Applied Data Science
Spring 2022
Instructors: Stanislav Sobolevsky,
Richard Vecsler
Teaching Assistants:
Devashish Khulbe, Mingyi He, Navpreet Singh, Divya Pathai
Class schedule:
Section A: Thursday, 2-4:30pm, 2 MetroTechCtr Rm 9.009
Section B: Wednesday, 2-4:30pm, 2 MetroTechCtr Rm 9.011
Section C: Wednesday, 5-7:30 pm, 2 MetroTechCtr Rm 9.009
Reach out with questions and support requests at cusp.ads22@gmail.com
and/or through student's Discord personal channel

Course Description and Objectives. This course builds upon the foundations provided within the Principles of Urban Informatics course and extends them with a spectrum of more advanced applied urban data analytics techniques, including Time Series Analysis, Dimensionality Reduction, Bayesian Inference, Network Analysis and Deep Learning. You will apply those techniques to urban analytics cases. The course will contain project-oriented practice in urban data analytics, 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 in person, but supplemented with pre-recorded video-lectures on select theoretic aspects to be reviewed by all the students. Also all the classes have an additional pre-recorded asynchronous videos, the students can review if they miss any aspect of the class. 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 Github/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.

Midterm is administered online as a take-home open-book assignment. Midterm is a combination of the overview multiple choice and/or open-ended 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 (at least 48hours will be given).

Class Communication. The class communication is facilitated through Discord (TBD). It has dedicated channels to discuss lab materials, homeworks, class projects. Common homework questions will be answered by the TAs and answers are available to the entire cohort. Those who may need additional support are encouraged to use the specified communication channel for questions and support requests, including scheduling individual or group zoom office hours with the teaching assistants (implementation and coding aspects) upon email request. Please notice that due to the large size of the student cohort this year, availability of the individual office hours with the instructors is limited to those matters that were attempted but could not be resolved as per above. "Open house" office hour webinars on specific subjects can also be offered from time to time and will be announced accordingly.

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 5-7 students. Each team will start from submitting 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) **project report** 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 report should be accompanied by a recorded team video presentation

[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

of up to 7 min (using Zoom for presentation recording is recommended). Presentations will be evaluated and the selected teams will be invited to present during the final class.

The milestones/deadlines

Project idea: submission - 3/13, noon
Midterm assignment administered online on 3/21-3/24
Final report submission – 5/3, noon, presentation – submission by 4/29, selected teams invited for the class of 5/4
Homeworks due – typically, Monday, noon, 11 days after assignment

The grading

Grading will be based on the following components:

I. Midterm assignment (20%)
II. Homework assignments (40%)
III. Course project report (30%) and presentation (10%)

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/
T. Mitchell. Machine Learning. McGraw Hill, 1997 http://www.cs.cmu.edu/~tom/mlbook.html
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
Network Analysis:
Barabási A.-L. Network Science, e-book: http://barabasilab.neu.edu/networksciencebook/
M.E.J. Newman, Networks – An introduction, Oxford Univ Press, 2010.

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
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 the 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.

ADS 2022 Spring Schedule[2]

Session	Topics	Class Date	Instructor	Homework due
1	Intro to the class. Tools for Big Data Analytics: Dask	1/26-27 (online)	Sobolevsky, Khulbe	2/7
2	Tools for Big Data Analytics II: MapReduce and Multiprocessing	2/2-3	Khulbe	2/14
3	Time Series Analysis 1	2/9-10	Vescler	2/21
4	Time Series Analysis 2: Models	2/16-17	Vescler	2/28
5	Dimensionality reduction 1: Linear PCA	2/23-24	Sobolevsky	3/7
6	Dimensionality reduction 2: Nonlinear (kernel) PCA. Non-linear model (random forest)	3/2-3	Sobolevsky	3/14
7	Midterm exam/assignment, project proposals due	3/9-10 (online)		Midterm and project proposals due 3/13
	Spring Break	3/14-19		
8	Bayesian Inference 1	3/23-24	Vescler	4/4
9	Bayesian Inference 2	3/30-31	Sobolevsky	
10	Network Analysis 1. Introduction. Node Centrality. Routing	4/6-7	Sobolevsky	4/18
11	Network Analysis 2. Community detection and network modeling	4/13-14	Sobolevsky	
12	Deep learning 1. Neural Networks and their implementation using Keras. Autoencoders	4/20-21	Vescler	5/2
13	Deep learning 2. Recurrent and Convolutional Neural Networks.	4/27-28	Vescler, Singh, Khulbe	
14	Final project deliverables/presentations	5/4-5		Presentation due 4/29 Reports due 5/3

Class lab repository

The lab materials for the class will be hosted on JupyterHub (primarily) as well as GitHub (backup). Please find the class GitHub repository: https://github.com/CUSP2022ADS. 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

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

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 to manage homework submission. You will be assigned a web-based GitHub environment which you will use to receive and submit homeworks.
3. In order to collect and submit homeworks, you will need to commit them through GitHub. Homework feedback and grades will be released by the TAs and available on Brightspace.
4. You may submit (commit) as many times as you need before the deadline. The deadline will be mentioned in the README file of each assignment repository, as well as in the homework notebooks. 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.

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.