Smridhi Patwari
smridhip
01/27/2026

95767 Cybersecurity for Artificial Intelligence & Machine Learning

Assignment #2

150 points

In this Assignment, you will complete the Hands-on exercise with the iris_extended dataset contained in the Notebook provided in the Files section of Canvas. Complete all sections of the Notebook. Then answer the following questions:

Deliverables

A completed Notebook with all sections completed as per the instructions in the Notebook.

Written answers to the below questions as a Word or PDF document.

1. Availability Poisoning Attack—Why does adding chaff data (random noise data) to the training dataset reduce the performance of the trained model?

   The chaff data adds noise to the training data which the model ultimately trains upon and outputs greater amount of incorrect predictions on the test data. This eventually reduces the model's performance.

2. Randomized Smoothing—In the first question, we see that adding random noise to the training dataset reduces model performance. Why then is randomized smoothing, adding random noise to query inputs, used as a mitigation technique? In 1-2 sentences, briefly describe how randomized smoothing can protect a ML model from one type of AML attack. Be sure to name the type of attack in your answer.

   The type of attack is an evasion attack. Randomized smoothing adds noise to test inputs at inference time and aggregates predictions over multiple noisy versions. This creates a protective zone around each input where small malicious changes cannot flip the prediction.

3. Functional Stealing Model Extraction Attack—If you completed this section of the notebook successfully, you (as a hypothetical adversary), now have a shadow model that provides similar predictions as the target model. Name another AML attack this shadow model could enable you to perform. In a few sentences, describe how the shadow model would be used in this attack.

   This shadow model can perform membership inference. In this attack the adversary tries to determine if a specific data point was part of a machine learning model's training dataset. The shadow model is used to learn the output

patterns for the 'in-training' vs the 'out-of-training' examples. By comparing the target model's output on a record to those patterns, the attacker can infer membership and thus reveal sensitive information about whether that individual's data was used.

4. Functional Stealing Model Extraction Attack—In the notebook, you trained a shadow model using a random forest classifier to emulate the logistic regression target model. Logistic regression and random forests do not have the same parameters or model architecture. Why is this not a problem for creating a shadow model?

   This is not a problem since the eventual goal of the shadow model is to mimic the functions of the actual model. So long as the shadow model is able to approximate the input-output behaviour of the query data, it can be used for downstream attacks even with a different internal architecture. Different models are still able to learn similar decisions even with different decision boundaries.

5. Suppose a customs enforcement team trained the original target model in the notebook and is using it to screen plants that are being exported. The customs enforcement team is barring *setosa* from being exported. You have a large delivery of *setosa* plants you want to export to make a large profit. You are unable to tamper with the target model or data because the customs enforcement team took Cybersecurity for Artificial Intelligence & Machine Learning. You know that one of your plants will be selected for screening by the customs enforcement team. Measurements will be taken on that plant and fed into their ML model to predict the plant's species. If they determine your plant is a *setosa* plant, your whole delivery will be destroyed. Describe what you could do to the plant that will be screened to attempt to evade the model. (You cannot replace it with another species).

   While it might not be the most feasible to implement this in real life especially if the plants are randomly chosen for inspection, but something that can be done would be to send in a batch of plants with similar dimensions of the leaves as setosa. If the selected sample has measurements that lies within the range of the species that are not setosa, it might be able to avoid getting detected by the ML model.