

## **Statistical Data Mining (ORIE 4740)**

### **Fall 2009**

“Data mining is the application of [statistical and machine learning] techniques to common business problems” (Two Crows Intro. to Data Mining). Data mining often involves data sets with many records and many variables. Frequently little is known about the distribution of any particular variable, or about the relationship between any subsets of two or more 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 a data mining technique or techniques to answer your question of interest, apply those techniques, compare them, and draw conclusions. In order to cement your understanding of the data mining techniques you will implement some simple techniques, and modify and extend implementations of some more complex techniques.

### **Prerequisites**

- **Class attendance Fri. 10/16 and Wed. 12/02** (exam times)
- Prerequisites: ORIE 2700 and 3500 (intro. statistics and probability) or equivalent. Simple linear regression. Probability for discrete-valued random variables: marginal probability, joint probability, conditional probability, Bayes' theorem (brush up with any basic probability or prob / stats textbook such as that by Sheldon Ross).
- Programming experience in R or Matlab. If you do not have experience in one of these languages but do have experience in a lower-level language like C or Java, come talk to me.
- Strongly recommended: Exposure to multiple linear regression and logistic regression

### **Course web site**

The course Blackboard site: <http://blackboard.cornell.edu>

Visit before the next class to access the course information and sign up for the course email list.

Also visit the ORIE intranet site to request a departmental account:

<http://intranet.orie.cornell.edu>

### **Instructors**

## **T.A.s**

Tuohua Wu (contact person for all questions regarding labs)

Contact: [tw226@cornell.edu](mailto:tw226@cornell.edu)

Office hours: Mon. 4:30-6 PM and Wed. 4-5:30 PM in Rhodes 431

Wei Chen (contact person for all questions regarding homeworks and grading)

Contact: [wc438@cornell.edu](mailto:wc438@cornell.edu)

Office hour: Fri. 3-4 in Rhodes 431

## **Prof. D. Woodard**

228 Rhodes Hall

**Office hours: Tu / Th 2-3 and by appointment (no drop-ins)**

**Email for emergencies only :**

*woodard@cornell.edu*

## **Lectures / Labs**

Lectures are MWF 1:25-2:15; Mon./Fri. lectures are in Olin Hall 165 and Wed. lectures are in Rhodes 471. Discussion sections are M 2:30-3:20, M 3:35-4:25, W 10:10-11:00, and W 11:15-12:05, in Rhodes 453. Lecture attendance is crucial since large parts of the material are not in the book and your only source for this information will be the lectures.

Course questions will be addressed during office hours and labs (not via 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 (2<sup>nd</sup> floor Rhodes, visit rm. 206 for directions), NOT by email, under door, etc.. You may discuss the content of the homeworks with other students in your 4740 class, but the final product must be your own. Your lowest homework grade will be dropped; this accommodates sickness, family emergency, or religious holiday without a formal process. If you miss a single homework for these reasons then it must count as your dropped assignment.

## **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 / Linux machine at:

<http://www.r-project.org/>

Good references for R and its sister language S-PLUS include:

- “An Introduction to R”, found at <http://www.r-project.org/>
- The “User’s Guide” or “Getting Started Guide” for S-PLUS
- The book “Data Mining with R”

## **Grading**

Grade allocation is: 10% homework, 35% final project and 55% exams. Class participation determines borderline cases. In case of a grading error you may resubmit the assignment or exam (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 is often lower due to us finding additional mistakes!

## **Exams**

There are two in-class midterm exams, the first midway through the semester (**Fri. 10/16**) and the second on the second-to-last day of class (**Wed. 12/02**). There will be no final exam; the final project is in lieu of a final 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 will work in teams of 2-3 students. Each team must write a project proposal (due 11/10), find the necessary data, carry out the project, and write a project report. The project report is due in lieu of a final exam, at the end of the time slot allocated for the final exam.

## **Textbooks**

Required: Hastie, Tibshirani, and Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Ed., Springer, NY.

There is one copy of the text on course reserve in the engineering library (Carpenter Hall)

## **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.

## Timeline

Week	Date	Topic	Reading	Lab/Homework
1	Aug 28	Intro. to data mining	Two Crows Intro, Chapter 1 in HTF	Review R / S-PLUS basics and discrete probability
2	Aug 31	Classification: Heuristic approaches Case Study: Olive oils data		Classification (olive oils data)
3	Sept 7	Naïve Bayes Case Study: Accidents data		Naïve Bayes on (accidents data)
4	Sept 14	ROC Curves and Graphical Models Case: Credit risk data		ROC curves (credit data)
5	Sept 21	Linear Regression Case: Cheese data		Graphical models (income data)
6	Sept 28	Linear Regression Case: Veteran's data		Linear regression (veteran's data)
7	Oct 6	Logistic Regression Case: Challenger data		Linear regression (baseball data)
8	Oct 14	Logistic Regression Case: Veteran's data		No Lab / HW: Midterm I
9	Oct 19	Clustering Case: Voting data		Logistic regression (targeted marketing)
10	Oct 26	Clustering Case: Utilities data		No Lab/ HW: Project groups meet
11	Nov 2	Principal Components Analysis (PCA)		Clustering (orthopedic data)
12	Nov 9	PCA. Case: Cereals data		Project proposal due
13	Nov 16	Classification and Regression Trees (CART)		PCA (customer demographics)
14	Nov 23	CART		No Lab/ HW
15	Nov 30	Midterm review / final project guidance		Midterm II
Final exam time slot				Final project due