Administrative Overview

Franziska Boenisch and Adam Dziedzic

Course on Trustworthy Machine Learning

April 17th, 2024

Organization

Flipped classroom:

- Lectures:
  - Published on Youtube around Wednesday
  - Please watch and prepare independently
- Questions:
  - Every student submits 2 questions weekly
  - Submission through CMS “Online-Test” (due Mondays 5PM)
  - Questions discussed during lecture hours
- Example:
  - Until Monday 22nd of April, watch the lecture on Privacy I and submit your questions.

Where and When?

Wednesdays from 4PM-6PM, CISPA, Lecture Hall Ground Floor (0.05)

24.04. Privacy I

08.05. Privacy II *(remote)*

15.05. Model Stealing

22.5. Defenses against Stealing

29.5. Robustness

05.06. Midterm Exam

05.06. Collaborative Learning I

12.06. Collaborative Learning II

19.06. Fairness & Bias

10.07. Explainability

17.07. Security & Governance

24.07. Summary *(remote)*

31.07. Final Exam

Accessing the material

- Lecture videos on Youtube
- Lecture notes and handouts on CMS
- All related work linked at the end of the presentations
- Homework assignments published on CMS
- Grades on CMS

Assignments

4 Programming assignments:

1. Implementing a membership inference attack 15.05.
2. Stealing a model behind an API 29.05.
3. Training a robust classifier 12.06.
4. Training a fair classifier 17.07.

Leaderboard for all assignments up on opening
Final submission of artefacts for evaluation
+Submission of code

Submission in groups of 2.

Grading

40% Assignment (10% per assignment)
10% Questions (10 weeks, 2 questions per week, graded on quality)
10% Midterm Exam
40% Final Exam

100%

Getting in Touch

Exchange between students: Discord
(link sent out to all students registered in CMS)

Reaching out to the instructors:

boenisch@cispa.de

dziedzic@cispa.de

Please include [TML24] in the subject line

Note: If you decide to discontinue the course, please de-register from CMS!

Thank you!

Franziska Boenisch and Adam Dziedzic
boenisch@cispa.de, adam.dziedzic@cispa.de
sprintml.com
Course on Trustworthy Machine Learning