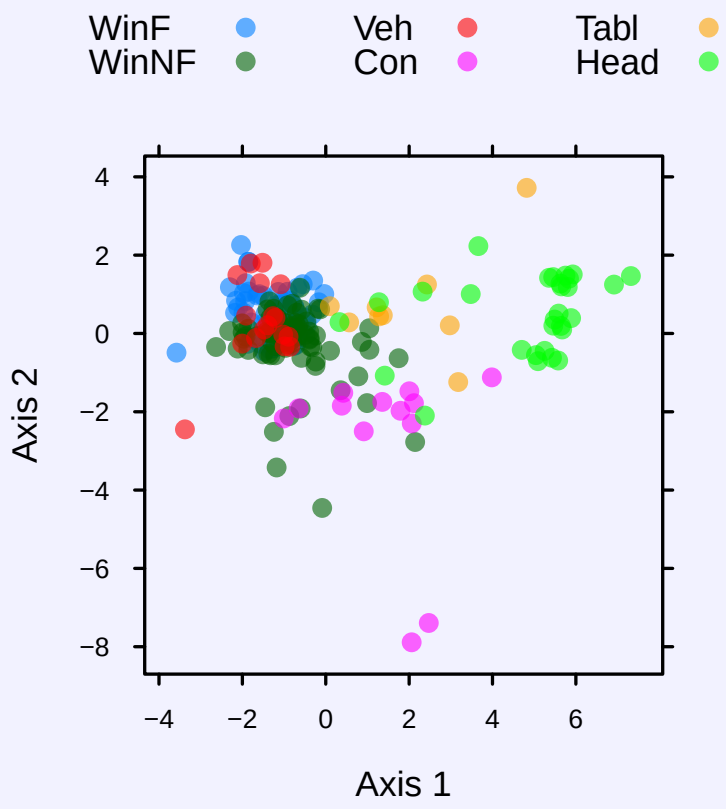
- From linear discriminant analysis, use the first two or three sets of discriminant scores
- Random forests yields proximities, from which relative distances can be derived.

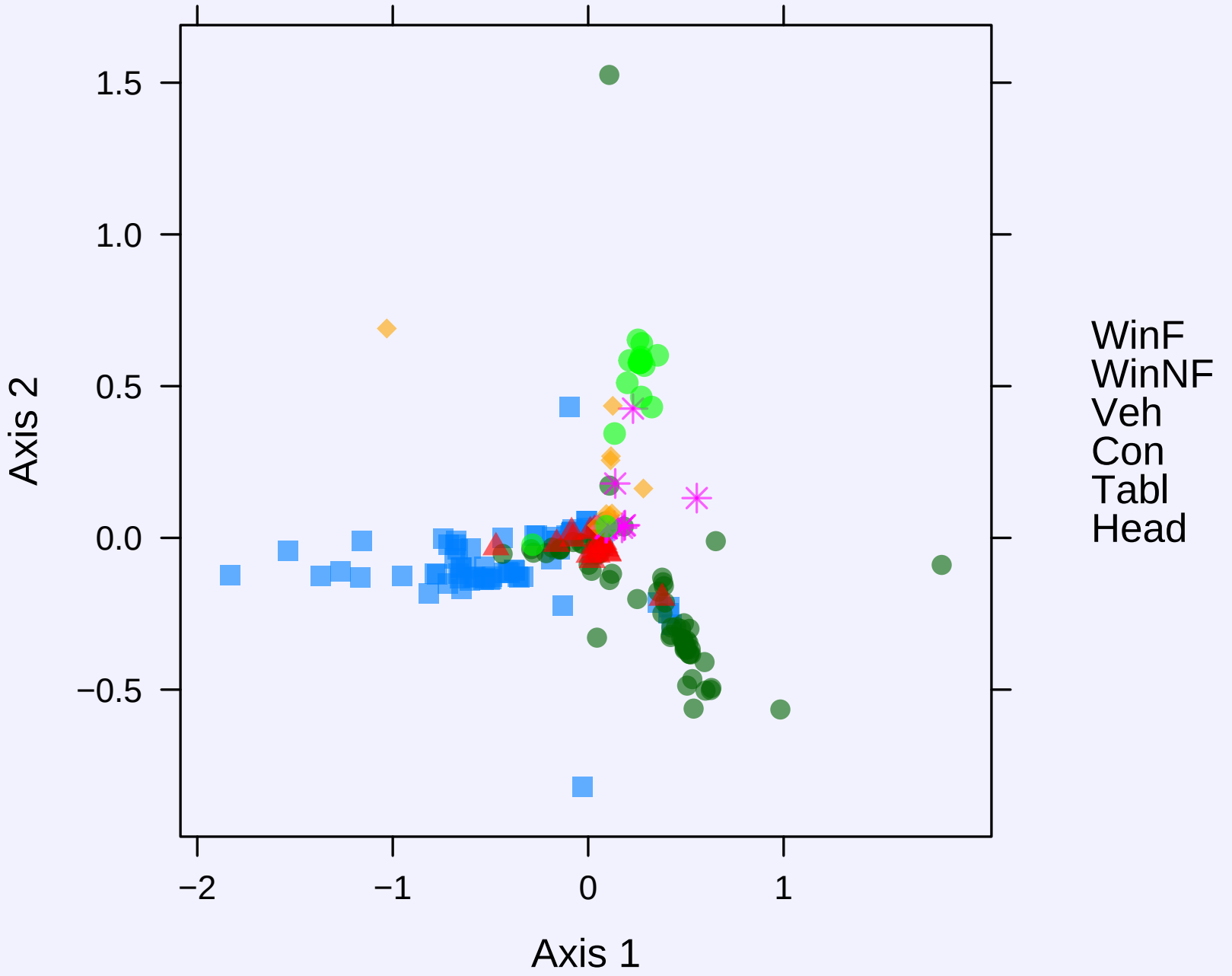
  Use semi-metric or non-metric multidimensional scaling (MDS) to obtain a representation in 2 or 3 dimensions.

