ISyE 8803 – Special Topic: Statistical and Probabilistic Methods for Data Science

Fall 2017
Tentative Syllabus

Statistics and probabilistic models are essential building blocks for data science and machine learning. In this course, we will highlight modern statistical and probabilistic approaches to data modeling and algorithms development, exploiting structures of the data. The course will combine theoretical insights with illustrative applications, on the following topics:

- Sparse signal models: lasso, compressed sensing, spares coding and ICA.
- Graphical and network models: Graph lasso, causality, and natural language processing.
- Low-rank models: matrix completion, covariance matrix estimation, sparse PCA, community detection, and collaborative filtering.
- Event data and time duration model: survival analysis and tree-based methods, Hawkes process, and healthcare applications.
- (Optional) High-dimensional test and change-point detection.
- (Optional) Introduction to Restricted Bolzmann machine (RBM) and belief networks.

Class Time and Location: Tuesday and Thursday, 12:00pm-1:15pm, College of Computing 101. From Aug. 22 to Dec. 5, 2017.

Instructor: Prof. Yao Xie, Groseclose #339, email: yao.xie@isye.gatech.edu
Instructor Office Hour: Tuesday, 1:30pm-2:30pm.

Class Website: T-square
Class material available on our website includes

- Announcements
- Course syllabus
- Homework assignments and solutions
- Slides and other lecture material
- Practical exams
- Your course grades on exams and homework
- Any important announcements

Class Mailing List: Registered students are automatically subscribed to the class mailing list.

Textbook: the course material will be based on lectures and slides posted on T-square.
References*:*
Statistical learning with sparsity: the Lasso and generalizations. Trevor Hastie, Rob Tibshrani, and Martin Wainwright.
High-dimensional probability with applications in data science. Roman Vershynin.
High-dimensional statistics: A non-asymptotic viewpoint. Martin J. Wainwright.
The elements of Statistical Learning: Data Mining, Inference, and Predictions, 2nd edition, Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
Machine learning: A probabilistic perspective, K. P. Murphy.

Prerequisites: ISyE 6416 or permission of the instructor.

Honor Code: For any question involving Academic Honor Code issues, please consult www.honor.gatech.edu

Software: MATLAB, R, and/or Python.

Grading Policy: Homework - 20%, Midterm – 30%, Project - 50%.

Homework: The homework should be handed in **before the end of the class on the due date**. Late Homework will NOT be accepted. Assignments will include both exercises and computer problems; the computer problems will ask you to carry out statistical analysis using computer statistical software. Keep in mind that you should not hand in raw computer output. Conclusions and interpretation of results are more important than good printouts. You are allowed to work together with other students on homework, as long as you write up and turn in your own solutions. You are also allowed (and encouraged) to ask me questions, although you should try to think about the problems before asking. Homework 1: out around Tuesday **August 29**. Homework 2: out around Tuesday **Sept. 12**.

Midterms: There will be an in-class midterm exams during the class, on Thursday **Oct. 5, in class**. The midterms are close notes (including assignment solutions) and close textbook but two and respectively, four two-sided pages with formulas will be allowed. Do not write homework solutions on the formula sheet. You are not allowed to use your cell phone. The notes have to be self-made. **No make-ups.**

Project: by group, each group consists of 1-2 students. Proposal: two-page write-up in NIPS style. Due: Thursday, **Oct. 26, 2017.** Final presentation: Tuesday **Dec. 1 and/or** Thursday **Dec. 5**, depending on the number of projects need to be presented. Final report: 7-8 write up in NIPS style. Due: Tuesday **Dec. 12, 2017**.