A Statistical Perspective on Data Mining

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

July 2, 2009

Content of the Talk

- Data Mining (DM); Machine Learning (ML)
  – Statistical Learning as an Over-riding Theme
- From regression to Data Mining
- The Tradition and Practice of Data Mining
- Validation/Accuracy assessment
- The Hazards of Size: Observations? Variables?
- Insight from Plots
- Comments on automation
- Summary

Computing Technology is Transforming Science

- Huge data sets[1], new types of data
  - Web pages, medical images, expression arrays, genomic data, NMR spectroscopy, sky maps, ...
- Sheer size brings challenges
  - Data management; some new analysis challenges
  - Many observations, or many variables?
- New algorithms; "algorithmic" models.
  - Trees, random forests, Support Vector Machines, ...
- Automation, especially for data collection
  - There are severe limits to automated analysis
- Synergy: computing power with new theory.

Data Mining & Machine Learning fit in this mix.
Both use *Statistical Learning* approaches.

[1] cf, the hype re *Big Data* in Weiss and Indurkha (1997)

DM & ML Jargon – Mining, Learning and Training

Mining is used in two senses:

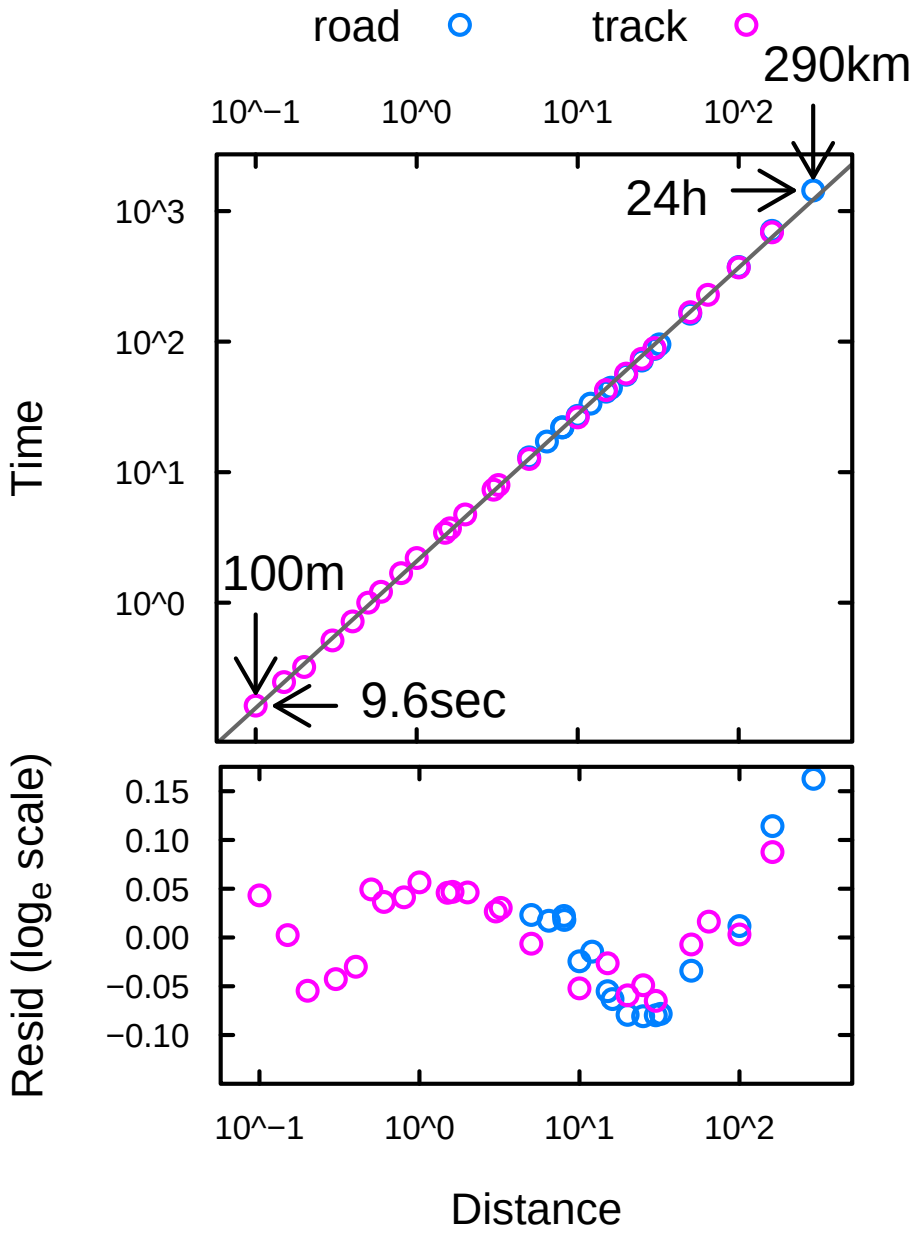
- Mining for data
- Mining to extract meaning, in a scientific/statistical sense.

