Assignments for the course on Trustworthy Machine Learning (Summer 2024)

Assignment 1 (privacy)

Implement a membership inference attack and achieve the highest attack success (TPR@FPR=0.05) and area under the curve (AUC).

- We give you a trained model (Resnet18) and a list of data points (PUB) (custom dataset containing fields: ids, imgs, labels (class labels), membership (1==is member, 0==non member)).
- You also get a second list of data points (PRIV OUT), with the membership field being filled with Nones. Your goal is to classify each sample as a member or nonmember, with a continuous score, not 1/0 (that's for the TPR@FPR metric to make sense).
- The underlying training dataset is undisclosed.
- Loading the model and the datasets is in the gist under "Loading & submitting example".
- Leaderboard on 30% samples from your submission is available at any time. It is an intermediate scoreboard; you can use it to evaluate your results and compare them to others while working on the solution.
- The final leaderboard on the remaining 70% of samples will be revealed once the deadline for the assignment passes. The same pattern will be true for future assignments. The leaderboard will be available under the same URL, just the contents for this task will (probably) change.

Task artifacts

1. Model: https://drive.google.com/file/d/1-rFEKopl4PZ4e3FR_dKcLbO_Y4pXOgLo/view?usp=sharing
2. PUB: https://drive.google.com/file/d/1OLZsYJteuUpnQnSoZTHC617zmSRPsSK0/view?usp=sharing
3. PRIV OUT: https://drive.google.com/file/d/1wGNkKdKRn2ZpQ-GtP3l8UCpNHUWBgHyN/view?usp=sharing

Scoreboard

Link: http://35.184.239.3:9090/score_board

The beginning values of your “scores” are always worse than for a random guess submission. Don’t get so nervous if you see -1000 TPR@FPR at the beginning. By running the example from the gist below you should get ~0.05 TPR@FPR=0.05 and AUC of ~0.5 (see the “debug” team).

Loading & submitting an example

In the following gist you can find an example of the code on how to load the model and the files pub.pt and priv_out.pt:
https://github.com/sprintml/tml_2024/blob/main/example_assignment_1.py

Replace TOKEN with the private token that was sent to you via email.

Note: you can only submit a csv with scores every hour. If your submission is malformed or wrong you should get a comprehensive error message. However, even if you fail you still get a 1h cooldown. :)

Note: the evaluation server can crash catastrophically at any moment. It is a good idea to store your solutions somewhere safe so that if that happens you can resubmit. This shouldn’t happen, but as we know, there are 2 types of people: those who do backups, and those who will be.

The initial directory structure:

Folder: Hack1

- 01_MIA_67.pt
- priv_out.pt
- pub.pt
- example_assignment_1.py