Principles of Urban Informatics
Spring 2021

Instructor:
Stanislav Sobolevsky, sobolevsky@nyu.edu
Teaching Assistants:
Minyi He mh5172@nyu.edu
Devashish Khulbe dk3596@nyu.edu

In-person sessions: Thursday, 5:30-7:00pm (on the specified days), 370 Jay st, Auditorium 1201
Webinars: Wed 8:30-10am
Weekly video lectures and lab materials published on NYU Classes by Monday
Office hours with instructor and TA's through zoom upon email request

Course Description and Objectives. This course builds the foundation of the skill set and tools necessary to address urban analytics problems with urban data. It starts with basic computational skills, statistical analysis, good practices for data curation and coding, and further introduces a machine learning paradigm and a variety of common supervised and unsupervised learning tools used in urban informatics, including regression analysis, clustering and classification. After this class you should be able to formulate a question relevant to Urban Informatics, locate and curate an appropriate data set, identify and apply analytic approaches to answer the question, obtain the answer and assess it with respect to its certainty level as well as the limitations of the approach and the data. The course will also contain project-oriented practice in urban informatics, including relevant soft skills – verbal and written articulation of the problem statement, approach, achievements, limitations and implications.

The course heavily relies on Python on the implementation end and understanding of probability concepts on the theoretic 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. An overview of the relevant foundations of Python coding and statistics is provided in the beginning, however some basic proficiency is expected as a prerequisite. 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, providing a layered learning approach for those who look for more). The limits of applicability of the considered techniques, diagnostic of the results as well as their interpretation will be also considered.

Course logistics. The course is delivered as a hybrid accommodating both – in-person and online students. All the lectures on the theoretic topics of the class are delivered through pre-recorded video-lectures to be reviewed by all the students. Implementation examples for the discussed techniques will be provided through commented iPython notebooks with reusable relevant code examples followed by coding practice through performing relevant homework assignments to be submitted as iPython notebooks through NYU classes.

In addition to formal lectures, implementation examples and homeworks, one of the key value propositions of the class is a series of experiential learning lab sessions (according to the schedule below) delivered through webinars with follow-up in-person sessions. A typical lab session includes an example of implementing an urban informatics problem illustrating approaches learned in the class introduced by the instructor and/or other CUSP researchers. The open discussion on how this or similar analytic

approaches can help the course projects of the students can follow. A Q&A session on this and other current class topics can be also included during the lab upon request.

The recorded lab webinar is offered on Wednesday 8:30-10am each week (besides those not having a lab component). The webinar is mostly a logistics discussion and a Q&A session as the lab notebook video introduction will be provided pre-recorded for asynchronous review. Both – in-person and online students are encouraged to join, so that they can participate in the discussion and Q&A live, however recorded instruction parts can be viewed on NYU classes at any later time and discussion forum will be available. In-person students are also offered the follow up Thursday (on selected dates-see schedule below) in-person sessions to review the lab and homework materials with the instructor and continue the discussion in person.

In-person attendance is expected for the students taking the class in-person. Students attending in-person sessions are required to review the webinar material and lab notebook first in order to make the most use of the session.

Those who may need additional consultations are encouraged to schedule individual or group zoom office hours with instructor (approaches, course project, general questions) and/or teaching assistants (implementation and coding aspects) upon email request. "Open house" office hour webinars can also be offered from time to time and will be announced accordingly.

Midterm, final exam and project presentations are administered online. Both are a combination of the overview multiple choice and/or open questions on the course concepts and coding assignments on urban data analytics. You can think of those as multi-topic homeworks just more constrained in time.

Course Requirements. While the course will include extensive coding and statistical practice, basic Python proficiency and understanding of the major probability and statistical concepts is a prerequisite. The only formal prerequisites for the course is the successful completion of the summer Urban Computing Skills Lab or having equivalent Python proficiency. Prior to the course, students should 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. While we will recap on the above skills in the beginning of the class, the value of the course to students without any coding skills and 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 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 the team 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

[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

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 project can be further continued within the Applied Data Science course being offered as a follow up class on more advanced topics from the necessary Urban Informatics skillset.

The milestones/deadlines

Project idea: submission - 3/24, noon, presentation - online students present on Wed, 3/24, 8:30-10am, in-person students present online on Thu 3/25, 5:30-7pm
Midterm assignment administered online on 3/25-3/29
First project report submission – 5/4, noon, presentation – 5/5-6
(online students present on Wednesday 8:30-10am, in-person students present online Thursday 5:30-7pm
Final exam assignment administered online on 5/5-9
Homeworks due – Friday, noon, one week after assignment

The grading

Grading will be based on the following components:

I. Midterm assignment (15%)
II. Final exam (25%)
III. Homework assignments (30%)
IV. Lab/discussion participation (5%)
V. Course project report (20%) and presentation (5%)

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/
Computing and coding: Beginning Python Visualization, 2009
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
Data analysis: Statistics in a Nutshell, S. Boslaugh, O'Reilly Media
Visualizations: Visualizations Analysis and Design, T. Munzer, 2014

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
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
Murray, S. Interactive Data Visualization, O'Reilly Media
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/)
Andrew Ng online course on Machine Learning

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/
https://github.com/SSobol/Python_bootcamp
https://www.youtube.com/channel/UCpCcKrQ-rpokHx0Ac2Hv_Gw
https://www.youtube.com/watch?v=bY6m6_IIN94&list=PLi01XoE8jYohWFPpC17Z-wWhPOSuh8Er-

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.

