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. 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 Attacks (15 points)

Write the code for the three known adversarial attacks : FGSM, PGD, CW. Then, test these attacks on the model you built in the first homework.

3 Defense (5 points)

Write the code for adversarial training and show how you can use it to defend your model. Show how you measure the robustness before and after applying the defense.