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.

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?

Exercise 12.3

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

Exercise 12.4

What are the main differences between:

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

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.

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.