CUSP-GX-5004.001: Applied Data Science

Fall 2015
Lecture: Mondays 5:30 pm – 8:20 pm
CUSP, 2 MTC, 8th Floor, MAGNET Room 820

Instructors:
Dr. Tim Savage, timothyhsavagephd@gmail.com or timothy.savage@nyu.edu
Dr. Stanislav Sobolevsky, sobolevsky@nyu.edu

Lab Instructor/ Lead Teaching Assistant:
Varun Adibhatla, va731@nyu.edu

Teaching Assistant:
Yash Chhajed, yashpchhajed@nyu.edu

Office Hours
Prof. Dr. Tim Savage: Mondays, 3 pm to 5 pm, 1 MTC, Office 1915
Prof. Dr. Stanislav Sobolevsky: Mondays, 2 pm to 4 pm, 1 MTC, Office 1910
Varun Adibhatla: Mondays, 1 pm to 3 pm, 1 MTC, Office 1947L

Course Description and Objectives:
This course introduces students to a wide variety of statistical learning tools currently used in applied data science. It is not a course in statistics, econometrics, or computer science *per se*. Rather, it is a synthesis of these disciplines with an eye toward the skeptical analysis of data, in particular when those data arise as the result of non-random human behavior. Students will also be introduced to the origins of analytic techniques where appropriate, including the distinction between or frequentist and Bayesian approaches to inference. A typical Foundations session will be two-thirds lecture and one-third lab.

The second half of the course will introduce students to another crucial component of modern urban informatics, network analysis and its applications to urban science. This is an important expertise, as many aspects of complex urban systems can only be understood through studying multiple layers of interacting urban actors. In addition, networks science see a diverse variety of applications in many other areas of science including but not limited to physics, biology, information technology, cognitive, and social systems. Network science is a new discipline that investigates the topology and dynamics of complex networks using modern methods of graph theory, computer science, and statistics.[1] Network science is not just an aggregate of such methods, however, but rather a synergetic synthesis leading to a novel way of thinking about complex networks. A typical network analysis session will be two-thirds lecture and one-third lab, with the exception of the first session being mainly a lecture, and the last one being mainly a lab.

[1] Prior experience in those areas could be helpful, but by no means is a necessary prerequisite for taking the course, which is designed to also accommodate the students starting from scratch.

Course Requirements

There are no formal requirements for the course, other than the successful completion of the summer boot camp. Prior to the course, students must be able to read in a real-world structured dataset in either R or Python (ideally both), to create basic graphical representations of the data, and to generate customary summary statistics, such as means and variances.[2] For the ADS module, students proficient in MATLAB are encouraged to use it as well if they wish to, although Python will be the primary language suggested. The value of the course to students without any undergraduate coursework in statistics, econometrics, computer science, or the physical sciences may be limited without considerable individual effort.

Grading:

Foundations Module:	50% of Course Grade (Foundations Project: 30%; Homework: 20%)
Applied Data Science Module:	50% of Course Grade (Network Analysis Project: 30%; Homework: 20%)

Foundations Project

The Foundations Module will culminate in a submission of a written paper that synthesizes the materials from the course. It aims to expose you to the task of original research using applied data analytics. At the end of the week prior to Fall Break, each student will submit a research proposal of one paragraph outlining a particular topic that you would like to explore. Students are permitted to work in groups of no larger than three, but each student must submit their own proposal. Topics can range from analysis of taxi data to the analysis of 311 data to the participation in Kaggle competitions. It is your call. In the proposal, you should address the hypotheses you would like to explore and how you might go about it. During the course, you will be taught a variety of techniques of statistical learning to apply to the data you propose to analyze. At the end of the course, you will submit a five-page, double-spaced paper that describes your research agenda, the data you have gathered, the hypotheses you explored, the methods you have used, and the results. Results can be both tabular and graphical. The paper size is independent of tables and visualizations. Each student must submit their own paper regardless of group size.

Network Analysis Project

The project aims for advancing student's data mining skills and building expertise in applying complex network analysis to the real-world challenges of urban science, utilizing various layers of big urban data. You will be able to pick up the problem of your interest within a broad scope of suggested objectives and ask a real research question to the urban data provided through CUSP's repository. Students are permitted to work in groups of no more than five, but each student must submit his/her own proposal.

You are expected to submit a paper describing your question, approach and findings or an application you've developed. Suggested paper size is five pages, double-spaced plus references, tables and plots. Each student must submit his/her own paper regardless of the group size, specifying student's individual contribution.

[2] R and Python are environments for computational statistics and data analysis that are free to users at the point of provision. RStudio is a popular version of R, while Anaconda is a popular version of Python. Both are freely available: https://www.rstudio.com/ and https://store.continuum.io/cshop/anaconda/.

NYU Classes:

You must have access to the class Blackboard site (http://newclasses.nyu.edu/). All announcements and class-related documents (supplemental and suggested readings, discussion questions, etc.) will be posted there.

Recommended/Suggested Readings for Foundations Module:

Hastie, Trevor *et al.*, The Elements of Statistical Learning, Data Mining, Inference and Prediction, 2nd Edition, Springer.
(Free: http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf)

Sheppard, Introduction to Python for Econometrics, Statistics, and Data Analysis, August 2014.
(Free: https://www.kevinsheppard.com/images/0/09/Python_introduction.pdf)

Zumel and Mount, Practical Data Science with R, 1st Edition, Manning Publications Company, March 2014. (Free select chapters: http://www.manning.com/zumel/)

Suggested Readings for ADS Module:

M.E.J. Newman, Networks – An introduction, Oxford Univ Press, 2010.

