Assignment 3: Adversarial attacks and defenses

Innopolis University

The security and interpretability of machine learning 2022 - Bachelors

1 General Instructions

In this assignment, you are going to attack and defend the model you built in the first homework with adapted attacks. You are required to submit your solutions via Moodle as a single ipynb file. Do not forget to include your name in the submitted document.

The source code should contain adequate internal documentation in the form of comments. Internal documentation should explain why and how you apply the instructions.

Plagiarism will not be tolerated, and a plagiarised solution will be heavily penalized for all parties involved. Remember that you learn nothing when you copy someone else's work, which defeats the exercise's purpose! You are allowed to collaborate on general ideas with other students as well as consult books and Internet resources. However, be sure to credit all the sources you use to make it clear what part of your solution comes from elsewhere.

2 Defense and Attacks (30 points)

The paper " Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods" explains how to attack ten proposed defenses. Implement one of the following defenses and attacks against them from the paper:

1. Adversarial Retraining (Adversarial samples detection): The defense proposes designing a detector that is trained to detect adversarial attacks (refer to the paper for the details)

2. Kernel Density Estimation : A statistical test is proposed as a defense (refer to the paper for the details).

3. Dropout Randomization : The defense shows how introducing randomness might increase the robustness of the model (refer to the paper for the details).

In any case you choose, you should replicate all the experiments showed in the paper on the defense (Zero-Knowledge Attack Evaluation, Perfect-Knowledge Attack Evaluation, Limited-Knowledge Attack Evaluation).

Please read the paper carefully since you might be asked about any other defenses in this paper. The results will be presented in the final presentations but you need to submit the assignment before for revision.