Statistical Learning and Data Mining Course —

ACSPRI 2011 Spring Program,

University of Technology Sydney, September 26-30

http://www.acspri.org.au/springprogram2011

Tentative Schedule

John Maindonald

Issues that will be emphasized throughout include:

• source/target, noting especially implications of temporal and spatial change,

• data accuracy and integrity,

low-dimensional visual summary,

realistic comparison of alternative methods,

limits on the meaningful interpretation of model parameters,

the limits of automation, noting in particular the crucial reliance of many of the methods on independence assumptions. (Time series data is an important special case where the independence assumption should be questioned.)

Monday Session 1:

Check R installations

Regression and classification examples that illustrate statistical learning ideas Overview of course content

 \triangleright Session 2

Introduction to R, using both the Rattle GUI and the command line

The Data Frame — the data structure commonly used for calculations with small to medium sized data sets

Distributions, and graphical display of distributions, with columns from datasets that will be used later in the course.

 \triangleright Session 3

Review of Sessions 1 and 2

Simple regression examples, used as a basis for further introduction to R, for revision of regression methodology, and for further exposure to R graphics

Regression in R – model formulae, and interpretation of R output.

 \triangleright Session 4

Regression calculation preliminaries – exploratory data analysis (this will at the same time be an introduction to lattice and other R graphics

Tuesday

 \triangleright Session 1:

Review of Monday

R data structures — overview and review Lattice graphics — review and further details

- \triangleright Session 2:
 - Populations and samples

Regression with a continuous outcome variable — revision of key ideas: interpretation of coefficients; regression diagnostics; leverage and why it matters; the very helpful (when you can use it) term plot; the use of transformations; comments on stepwise regression (use only if you know the traps and can avoid them!) and other variable selection techniques; multicollinearity.

 \triangleright Session 3:

• Detecting where regression smoothing terms may be required.

Cross-validation, bootstrap and test set accuracy estimates.

Regression with very large datasets — lots of data may not bring the large benefits that are often claimed, and lots of data does bring traps for the unwary

- \triangleright Session 4:
 - Logistic regression (with a binary outcome variable)

Poisson regression (with an outcome variable that is a count).

All the same questions that we asked about regression with a continuous outcome variable carry across to this context.

Wednesday

Session 1

- Review of Tuesday
- Splines, and spline terms in regression
- NB: Spline fits should mostly be a last recourse, after possibilities for transformation to linearly related covariates have been investigated and found inadequate to the task.

Smooth terms in regression models, with automatic choice of smoothing parameter – examples with a continuous outcome variable.

The importance of independence — what happens if there is temporal dependence? \triangleright Session 2

- Smooth terms in regression models examples with binary and count outcome variables.
- Further comments on regression methods.

 \triangleright Session 3

• Review of sessions 1 and 2

Accuracy and other performance criteria

The limitations of \mathbb{R}^2 as a performance measure, and alternatives

Training set, validation set, and test set

Cross-validation and bootstrap accuracy measures.

 \triangleright Session 4

- Logistic regression as a classification method when there are two outcomes
- Extending logistic regression to the multinomial case multinom() in the *nnet* package, mlogit() in the *mlogit* package and (now set in a Bayesian framework) lda() in the *MASS* package.

Thursday

 \triangleright Session 1

- Review of Wednesday, noting especially ideas of training/test set, cross-validation and bootstrap accuracy measures.
- Linear and quadratic discriminant analysis (lda() and qda(), both from the MASS package)

Plots derived from lda()

 \triangleright Session 2

- Linear and quadratic discriminant analysis, continued
- Tree-based regression (using rpart() from the *rpart* package)

Random forests (using randomForest() from the *randomForest* package)

- \triangleright Session 3
 - An overview of other popular methods limiting attention however to those that are available in the *rattle* package.
- \triangleright Session 4

• Distance measures

Ordination (low-dimensional views, usually 2 or 3 dimensions, of high-dimensional data)

Friday

 \triangleright Session 1

- Review of Thursday
- Ordination methods classical multi-dimensional scaling (metric extensions of principal components, implemented using cmdscale()) and non-metric scaling (using isoMDS() from the MASS package)

Note that lda () provides an approach to ordination that takes account of prior knowledge. Use of prior knowledge can lead to a much better end result.

- Use of ordination methodology to obtain a low-dimensional representation from output from any classification method
- \triangleright Session 2
 - Classification Multiple observations on individual units what are the implications for the potential for over-fitting and for predictive accuracy? The Vowel dataset in the *mlbench* package will be used for illustration.
 - Commentary: This (hierarchical of variation) is one of a number of types of structure that may be found in data, and that have applications for modeling. Other common types are: temporal (time series), and spatial. These different types may be combined. In each instance, there are implications both for modeling and for assessment of predictive accuracy.
- \triangleright Session 3
 - Text mining using the *tm* package (only if there is sufficient interest)
- \triangleright Session 4
- Overview of topics left over from earlier sessions

Further examples, if time permits

Wrap-up and review.

B: Further packages that may need to be installed:

SMIR : Required to give access to the bronchit dataset (Subsection 3.4.3 and Section 6.1)

forecast : Required to allow use of the function **forecast()** and related functions in Subsection 7.2.1. If installed from the menu, the dependency *fracdiff* should be installed at the same time. Otherwise, it may need to be installed separately.

DAAGbio : Required to give access to the *DAAGbio* packages that features in Subsection 10.3.1. If necessary, install also the dependency *limma*, and *Biobase* which is a *limma* dependency.