NUS CS-CS5562: Trustworthy Machine Learning November 14, 2023

Assignment 6

Lecturer: Reza Shokri *Student:*

The objective of this assignment is to let you understand and implement federated learning algorithms. The assignment contains the following parts:

1. **Warm up: Running federated learning:** Implement the federated learning framework using an existing library to get familiar with the procedure for the FL model with multiple clients.
2. **Robustness in federated learning** Design and implement data poisoning attacks and model poisoning attacks in FL. Reason about the performance difference between the two attacks. Observe how the robust aggregation algorithm works under naive attacks and design adaptive attacks to bypass robust aggregation.

You will need to implement the code in `Assignment06_Task01_Warmup.ipynb` and `Assignment06_Task02_Robustness.ipynb`. You will also need to write a report about the tasks. Details about the exact items to be reported are given in the notebook. You write your answers within the notebook.

1 Warm up: Get familiar with federated learning

You are required to train a global model using the Flower federated learning library. After the model is trained, observe the accuracy of each client.

2 Robustness Issues of Federated Learning

We provide code and detailed information in `Assignment06_Task02_Robustness.ipynb`. Specifically, you are asked to implement the poisoning attacks against FedAvg and observe the performance of your attack when replacing the weighted average with the element-wise median. Lastly, you need to design your own adaptive attack for bypassing the median aggregation.

2.1 Task 1: Poisoning attacks

In the first subtask, you are asked to implement the label-flipping* data poisoning attack (finish the `train_poisoning_data` function in the notebook), in which the adversary can only modify the labels of the training data of the party.

In the second subtask, you are asked to implement the model poisoning attack (i.e., complete the function `train_poisoning_models`), in which the adversary can modify the local updates during the federated learning. In addition, you need to answer the questions in the notebook (`Report` cell). Details are provided in the notebook.

2.2 Task 2: Robust aggregation

The server may apply median instead of mean to aggregate the local models. In the first subtask, evaluate your attack designed in Task 1 under the median aggregation and observe how the performance of the model changes when you increase the number of malicious parties. In the second subtask, design your adaptive attack to bypass median aggregation and explain your attack strategy. More detailed information is provided in the notebook. The goal is to find an attack that makes the model perform worse than the non-attacked median model (this could be your initial attack). If you

*You are only allowed to change the labels and you can use any changing strategy you like as long as there is some effect and you change the labels.

need inspiration Baruch et al. [2019] might be a good place to start.

References

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.