Special Topics: Visualization for Machine Learning

Spring 2024

Instructor: Claudio Silva (csilva@nyu.edu); [webpage](#)

Section Leader: Erin McGowan (erin.mcgowan@nyu.edu)

Teaching Fellow: Vitoria Guardieiro (vitoria.guardieiro@nyu.edu)

Grader: Rithvik Guruprasad (rg4361@nyu.edu)

Instruction Mode: In-Person

Dates: Spring 2024

Meeting Times:

DS-GA 3001.001 (Lecture) Thursdays 6:45pm-8:25pm

Classroom: 31 Washington Pl (Silver Ctr) Room 520

DS-GA 3001.002 (Lab) Fridays 10:15am-11:05am

Classroom: 31 Washington Pl (Silver Ctr) Room 520

Class Discord: https://discord.gg/6zTAUXEEYk

Course Prerequisites

You should have solid programming expertise, of the level expected from a first-year graduate student in computer science or data science.

The coursework includes extensive programming with JavaScript, the D3.js library, and web technologies (CSS, SVG, etc.). While previous knowledge of these technologies is not required, being proficient and comfortable with extensive programming is a fundamental prerequisite for this course. If you are not comfortable with programming please contact the instructor before enrolling.

We will also expect students to be able to program in Python.

We expect that you have a solid foundation in *either* data visualization or machine learning. If you have no knowledge of machine learning, this course might not be appropriate for you.

Course Description

The material for this course is part of a fast-changing field of computing. It is a research-oriented course on topics related to visualization for machine learning, and all the students will be expected to work on a guided research project.

Our course is based on foundations of visual analytics, which is an area of data visualization that is concerned with improving a human analytic process, or how one makes sense of data for a given problem: understanding, reasoning, and making decisions about a provided dataset, and a given problem domain. Visual analytics is concerned with combining automated processes with human-driven processes that are built around data visualization: visual representations of data, and ways to interact with data. Given the rapid growth in machine learning in the last decade, research in visual analytics has witnessed similar growth in leveraging machine learning in a variety of ways.

Course History

This course started as a variation of the Visual Analytics and Machine Learning course designed by Professor Matthew Berger (Vanderbilt University). We first offered it at Tandon CSE on Spring 2020.

For the second offering in Spring 2021, the content of the course was updated quite substantially, in particular with more practical material aimed at enabling students to experience data analysis tasks through visualization.

The course offered in Spring 2022 was a further refinement of the material. Borrowing ideas from Professor Chris Manning (Stanford) in his course Natural Language Processing with Deep Learning, we provided *default* projects to help those students that are not already engaged in research. Also, we limited the assignments to the first part of the course before the project-related deadlines start.

The Spring 2023 was based on the Spring 2022 course, with updated materials and lectures.

Course Objectives

This course is designed to sharpen a student's knowledge of visualization and machine learning, and how the two areas interact. It is expected that the student will be a more effective data scientist by being fluent on the connections between the two areas. It is also designed around a major project, which will help the student develop research skills.

Course Structure

The course include lectures and labs. We will strive to have hands-on sessions to complement theoretical materials.

The course starts with a short primer on visualization. We will introduce machine learning concepts as they are needed in the class. We will cover visualizations for model assessment, while-box and black-box machine learning explainers. After that, we will continue with dimensionality reduction (clustering) techniques (e.g., PCA, t-SNE, UMAP).

After this initial set of lectures, we will continue with more advanced and specialized topics. We will cover Topology Data Analysis, followed by multiple lectures on visualizing deep neural networks.

Reading Material

There is no textbook for the course - most lectures will be based on recent technical papers, which have not yet been incorporated into textbooks. We will have suggested reading materials for each class. It is expected that, prior to the lecture, you have read the corresponding papers.

Here are supplemental readings to be used as reference material:

1. Data Visualization Curriculum, Jeff Heer, link
2. A Course in Machine Learning, Hal Daume, link
3. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Christoph Molnar, link
4. Introduction to Machine Learning, Etienne Bernard, link
5. Deep Learning, Ian Goodfellow, Yoshua Bengio and Aaron Courville, MIT Press, 2016 link

Research Project

This project includes a substantial research project. Please see the project section of the course for more details. As part of the project, you will be expected to reproduce prior work or implement a proposed research idea of your choosing (requirement details will be forthcoming and discussed in class). Moreover, you will be expected to demonstrate both the prior work, and your final research project, to the class during lectures. Again, please see project for additional details. Projects are expected to be pursued in groups of 2-3, although you can optionally pursue your project by yourself. Once the group is finalized, students cannot change or separate their groups throughout the semester.

Course Assessment

- Assignments (50%)
- Project Proposal (4-page writeup): 10%
- Project Updates (1-page writeup): 10%
- Full Project (8-page writeup): 25%
- Class Participation: 5%

Late Submissions

Late submissions of assignments will be penalized as follows: - A standard deduction rate of 20% per day.

It means that after 5 days of being late, your assignment will have a maximum grade of 0 (zero).

Course Schedule (tentative)

The course schedule is tentative and might need to be adjusted along the way.

Lecture 1: Introduction to Visualization – Part I

Lecture 2: Introduction to Visualization – Part II

Lecture 3: Model Assessment

Lecture 4: White Box Methods

Lecture 5: Black Box Methods

Lecture 6: Dimensionality Reduction

Lecture 7: Project and Research Discussion

Lecture 8 and 9: Topological Data Analysis

Lecture 10: Reserved for Invited Lecture

Lecture 11, 12, and 13: Deep Learning (incl. LLM, Convolutional Nets)

Lecture 14: Advanced Topics

Moses Center Statement of Disability

If you are a student with a disability who is requesting accommodations, please contact New York University's Moses Center for Students with Disabilities (CSD) at 212-998-4980 or mosescsd@nyu.edu. You must be registered with CSD to receive accommodations. Information about the Moses Center can be found at www.nyu.edu/csd. The Moses Center is located at 726 Broadway on the 3rd floor.

Academic Integrity

All students are expected to do their own work. Students may discuss assignments with each other, as well as with the course staff. Any discussion with others must be noted on a student's submitted assignment. Excessive collaboration (i.e., beyond discussing the assignment) will be considered a violation of academic integrity. Questions regarding acceptable collaboration should be directed to the class instructor prior to the collaboration. It is a violation of the honor code to copy or derive solutions from other students (or anyone at all), textbooks, previous instances of this course, or other courses covering thesame topics. Copying solutions from other students, or from students who previously took a similar course, is also clearly a violation of the honor code. Finally, a good point to keep in mind is that you must be able to explain and/or re-derive anything that you submit. This is particularly important if you should adapt solutions from online sources.

Here is a link to the GSAS statement on Academic Integrity.

AI policy

We live in the age of viable generative AI. Banning these tools is neither realistic, nor desirable. In fact, learning to use these tools is an emerging skill. Note that AI tools do not always produce correct or accurate results. In addition, it is unwise to rely on them too much. There are situations where you won't have access to these tools, for instance during technical interviews. In addition, there are also skills someone with an advanced degree in Data Science is just expected to have on tap - without AI assistance or looking anything up. To integrate both considerations, you can use generative AI tools to do the assignments in this class. If you use an AI to guide you in completing an assignment, you have to disclose which parts were generated by the AI.

NYU Academic Calendar

link to NYU Academic Calendar

This course does not have a final exam.

Also, please pay attention to notable dates such as Add/Drop, Withdrawal, etc.