Stat 133: Concepts in Computing with Data

Aug 30, 2004

This course will introduce you to various aspects of computing to conduct statistical analysis. It will NOT focus on computational aspects of statistical methods.

**DATA Technologies** – You will need to work to get data, work with data, and work with the data “owners”. Learn how to think about the *data process*

- data acquisition – Input/output, regular expressions
- data cleaning and verification – graphics, exploratory data analysis
- data organization – data frames, XML, databases
- data as a pseudo population – bootstrap, cross-validation,

**Statistical Computing Toolbox** You will also learn many skills.

- R - statistical software
- SQL - relational database language
- XML - Extensible Markup language
- Unix commands
- Programming concepts - e.g. loops, recursion, trees
- Regular expressions
- Grammar of graphics
This course is NOT:

- A course in **Computational Statistics**,  
  For example, we will not cover various algorithms for computing least squares solutions and inverting matrices

- A course in **Applied Statistics**  
  For example, we will not learn a battery of statistical methods such as ordinary least squares, weighted least squares, and general linear models, etc.

- A course in **Mathematical Statistics**  
  For example we will not cover the expectation, variance, and large sample properties of least squares estimators.

1 An Example: SPAM

- SPAM = Unsolicited, mass, junk email.
- > 50% of electronic mail is SPAM
- Offensive, time-consuming
- Want to be able to identify SPAM before we read it.

Here’s one email message:

Return-Path: whisper@oz.net  
Delivery-Date: Fri Sep 6 20:53:36 2002  
From: whisper@oz.net (David LeBlanc)  
Date: Fri, 6 Sep 2002 12:53:36 -0700  
Subject: [Spambayes] Deployment  
In-Reply-To: <LNBBLJKPBFEHDFALCIEJABCA.tim@comcast.net>  
Message-ID: <GCEDKONBLEFPPADDJOCOECEHJENAA.whisper@oz.net>

You missed the part that said that spam is kept in the "eThunk" and was viewable by a simple viewer for final disposition?

Of course, with Outbloat, you could fire up PythonWin and stuff the spam into the Junk Email folder... but then you loose the ability to retrain on the user classified ham/spam.

David LeBlanc  
Seattle, WA USA
Goal is to use statistical methodology to filter our mail.

- Get sample, classified messages
- Convert to response, predictors data frame
- Fit statistical model to data
  - use information from mail headers (i.e. sender, routing information, date, return address, etc.)
  - use information in the content of the message body
- Tune Type I and Type II errors
  - Type I error: reject regular message as SPAM
  - Type II error: accept SPAM as regular message.
- Deploy classifier as filter.

2 Another Example: Web Logs

- Each request to a Web site for a page is a hit
Each hit recorded by Web server in a Web log

Provides information about site
- popular pages
- common navigation paths through sites
- busy times
- resource usage and allocation
- pre-fetching of pages
- efficient proxy information

DATA

ncass03.telenet-ops.be - - [28/Dec/2003:06:33:55 -0600] "GET / HTTP/1.1" 200 718
ncass03.telenet-ops.be - - [28/Dec/2003:06:34:03 -0600] "GET /R.css HTTP/1.1" 200 658
ncass03.telenet-ops.be - - [28/Dec/2003:06:34:13 -0600] "GET /logo.html HTTP/1.1" 200 244
ncass03.telenet-ops.be - - [28/Dec/2003:06:34:23 -0600] "GET /navbar.html HTTP/1.1" 200 1418
ncass03.telenet-ops.be - - [28/Dec/2003:06:34:23 -0600] "GET /banner.shtml HTTP/1.1" 200 3185
ncass03.telenet-ops.be - - [28/Dec/2003:06:34:33 -0600] "GET /Rlogo.jpg HTTP/1.1" 200 8793
pcp07845748pcs.wilmsc01.tn.comcast.net - - [28/Dec/2003:06:36:06 -0600] "GET /contrib/extra/dcom HTTP/1.1" 301 342
pcp07845748pcs.wilmsc01.tn.comcast.net - - [28/Dec/2003:06:36:06 -0600] "GET /contrib/extra/dcom/ HTTP/1.1" 200 1404
pcp07845748pcs.wilmsc01.tn.comcast.net - - [28/Dec/2003:06:36:06 -0600] "GET /icons/blank.gif HTTP/1.1" 200 148
pcp07845748pcs.wilmsc01.tn.comcast.net - - [28/Dec/2003:06:36:06 -0600] "GET /icons/blank.gif HTTP/1.1" 200 148
pcp07845748pcs.wilmsc01.tn.comcast.net - - [28/Dec/2003:06:36:06 -0600] "GET /icons/compressed.gif HTTP/1.1" 200 246
pcp07845748pcs.wilmsc01.tn.comcast.net - - [28/Dec/2003:06:36:06 -0600] "GET /icons/back.gif HTTP/1.1" 200 216
pcp07845748pcs.wilmsc01.tn.comcast.net - - [28/Dec/2003:06:36:15 -0600] "GET /contrib/extra/dcom/ReadMe.txt HTTP/1.1" 200 12247
ncass03.telenet-ops.be - - [28/Dec/2003:06:37:15 -0600] "GET /bin/windows HTTP/1.1" 301 335
ncass03.telenet-ops.be - - [28/Dec/2003:06:37:32 -0600] "GET /bin/windows/ HTTP/1.1" 200 1353
ncass03.telenet-ops.be - - [28/Dec/2003:06:39:28 -0600] "GET /bin/windows/base HTTP/1.1" 200 2167
ncass03.telenet-ops.be - - [28/Dec/2003:06:39:47 -0600] "GET /bin/windows/base/ HTTP/1.1" 200 2167
ncass03.telenet-ops.be - - [28/Dec/2003:06:41:28 -0600] "GET /bin/windows/base/README.rw1081 HTTP/1.1" 200 11036
p38.nas2.is4.u-net.net - - [28/Dec/2003:06:43:40 -0600] "GET / HTTP/1.1" 200 718
p38.nas2.is4.u-net.net - - [28/Dec/2003:06:43:40 -0600] "GET / HTTP/1.1" 200 718
lj1229.inktomisearch.com - - [28/Dec/2003:06:45:34 -0600] "GET /robots.txt HTTP/1.0" 404 283
Steps

