Fall 2023

DS 789-001: Trustworthy AI

Hai Phan

DS 789 Trustworthy AI

Course Description: As machine learning (ML) systems are increasingly being deployed in real-world applications, it is critical to ensure that these systems are behaving responsibly and are trustworthy. That will lead to a wider adoption of ML in real world applications in practice. This course will provide a deep understanding of state-of-the-art ML methods designed to make AI more trustworthy to unforeseen faults, to adversarial manipulation, and to violations of ethical norms in privacy and fairness. Students will gain understanding of and experience in using a set of methods and tools for deploying transparent, ethically sound, and robust machine learning solutions. The course is also an excellent opportunity to conduct research on the security/privacy/trustworthiness in ML and find research topics for Ph.D. and M.S. theses.

Learning Goals:

1. Obtain a good foundation of recent developments in trustworthy machine learning, data science, and AI approaches.

2. Learn and implement state-of-the-art deep learning models, addressing emergent challenges in real-world applications, i.e., privacy and security, image processing, NLP.

3. Gain hands-on multidisciplinary machine learning research experiences through Open source toolkit, which is either publicly available or developed by our team.

4. Design and communicate project deliverables to solve specific deep learning-based problems and applications, i.e., adversarial learning, generative adversarial networks, interpretable deep learning, privacy-preserving machine learning.

Term: Fall 2023

Pre-requisites: DS 675 (Machine Learning), OR approval of instructor.

Instructor: Hai Phan

Office: GITC 3901B

Email: phan@njit.edu

Wegpage: https://sites.google.com/site/ihaiphan/

Office Hours: Wed 1-2 PM.

Canvas: Additional material and resources will be found on the class website on Canvas. It will be modified and updated as the course progresses and will contain the most recent information.

Schedule: The following is a tentative schedule and subject to change. Refer to class web page for most recent information.

Course outline (15-week schedule)

Weeks

1. Introduction to AI and Machine Learning
2. Introduction to Deep Learning and TensorFlow
3. Introduction to Trustworthy Machine Learning
4. Topics in Trustworthy Machine Learning
5. Deep Learning Term Projects - Case studies
6. Interpretability and Explainability in Machine Learning
7. Fairness in Machine Learning
8. Adversarial Machine Learning
9. Adversarial Robustness: Theory, Practice, and Applications (NLP, Graph, Malware)
10. Differential Privacy and Machine Learning
11. Generative Adversarial Networks
12. Privacy and Security in Federated Learning: Attacks and Defenses
13. Term Projects Review
14. Term Projects Presentations
15. Final Exam

Laboratory Sessions: This course does not have a separate laboratory session. However, some class meeting time throughout the semester will be dedicated to hands-on laboratory assignments. This work will be done using the computers in the classroom (if not, please bring your laptop). If necessary, laboratory assignments should be worked on outside the class time.

Credit: 3

Grade: Final Grades will be based on:
Class participation – 5%
Assignments – 15%
Quiz – 10%
Paper presentation – 23%
Term Project – 27%
Final – 20%

The final letter grades for the semester are based solely on the points you earn according to Table 2.

Grade	Points
A	90-
B+	86-90
B	80-85
C+	70-79

C	60-69
D	50-59
F	0-49

Table 2: The final letter grade converting table

POLICIES:

Assignments (Homework and Project)

Homework for this class is usually due about one week after being issued. Their purpose is to help you keep up with the material and assess your readiness for the midterm and final.

Homework is due before midnight (11:55pm) on the due date specified on the schedule. It will be submitted via Canvas electronically. Late homework will be penalized 10% of the available points (, and another 10% will be deducted for every 24-hour period after the original due date), unless there is a reason beyond your control.

Makeup Tests

Requests for makeup tests must be made in advance with the instructor and will only be approved if the reason is beyond your control.

Academic Integrity Policy

The NJIT academic honor code is located at: http://integrity.njit.edu/index.html. This honor code applies in its entirety to this class. Violations will not be tolerated. In addition, students should familiarize themselves with NJIT's "Best Practices related to Academic Integrity" which is developed and published on the Provost's website (on the policies page).

Disabilities

If you have a disability that may require some modification of seating, testing, or any other class requirement; please let the Professor know so that appropriate arrangements can be made. Similarly let the Professor know if you have any emergency medical information about which to be aware, or if you need special arrangements in the event of building evacuation. See the Professor after class hours or schedule an appointment. Assistance is available from the Office of Student Disability Services (205 Campbell Hall; 973-596-3420). Be sure and fill out appropriate paperwork with this office during the first week of class.