CUSP-GX-5004: Applied Data Science
Fall 2019
Location and Lecture Times:
Tuesdays 12:25-2:55 (group 001)
Wednesdays 6:00-8:30pm (group 002),
Location: 2MTC 820

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

Course Assistants
Yusu Qian yq729@nyu.edu, Soniya Chawla sc7221@nyu.edu, Hong Jiang hj1274@nyu.edu

Office Hours
Stanislav Sobolevsky: Wednesdays, 3:30-5:30pm, 370 Jay St, Brooklyn, NY, Office 1310, by appointment

Course Description and Objectives
This course equips students with the basic skills and tools necessary to address urban data science problems. It introduces to a wide variety of tools currently used in applied data science, from basic regression analysis, clustering and classification to more advanced topics, such as deep learning and network science and will heavy rely on Python on the implementation end. However, it is not a course in programming, 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, illustrated by iPython notebooks. Students will be introduced to the origins of analytic techniques where appropriate with necessary minimum of the theoretic material provided (more advanced theory could be included in the notes, as references or discussed in separate sessions upon request). The limits of applicability of the considered techniques, diagnostic of the results as well as their interpretation will be also considered.

A typical 2.5-hour session will be 1-1.5h of an overview lecture and 1-1.5h of 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. 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.

Course Project
The course will culminate in a submission and presentation of an urban data science project that synthesizes the considered materials and techniques. It aims to expose the students to the task of

[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

original research using urban data analytics. The projects are done in teams of 3-5 students. Each team will start from submitting and then presenting (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. The topic is your call. In the proposal, you should address what hypotheses you would like to explore, the data and methods you are going to use. 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, each team will submit a 5-8 page (up to 2500 words, excluding tables, graphics and references and any appendixes, presented in the end) paper that describes your research question/hypothesis, its importance and context, key takeaways from the literature, the data you have gathered, the methods you have used, the results and their interpretation. Proposed structure is: abstract (up to 150 words), introduction, literature review, data, methods, results, discussion (optional), conclusions, tables and figures, references, appendixes). While joint team submissions are allowed, individual roles and contributions should be clearly outlined in one of the appendixes. Typically the team members providing fair contribution get the same project grades, but this may vary depending on the scope of contribution.

The grading

Grading will be based on three components:

I. Midterm exam (20%)
II. Homework assignments (40%)
III. Attendance and in-class participation (10%)
IV. Final project and presentation (30%)

The deadlines

Project proposals submission – 12pm (noon), October, 21, 2019
Final project paper submission – 12pm (noon), Dec, 16, 2019
Homework submission – 12 days from each assignment (posted on Wednesdays) 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
A byte of Python https://python.swaroopch.com

Other recommended readings

McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012.
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.

Further resources

Introductions to statistics:

https://pdfs.semanticscholar.org/5777/2c52696be0881728ebde18eb84c8397309b8.pdf

https://faculty.washington.edu/ezivot/econ424/probreview.pdf (Section 1.1.1, 1.1.2, 1.1.6, 1.2)

http://www.cim.mcgill.ca/~paul/StIEs43z.pdf

Data mining/analysis:

Data Mining Concepts And Techniques http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Introduction to Data Mining https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Python tutorials:

https://pythonprogramming.net/machine-learning-tutorial-python-introduction/

Course Schedule[2]

Date	Session	Topics	Assignment
9/3-4	Session 1	Introduction to Urban Data Science. Use cases. Practical case with urban data	
9/10-11	Session 2	Basic Machine Learning concepts. Practical case with supervised/unsupervised learning	Homework 1
9/17-18	Session 3	Linear regression and its applications	Homework 2
9/24-25	Session 4	Unsupervised learning: clustering techniques. K-means and Gaussian Mixture clustering	Homework 3
10/1-2	Session 5	Issues with multivariate linear regression: multicollinearity and overfitting. Regression diagnostics and hypothesis testing. Out-of-sample evaluation, cross-validation.	Homework 4
10/8-9	Session 6	Notes on project team work, efficient academic writing and presentations. Brainstorming on possible project ideas. Overview lab	
10/15-16	no classes	Fall break	
10/22-23	Session 7	Presentations and discussion of ADS project ideas	
10/29-30	Session 8	Regularized regression – Lasso and Ridge. Classification through Logistic regression	Homework 5
11/5-6[3]	Session 9	Dimensionality reduction through Principle Component Analysis	Homework 6
11/12-13	Session 10	Midterm exam	***Midterm exam***
11/19-20	Session 11	Time-series analysis	Homework 7
11/26-27	no classes	Thanksgiving break	
12/3-4	Session 12	Advanced topic I: Introduction to Network Analysis	Homework 8 (choice)[4]
12/10-11	Session 13	Advanced topic II: Introduction to Deep Learning	Homework 8 (choice)[4]
12/17-18	Session 14	Project final presentations	

[2] Instructor reserves right for adjustments of the course schedule, subject to prior notice

[3] Tuesday class of Nov,5 may be rescheduled to Fri, Nov,8 – please stay tuned for announcements

[4] Students can choose one of the advanced topics for homework 8. Doing both would give homework extracredit

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.