Pre-analysis data extraction & processing may rely heavily on computer technology. Additionally, design of data collection issues require attention.
In mining to extract meaning, statistical considerations are pervasive.

Learning & Training

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

Example 1: Record Times for Track and Road Races



Is the line the whole story?

Ratio of largest to smallest time is ∼3000.

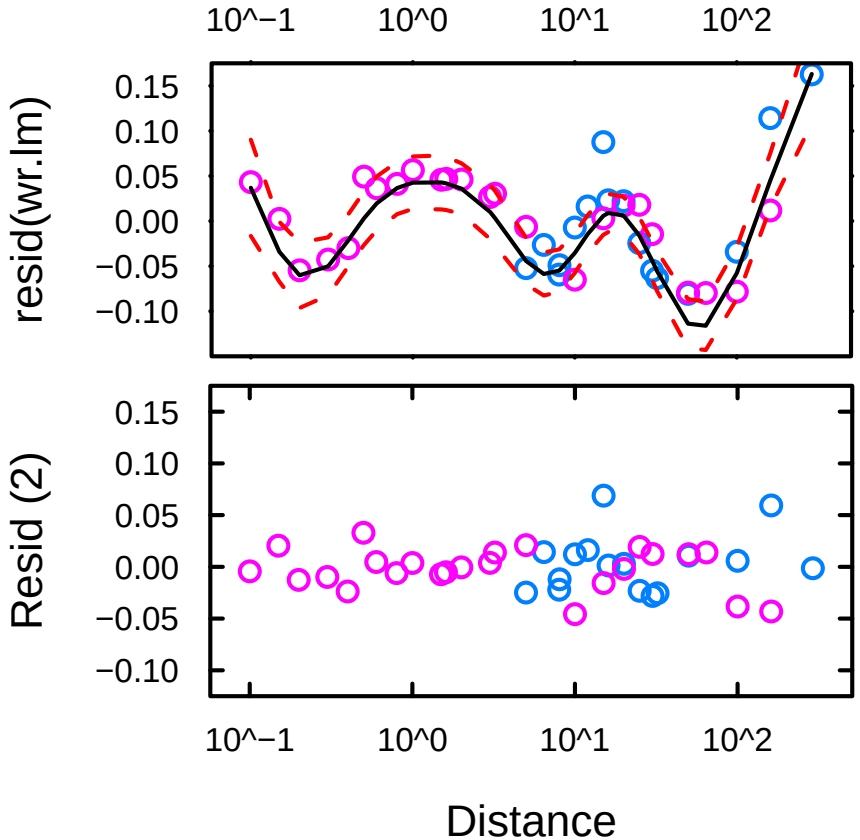
A difference from the line of >15%, for the longest race, is not visually obvious!

The Plot of Residuals

Here, differences from the line been scaled up by a factor of ∼20.

World record times – Learning from the data

Fit a smooth to residuals.
('Let the computer rip'!)
The lower panel shows
residuals from the smooth.



Is the curve 'real'?

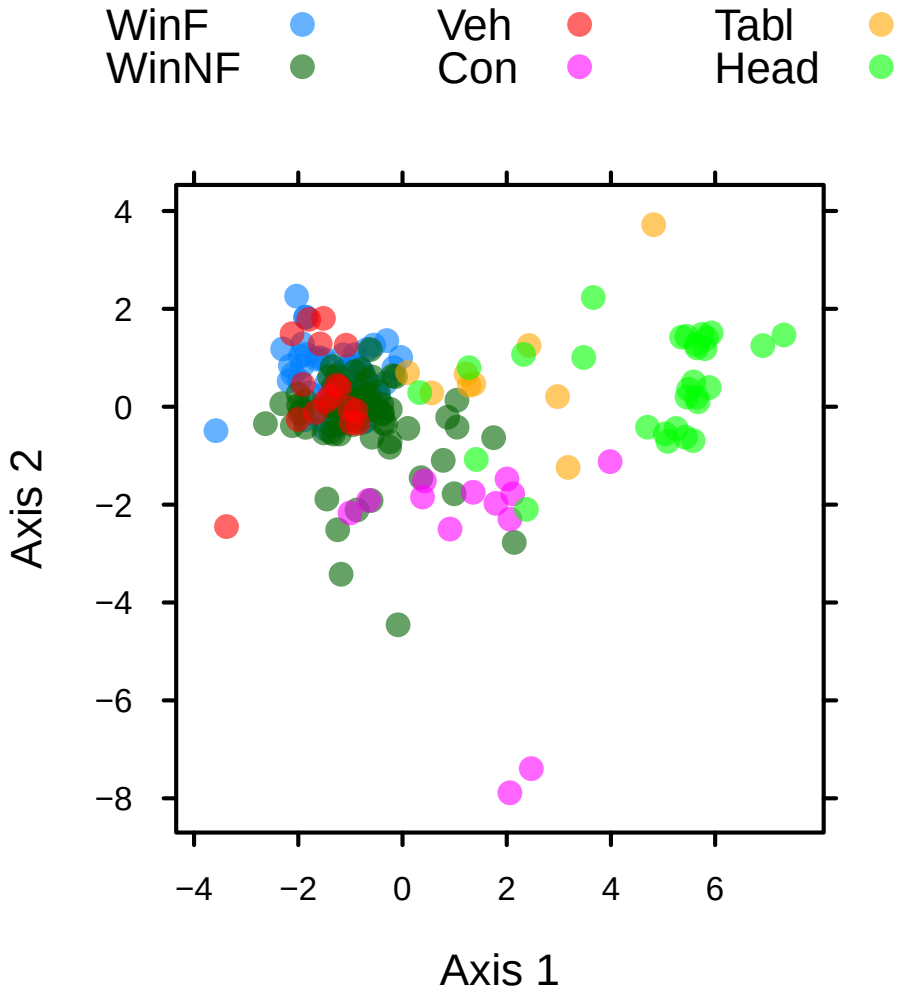
The algorithm says "Yes".
Questions remain:

