CUSP-GX-5004: Applied Data Science
Fall 2017
Location and Lecture Times:
Tuesdays 2:00-4:50pm (afternoon group) and
Wednesdays 5:30-8:20pm (evening group),
Building 2MTC, Rm 820

Instructor
Dr. Stanislav Sobolevsky, sobolevsky@nyu.edu, 646.997.0527

Course Assistants
Tushar Ahuja, ta1302@nyu.edu, 929.423.0639,
TBD

Office Hours
Stanislav Sobolevsky: Wednesdays, 2-5pm, 1MTC, 1910

Course Description and Objectives
This course introduces students to a wide variety of tools currently used in applied data science. It is not a course in statistics, econometrics, or computer science *per se*. Rather, it is a practice-oriented synthesis of these disciplines with strong urban focus — concepts and techniques are motivated and illustrated by applications to urban problems and datasets. Students will also be introduced to the origins of analytic techniques where appropriate with necessary theoretic material provided. Certain important additional fundamental topics such as Bayesian Inference and Network Analysis will be considered in the end of the program.

A typical 3-hour session will be half lecture and half interactive lab, where students are provided with examples of the code (through IPython notebooks) implementing the considered techniques and are asked to implement similar assignments on their own under the instructor's supervision.

Course Requirements
The only formal pre-requisites for the course is the successful completion of the summer Urban Computing Skills Lab. Prior to the course, students must be able to read structured datasets in Python [1], to create basic graphical representations of the data, and to generate customary summary statistics, such as means, variances as well as the distributions. Students proficient in MATLAB or R are encouraged to use it as well if they wish to, although Python (through IPython environment) 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.

[1] Python and R 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/. In the class we'll be mostly using IPython environment https://ipython.org

Course Project

The course will culminate in a submission of a written paper that synthesizes the considered materials and techniques. It aims to expose you to the task of original research using urban data analytics. Each student will submit and then present (a short 5 min talk) a 1-2 page long research proposal outlining a particular urban analytics topic that she/he would like to explore. Question/hypothesis-driven research topics are particularly encouraged. The project is supposed to utilize urban data, ideally open data such as census, taxi, 311, LEHD, weather data or other. The topic is your call. In the proposal, you should address what hypotheses you would like to explore and how you might go about it. During the course, you will be taught a variety of techniques that you should be able to apply to the data you propose to analyze. At the end of the course, you will submit a five-page (excluding tables, graphics and references, presented in the end), 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. Students are permitted end encouraged to work in groups of no larger than 3-5, but each student must submit their own proposal and project paper. Submissions of the team members may overlap, however individual roles and contributions should be clearly outlined in the submissions.

The grading

Grading will be based on three components:

I. Midterm exam (20%)
II. Six homework assignments, typically every second week (40%)
III. Final project and presentation (40%)

The deadlines

Project proposals submission – 12pm (noon), October, 9, 2017
Final project paper submission – 12pm (noon), Dec, 10, 2017
Homework submission – 12 days from each assignment by Monday noon

Suggested Readings

Hastie, *et al*., THE ELEMENTS OF STATISTICAL LEARNING, DATA MINING, INFERENCE AND PREDICTION, 2nd Edition, Springer. http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf
http://statweb.stanford.edu/~tibs/ElemStatLearn/
Sheppard, INTRODUCTION TO PYTHON FOR ECONOMETRICS, STATISTICS, AND DATA ANALYSIS, August 2014.
https://www.kevinsheppard.com/images/0/09/Python_introduction.pdf

Other recommended readings

Alpaydin, E.. Introduction to Machine Learning, Second Edition
http://cs.du.edu/~mitchell/mario_books/Introduction_to_Machine_Learning_-_2e_-_Ethem_Alpaydin.pdf
Barabási A.-L. Network Science, e-book: http://barabasilab.neu.edu/networksciencebook/
Bishop, C.M. PATTERN RECOGNITION AND MACHINE LEARNING. Springer, 2006
T. Mitchell. Machine Learning. McGraw Hill, 1997 http://www.cs.cmu.edu/~tom/mlbook.html
Murphy, K.P. MACHINE LEARNING. A PROBABILISTIC PERSPECTIVE. The MIT Press, 2012
Provost, F. and Fawcett, T. Data Science for Business. O'Reilly
Zumel and Mount, PRACTICAL DATA SCIENCE WITH R, 1st Edition, Manning Publications Company, March 2014. (Free select chapters: http://www.manning.com/zumel/)
M.E.J. Newman, Networks – An introduction, Oxford Univ Press, 2010.

Course Schedule

Date	Session	Topics	Assignment
9/12-13	Session 1	Introduction to Urban Data Science. Basic Machine Learning concepts	
9/26-27	Session 2	Single-attribute linear regression and its applications	Homework 1
9/29	Session 3	Lab practicum session – handling urban data in python, performing basic regression analysis and visualizations	
10/3-4	Session 4	Multivariate linear regression. Multicollinearity and overfitting	Homework 2
10/10-11	Session 5	Regression diagnostics and hypothesis testing. Confidence intervals	
10/13	Session 6	Presentations and discussion of ADS project ideas	Homework 3
10/17-18	Session 7	Dealing with multicollinearity and overfitting. Dimensionality reduction through Principle Component Analysis	
10/24-25	Session 8	Unsupervised learning: clustering techniques. K-means, k-medians. Classification through Logistic regression. Max-likelihood estimate (theory only)	Homework 4
10/31-11/1	Session 9	Clustering and classification lab session. Midterm quiz	***Midterm quiz***
11/07-08	Session 10	Introduction to Bayesian Inference. Linear regression revisited	Homework 5
11/14-15	Session 11	Introduction to Network Analysis-I	Homework 6
11/21-22	no classes	Thanksgiving break	
11/28-29	Session 12	Introduction to Network Analysis-II	
12/04-05	Session 13	Spatial regression	
12/11-12	Session 14	Project final presentations	

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
    a) Prior documented approval from instructor and
    b) 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.