Software for multivariate classification

Ilya Narsky, Caltech

What is available

- R – lots of methods (but slow)
- WEKA (Java) – lots of methods
- MatLab modules – lots of methods (?)
- various Neural Net implementations
  - Stuttgart Neural Network Simulator (standard backprop, RBF, Kohonen maps, DVQ etc)
  - JetNet (?)
  - ….
- StatPatternRecognition (C++)
  - DT, boosting, RF, arc-x4, BH, LQDA, interfaces to Stuttgart NN's
- TMVA (Root module)
  - BDT, LH, NN (2 impls), nearest neighb. method, LDA, rect. cuts
- m-boost (C) – BDT and RF
- Lots of others packages I don't know much about

How analysts choose software

- Analyst knows little but has some basic requirements
- Criterion most often exercised in practice:
  - use whatever your officemates use
  - …because they have experience with this software and can help you out
- If you want to start from scratch, you need to:
  - install the package
  - run “Hello, world” application
  - learn more advanced features
  - …and then discover that the package can only handle small datasets within an infinite amount of time

We need edmunds.com for software

- Select several packages for multivariate classification (open to suggestions)
- …and compare these features
  - versatility (what methods are implemented)
  - ease of installation
  - quality of manuals and documentation
  - CPU speed and how it scales vs sample size and dimensionality, both for training cycle and post-training classification; also max sample size and dimensionality
  - types of inputs that can be handled (real, integer, categorical, mixed etc)
  - graphics interface, both for input (GUI) and output
  - interactive analysis or batch jobs
  - ease of integration in the C++ framework

Compare classification error?

- I am willing to make an assumption that as long as a certain method is implemented, it is implemented in a more or less reasonable way
  - For example, I optimized SPR Random Forest on ~80k events in 4D, then plugged the optimal params into RF implementation in R
  - Outcome: R gives me a ~20% worse performance than my RF, although I'd expect that it be identical (and also is slower by an order of magnitude)
- But I don't care about 20% effects! What would you do – publish 2 analyses that are 20% suboptimal than one that mastered the ultimate optimality?
- In the first approximation – don't bother as long as there are no obvious issues.

Plea for manpower

- This is a project for a fresh graduate student or an undergrad
- I already have an undergrad working on StatPatternRecognition and I was promised a grad student starting in October. But they are in for long-term higher priority projects.
- This is a short-term project and can be a fun exercise. Could be presented at PhystatXX.