HS616_Syllabus

Robert Horton

October 15, 2014

Statistical Computing for Biomedical Data Analytics

Description

This is an intensive introduction to statistical computing with R. Programming assignments will draw from a wide range of computational and applied mathematical concepts required for biomedical data analytics, including probability, statistics, linear algebra, optimization, data manipulation, visualization, linear modeling and model diagnostics.

Overview

This course is designed to develop core computing skills, to complement traditional coursework in mathematical statistics, and to lay the groundwork for further study of data analytics. Students will learn to use the R statistical programming language and environment by solving structured problems in a wide range of application areas relevant to biomedical data analysis.

The programming exercises will emphasize interaction between R and other computational resources, particularly SQL databases, networked resources, the Linux environment, and Python. Special emphasis will be placed on approaches that scale to large data sets.

Objectives

On completing this course, students will be able to:

- take advantage of the large number of software modules available from the R language
- merge, filter, and arrange data into formats suitable for regression or machine learning analysis
- create sophisticated graphs of functions or data
- use linear algebra approaches to find optimal solutions for systems of equations
- be able to run advanced SQL queries (including GIS extensions, stored procedures, and in-database analytics)
- develop and deploy interactive visualizations and simple data products
- use scripting and literate analysis tools to create fully documented and reproducible data analysis workflows

Recommended texts

- R in Action: This is a good introduction to using the R environment for doing analysis; it does not focus on programming. Note that there will be a new edition coming out this spring.
- An Introduction to R: The "sample session" in Appendix A is a useful tutorial; otherwise, this book is mostly good for reference.
- Advanced R: also available online. This is a new text written by the author of many of the packages we will focus on (ggplot2, tidyr, dplyr, etc). It is not an introductory book, but we will probably use a few sections.

Websites

- For asking questions
  - StackOverflow
  - RSeek: interface to Google that focuses on R.
- Reference
  - Reference Card
- Getting code
  - CRAN The Comprehensive R Archive Network is the main repository for R packages.
  - Bioconductor is a separate repository for bioinformatics packages
  - GitHub, especially the course repository and Hadley Wickham's site ## Syllabus

This is a rough draft. It may make more sense to change the order of some lessons; in particular, we may want to go back and forth between visualization and analysis, and tease them with enough analysis to motivate data munging. I'm planing to informally introduce classification examples during the data manipulation section (probably starting with a simple classifier like kNN to minimize the magic). We may also need to move some discussion of analysis earlier in the course to support the students taking the additional project credit.

Section 1: R Programming Idioms

- Week 1: Idioms matter
  - How to baffle R: the infamous "loop that allocates memory" ploy
  - Profiling tools: plotting profiling results
    - vectorization
    - parallelization
    - lists vs. hash tables
  - Set up github accounts
- Week 2: Probability
  - Pseudorandom numbers
  - Rolling dice
    - simulating the Binomial distribution
  - Curve fitting
    - using lm and formulas to fit a polynomial to a set of points
    - parameter scanning to fit bell curve to Binomial simulation results
  - Sensitivity and specificity exercise
    - structuring code with functions
    - organizing data into dataframes
- Week 3: Calculus and linear algebra review
  - Checking derivatives
    - functions as parameters
  - Numerical integration
    - Area under (part of) a bell curve
  - Matrix manipulation
    - Gaussian elimination

- Week 4: Hypothesis testing
  - Distributions under null and alternative hypotheses; alpha, beta
    - Interactive graphs using "manipulate"
  - Multiple testing adjustments
    - Bonferroni, Benjamini-Hochberg

Section 2: Loading and Munging Data

- Week 5: Spreadsheets and CSV files
  - Reading into data frames
- Week 6: Using SQL from R
  - sqldf: bidirectional promoters in a mammalian genome
    - self-join on a table of transcription start sites.
  - Direct database access
- Week 7: Tidy Data: rows are examples (with outcome), columns are attributes
  - Classification example
  - Introduction to dplyr & tidyr
- Week 8: Merging and aggregating with dplyr & tidyr
  - Another classification example (from merged data)
  - **Midterm Exam**

Section 3: Visualization and User Interaction

- Week 9: ggplot2
- Week 10: Clustering and Graphs
  - Heat maps
  - Graphing graphs
- Week 11: Shiny
  - Interactive ODE model
  - Interactive analysis
- Week 12: Visualizing Regression
  - regression as an optimization problem
  - interactive visualization of line, data, and cost function in parameter space.

Section 4: Multivariate Analysis

- Week 13: Linear models
  - Feature selection
  - Performance profiling revisited
  - Generalized linear models
  - Regularization
- Week 14: Interactions
  - Simpson's paradox

  - categorical inputs
  - Dimension Reduction
- Week 15: Diagnostics
  - Model diagnostics
  - Prediction diagnostics
    * ROC curves: sensitivity/specificity tradeoff
    * Learning curves: bias/variance tradeoff
  - Hypersonic survey of Machine Learning algorithms
- Week 16: Finals Week
  - **Final Exam**
  - Project presentations