CS 442 - Trustworthy Machine Learning (Spring 2023)

Instructor
Bo Li (https://aisecure.github.io), lbo@illinois.edu (mailto:lbo@illinois.edu), 4310 Siebel Center

Lectures
3101 Sidney Lu Mech Engr Bldg
Zoom (https://illinois.zoom.us/j/89959281193?pwd=alVReHlaS0NYVENUeExBVTJNOEttQT09)

Teaching Assistant
Zijian Huang, Zoom (https://illinois.zoom.us/j/6254729767?pwd=M2pJREtjMWI4cytwYkVKWDRUQlRuUT09)

Office Hours
Bo Li: 4:45pm-5:30pm Monday / Wednesday

Zijian Huang: 4:30pm-5:30pm Tuesday

Forums
Canvas (https://canvas.illinois.edu/courses/35103)

Course Overview

This course will introduce basic knowledge about machine learning, security, privacy, adversarial machine learning, and game theory. Students will understand different machine learning algorithms, practice their implementations, and analyze their security vulnerabilities through a series of homeworks and projects.

Please contact the instructor if you have questions regarding the course materials. The syllabus is here (./442_aml_syllabus.pdf).

Course Schedule

The following table outlines the schedule for the course. We will update it as the quarter progresses.

Date	Lecture	Content	Materials
1/23	**Course Intro**		Slides (https://uofi.box.com/s/at5ec5gmpkjz2hc2nc7670i6nn69xzx4) reading 1 (https://oaklandsok.github.io/papers/papernot2018.pdf) reading 2 (https://arxiv.org/abs/1707.08945)
			Slides (https://uofi.box.com/s/u9o158zedn9c48e8jsf1sz5mpx25csbz)

1/25	Supervised Learning I	Regression, classification, Gradient	reading 1 (https://arxiv.org/abs/1412.6572) reading 2 (https://arxiv.org/abs/1801.02610) reading 3 (https://arxiv.org/pdf/1706.04599.pdf)
1/30	Supervised Learning II	PAC Learnability, supervised learning in Adversarial Settings	Slides (https://uofi.box.com/s/37pzw0gzhl5awebj3flnsqtj65knzjtw) reading 1 (https://arxiv.org/abs/1801.02612) reading 2 (https://arxiv.org/abs/1904.06347)
2/1	Unsupervised Learning I	Clustering, PCA, Matrix completion	Slides (https://uofi.box.com/s/ts4v9s6cm7hb0w9azsvliyyllj9u6inz) reading 1 (https://arxiv.org/abs/2106.06235) reading 2 (https://arxiv.org/abs/2003.00120)
2/6	**Homework 1 Walkthrough, Q&A**		Notes (https://uofi.box.com/s/d71crnzcxlb8i81sn8ztwt6tv911bqwr)
2/8	**Talk: From Heatmaps to Structural and Counterfactual Explanations**		Talk abstract and bio (https://uofi.box.com/s/d8envgfhit7n2jxwhz0qjgwtwouprqmf)
2/13	Unsupervised Learning II	Unsupervised learning in Adversarial Settings, Categories of Attacks on Machine Learning	Slides (https://uofi.box.com/s/z2c2q4jnvueawc3tq00gxpo8occh6c1c) reading 1 (https://arxiv.org/abs/1712.05526)
2/15	Attacks at Decision Time	Evasion Attacks, Anomaly Detection	Slides (https://uofi.box.com/s/xsocpnzp9vrh7o07cgor6z34zi4otxon) reading 1 (https://papers.nips.cc/paper/2014/hash/8597a6cfa74defcbde3047c8Abstract.html) reading 2 (https://arxiv.org/abs/1810.05162)
2/20	Modeling Decision-time Attacks		Slides (https://uofi.box.com/s/rlh9ef6jwptr80cksbwjn4irnxnqqzh3) reading 1 (https://arxiv.org/abs/1712.09491) reading 2 (https://arxiv.org/abs/1904.02144)
2/22	White-box/black-box Decision-time Attacks and Physical attacks		
	Homework 2		

2/27	**Walkthrough, Q&A**		Notes (https://uofi.box.com/s/gc6bkfxxalmbegnmwyu1lj4pmm8484cc
3/1	Defending against decision-time attacks I	Optimal evasion-robust classification	
3/6	Defending against decision-time attacks II	Feature level protection, randomized smoothing	Slides (https://uofi.box.com/s/6axag6j236l645smhhjuxv4jxmelmy11) reading 1 (https://arxiv.org/abs/2005.14137)
3/8	**Midterm Exam**		
3/20	Knowledge enriched robust learning models		Slides (https://uofi.box.com/s/ybjwrvmx5hv8upvoofc0j6il33xqs666) reading 1 (https://arxiv.org/abs/2209.05055)
3/22	Defending against decision-time attacks III	Adversarial retraining	Slides (https://uofi.box.com/s/ku9zxhuus8who6z9i3l9llqytsrg7x9j) reading 1 (https://arxiv.org/pdf/1902.02918.pdf)
3/27	Data Poisoning attacks	Binary classification, SVM, unsupervised learning, Matrix factorization, general framework	Slides (https://uofi.box.com/s/u59fslxwkhijme046u6meb4bqfeo5plo) reading 1 (https://arxiv.org/pdf/2103.06624.pdf)
3/29	Defending against poisoning attacks	Data sub-sampling, outlier removal	
4/3	Trustworthy Federated Learning	Certifiably robust federated leanring against training-time adversaries	Slides (https://uofi.box.com/s/3kpwckn5a4b2veagieph7crba2p3j54p) reading 1 (https://arxiv.org/pdf/2106.08283.pdf)
	Privacy in trustworthy	Membership attacks,	Slides (https://uofi.box.com/s/69cqk18vnaiqge4ksazq6oj2w01c97ia)

