Reading 8 : AI ML Challenges

Name: Pushpak Vijay Katkhede **Date: 03/06/2023**

---

1. *What are three kinds of security violations that the article says an attacker might try to achieve (see section 3.1, paragraph 2)? Explain each one in your own words.?*

Ans: The three types of security violations that the article talks about are (i) Integrity violations, (ii) Availability violations, and (iii) Privacy violations:

1. <u>Integrity Violation</u> → This type of attack can make the system to perform incorrect detections due to the attack. For example, instead of detecting class A the classifier detects class B due to various factors like introduced bias or data poisoning done by the attacker. Overall, the motive of the attacker is to make the classifier produce incorrect results without obstructing the normal working of the model or hindering the mechanics behind it.
2. <u>Availability Violation</u> →This kind of attack is like a standard Denial of Service (DoS) attack on machine learning models to make them unavailable to the users who are original users of the system. This kind of violation can be introduced by tampering with the system's accessibility features to the end users or by introducing the middle system to absorb the full bandwidth available for the end users. The effect of this will make the system inaccessible to the end users.
3. <u>Privacy Violation</u> → In this type of violation, the attacker gets access to the internal information about the model, for an instance hyperparameters of the classifiers. This information can also be about end users like personally identifiable information. The data can be breached when it is stored on the server, or the datasets are exposed to public domains. All of these can be used to identify the individuals and hence, violate the privacy of the system, user, or data.

2. *(a) In Figure 3, why might you prefer to use C1 versus C2, even though C2 has higher accuracy?*

Ans: At the initial condition, where there is no attack, the accuracy of the C2 is slightly higher than C1, but as we simulate the attack strength in a increasing manner we see that the curve for C2 accuracy is falling down drastically as compared to that of the C1. Which means that the classifier C2 is less robust than C1 to adversarial attacks. Hence, assuming the tradeoff between c2 and c1, given an attack happens c1 will perform better than c2 in even the smallest strength attack happened. This is evident from the curve for c2 getting dropped below c1 suddenly after small attack strength.

(b) What is the difference between a "poisoning" and an "evasion" attack, in your own words?

Ans: Poisoning → In poisoning the input data is tempered by the attacker by putting malicious data in the training data so the model gets trained on false data. Hence, the classifier will produce false outputs upon deployment. The data is usually incorrect class labels or giving sense of false pattern to divert the classifier from true labels while training. Ultimately, the goal is to produce biased results in live setting. This causes the system to lose its important value of integrity. This is usually happened in the training phase of the data.

Evasion → Evasion attack is done by manipulating the input data so that the output will be incorrect. Usually , this is noise data and make the model predict incorrect outcomes. This attack is just to tamper the current testing results by adding a small disorder in the data being feed. This usually occurs in the deployment phase of machine learning.

3. *What is one concept or term in this reading that was new to you? What does it mean, in your own words?*

Ans: The concept of determining the attack in terms of available information and presenting it as a model was very foreign to me.

White box attack → When attacker have full information of the machine learning system.

Grey Box attack → When attacker have partial information of the machine learning system.

Black Box attack → When the attacker is assumed to have no information about the working principle or data in the system, but the attack is done on the basis of feedback mechanism. However, this assumption does not hold true fully as the attacker may partially know information about the feature or confidence scores, etc.

4. *Reflect on your work for the course over the past week. What did you do that was effective and increased your knowledge? What could you do or change to increase what you gain from this course? Is there anything about this course you are anxious / worried about? (There are no wrong answers here; this is your chance to maximize what you get out of the course and to let me know about any concerns.)*

Ans: This week, I have worked on the draft of the project. I explored various papers and techniques for classifiers and hyper-tuning the classifiers to attain better accuracy. The concept of calibration was also intriguing for me, so I spent a couple of hours reading and watching videos about that and specifically the tradeoff between accuracy and calibration. This helped me understand the topic in greater detail. Currently, I am not tensed or worried about anything and just follow the process to complete the assignments and projects.