**CUSP-GX 5003: Machine Learning for Cities**

**Spring 2022**

**Class Meetings (all times Eastern/NYC):**
Thursdays 11am-1:30pm
5 MetroTech Center (Dibner Hall), Room LC400, Brooklyn

**Instructor**:
Professor Daniel Neill, daniel.neill@nyu.edu

**Course Assistants:**
Tamanna Manek (tm3734@nyu.edu)
Valay Shah (vs2393@nyu.edu)
Shreyas Umare (ssu4614@nyu.edu)

**Office Hours:**
Professor Daniel Neill, daniel.neill@nyu.edu
In person office hours: Thursdays 2:45-3:30pm, 370 Jay Street, 13th fl., #1301D, Brooklyn
Virtual office hours: Mondays 9am-9:45am via Zoom

Tamanna Manek (tm3734@nyu.edu)
Virtual office hours: (*) Tuesdays 4:00-4:45pm and Fridays 3:00-3:45pm via Zoom

Valay Shah (vs2393@nyu.edu)
Virtual office hours: Tuesdays 6:00-6:45pm and (*) Thursdays 6:00-6:45pm via Zoom

Shreyas Umare (ssu4614@nyu.edu)
Virtual office hours: Tuesdays noon-12:45pm and (*) Wednesdays 5:00-5:45pm via Zoom

(*) The starred office hour times are likely to be moved to in-person office hours later in the semester.

**Course Description and Objectives**
The objective of this course is to familiarize students with modern machine learning techniques and demonstrate how they can be effectively applied to urban data. The course is practice-oriented: concepts and techniques are motivated and illustrated by applications to urban problems and datasets. For that reason, it involves a significant programming component, with Python as the primary programming language. Topics include a variety of supervised and unsupervised learning methods, such as support vector machines, clustering algorithms, ensemble learning, Bayesian networks, Gaussian processes, and anomaly detection. Strategies for effective machine learning and discussion of the limitations of ML as well as a variety of supplementary techniques are also considered.

The primary objectives of the course are to enable students to:

1) Understand the motivation behind different machine learning methods and their applicability in a given practical context.

2) Be able to implement methods adapted to the problem at hand using existing software libraries.

3) Know how to interpret the results appropriately.

Necessary background theory is provided; however, the course does not focus on theory. That said, students who wish to engage more with the theory behind machine learning methods are encouraged and supported.

## Course Requirements
Principles of Urban Informatics (CUSP-GX 5002) and Applied Data Science (CUSP-GX 6001) or equivalent. Alternatively, undergraduate coursework in computer science and statistics or some equivalent experience (including basic proficiency in Python) could be sufficient for following the course. Students who are unsure if they satisfy the prerequisites should check with the course instructor.

## Course Structure
Class sessions will typically include ~1.5 hours of slide-based lecture presentations, followed by a ~1 hour interactive coding session.

## Grading:
1) Four homework assignments (each 10% of the grade, for a total of 40%) – due 2/17, 3/3, 4/14, 4/28.
2) Midterm exam (15%) – March 24th
3) Project checkpoints (5%) – March 31st
4) Final project report (30%) and presentation (10%) – May 5th

## Readings:
This course will primarily be based on the instructor's course notes, supplemented by articles or book chapters when necessary.  There is no required textbook.  Recommended *reference* textbooks include:

1)  "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman (available for free download at http://statweb.stanford.edu/~tibs/ElemStatLearn/)
2) "Pattern Recognition and Machine Learning" by Christopher M. Bishop
3) "Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy
4) "Introduction to Machine Learning, Second Edition" by Ethem Alpaydin
5) "Machine Learning" by Tom Mitchell
6) "Data Science for Business" by F. Provost and T. Fawcett

## Software:
This course will use a variety of software tools and packages. Python (usually through ipython notebooks including packages like pandas, numpy and sklearn) will be the primary programming language. There will be a significant programming component and basic exploratory, modeling, and visualization abilities will be assumed.

## Late Work Policy:
You are expected to turn in all work on time (at the start of class on the due date) via NYU Brightspace. Because we understand that exceptional circumstances may arise, each student will be permitted to turn in one assignment up to 48 hours late with no penalty. Any other late assignments will not be accepted.

NOTE: Assignments turned in more than **five (5)** minutes after class starts will be counted as "late" and treated according to the Late Work Policy above.

**E-mail Questions Policy:** To balance the workload fairly among the three course assistants, and to ensure a reasonable response time for questions received via e-mail, each student has been assigned one course assistant as a "first contact". This is the person that you should e-mail first with any questions that you might have regarding the course material, policies, etc. We will do our best to answer all questions within 24 hours. If you do not receive a response within this time, or if your first contact is unable to resolve your question, then you should feel free to e-mail the instructor or another course assistant. First contacts are assigned based on the first letter of your last name. For A-Hsu, e-mail Tamanna (tm3734@nyu.edu). For Hu-P, e-mail Valay (vs2393@nyu.edu). For Q-Z, e-mail Shreyas (ssu4614@nyu.edu). Note that this policy applies to questions submitted via e-mail only – please feel free to attend any of our (in-person or virtual) office hours that are most convenient for your schedule!