- 'Real', in what sense?
  - 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?

Martians may have a somewhat different curve!

Example 2: the Forensic Glass Dataset (1987 data)

Find a rule to predict the type of any new piece of glass:



Graph is a visual summary of a classification.

- 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. (214 rows × 10 columns.)

NB: These data are well past their "use by" date!

Statistical Learning – Commentary

1. Continuous Outcome (eg, Times vs Distances)

- Allow data to largely choose form of response.
- Often, use methodology to refine a theoretical model (large datasets more readily highlight discrepancies)

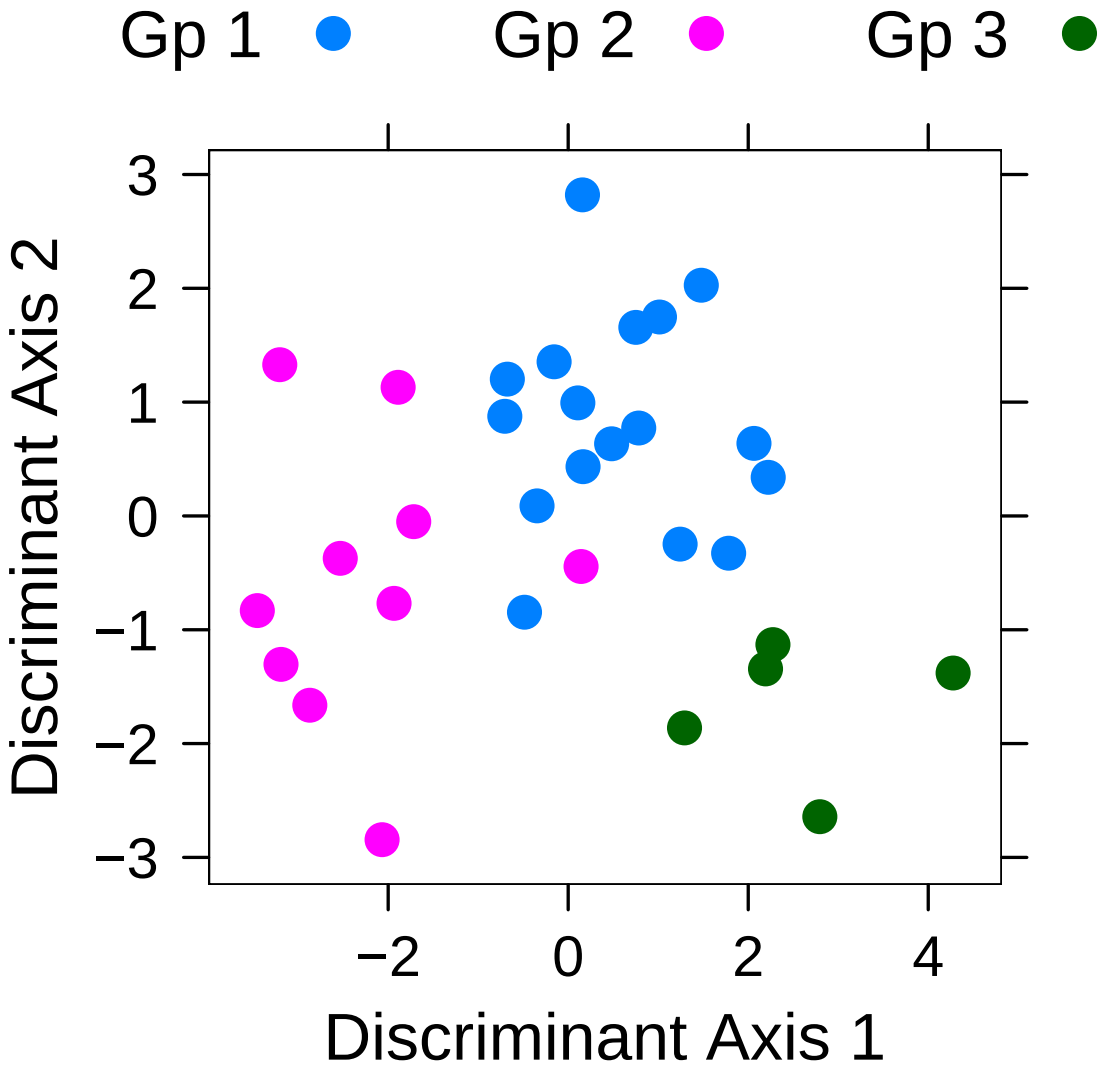
2. Categorical Outcome (eg, forensic glass data)

