PLA6113 Exploring Urban Data with Machine Learning

Instructor: Mauricio Rada mer2245@columbia.edu
Spring 2023
202 Fayerweather (UP Lab)
Monday 09:00 am – 11:00 am

Office hour: Wednesday 10:00 am – 11:00 am or by appointment (Schedule)

Syllabus developed by Professor Boyeong Hong. Schedule, and selection of assignments and readings have been modified by the instructor.

Course description and objectives

Data analytics and data-driven processes have been used to make urban planning decisions and to improve related city service operations. The most benefit of civic analytics is not only an in-depth understanding of urban phenomena but also predicting and preparing for future scenarios in cities composed of complex systems. There are immense opportunities with big data and analytic capacities to support responsive and effective urban systems ultimately aiming at sustainable and livable cities through a problem-driven analytic approach.

This course will engage the role of technologies and quantitative methods in the planning process. The main objective of this course is to familiarize students with modern machine learning techniques and demonstrate how they can be applied to urban data and real-world problems alongside the planning perspectives. Students will learn to apply the skills and techniques necessary to (1) understand the motivation behind different machine learning methods and their applicability in a given practical context, (2) implement and develop methodological framework, (3) model algorithms, (4) interpret and evaluate results appropriately, and (5) deliver insights with respect to urban planning perspectives and real-world problems.

Course structure

The course is practice-oriented class, learning concepts and techniques are motivated and illustrated by applications to urban problems and datasets. The class will include a mix of lectures and interactive coding lab sessions. The relevant theoretical background is provided. However, the course will not go into every detail of each technique. That said, students who wish to engage more with the theory behind machine learning methods are encouraged and supported through discussions and further readings. In order to help students understand applications of machine learning, practical examples will be introduced.

Textbooks and Online Resources

This course will use a combination of articles, book chapters, and instructor notes. There is no required textbook, but the followings are recommended reference textbooks:

- "Introduction to Machine Learning with Python" *by Andreas C. Müller, Sarah Guido*
- "The Elements of Statistical Learning" *by Hastie, Tibshirani, and Friedman (available for free download at http://statweb.stanford.edu/~tibs/ElemStatLearn/)*
- "Pattern Recognition and Machine Learning" *by Christopher M. Bishop*
- "Machine Learning: A Probabilistic Perspective" *by Kevin P. Murphy*
- "Introduction to Machine Learning, Second Edition" *by Ethem Alpaydin (available for free download at https://ieeexplore.ieee.org/book/6267367)*
- "Machine Learning" *by T. Mitchell*
- "Data Science for Business" *by F. Provost and T. Fawcett*
- "Applied Predictive Modeling" *by Max Kuhn and Kjell Johnson (https://link-springer-com.ezproxy.cul.columbia.edu/book/10.1007%2F978-1-4614-6849-3)*
- "Online Statistics Education" *developed by Rice University, University of Houston Clear Lake, and Tufts University (http://onlinestatbook.com/2/index.html)*
- "Machine Learning with Python Cookbook" *by Chris Albon*

Software

This course will use a variety of software tools and packages. *Python* (usually through *Jupyter notebooks* including packages like pandas, numpy and sklearn) will be the primary programming language. There will be a significant programming component and basic analysis and visualization abilities will be assumed. Students who don't have programming experience are welcome, but those should have a strong willingness to learn and build up related skills. No programming experience cannot be a hurdle of this course!

Assignments

There will be eight ***weekly assignments***, consisting of problem sets and/or programming that reinforce and propel topics covered in class. The assignments will be an extension of a lab session. In addition to regular assignments, students will be asked to participate in a group contemporary case study once per semester. This assignment deals with how topics in class are being used in practice or applied to urban issues. Details will be announced during the first lecture.

Late Assignments

Assignments will be deducted 10% for each day a submission is late unless there is a legitimate reason that the instructor is informed of in advance. Assignments later than a week will not be accepted.

Readings

There will be weekly readings assigned and uploaded on Canvas.

Midterm and Final Deliverables

Students, as groups, are asked to work on a final project to apply newfound technical savviness to analyze and synthesize urban data around a research question to deliver meaningful planning insights. It requires i) project proposal, ii) midterm presentation (exploratory analysis), iii) final presentation, iv) final report, and v) code documentation. Details will be provided in due course.

