Statistical Data Mining (ORIE 4740) **Spring 2014**

"[Data mining is] the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data…[it] employs pattern recognition technologies, as well as statistical and mathematical techniques" (The Gartner Group). Data mining often involves datasets with many records and many variables. Frequently little is known about the distribution of any particular variable, or about the relationships between variables. Desirable approaches have few assumptions or are robust to the violation of those assumptions. They also must be computationally tractable on large data sets.

By the end of this course, you will be able to take a large commercial or governmental data set, decide on data mining techniques to answer your question of interest, apply those techniques, compare them, and draw conclusions. In order to cement your understanding you will implement some simple techniques, and modify implementations of some more complex techniques.

Exams are Mon. 3/03 in class (first exam) and Tues. 4/15, 7:30-9:30 PM (second exam).

Prerequisites

- **ORIE 2700 and 3500** (statistics and probability) or equivalent. Point and interval estimation, hypothesis testing, p-values. Simple linear regression. Marginal probability, joint probability, conditional probability, Bayes' theorem (refs include Ross, 2006 and Freedman, Pisani, and Purves, 1998).
- **Math 2940** (linear algebra) or equivalent.
- **A programming course (2+ credits)** in R, Matlab, C, Java, or similar.
- Strongly recommended: Background in **multiple linear regression and logistic regression**

First Steps

(1) Visit the course website at http://blackboard.cornell.edu to **access the course information**.
(2) **Register your iClicker at** http://atcsupport.cit.cornell.edu/pollsrvc/. For general information see http://www.it.cornell.edu/services/polling/howto-students.cfm.

Instructors

Prof. D. Woodard
Office hours: Tu 3-4 & Thurs 12-1 or by appointment, in Rhodes 228. You can also ask me questions immediately after class.
Email: woodard@ cornell.edu

Jian Wu, TA (jw926; Mon Sections)
Office hours: Th 2-3 and F 3-4 PM in Rhodes 293

Hao Ran Lee, TA (hl859; Tues Sections)
Office hours: M 12-1 and M 5-6 in Rhodes 431

Lectures / Labs

Lectures are MWF 8:40-9:55 AM; Mon./Fri. lectures are in Olin Hall 255 and **Wed. lectures are in Rhodes 471**. **Lectures end Friday, April 18 although the final project will continue through final exam week.** Print lecture notes off Blackboard and bring them to class. Labs are on Monday afternoon or Tues morning, in Rhodes 453. **Lab participation is crucial to prepare you for the final project!** Questions are best addressed during office hours and labs (instead of email), so make sure that several of the office hours are at times that work for you.

Homework

There will be about 8 homework assignments. Homework is due at 12 noon on Tuesday a week after it is given out, and must be submitted to the course mailbox (2nd floor Rhodes, across from rm. 206), NOT by email, under door, etc.. You may discuss the content of the homework with other students in your 4740 class, but the final product must be your own. Your lowest 2 homework grades will be dropped; this accommodates sickness, family emergency, or religious holiday without a formal process. If you miss an assignment for these reasons then it must count as one of your dropped assignments.

Software

We will use the statistical software package R, latest version. This is on the Windows machines in the ORIE labs, and students can obtain a free copy for their personal Windows / Mac / Linux machine at http://www.r-project.org/

Good references for R include:

- “An Introduction to R”, found at http://www.r-project.org/
- “R Reference Card”, at http://cran.r-project.org/doc/contrib/Short-refcard.pdf
- The book “Data Mining with R”

Grading

Grade allocation is: 10% homework, 34% final project and 54% exams, and 2% class participation (clicker participation & other). Responding to 80% of clicker questions in class gives 100% clicker grade. In case of a grading error you may resubmit the assignment (to your TA, with permission) or exam (to Prof., with permission) within one week of when it was returned to you, with a written explanation of the grading error. The entire assignment or exam is carefully regraded, so the final grade may be lower due to our finding additional mistakes.

Exams

There is one in-class exam (**Mon. 3/03 in class**) and a longer evening exam (**Tues. 4/15, 7:30-9:30 PM**). There will also be a final project that will most likely be due during the final exam slot. Request for special accommodation must be made at least 2 weeks prior to each exam.

Final project

In the final project, the techniques taught in the class are used to analyze a large business or engineering data set. Students work in teams of 2-3 students. Each team writes a project proposal, finds the necessary data, carries out the project, and writes a project report. A late project loses 10% credit per day.

Textbooks (required)

iClicker

Shmueli, Patel, & Bruce (2010). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner.* Second Ed., Wiley: NJ.
One copy on reserve in Uris library. Data available at http://www.dataminingbook.com

James, Witten, Hastie, and Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R.* Springer: NY. Freely available at:
http://www.stanford.edu/~hastie/local.ftp/Springer/ISLR_print1.pdf

Academic integrity

Violations of the Cornell’s Code of Academic Integrity are punished at minimum with failure of the course. There is a link to this code on the course Blackboard page.