4/5	machine learning	model inversion attacks	reading 1 (https://arxiv.org/pdf/1610.05820.pdf)
4/10	**Guest lecture**		
4/12	**Homework 3 Walkthrough, Q&A**		Notes (https://uofi.box.com/s/f9akf8rsw7cyrf7tmn9nh2r42tzymtv1) CMU SafeBench AWS Instruction (https://uofi.box.com/s/hqn66a3qrx92z78iragmk5mse61ib5f1)
4/17	Differentially private machine learning	Differentially private data generative models	Slides (https://uofi.box.com/s/zfjx83ccd2dvtj28l5adl0sd7efwfmlo) reading 1 (https://arxiv.org/abs/1607.00133)
4/19	**Guest lecture**		Slides (https://uofi.box.com/s/os446ofjqmn67t9lq4g0dxx07ggsamot)
4/24	Trustworthy generative models	Trustworthy diffusion models against training and testing time adversaries	Slides (https://uofi.box.com/s/wuxr9wr66d6jxlngr1ruceclanddr98l) reading 1 (https://arxiv.org/abs/2011.13456)
4/26	Trustworthy foundation models	Training foundation models, and diverse trustworthy issues of foundation models	Slides (https://uofi.box.com/s/532x17hr69brvc7ra3b6hdhvi6ozto32) reading 1 (https://arxiv.org/abs/1706.03762)
5/1	**Final review, final project presentation**		
5/3	**Final Exam**		

Grading

The course will involve 3 programming homework, a midterm, and a final. Unless otherwise noted by the instructor, all work in this course is to be completed independently. If you are ever uncertain of how to complete an assignment, you can go to office hours or engage in high-level discussions about the problem with your classmates on the Piazza boards.

Grades will be assigned as follows:

- **Homework:** Homework accounts for 30% of the grade. Homework will be

assigned sparingly.
- **Midterm:** The midterm will be worth 25% of your grade, date/time TBD.
- **Final:** A final exam will be worth 35% of your grade, 5/3 (3 credit). For the 4 credit registration, a final project is required, which will take 30% of the final score after normalizing the other parts of performance.
- **Participation:** Accounts for 10%. Includes quizzes and class discussion (either in person or online).
- **Bonus Points:** If you want to do a class project, you could potentially get extra 20% bonus points depending on the quality of the project.

Course Expectations

The expectations for the course are that students will attend every class, do any readings assigned for class, and actively and constructively participate in class discussions. Class participation will be a measure of contributing to the discourse both in class, through discussion and questions, and outside of class through contributing and responding to the Piazza forum.

More information about course requirements will be made available leading up to the start of classes.

Ethics Statement

This course will include topics related computer security and privacy. As part of this investigation we may cover technologies whose abuse could infringe on the rights of others. As computer scientists, we rely on the ethical use of these technologies. Unethical use includes circumvention of an existing security or privacy mechanisms for any purpose, or the dissemination, promotion, or exploitation of vulnerabilities of these services. Any activity outside the letter or spirit of these guidelines will be reported to the proper authorities and may result in dismissal from the class and possibly more severe academic and legal sanctions.

Academic Integrity Policy

The University of Illinois at Urbana-Champaign Student Code should also be considered as a part of this syllabus. Students should pay particular attention to Article 1, Part 4: Academic Integrity. Read the Code at the following URL: http://studentcode.illinois.edu/ (http://studentcode.illinois.edu/).

Academic dishonesty may result in a failing grade. Every student is expected to review and abide by the Academic Integrity Policy: http://studentcode.illinois.edu/ (http://studentcode.illinois.edu/). Ignorance is not an excuse for any academic dishonesty. It is your responsibility to read this policy to avoid any misunderstanding. Do not hesitate to ask the instructor(s) if you are ever in doubt about what constitutes plagiarism, cheating, or any other breach of academic integrity.

Students with Disabilities

To obtain disability-related academic adjustments and/or auxiliary aids, students with disabilities must contact the course instructor and the as soon as possible. To insure that disability-related concerns are properly addressed from the beginning, students with disabilities who require assistance to participate in this class should contact Disability Resources and Educational Services (DRES) and see the instructor as soon as possible. If you need accommodations for any sort of disability, please speak to me after class, or make an appointment to see me, or see me during my office hours. DRES provides students with academic accommodations, access, and support services. To contact DRES you may visit 1207 S. Oak St., Champaign, call 333-4603 (V/TDD), or e-mail a message to disability@uiuc.edu (mailto:disability@uiuc.edu). Please refer to http://www.disability.illinois.edu/ (http://www.disability.illinois.edu/).

Emergency Response Recommendations

Emergency response recommendations can be found at the following website: http://police.illinois.edu/emergency-preparedness/ (http://police.illinois.edu/emergency-preparedness/). I encourage you to review this website and the campus building floor plans website within the first 10 days of class: http://police.illinois.edu/emergency-preparedness/building-emergency-action-plans/ (http://police.illinois.edu/emergency-preparedness/building-emergency-action-plans/).

Family Educational Rights and Privacy Act (FERPA)

Any student who has suppressed their directory information pursuant to Family Educational Rights and Privacy Act (FERPA) should self-identify to the instructor to ensure protection of the privacy of their attendance in this course. See http://registrar.illinois.edu/ferpa (http://registrar.illinois.edu/ferpa) for more information on FERPA.

TBD (mailto:)