

COM 402 exercises 2023, session 12:

Machine Learning Security and Privacy

Exercise 12.1

- Are the following statements true or false? Justify.
 1. Stealing non-linear models is impossible because models are too complex.
 2. As a defender of a machine learning model you should be more worried about black-box effective attacks than white-box effective attacks.
 3. Privacy problems in machine learning stem solely from the need for data to train models.
 4. Poisoning attacks can be used to increase vulnerability to adversarial examples.

Solution 12.1

1. False. Stealing non-linear models is more costly than stealing linear models, but can be done. Linear models can be stolen by solving a simple system of linear equations, which is not possible for non-linear functions. However, one can steal the model by using the target as a "labeler" in order to train a new model that performs similarly to the target itself.
2. True. An adversary performing a black-box attack needs much less resources and capabilities than a white-box adversary. This is much more dangerous, as the adversary only needs the ability to interact with the model.
3. False. Data collection for training is one of many privacy attack vectors in machine learning. There exist attacks on models and outputs; and naturally exposing data for test is a risk in itself.
4. True. By providing poisoning inputs, the adversary gets to shape the boundaries of the model. Thus, she can carve this boundary to facilitate classification errors. In fact, you can understand a backdoor attack as a particular instance of an adversarial example.

Exercise 12.2

- A typical approach to avoid the processing of individual's personal data is aggregation. Discuss whether this is a good technique to avoid privacy risks when collecting data for training machine learning models.

Solution 12.2

Aggregation is a poor choice to enable privacy-preserving training of machine learning models. Three main issues:

1. Where / when do you do the aggregation? To aggregate you still need to collect the data. How to aggregate in a privacy-preserving way is also a hard problem as we explained in the next lectures. Also, on what groups should one aggregate? Depending on the task it may be better to aggregate on some users or on others. Deciding on which patients and how often to aggregate may affect both the privacy properties and utility of the aggregation (see the following two points).
2. The privacy provided by aggregation depends on the adversary's knowledge. We can learn membership/attributes from aggregates (think of the aggregates as a very, very simple machine learning model). Also, aggregates only protect when there is something to aggregate.

Imagine a situation in which all samples in a dataset have cancer. Aggregation will not protect the privacy of these patients.

3. Aggregation has great impact on utility, in particular for personalization-oriented tasks.

Exercise 12.3

- Can we prevent adversarial examples using encryption?
- And poisoning attacks?

Solution 12.3

No. None of them can be solved by encryption. The problem is derived from weaknesses in the model introduced by training or testing samples. Encrypting these samples does not change