- Collapse sequence of lines into top-level page request
- Find included pages and read HTML.
- Find links in pages to get connectivity.
- Analyze connectivity and hits.
- Explore data graphically - mining.
- Pre-fetching model: fit 2-step Markov models for transitions between pages.
- Evaluate fit.
- Deploy pre-fetching model: Client or Server?

3 Goals of the Course

- Focus: use existing software and functionality for context-specific analyses.
- Learn about: box of tools and how to use them to create things, and even build new tools. Learn about currently emerging technologies
- De-emphasize: understanding the existing algorithms. Be able to intelligently discuss different technologies and tools, knowing when to use them and what are the trade-offs
- Understanding fundamental algorithms is important if you need to
  - recreate them in a new language
  - use them in new ways when developing new algorithms.
- Practical: how statistical methodology is used in Industry, Laboratory, Research
- Focus: overall task not just on the application of specific statistical methodology but on how to think about approaching problems related to computing on data
4 General Information

• Instructor
  – Deborah Nolan
  – Office: 395 Evans
  – Email: nolan@stat.berkeley.edu
  – Office Hours: Mon 4:00-5:00, Fri 1:00-2:00

• GSI
  – TBA
  – Office:
  – Email:
  – Office Hours:

• USI
  – Erica Christenson and Sandy Nathan
  – Email: erica@stat sandy@stat

• Lab meets on Fridays: 2-3 or 4-5 in 342 Evans

Course Materials

• There is NO textbook for the class.

• I’ll prepare detailed notes or chapters that go into more details than in class.

• Distributed via the Web at www.stat.berkeley.edu/users/nolan/stat133/Fall04

• Many links to resources on the Web as you need them.

• We will use R as the primary computational environment

• R manuals
  – An Introduction to R cran.r-project.org/doc/manuals/R-intro.pdf
  – R Data Import/Export cran.r-project.org/doc/manuals/R-data.pdf
  – R Language Definition cran.r-project.org/doc/manuals/R-lang.pdf
  – On-line user guides to R, on-line help
Computing Resources

- Statistical Computing Facilities (SCF) – networked computers running Unix
- Undergraduate computer laboratory
  - 342 Evans and 432 Evans
  - Open 8am to 6pm Monday through Friday.
  - Remote access through ssh
- Account from GSI on Friday
- Mailing lists: archive, post.
  - Class mailing list

Grading: Grades will be based on a combination of individual and group work.
Participation 5% Class mailing list and in class
Homeworks 40% 8-10 Short computing assignments
Projects 40% 2 Parts each worth 20%
  - Must be done in groups of 2 or 3
  - SPAM, Web site access, Network intrusion detection
  - E-mail connectivity and social networks
Oral 15% Short 10-minute Q&A during exam week

Expectations: Although there is no computing, probability, or statistics prerequisite for this course, there is an expectation that you have the Curiosity, Initiative and Motivation to Explore on your own and Learn as needed.
There will be the opportunity to learn and receive help from many sources – instructor, GSI, student assistants, fellow students.

5 Academic Integrity

- It is okay to discuss your work with instructor, GSI, SAs, fellow students
- But keep your code to yourself
- Make a significant contribution to your group’s work
- If you have questions about academic integrity, ask the instructor for clarification
Writing a program is like writing a paper – your code should be your original work.
The following is excerpted from the Code of Student Conduct
http://students.berkeley.edu/sas/rights.shtml

Cheating  Fraud, deceit, or dishonesty in an academic assignment

• Providing answers to or receiving answers from others for any academic assignment.

• In group assignments it is the responsibility of the student to ascertain from the instructor to what degree the work must be done exclusively by the student or may be done in collaboration with others;

• Improperly obtaining or using improperly obtained information about an assignment or assisting others in doing so;

• Putting one’s name on another student’s assignment

• Altering previously graded work for the purpose of seeking a grade appeal

Plagiarism  The use of intellectual material produced by another person without acknowledging its source

• Copying from the writings or works of others into one’s academic assignment without attribution

• Submitting work of others as if it were one’s own

• Using the views, opinions, or insights of another without acknowledgment

Other violations:

• Writing an exam, paper, assignment for another student

• Representing oneself as another person

• Representing, explicitly or implicitly, that work obtained from another source was produced by oneself

• Failure to comply with the instructions or directives of the course instructor