City College

City University of New York

CSC I4490: Adversarial AI

Course Information

Tuesday 6:30pm – 9pm, Hybrid (lectures will be in person in NAC 5/102 and simultaneously via zoom)

Online office hours: Sundays 3-5 pm or by appointment (in zoom)

Instructor information: Dr. Allison Bishop, abishop@ccny.cuny.edu, or allibishop@gmail.com

Course Description

This course will explore the nascent field of adversarial machine learning, which seeks to understand and improve the behavior of machine learning algorithms in the presence of adversarial interactions. It will cover some topics in the related fields of robust statistics (where some of the training data is adversarially perturbed). In other words, we will interrogate each step of a machine learning process, from the input data itself, to the training process, to the final model performance and analyst queries, and ask: what happens if things go wrong? How might we make choices that are less vulnerable to mistakes and manipulation?

Course Objectives & Learning Outcomes

Students will learn how to systematically delineate threats to the integrity and performance of machine learning models, and what kinds of approaches are being developed to address them. Students will learn about several real examples of statistical models failing to behave as intended in practice, and what can be done to anticipate and mitigate such issues. Students will also see and work through several practical examples. By the end of the course, students will have completed and evaluated a theoretical design for a machine learning application in terms of its susceptibility to adversarial attack. (This project will be completed as a thought experiment in several steps over the course of the semester.)

Textbooks/Materials/Resources

There are no textbooks associated with this course. Reading assignments and supplemental recommended resources will be provided as pdfs and links.

Prerequisite Courses/Knowledge

There are no specific prerequisite requirements. Students who lack a working knowledge of probability theory, linear algebra, or basic statistics may struggle more than others in the course, and are encouraged to reach out to the instructor for further resources and guidance.

Assignments/Grading Policy

Each week at the end of class on Tuesday, two short assignments will be announced: a problem set containing three problems, and a second assignment that is either an assigned reading, or an assigned step of a semester-long personalized thought experiment. These assignments will also be posted in electronic form immediately after class.

Each week before class begins on Tuesday evening, each student will be expected to turn in 2 short documents:

1. Proposed solutions to the three problems
2. A one-page response to the second assignment (which is either a reading summary or a step of the thought experiment).

The final grade will be based on: 35% problem solutions, 30% reading summaries, 35% thought experiment responses. For problem sets and reading summaries, the lowest 2 grades will be dropped from the computation. For the thought experiment, all assigned steps will be included.

Late assignments will not be accepted for problem solutions or reading summaries. The 2 dropped grades are intended to give you enough flexibility so that extensions and late submissions are not necessary. Note that problem solutions will be provided in class on Tuesday, so extensions are not really possible.

For thought experiment responses, you can have a total sum of 14 days late for late submissions across all submissions. (e.g. one submission 14 days late, or one submission 10 days late and another 4 days late, etc.)

Beyond those 14 days, each additional day late will result in a – 10 % penalty for the grade of that individual response.

Academic Integrity/Honesty Policy

Academic dishonesty is prohibited in The City University of New York. Penalties for academic dishonesty include academic sanctions, such as failing or otherwise reduced grades, and/or disciplinary sanctions, including suspension or expulsion.

Other Classroom Policies

It is allowed for students to freely discuss and collaborate on assignments, though each student must ultimately write up their own assignment alone and in their own words.

Video of the lecture will be made available following each evening class. Lecture slides will typically be made available a few hours before each class.

Problem solutions will be discussed in class each week and posted electronically after class.

An example thought experiment response for each step will be posted electronically (ahead of the assigned step being due, to serve as a sample).

There is no final exam for this course.

Sample of Topics (subject to change and some topics may span more or less than 1 week)

1. What is machine learning/statistics/AI? How is it supposed to work in an ideal world?
2. Threat-modeling – what are the different ways that things can wrong in building and using a statistical model?
3. Basic attacker strategies and examples
4. Some relevant foundations of statistics and probability theory
5. The assumptions underlying statistical significance and how they can be wrong in practice
6. Robust statistics I – basic definitions and techniques for understanding and mitigating the influence of outliers
7. Errors in machine learning classification models – examples and emerging theoretical frameworks
8. Privacy leaks through analyst queries – examples and proposed mitigations
9. Robust statistics II – more advanced techniques for high dimensional data
10. The parable of "InstaHide" – an example of a real and broken system
11. Spam filters
12. Challenges of performing robust statistics in financial data