- Theory is rarely much help in choosing the form of model.
- Theoretical accuracy measures are usually unavailable.

Both Types of Outcome

- Variable selection and/or model tuning wreaks havoc with most theoretically based accuracy estimates. Hence the use of empirical approaches.

Selection – An Aid to Seeing What is not There!



Method

Random data, 32 points, split randomly into 3 groups .

(1) Select **15** 'best' features, from **500**.

(2) Show 2-D view of classification that uses the 15 'best' features.

NB: This demonstrates the usefulness of simulation.

Selection Without Spurious Pattern – Train/Test

Training/Test

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

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.

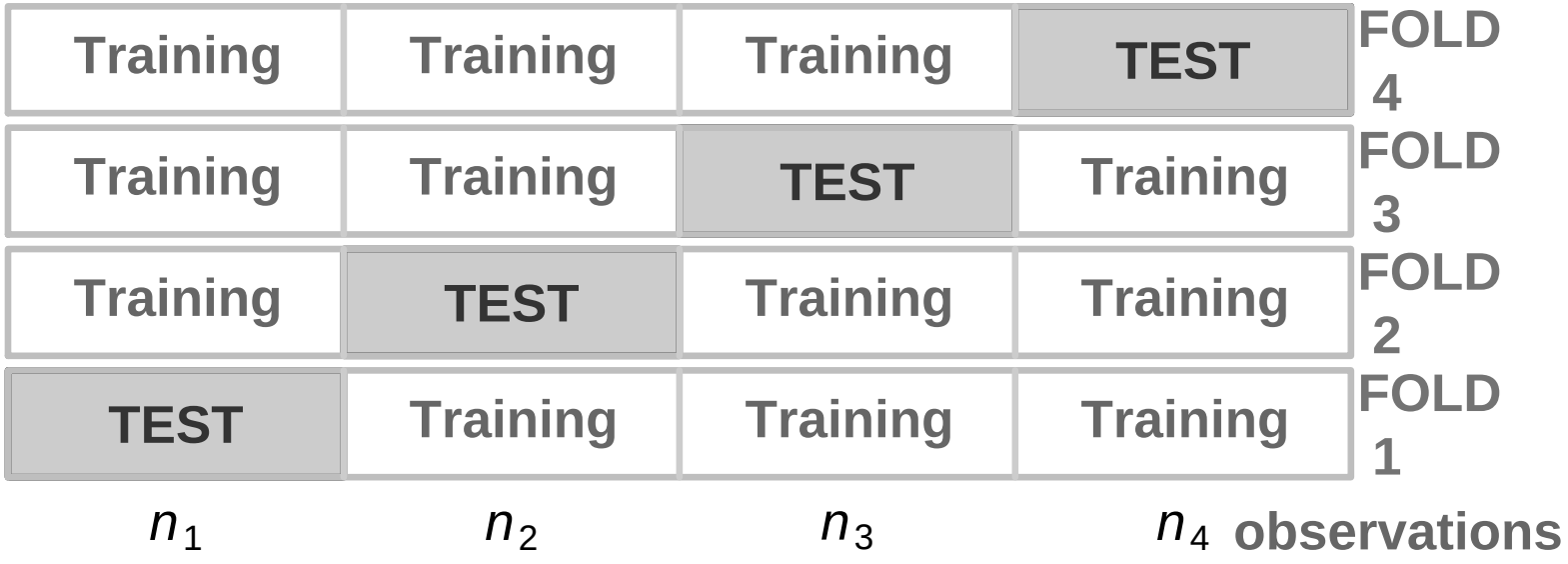
Warning – What the books rarely acknowledge

- All methods give accuracies based on sampling from the source population.
- For observational data, target usually differs from source.

