CS 8803 Explainable AI (XAI)

Fall 2023 Syllabus

Instructor: Sonia Chernova
chernova@gatech.edu

Monday/Wednesday 3:30-4:45pm
College of Computing Building (CCB) 103
Course Website on Canvas

Course Description

This course focuses on Explainable AI (XAI), exploring the leading research, design principles and technical challenges developers face in developing and deploying interpretable machine learning methods for practical real-world applications. The course will cover a range of multidisciplinary topics, including AI and machine learning techniques, human factors, statistical methods for data analysis, aspects of social cognition, as well as ethical and societal considerations. These topics will be pursued through independent reading, class discussion, programming assignments and a project.

The course will meet in person, at the above-listed time and location. Much of the benefit from the course comes from participation in in-class discussions, and no content will be recorded and available online.

Learning Outcomes

Outcomes of this course include understanding of the challenges that machine learning and artificial intelligence fields face is developing interpretable models that enable users to understand the key factors that influenced the model's decision making, what current techniques have been established in the field, the strengths and limitations of current methods, how to develop new techniques, and systematic and effective ways of evaluating such systems.

Textbook

None to be purchased. Selected texts from conference and journal articles will be uploaded and assigned as reading throughout the course.

Assignments and Grading:

The course will include the following graded assessments:

- **One (1) "IML Insights" Presentation** – 5%, one submission
- **Two (2) Homework Assignments** – 10%, multiple submissions
- **Nine (9) Reading Reflections** – 35%, multiple submissions
- **Class Project** – 40%, cumulative from pitch, proposal, checkpoint and final submission
- **Class Participation** – 10%, multiple factors

- *"IML Insights" Presentations:* Each class meeting will include two 15-minute student presentations, called "IML Insights", where IML stands for interpretable machine learning, another common term in the XAI field. Each student will make one IML Insight presentation during the semester. Presentations can be made individually or in groups of two. The topic of the presentation is a research paper that is not covered as part of the course reading list to date. See Canvas for more details.
- *Homework assignments:* We will have two homework assignments at the beginning of the term focused on usage of established XAI methods and toolboxes.
- *Written paper reflections:* Weekly reading assignments will be posted on Canvas, which we will then be followed up with class discussion. To facilitate productive analysis and discussion, students will be required to complete a reading reflection of the assigned paper following the provided template. More details on the content of the reflection are provided in a separate document. You must submit a total of 9 writeups over the course of the semester to receive full credit, out of the 12+ readings that will be assigned in total, so you can skip some writeups without penalty.
- *Final project:* Projects will be conducted individually or in small groups. Project topics will be determined in September, and final project submissions will be at the end of the semester. More details about the projects will be available in a separate document.
- *Participation:* The participation grade totals 10% of the final grade and will be determined based on active participation in class discussions and poster sessions. Peer review from teammates will also be considered.

Late submission policy: Projects will not be accepted late except by prior permission. IML Insight dates are selected by the student. In case of illness or last minute changes, the student is responsible for either rescheduling or finding another student group to switch with. Written case study descriptions will not be accepted late except by prior permission since the assignment will be discussed in detail in class on the due date. Homework assignments will incur a 10% late penalty for up to 4 days; after 4 days the assignment can no longer be submitted.

Prerequisites: This is a graduate course meant for students interested in XAI research. It will be assumed that students have some background in AI, Machine Learning, and/or HCI, and an interest in all three. No formal prerequisite exists.

Course Policies

The course schedule and policies mentioned in this syllabus may change at any time during the term, but all changes will be clearly documented and announced.

Student Disability Services: If you need course adaptations or accommodations because of a disability, or if you have medical information to share with the instructor, please make an appointment or stop by to speak with Prof. Chernova within the first week of classes.

Academic Honesty Policy: Review Georgia Tech's Academic Honor Code. Any work you present as your own should represent your own understanding of the material. When external sources were used as significant points of information (sample code, etc.), the source must be referenced in your submission. Following Georgia Tech's guidelines, all suspected cases of academic dishonesty will be forwarded for review by the Office of Student Integrity.

Tentative Schedule

DATE	TOPIC	NOTES
Mon Aug 21	Course Introduction	
Wed Aug 23	Established Methods and Toolkits for XAI/IML	
Mon Aug 28	Established Methods and Toolkits for XAI/IML	IML Insights 1, 2
Wed Aug 30	Established Methods and Toolkits for XAI/IML	IML Insights 3, 4 / Hwk 1 Due
Mon Sept 4	*Labor Day, no class*	
Wed Sept 6	Human Factors in XAI	IML Insights 5, 6
Mon Sept 11	Project Pitches, individual meetings during class	
Wed Sept 13	Project Pitches, individual meetings during class	Hwk 2 Due
Mon Sept 18	Guest Lecture – Mark Riedl	IML Insights 7, 8
Wed Sept 20	Human Factors in XAI	Project Proposals Due
Mon Sept 25	Inherently Interpretable Models	IML Insights 9, 10
Wed Sept 27	Inherently Interpretable Models	IML Insights 11, 12
Mon Oct 2	Counterfactual Explanations	IML Insights 13, 14
Wed Oct 4	Counterfactual Explanations	IML Insights 15, 16
Mon Oct 9	Concept-Based Explanations	IML Insights 17, 18
Wed Oct 11	XAI for Sequential Decision Making	IML Insights 19, 20
Mon Oct 16	*Fall Break, no class*	
Wed Oct 18	XAI for Sequential Decision Making	IML Insights 21, 22
Mon Oct 23	Project Checkpoint, posters	
Wed Oct 25	Project Checkpoint, posters	
Mon Oct 30	User Study Design	IML Insights 23, 24
Wed Nov 1	Statistical Testing Best Practices	IML Insights 25, 26
Mon Nov 6	TBD	IML Insights 27, 28
Wed Nov 8	Guest Lecture – Devleena Das	IML Insights 29, 30
Mon Nov 13	Interactive Explanations	IML Insights 31, 32
Wed Nov 15	Interpretable Generative Models	IML Insights 33, 34
Mon Nov 20	Adaptive Explanations, Theory of Mind	IML Insights 35, 36
Wed Nov 22	*Thanksgiving Holiday, no class*	
Mon Nov 27	Accuracy versus Explainability, an Ethical Perspective	IML Insights 37, 38
Wed Nov 29	Final project posters	
Mon Dec 4	Final project presentations	
Fri Dec 8	Final Project Report Due, no class (we are not using our final exam slot)	