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.