Cross-Validation

Steps

- Split data into k parts (below, 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.



Selection Without Spurious Pattern – the Bootstrap

Bootstrap Sampling

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)

Bootstrap Sampling – Putting it to Use

- 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.

JM's Preferred Methods for Classification Data

Linear Discriminant Analysis

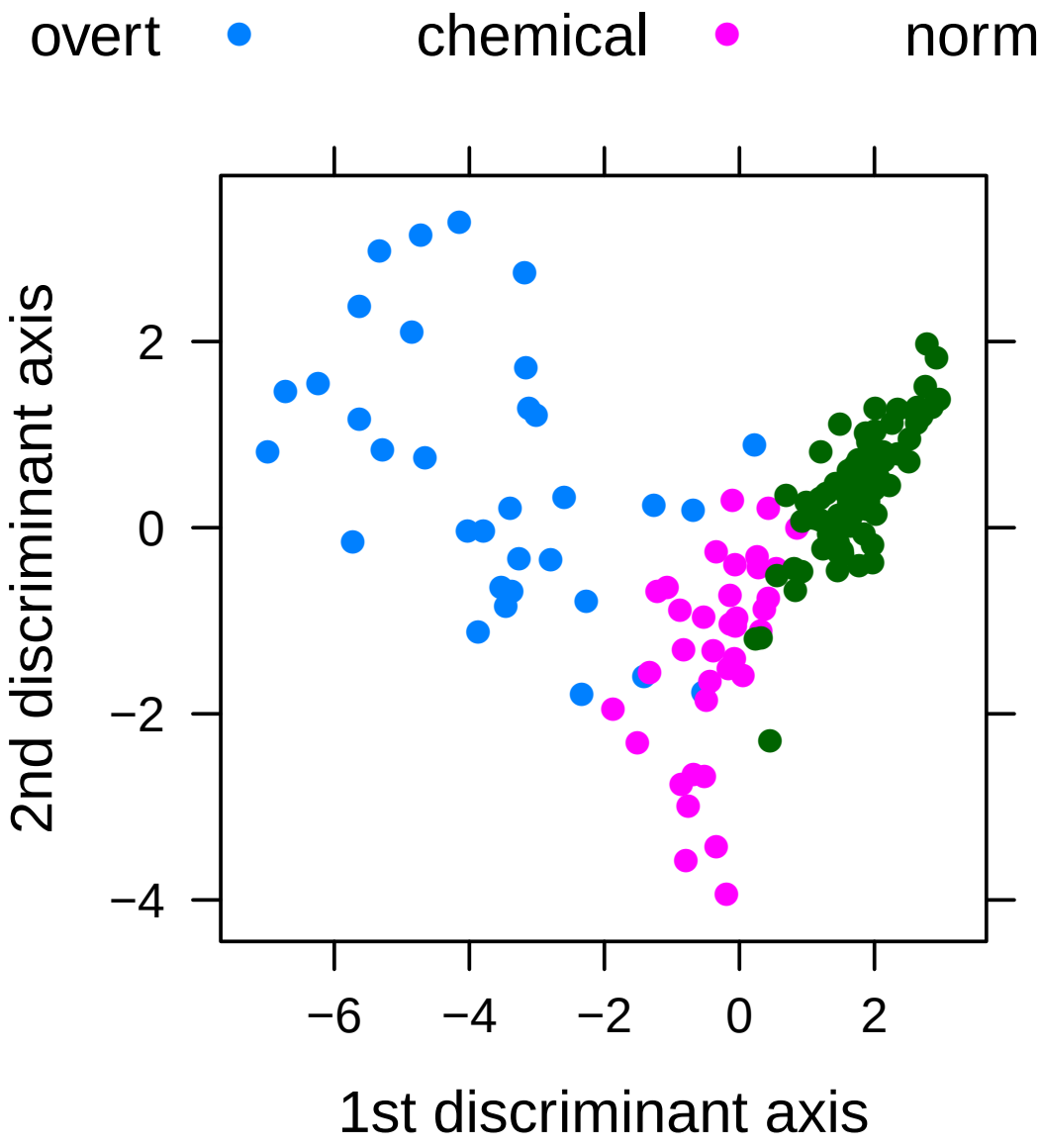
- It is simple, in the style of classical regression.
- It leads naturally to a 2-D plot.
- The plot may suggest trying methods that build in weaker assumptions.