Albert-László Barabási. Network Science, e-book: http://barabasilab.neu.edu/networksciencebook/, http://www.openrev.org/group/complex-networks

U. Brandes and T. Erlebach (Eds.), Network Analysis: Methodological Foundations, Springer, 2005.

Other useful references:
A.-L. Barabási. Linked. The New Science of Networks, Perseus Publ, 2002
A. Barrat, M. Barthelemy and A. Vespignani, Dynamical Processes on Complex Networks, Cambridge Univ Press, 2008.
G. Caldarelli & M. Catanzano Networks: A Very Short Introduction. Oxford, 2012
R. Cohen and S. Havlin, Complex Networks – Structure, Robustness and Function, Cambridge Univ Press, 2010.
R. Diestel, Graph Theory (4th edition), Springer, 2010.
D. Easley and J. Kleinberg, Networks, Crowds and Markets , Cambridge Univ Press, 2010.
E. Estrada. The Structure of Complex Networks. Oxford. 2011
R. Hanneman and M. Riddle. Introduction to Social Network Methods.e-book. http://faculty.ucr.edu/~hanneman/nettext/
M.O. Jackson, Social and Economic Networks , Princeton Univ Press, 2008.
E. Kolaczyk, Statistical analysis of network data, Springer, 2009.
P. Van Mieghem, Graph Spectra for Complex Networks, Cambridge Univ Press, 2011.
M. Nowak, Evolutionary Dynamics: Exploring the Equations of Life, Belknap Press, 2006.
S. Strogatz, Nonlinear Dynamics And Chaos: With Applications To Physics, Biology, Chemistry, And Engineering, Westview Press, 2001.
Duncan Watts. Six Degrees: The Science of a Connected Age. W. W. Norton & Company. 2004

Course Schedule:

DATE	MODULE	TOPICS	ASSESSMENT
9/14	Foundations Session 1	• Course introduction • Optimization, probability and random variables • Hypothesis testing and confidence intervals • An introduction to the bivariate linear model	Homework 1
9/21	Foundations Session 2	• Bivariate model and regression diagnostics • Multivariate linear regression model • Introduction to probability models and classification	Homework 2
9/28	Foundations Session 3	• Probability models and classification • Model selection and overfitting • Violations of classical assumptions • Omitted variable bias	Homework 3
10/5	Foundations Session 4	• Violations of classical assumptions (cont.) • Correlation, causation, and identification • Non-experimental data and inference • Natural and quasi-natural experiments	Homework 4
10/13	Foundations Session 5	• Introduction to Bayesian inference	Homework 5 and Foundation Project Proposal
10/19	Foundations Session 6	• Review of materials to date • Topics in machine learning	Homework 6
10/26	Foundations Session 7	• Time series analysis I • Markov processes in detail • White noise processes as a building block	Homework 7
11/2	Foundations Session 8	• Times series analysis II • The autoregressive model and its linkages to Markov processes • Moving averages • ARIMA	Homework 8
11/9	ADS Module Session 1	• Concept of a network, its basic properties. • Weighted and directed networks.	Homework 9

		• Bi-partite and multi-layered networks. • Notion of a complex network. Small-world phenomenon. • City through complex networks: examples and approaches, network representation of urban data	
11/16	ADS Module Session 2	• Trees and hierarchical properties of the network. • Degree distributions and other statistical properties of the networks. • Random networks. Idea of scale-free complex networks and real-world examples. • Network models. • Node centrality – network hubs and measure of node significance. Pagerank approach.	Homework 10 Network Analysis Project Proposal
11/23	ADS Module Session 3	• Connectivity. Shortest path problem and Dijkstra's algorithm. • Travelling salesman problem. Concept of an NP-hard problem.	Homework 11
11/30	ADS Module Session 4	• Structure of networks. Motifs. Communities. • Clustering and betweenness. • Modularity. Community detection methods.	Homework 12
12/7	ADS Module Session 5	• Network dynamics. Information diffusion and epidemic spread perspective. Social networks. • Behavior and opinion propagation in social networks. • Random walker. • Network resilience.	Homework 13
12/14	ADS Module Session 6	Network practicum: • Network approach for transportation optimization; • Resilience of infrastructural networks; • Social network analysis.	Foundations Project Due. Network Analysis Project Due.
12/21		Project Presentations	Final Grading

Statement of Academic Integrity

NYU-CUSP values both open inquiry and academic integrity. Full and Part-Time graduate programs and advanced certificate programs are expected to follow standards of excellence set forth by New York University. Such standards include but are not limited to: respect, honesty and responsibility. The program has zero-tolerance for violations to academic integrity. Such violations are deemed unacceptable at NYU and CUSP. Instances of academic misconduct include but are not limited to:

- Plagiarism
- Cheating
- Submitting your own work toward requirements in more than one course without
  (1) Prior documented approval from instructor and
  (2) Proper citation
- Forgery of academic documents with the intent to defraud
- Deliberate destruction, theft, or unauthorized use of laboratory data, research materials, computer resources, or University property
- Disruption of an academic event (lecture, laboratory, seminar, session) and interference with access to classroom, laboratories, or academic offices or programs

Students are expected to familiarize themselves with the University's policy on academic integrity and CUSP's policies on plagiarism as they will be expected to adhere to such policies at all times – as a student and an alumni of New York University.