CENTER FOR URBAN SCIENCE+PROGRESS

NEW YORK UNIVERSITY

Principals of Urban Informatics
Fall 2020
Instructor:
Stanislav Sobolevsky, sobolevsky@nyu.edu
Teaching Assistants:
Minyi He mh5172@nyu.edu
Devashish Khulbe dk3596@nyu.edu

In-person sessions: Wed, 6-7pm and 7:15-8:15pm (on the specified days)
Webinars: Tue 8:30-10am (on the specified days)
Weekly video lectures and lab materials published on NYU Classes by Monday
Office hours with instructor and TA's through zoom upon e-mail request

Course Description and Objectives. This course builds the foundation of the skillset 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 Tuesday 8:30-10am each week (besides those not having a lab component). 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 can choose to attend the follow up Wednesday (on selected dates) in-person sessions offered (see schedule below) to review the lab and homework materials with the instructor and continue the discussion in person.

In-person attendance is not mandatory and due to limited class capacity (no more than 15 students allowed at a time due to COVID restrictions) RVSP will be required. First 15 RSVPs will be accommodated from 6 to 7pm, next 15 – from 7:15 to 8:15pm. RSVP for the in-person sessions will be made available online at 6pm on Wednesday a week before. Preference for either session can be expressed and accommodated when possible but not guaranteed (e.g. if less than 15 students RSVP we start at 6pm with everyone). Attendance requests above those limits if any won't be accommodated at the given time but will be given priority for the next in-person session. 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 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,

[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

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 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 proposals submission – 10/26, noon
Project idea: submission - 10/26, noon, presentation - online students present on Tue, 10/27, 8:30-10am, in-person students present online on Wed 10/28, 6-8pm
Midterm assignment administered online on 10/27-28
First project report submission – 12/14, noon, presentation – 12/15-16
(online students present on Tuesday 8:30-10am, in-person students present online Wed 6-8pm)
Final exam assignment administered online on 12/15-16
Homeworks due – Wed, noon, one week after assignment; can be extended during the breaks

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/
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
Computing and coding: Beginning Python Visualization, 2009
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 2020 Fall 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	9/2-9/7	9/8	-	9/16
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	9/14	9/15	9/16	9/23
4	Data acquisition through APIs. Exploratory data analysis. LAB: Transportation route, travel time info. Discussion on course projects	9/21	9/22	9/23	9/30
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.	9/28	9/29		10/7
6	Hypothesis testing. LAB	10/5	10/6	10/7	10/21
	Fall break	10/12-18			
7	Introduction to Machine learning. Quiz. Midterm and project discussion webinar	10/19	10/20		10/28 (quiz only)
8	Presentations/discussion of course team project ideas. Midterm exam/assignment		10/27, 8:30-10am for online, 10/28 6-8pm for in-person cohort		
9	Linear regression. LAB: Real Estate Prices	11/2	11/3		11/11
10	Multicollinearity and overfitting. Regression diagnostics and hypothesis testing. LAB: Urban analytics cases review. Discussion on ongoing course projects	11/9	11/10	11/11	11/18

[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	11/16	11/17		12/2
	Thanksgiving break	11/23-29			
12	Classification. Logistic regression, SVM. Multi-class classification LAB: Urban classification case review. Discussion on ongoing course projects	11/30	12/1	12/2	12/9
13	Out-of-sample evaluation, cross-validation. Regularized regression. LAB	12/7	12/8		12/16 (optional, extra-credit)
14	Presentations/discussion of course team project report. Final exam		12/15, 8:30-10am for online, 12/16 6-8pm for in-person cohort		