Random forests

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

Use for illustration data with 3 outcome categories

Example (3) – Clinical Classification of Diabetes

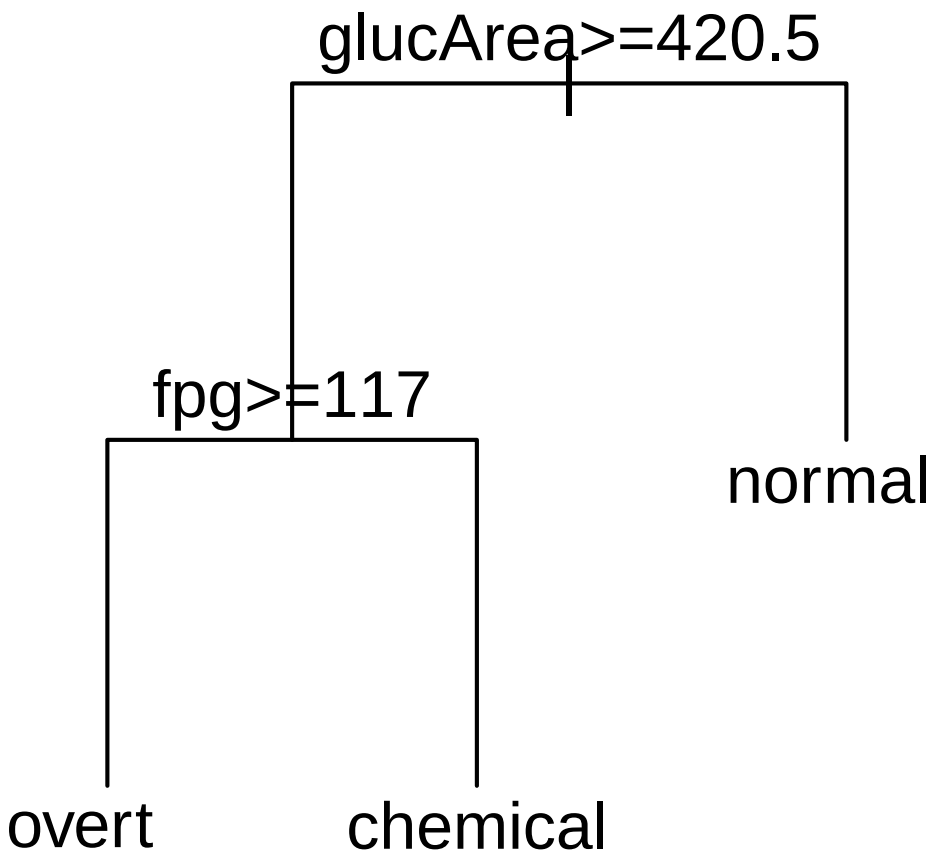


Predict class (overt, chemical, or normal) using relative weight, fasting plasma glucose, glucose area, insulin area, & SSPG (steady state plasma glucose).