**Course Schedule (subject to change):**

| Date | Lecture | Topics |
| --- | --- | --- |
| 1/27 | 01 | **Introduction to Machine Learning for Cities**<br>Course structure and syllabus<br>Course overview- methods and example applications<br>Supervised and unsupervised learning<br>Common ML problem paradigms- classification, regression, clustering, modeling, detection<br>Python review part 1: basics. |
| 2/3 | 02 | **Interpretable Classification (and Regression) using Decision Trees**<br>The prediction problem (classification and regression)<br>Overview of classification methods; discussion of interpretable vs. black-box predictors<br>Interpretable prediction: Rule-based, instance-based, and model-based<br>Rule-based Learning: Decision trees for classification and regression<br>Python review part 2: data science with Python. |
| 2/10 | 03 | **Ensemble Methods and the Accuracy vs. Interpretability Tradeoff**<br>Overview of ensemble methods (stacking, boosting, bagging, random forests)<br>From trees to forests: detailed discussion of random forests<br>Revisiting accuracy vs. interpretability<br>Decision trees and random forests in Python |
| 2/17 | 04 | **From Linear to Non-Linear Classifiers with Support Vector Machines**<br>**\*\*\* Homework Assignment 1 (Trees & Forests) due 11am \*\*\***<br>Linear decision boundaries.<br>Support vector machines (SVMs) for classification. Why maximize margin?<br>Moving from linear to non-linear decision boundaries with kernel SVM.<br>SVM in Python. |
| 2/24 | 05 | **Bayesian Methods for Supervised, Unsupervised, and Semi-Supervised Learning**<br>Comparison of supervised, unsupervised, and semi-supervised learning<br>Applications of clustering for modeling group structure<br>Naïve Bayes (NB) classification. A simple supervised learning algorithm.<br>Naive Bayes vs. logistic regression.<br>Expectation-maximization (EM) for clustering and semi-supervised NB classification.<br>NB and EM in Python. |

| 3/3 | 06 | **More clustering algorithms**<br>**\*\*\* Homework Assignment 2 (SVMs, Naïve Bayes, EM) due 11am \*\*\***<br>Brief digression: clustering as a search problem<br>Hierarchical clustering: bottom-up and top-down<br>K-means clustering<br>Leader clustering for massive streaming data<br>Clustering in Python. |
|---|---|---|
| 3/10 | 07 | **Fairness, Accountability, and Transparency in Machine Learning**<br>Definitions of fairness and bias<br>Fairness-aware algorithms<br>Automatically detecting biases in classification<br><br>**[NO LAB TODAY]**<br><br>**(Spring break 3/14-3/18: No classes, no office hours!)** |
| 3/24 | 08 | **Midterm Exam [NO LAB TODAY]** |
| 3/31 | 09 | **Bayesian Networks**<br>**\*\*\* Project Checkpoints due 11:55pm today \*\*\***<br>Bayes Nets: a tool for modeling the relationships between multiple variables<br>Other uses of Bayes Nets (such as classification and anomaly detection)<br>Efficient representation with Bayes Nets<br>Building and interpreting Bayes Nets<br>Inference of conditional dependencies with Bayes Nets<br>Learning Bayes Net Parameters from Data<br>Learning Bayes Net Structure from Data<br>Bayes Nets with Python |
| 4/7 | 10 | **Causal Structure Learning with Bayesian Networks**<br>Causal structure learning with the PC algorithm<br>Assumptions and comparison to econometric methods<br>Extensions and variants<br>Causal Orientation methods |
| 4/14 | 11 | **Gaussian Processes**<br>**\*\*Homework Assignment 3 (Clustering & Bayes Nets) due 11am \*\***<br>Modeling spatial and temporal dependencies and other non-iid data<br>GP regression<br>Scaling up GP regression; long range forecasting<br>Other uses of GPs: causal inference and change point detection |
| 4/21 | 12 | **Anomaly and Pattern Detection Part 1**<br>Model-based anomaly detection<br>Detecting anomalies using Bayesian networks<br>Distance-based, cluster-based, and density-based anomaly detection |
| 4/28 | 13 | **Anomaly and Pattern Detection Part 2**<br>**\*\*\* Homework Assignment 4 (GPs & Anomaly Detection) due 11am \*\*\***<br>Detecting patterns of anomalies<br>Spatial cluster detection and event detection<br>Applications to disease surveillance |
| 5/5 | 14 | **Final Project Presentations in Class [NO LAB TODAY]**<br>**\*\*\* Final Project Reports due 11:55pm today \*\*\*** |

**Acknowledgements**
Some of the material for this course has been borrowed from earlier iterations, and is used with permission. Thanks very much to Drs. Stanislav Sobolevsky, Martin Jankowiak, and Ravi Shroff for their contributions to the course content.

**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.

**Inclusion Statement**
The NYU Tandon School values 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.