Grading

Grading will be performed through a numerical assessment of students' submitted work. A final score will be translated into the GSAPP grading system. The breakdown is as such:

- Group case study presentation - 3 students per week (10%)
- Lab sessions and assignments, typically every week, unless specified otherwise (30%)
- Project proposal (10%)
- Midterm packet (20%)
- Final packet (30%)

"High Pass" will be offered to the top 20% of students based on their numerical score and level of participation. "Pass" will be given to all final scores above 75. "Low Pass" will be 60-75, or automatically offered as a maximum if any major assignment is missing.

GSAPP Honor System and Plagiarism

Students must adhere to the principles of academic honesty (https://www.arch.columbia.edu/honor-system) and ensure that all work submitted is fully theirs and adhere to the GSAPP Plagiarism Policy (https://www.arch.columbia.edu/plagiarism-policy) set forth. Students found guilty of plagiarism or academic dishonesty will be subject to appropriate disciplinary action.

Collaboration and quoting policy*

Coding has unique challenges when it comes to collaboration and plagiarism, so please familiarize with this section.

Firstly, all the work you turn in must be your own (as an individual or as teams, as appropriate). However, you are welcome to discuss course materials, ideas, and assignments with others. When working through code with others, you must not discuss specific code—what you are going to implement within the computer that will be compiled—but you may discuss resources logic, structure and/or pseudo code with others. Nor may you provide or make available solutions to assignments to individuals who take or may take this course in the future. You may not directly use code found on the internet (cut-copy'ing) for assignments.

For the project, you may "quote" from resources online. You must acknowledge any source code that was not written by you by mentioning the original author(s) directly in your source code (comment or header). You can also acknowledge sources in a README.txt file if you used whole classes or libraries. Do not remove any original copyright notices and headers. However, you are encouraged to use libraries, unless explicitly stated otherwise by copyright, the code author or the teaching team! Although you may be using code found elsewhere, it is expected that your final projects are of substantive originality in concept and implementation.

** Developed by Professor Anthony Vanky*

Writing and Technical Assistance

The strength of GSAPP and the urban planning program is the diversity of experiences among its community members. However, with the diversity of languages, academic writing in English is a difficult art to master. While you will gain practice in communicating to diverse audiences in this class, 1) the writing center is a great resource that you should feel welcome to take advantage of: https://www.college.columbia.edu/core/uwp/writing-center and 2) a doctoral student will be a mentor providing support for academic writing specialized in the urban planning context. The Teaching Assistant of this course is Sean Chew, who will help with general data and coding related problem solving.

Schedule

Week	Date	Topic	Task Due
01	23 Jan	Course preview From data to Machine Learning, and Urban Planning Machine Learning Fundamentals Lab 01 - Intro to Python for ML	
02	30 Jan	Exploratory data analysis Probability theory Lab 02 - Exploratory analysis and basic ML practice	
03	06 Feb	Supervised learning 1 Linear models Example case - Predicting real estate prices Lab 03 - Linear regression modeling	Assignment 01
04	13 Feb	Supervised learning 2 Linear models 2 Probability models and classification Example case - Disparities in 311 usage Lab 04 -Logistic regression/Naive Bayes classifier modeling	Assignment 02
05	20 Feb	Unsupervised learning 1 Feature engineering and Dimensionality reduction Example cases - How to apply ML to urban problems Lab 05 - Principle component analysis (PCA)	Assignment 03
06	27 Feb	Unsupervised learning 2 Clustering fundamentals Example cases -1) K-Means clustering and urban resilience and 2) homelessness in NYC Lab 06 - K-Means clustering	Assignment 04
07	06 Mar	Text data and Natural Language Processing (NLP) Image data and processing Example case - Building permits and NLP, 311 comparative study	Assignment 05
08	13 Mar	Spring break - No class	Project proposal
09	20 Mar	Guest lecture - Cities, data, and machine learning Project brainstorm session	

10	27 Mar	Midterm presentation	Midterm packet
11	03 Apr	Unsupervised learning 3 Advanced clustering Example cases -1) Urban land-cover clustering and 2) similarity of structured urban open data Lab 07 - Agglomerative, GaussianMixture, and DBscan	
12	10 Apr	Supervised Learning 3 Support Vector Machine (SVM) Lab 08 - Classification using SVM	Assignment 06
13	17 Apr	Supervised learning 4 Decision Tree and Ensemble models Example case - Severe living condition building detection in NYC Lab 09 - Decision trees and Random forests	Assignment 07
14	24 Apr	Supervised Learning 5 Introductory Neural Network (NN) - time series and forecasting (LSTM) Lab 10 - NN application	Assignment 08
15	01 May	Final presentation	Presentation
16	05 May		Final submission