Linear Discriminant Analysis

CV acc = 89%

Tree-based Classification

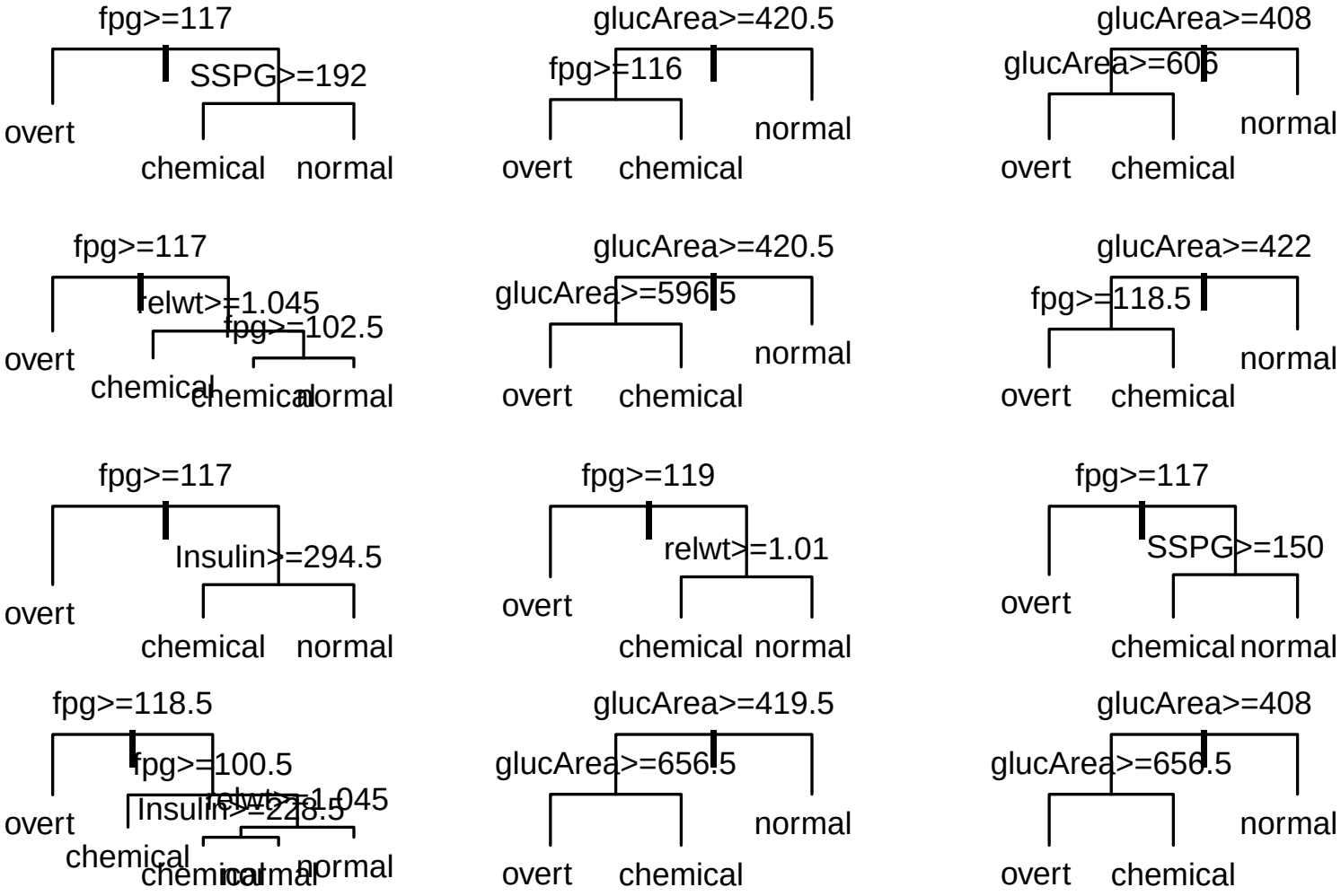


Tree-based Classification

CV acc = 97.2%

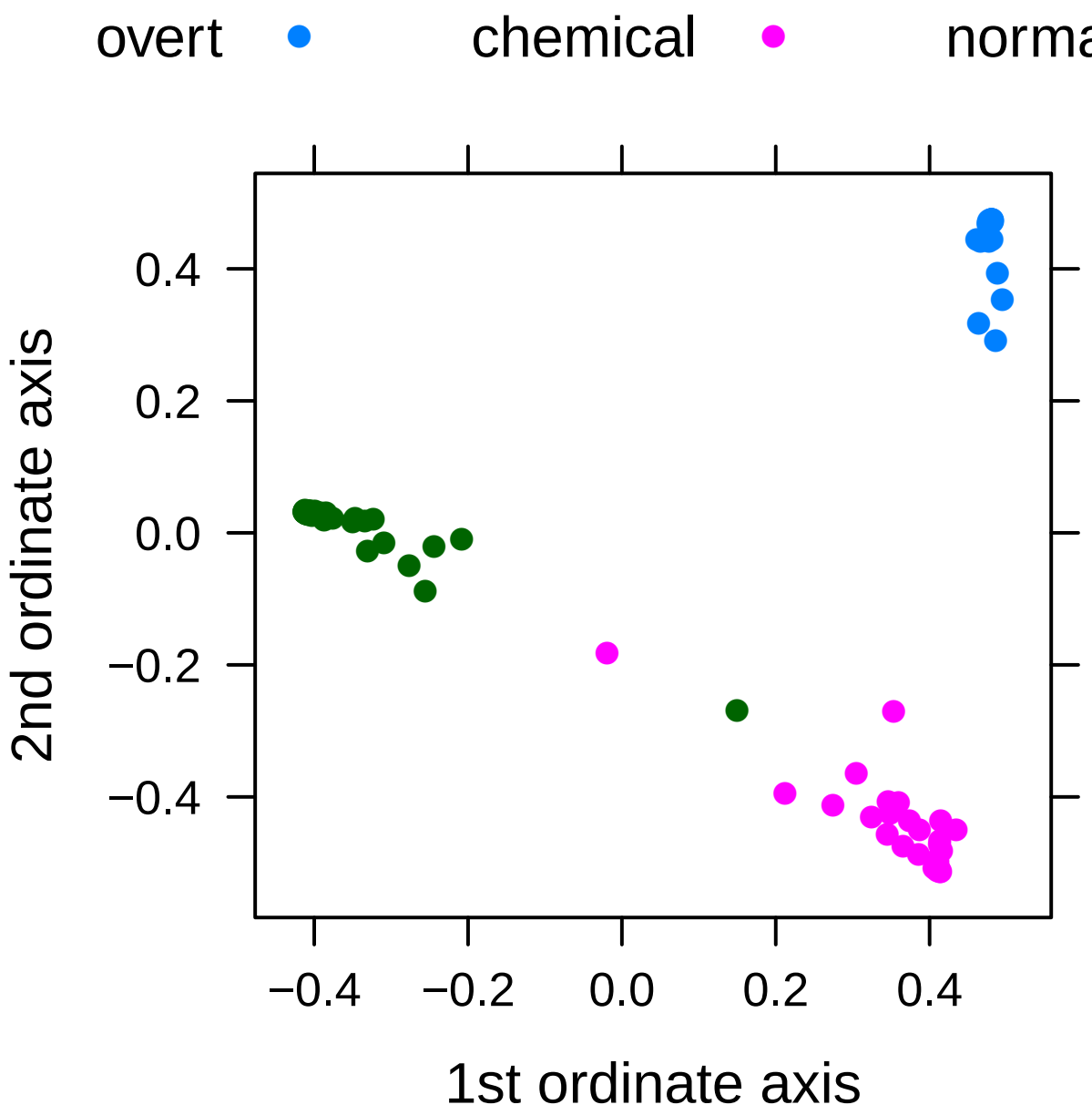
Random forests – A Forest of Trees!

