COM 402 exercises 2024, 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.

Exercise 12.2

You're using an API that provides a machine learning model for classifying cat or dog images. You think that the model might be using a simple linear classifier. However, you don't have access to the model weights, but you can query the model with any image you want.

1. Are there any attacks you can perform to steal the model? If so, how would you do it?
2. How would you protect the model from such attacks?

Exercise 12.3

What are the main differences between:

- Opaque-box attacks
- Grey-box attacks
- Clear-box attacks

Exercise 12.4

- 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.

Solutions to the Exercises

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.

Solution 12.2

Assuming the model performs a "simple linear classification" on the image, we can *model* the model as such:

$f(I) = b + w_{0,0} \cdot I_{0,0} + w_{0,1} \cdot I_{0,1} + \ldots + w_{W,H} \cdot I_{W,H}$

where $I \in \mathcal{R}^{3 \times W \times H}$ is the RGB matrix of the image of width W and height H, $w_{i,j} \in \mathcal{R}^3$ are the model's weights ($b \in \mathcal{R}$ is a bias), and $\cdot$ denotes typical vector dot product.

The classic model stealing attack would then work, by simply inputting carefully crafted images to the model (altering the pixels one by one, color by color), to infer the value of one weight's component each query. This would thus require $3 \times W \times H + 1$ queries.

To protect the model, we can limit the query rate, add noise to the output, or even slightly noisen the input itself before throwing it at the model (this way, the attacker won't know what exactly the model is "answering" to).

Solution 12.3

- Opaque-box attacks: Model architecture and parameters unknown. Can only interact blindly with the model.
- Grey-box attacks: Model architecture known, parameters unknown. Can only interact with the model, but has information about the type of model
- Clear-box attacks: Known architecture and parameters. Can replicate the model and use the model's internal parameters in the attack

Solution 12.4

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-oriened tasks.