Statistical Learning and Data Mining — Content and Emphases

John Maindonald

The course will emphasize

- source/target considerations, noting especially implications of temporal and spatial change,
- data accuracy and integrity,
- low-dimensional visual summary,
- realistic comparison of alternative methods,
- limits on the meaningfully 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.)

The main methods will be

- classical regression with a continuous outcome variable,
- logistic & related Generalized Linear Models,
- multinomial logistic regression,
- linear and quadratic discriminant analysis,
- trees,
- random forests.

There will be briefer discussion of

- Support Vector Machines,
- Boosting methods, notable AdaBoost.

This course will not try to cover a wide range of methodologies; rather it will emphasize careful and critical use of whatever methodologies are applied. (For what it is worth, my experience has been that where linear or quadratic discriminant analysis does not work well, random forests is hard to beat. Note however that this judgement is mostly based on cross-validation evidence or its training/test equivalent, not on the really crucial test of accuracy on the real target!).

Data mining techniques requires the same carefulness in data preparation, the same attention to methodological care, and the same critical scrutiny, as any other style of

data analysis. A high level of automation is sometimes possible and reasonable, without too much attention to statistical niceties. The circumstances must however be favorable, and the limitations should be understood and taken on board.

Participants will be encouraged to bring to the discussion any experience of their own of methodologies that have proved effective, explaining the evidence on which that conclusion is based.