Using R for Predictive Analytics

Szilárd Pafka

Predictive Analytics World / DC useR Group

October 20, 2009
Introduction and agenda

Introduction:
- physics, finance, statistical modeling (PhD)
- data mining at a credit card processor (Chief Scientist)
- co-organizing the Los Angeles area R users group

Agenda:
- using R for predictive analytics
- benefits from R users groups
R and predictive analytics

R:

- data manipulation
- statistical analysis, exploratory data analysis
- computations, simulations
- modeling
- visualization

Predictive analytics (data mining for prediction problems):

- understanding the domain problem and related questions
- data exploration and cleaning
- building predictive models (supervised learning)
- model diagnostics, selection and evaluation
- deployment
R vs alternatives

R:
- open source
- packages
- latest cutting-edge methods
- command line interaction (vs GUI)
  - flexible
  - results easier to reproduce
- learning curve
  - books
  - R-help
- cost, multiple platforms

alternatives:
- spreadsheets
- programming languages
- data mining software
- similar software environments
Data exploration

- data frames
- import from various sources (csv files, database connection)
- flexible data manipulation functionalities
  - subscripting
  - numerous transformations
  - text manipulation (e.g. regexp)
  - data aggregation
  - specialized packages (reshape, plyr etc.)
- graphics
- sample R code
Visualization

- powerful tool for data exploration, diagnostics etc.
- R graphics: base, lattice, ggplot2
- visual perception, principles of visual design
- ggplot2
  - based on the grammar of graphics
  - nice defaults
  - elements: mappings, geoms, stats, scales, facets
  - very expressive (sample code)
Building predictive models

- supervised learning: \( y = f(x_1, x_2, \ldots, x_p) \)
- regression (numeric \( y \)), classification (categorical \( y \))
- spam detection example: classification (2 classes, Y/N)
- numerous methods: logistic regression, LDA, nearest neighbors, decision trees, bagging, boosting, random forests, neural networks, support vector machines etc.
- example: neural networks with \texttt{nnet} (sample code)
  - weight decay as regularization
  - helps convergence and against overfitting
Model evaluation

- package caret: unified interface/wrapper for
  - training models, tuning (complexity parameters)
  - diagnosing, selecting, evaluating models
- model deployment
- **Conclusion**: R is a viable tool for predictive analytics
R users groups

- R users groups around the world, in the US: SF, NY, LA, DC...
- Los Angeles: 150+ members, 4 meetings so far

  - talks, Q&A, discussions, exchange of knowledge
    - pointing people where to look for (packages, docs etc.)
    - also more generally for statistical methodology issues
  - networking opportunities