Each tree is for a different random with replacement ('bootstrap') sample of the data, and sample of variables



Each tree has one vote; the majority wins

Plot derived from random forest analysis



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.

When are crude data mining approaches enough?

Often, rough and ready approaches, ignoring distributional and other niceties, work well!

In other cases, rough and ready approaches can be highly misleading!

Some Key Issues

- Watch for source/target (eg, 2008/2009) differences.
- Allow for effects of model and variable selection.
- Do not try to interpret individual model coefficients, unless you know how to avoid the traps.

Skill is needed to differentiate between problems where relatively cavalier approaches may be acceptable, and problems that demand greater care.

Different Relationships Between Source & Target

Source versus target	*Are data available from target?*
1: Identical (or nearly so)	Yes; data from source suffice
2: Source & Target differ	Yes
3: Source & Target differ	No. But change can be modeled (cf: multi-level models; time series)
4: Source & Target differ	No; must make an informed guess.

Where source & target differ, it may be possible to get insight from matched historical source/target data.

There are major issues here, which might occupy several further lectures!

The Hazards of Size: Observations? Variables?

Many Observations

- 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.

Many variables (features)

Huge numbers of variables spread information thinly!

This is a challenge to the analyst.

The Hazards of High Dimensional Data

- Select, or summarize?
  - For selection, beware of selection effects – with enough lottery tickets, some kind of prize is pretty much inevitable
  - A pattern based on the 'best' 15 features, out of 500, may well be meaningless!
  - Summary measures may involve selection.
- Beware of over-fitting
  - Over-fitting reduces real accuracy.
  - Preferably, use an algorithm that does not overfit with respect to the source population.
  - Unless optimized with respect to the target, some over-fitting may be inevitable!
  - Any algorithm can be misused to overfit! (even those that do not overfit!)

Why plot the data?

- Which are the difficult points?
- Some points may be mislabeled (faulty medical diagnosis?)
- Improvement of classification accuracy is a useful goal only if misclassified points are in principle classifiable.

What if points are not well represented in 2-D or 3-D?

One alternative is to identify points that are outliers on a posterior odds (of group membership) scale.

New Technology has Many Attractions

Are you by now convinced that

- Data analysis is dapper
- Data mining is delightful
- Analytics is amazing
- Regression is rewarding
- Trees are tremendous
- Forests are fascinating
- Statistics is stupendous?

Some or all of these?

Even however with the best modern software, it is hard work to do data analysis well.

The Science of Data Mining

- Getting the science right is more important than finding the true and only best algorithm! (There is none!)
- Get to grips with the statistical issues
  - Know the target (1987 glass is not 2008 glass)
  - Understand the traps (too many to talk about here[2])
  - Grasp the basic ideas of time series, multi-level models, and comparisons based on profiles over time.
- Use the technology critically and with care!
- Do it, where possible, with graphs.

[2] Maindonald, J.H. (2006) notes a number of common traps, with extensive references. Berk (2006) has excellent advice.

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)'[3]

Maindonald, J. H. and Braun, W. J. 2007. *Data Analysis and Graphics Using R – An Example-Based Approach.* 2nd edn, Cambridge University Press.[4]
[Statistics, with hints of data mining!]

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

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

Web Sites

http://www.sigkdd.org/
[Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining.]

http://www.amstat.org/profession/index.cfm?fuseaction=dataminingfaq
[Comments on many aspects of data mining.]

http://www.cs.ucr.edu/~eamonn/TSDMA/
[UCR Time Series Data Mining Archive]

http://kdd.ics.uci.edu/ [UCI KDD Archive]

http://en.wikipedia.org/wiki/Data_mining
[This (Dec 12 2008) has useful links. Lacking in sharp critical commentary. It emphasizes commercial data mining tools.]

The R package *mlbench* has "a collection of artificial and real-world machine learning benchmark problems, including, e.g., several data sets from the UCI repository."