# PLA6113 Exploring Urban Data with Machine Learning

Instructor: Boyeong Hong
Spring 2020 Tuesday 7-9 pm
202 Fayerweather
Office hour: Tuesday 6-7pm or by appointment via email

## 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, although R and other tools may also be used. There will be a significant programming component and basic exploratory, modeling, 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.

## Assignments

There will be weekly assignments, consisting of problem sets and/or  that reinforce and propel topics covered in class. The assignments will be an extension of a lab session.

## 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.

## Midterm and Final Deliverables

There will be a midterm project and a final project, requiring i) project proposal, ii) project report, iii) presentation, and iv) code documentation.

## Grading

Grading will be based on four components:

- Weekly idea writing and 3 minutes pitch - 1 or 2 students per week (10%)
- Lab sessions and assignments, typically every week, unless specified otherwise (20%)
- Midterm packet (35%)
- Final packet (35%)

## 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.

## Lecture schedule

| Week | Date | Topics | Deadlines* |
|---|---|---|---|
| 01 | 01/21/2020 | Preview of the course<br>From data to Machine Learning, and Urban Planning<br>Lab 01 - Intro to Python for ML 1 | |
| 02 | 01/28/2020 | Types of ML<br>About data 1; Data munging, cleaning, and processing<br>Example case - Heating issues of multiple dwellings in NYC<br>Lab 02 - Intro to Python for ML 2 | Assignment 01 |
| 03 | 02/04/2020 | About data 2; Exploratory data analysis<br>Descriptive statistics and visualization<br>Example case - Waste generation in NYC<br>Lab 03 - Data exploration and basic ML practice | Assignment 02 |
| 04 | 02/11/2020 | Supervised learning 1<br>Linear models<br>Example case - Predicting real estate prices<br>Lab 04 - Linear regression modeling | Assignment 03 |
| 05 | 02/18/2020 | Supervised learning 2<br>Linear models 2<br>Probability models and classification<br>Example case - Disparities in 311 usage<br>Lab 05 - Lasso/Logistic regression/Naive Bayes classifier modeling | Assignment 04 |
| 06 | 02/25/2020 | Unsupervised learning 1<br>Feature engineering and Dimensionality reduction<br>Example cases - How to apply ML to urban problems<br>Lab 06 - Principle component analysis (PCA) | Assignment 05 |
| 07 | 03/03/2020 | Unsupervised learning 2<br>Clustering<br>Lab 07 - K-Means clustering<br>Guest lecture 1 - applications of ML for cities and urban planning | Assignment 06 |
| 08 | 03/10/2020 | Guest lecture 2 - applications of ML for cities and urban planning<br>Midterm presentation preparation and technical session | Project proposal |
| 09 | 03/17/2020 | Spring break - No class | |
| 10 | 03/24/2020 | Midterm presentation | Midterm packet |
| 11 | 03/31/2020 | Unsupervised learning 3<br>Clustering - more algorithms<br>Example cases -1) Urban land-cover clustering and 2) similarity of structured urban open data<br>Lab 08 - Agglomerative, GaussianMixture, and DBscan | Assignment 07 |
| 12 | 04/07/2020 | Supervised Learning 3<br>Support Vector Machine (SVM)<br>Lab 09 - Classification using SVM | Assignment 08 |
| 13 | 04/14/2020 | Supervised learning 4<br>Decision Tree and Ensemble models<br>Example cases - Severe living condition building detection in NYC<br>Lab 10 - Decision trees and Random forests | Assignment 09 |
| 14 | 04/21/2020 | Supervised Learning 5<br>Introductory Neural Network (NN) - timeseries and forecasting (LSTM)<br>Text data and Natural Language Processing (NLP)<br>Example case - Building permits and NLP, 311 comparative study<br>Lab 11 - NN and NLP applications | Assignment 10 |
| 15 | 04/28/2020 | Final presentation 1 (5 groups) | |
| 16 | 05/05/2020 | Final presentation 2 (4 groups) | Final packet |

** Weekly assignments will be due at the start of class (Tuesday 7pm) and midterm/final projects will be specified.*