  The *MASS* package has `sammon()` (semi-metric) and `isoMDS()` (non-metric) MDS.

The next two slides give alternative two-dimensional views of the forensic glass data, the first on linear discriminant scores, and the second based on the random forest results.[2]

[2] Code for these graphs will be placed on JM's webpage.


WinF
WinNF
Veh
Con
Tabl
Head
Axis 2
Axis 1
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6



Distances were from random forest proximities

Issues for Large Datasets

Distinguish the two cases:

- Numbers of points may be the challenge
  - There are effective ways to handle this problem (density plots, contour plots, perhaps with outliers shown separately), but implementation lags behind insight on what is required.
- High dimensional data (many variables) is a huge challenge
  - Different applications have different demands (e.g., variation in the number of dimensions that carry useful information).

There is a huge amount of work that aims to address these issues.

Advice to Would-be Data Miners – the Technology

Four classification methods may be enough as a start

- Use linear discriminant analysis (`lda()` in the *MASS* package) as a preferred simple method. The first two sets of discriminant scores allow a simple graphical summary.
- Quadratic linear discriminant analysis (`qda()` in the *MASS* package) can perform excellently, if the pattern of scatter is different in the different groups.
- Random forests (`randomForest()` in the *randomForest* package) can be used in a highly automatic way, does not overfit with respect to the source data, and will often outperform or equal all other common methods.
- Where complicated (but perhaps clearly defined) boundaries separate the groups, SVMs (`svm()` in the *e1071* package) may perform well.

Further Comment on the Forensic Glass Dataset

The random forest algorithm gave a rule for predicting the type of any new piece of glass. For glass sourced from the same "population", here is how the rule will perform.

	WinF	WinNF	Veh	Con	Tabl	Head	CE[3]
WinF (70)	63	6	1	0	0	0	0.10
WinNF (76)	11	59	1	2	2	1	0.22
Veh (17)	6	4	7	0	0	0	0.59
Con (13)	0	2	0	10	0	1	0.23
Tabl (9)	0	2	0	0	7	0	0.22
Head (29)	1	3	0	0	0	25	0.14

The data consist of 214 rows $\times$ 10 columns.

WinF = Window float WinNF = Window non-float
Veh = Vehicle window Con = Containers
Tabl = Tableware Head = Headlamps

[3] Classification Error Rate (cross-validation)

Questions, Questions, Questions, . . .

- How/when were data generated? (1987)
- Are the samples truly representative of the various categories of glass? (To make this judgement, we need to know how data were obtained.)
- Are they relevant to current forensic use? (Glass manufacturing processes and materials may have changed since 1987.)
- What are the prior probabilities? (Would you expect to find headlamp glass on the suspect's clothing?)

These data are not a good basis for making judgements about glass fragments found, in 2008, on a suspect's clothing. Too much will have changed since 1987.

But surely all this stuff can be automated?

Advice often credited to Einstein is:
"Simplify as much as possible, but not more!"

In the context of data analysis, good advice is:
"Automate as much as possible, but not more!"

Efforts at building automation into data mining software continue. But the extravagant promises of the early years of computing are still a long way from fulfilment:

> *1965, H. A. Simon: "Machines will be capable, within twenty years, of doing any work a man can do" !!!*

Even for such "easy" tasks as Word Sense Disambiguation (WSD), or automatic language translation, this has not happened! (See the relevant Wikepedia articles.)

Think about what automation has achieved in the aircraft industry, and the effort it required!

Analytics on autopilot?

"...analytical urban legends..." (D & H)

Sometimes, autopilot can work and is the only way!

. . . but there is a massive setup and running cost

Sources of Practical Insight – Observation vs Experiment

Much can be learned from comparing observational study results with randomized trial results in clinical medicine and elsewhere. An interesting paper is:

Hernn MA, et al. 2008. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease (with discussion). Epidemiology 19(6): 766-779

Question: Does Hormone Replacement Therapy (HRT) increase the risk of coronary heart disease (CHD)?

- Randomized study (WHI): Yes, HRT increases the risk.
- Observational study (NHS): No, HRT reduces the risk.

Hormone Replacement Therapy – Why the discrepancy?

Why the discrepancy?

1. Explanation 1: In the NHS observational study, users differed in some systematic way from non-users.
   - Consider, e.g., differences in socioeconomic status at birth.
2. Explanation 2: The populations were different. ✔

A recent reanalysis suggests explanation 2:

- In the randomized trial HRT was initiated, on average, a much longer time after menopause.
- NHS Women had on average used HRT for much longer.

The two sets of results are relevant to different **targets**!

Observe that the WHI randomized study gave the right answer, on average, for a group of women who are atypical (most started HRT subsequent to menopause).

Data Mining and R

- Data mining requires a powerful application-oriented language, supported by GUIs, that give access to standard and new methods. For this, R is unsurpassed!
- The R project is the ideal platform for supporting the analysis, graphics and software development activities of data miners and machine learners
  - Note the data mining community's much less ambitious Weka system. (Both Weka and R started in New Zealand!)
  - There is an R interface to Weka! In time, pretty much all non-commercial data analysis and graphics software of note gets incorporated into R!
- Note the *rattle* GUI interface that is aimed at data mining applications, developed by Graham Williams (JM's colleague).

References

Berk, R. 2008. *Statistical Learning from a Regression Perspective.*