COM 402 exercises 2025, 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

- You are the new VP for Education at EPFL. Your team tells you that they want to install a new plagiarism detection mechanism. They propose to buy a tool called YouAreCaught for Master theses. In the specifications of this tool they promise that:
  YouAreCaught misses 10% of the True plagiarism cases
  YouAreCaught makes mistakes on 3% of the False plagiarism cases, flagging them as plagiarism

  You know that at EPFL students are very honest, i.e., only 5 in 1000, plagiarise in their Master thesis. Is YouAreCaught a good tool for you? Justify.
- What percentage of students need to be cheating for YouAreCaught to provide good performance?

Solution 12.2

- Pr[Real plagiarism | YouAreCaught says plagiarism] = 0.9 * 0.005 / (0.9 * 0.005 + 0.03*0.995) = 0.13

  Not a good tool. Only finds 4% of the plagiarism cases, the rest are false positives!

- Pr[Real plagiarism | YouAreCaught says plagiarism] = 0.9 * cheat / (0.9 * cheat + 0.03*(1-cheat))

  40% or more to be on 95%, or 80% to be above 99%. This tool is just not good.

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

Exercise 12.4

What are the main differences between:

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

Solution 12.4

- 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

Exercise 12.5

- 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 12.5

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.

Exercise 12.6

1. You are tasked with designing a new algorithm for assigning lockers at EPFL. Of course, you decide to use Machine learning (what else!). Given the diversity in the EPFL student population, what should you pay attention to if you want your system to be fair? How many subpopulation do you need to take into account?

2. To avoid any type of problems regarding fairness with respect to demographic algorithms, you decide to only consider grades from last year as training, and this semester grader as input to the final decision. Discuss whether you should be worried about adversarial examples and poisoning attacks.

Solution 12.6

1. Fairness, as privacy, is a hard concept to define. You should take care that your algorithm does not discriminate people depending on *none* of their attributes (gender, race, sexual orientation, ...). The number of subgroups to take into account is potentially infinite. There is no clear answer to this question. The rationale of this exercise is to make you reflect upon how difficult it is to create an algorithm that is fair in all possible dimensions.

   More info - Arvind Narayanan. *Tutorial: 21 fairness definitions and their politics*: `https://www.youtube.com/watch?v=jIXIuYdnyyk`

2. Poisoning is not a concern, as data from the past cannot be changed. In principle, one would say that adversarial examples are not a problem either, as the cost associated to changing grades is most likely too high compared to the benefits from having a locker in campus. However, if you are a really paranoid security designer (or you really value the availability of lockers in campus), you can consider this attack in your threat model.