PUI 2021 Spring Schedule[2]

Session	Topics	Materials published	Webinar	Date in-person	Homework due
1-2	Course overview and logistics. Intro to Urban Informatics and Data Science - overview. Urban Data Analytics Project structure and case example. Installing iPython and using CodeLab. Python Code Conventions. Using GitHUB LAB: Loading different types of urban data, basic descriptive analysis	1/28	2/3	-	2/12
3	Data curation: common data issues and ways to address them LAB-1: Data curation and descriptive analysis LAB-2 (Urban data practicum): Real-estate market analysis, merging datasets, spatial visualization	2/8	-	2/11	2/19
4	Data acquisition through APIs. Exploratory data analysis. LAB: Transportation route, travel time info. Discussion on course projects	2/15	2/17		2/26
5	Introduction to probabilities and statistics, Gaussian and other basic distributions, PDF and CDF, percentiles. Correlation vs causality. LAB: Urban analytics case review –exploratory correlation analysis.	2/22	2/24	-	3/5
6	Hypothesis testing. LAB. Midterm and project discussion	3/1	-	3/4	3/12
7	Introduction to Machine learning. Quiz. Midterm and project discussion	3/8	3/10		
	Class Spring Break	3/15-19			
8	Presentations/discussion of course team project ideas. Midterm exam/assignment	3/22	3/24, 8:30-10am for online, 3/25 5:30-7pm for in-person cohort		Midterm-administered before 3/25, due 3/29
9	Linear regression. LAB: Real Estate Prices	3/29	3/31		4/9
10	Multicollinearity and overfitting. Regression diagnostics and hypothesis testing. LAB: Urban analytics cases review. Discussion on ongoing course projects	4/5	-	4/8	4/16

[2] Topics and timeline is subject to adjustment throughout the semester. Any important updates will be announced

11	Clustering. K-Means, Gaussian Mixture. LAB: Clustering of NYC locations, urban zoning and partition	4/12	4/14		4/23
12-13	Classification. Logistic regression, SVM. Multi-class classification LAB: Urban classification case review. Out-of-sample evaluation, cross-validation. Regularized regression. LAB	4/19	4/21		4/30
13	Material posted previous week Discussion on completing the course projects and finals		4/28	4/29	5/3 (optional, extra-credit)
14	Presentations/discussion of course team projects final report. Final exam	5/3	5/5, 8:30-10am for online, 5/6 5:30-7pm for in-person cohort		Final exam administered before 5/6 due 5/10

Class GitHub

Please find the class GitHub repository: https://github.com/CUSP2021PUI (to be activated). It includes class public materials such as tutorials (please feel free to review) and data (please feel free to review the NYC_open_data_introduction for useful links). It will also include private materials - homework assignments, labs and released homework solutions to be posted there under the private Labs_Solutions repository. In order to get access please follow the instructions:

1. Signup for GitHub, and please provide your GitHub account information in Class SignUp form (to be provided)
2. We are using GitHub Classroom to manage homework submission. You will receive an invitation link for each assignment, by accepting the invitation, you will create a new private repository under CUSP2021PUI which is only visible to yourself and class instructor and teaching assistants.
3. Here is the invitation Homework0_GitHub (to be provided) to your first assignment - a trial pass/fail homework to practice github and homework submission logistics. This trial homework is due by Feb,5 noon.
4. In order to submit a homework, you only need to commit to your own private assignment repository. You may submit as many times as you need before the deadline. The deadline will be mentioned in the README file of each assignment repository. Please keep in mind that late submissions will encounter late penalties or may not be accepted. And please check that your submission got uploaded timely and correctly, as there is little we can do if any issue is revealed after the grading is complete and sample solutions are released.
5. If you have any questions, please do not hesitate to contact Mingyi He mh5172@nyu.edu or Devashish Khulbe dk3596@nyu.edu

Inclusion statement

The NYU Tandon School and CUSP value an inclusive and equitable environment for all our students. I hope to foster a sense of community in this class and consider it a place where individuals of all backgrounds, beliefs, ethnicities, national origins, gender identities, sexual orientations, religious and political affiliations, and abilities will be treated with respect. It is my intent that all students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. If this standard is not being upheld, please feel free to speak with me.