AI in Medicine I
WiSe 2022/23

Technical University Munich
Department of Informatics

I31 - AI in Medicine and Healthcare
I32 - Computational Imaging in AI and Medicine

Theoretical exercise 13 **23. Jan. 2022**

Privacy-Preserving ML

The solutions will be discussed in the tutorial session

26. Jan 2022, 4-6 p.m. in lecture hall 5901.EG.051

For questions regarding this exercise sheet, please contact: `g.kaisis@tum.de`
For general questions, please contact: `course.aim-lab@med.tum.de`

1. Question 1
   (a) Recall the definition of the term "privacy" we gave in the lecture.
   (b) Sketch the communication model we discussed and title each component.
   (c) Why are immunity to post-processing and choice of prior important characteristics of a good privacy definition?

2. Question 2 Fill in the gaps in the following text.

   Federated Learning is a collaborative learning protocol between a (a) and several (b) whereby a (c) is trained at every one of the (b) and then aggregated at the (a). To ensure the safety of the protocol, it is important that the (a) and the (b) are mutually (d). Moreover, input privacy methods like (e), (f) and (g) and output privacy methods like (h) should be used.

3. Question 3
   (a) Define Differential Privacy (without mathematics) in terms of relative risks an individual faces when contributing their data to a scientific study.
   (b) Provide the formula for ε-Differential Privacy in terms of probabilities of events.
   (c) Define the terms *sensitivity*, *adjacency* and *symmetry* as they pertain to Differential Privacy.
   (d) Reflect on the similarity of the two definitions. Is Differential Privacy a "strong" privacy definition in the sense of Question (1) above?

4. Question 4 **(Thursday content)**

   A threat model of a learning setting can be summarised as follows:

   - Attack time: *inference* or *train*
   - Model access: *white-*, *gray-* or *black-box*
   - Adversarial intentions: *malicious*, *honest-but-curious*
   - Adversarial access: *participant*, *central entity*, *off-path observer*

Describe what threat model(s) correspond to:

1. Gradient-based model inversion attacker
2. Data poisoning attacker
3. API-only-access membership inference attacker

5. Question 5 **(Thursday content)**

   A federation of hospitals (between 7 and 10) are planning to run a collaborative training task on chest X-ray images as well as clinical records. Every participant trains the model locally and sends the model update to the central server (which is assumed to be trustworthy).

   You are an attacker who is taking part in this training protocol: Propose 3 concrete attacks you can conduct with corresponding threat models and the defences which can be used to mitigate your influence.

   The attacks can have the same formulation, but different implementations (e.g. you can run passive white-box membership inference from the shared updates or actively corrupt your updates to make the clients reveal more of their data and then run the inference attack; needless to say, you cannot use these examples in your response).