Week 3 Workshop

FIT5230 Malicious AI

Adversarial Machine Learning II

Adversarial attacks

1. **Poisoning** Attacks: Mislead the model by corrupting the training data.
2. **Evasion** Attacks: Craft adversarial examples to look genuine to humans but are misclassified by the model.
3. Backdoor Attacks: Embeds a hidden pattern in the training data that triggers malicious behavior.
4. **Model Extraction** Attacks: Use a series of queries to reconstruct the model.
5. **Inference** Attacks: Analyze the model's outputs to infer sensitive information.
6. **Transfer** Attacks: Use adversarial examples generated for one model to attack another model

Activity 1

What you need to do: (30 mins)

1. **Understand** the attacks from the list.
2. **Explain** how it works, and its potential impact: https://shorturl.at/a8iyn
3. Teams will be randomly selected to **present** their ideas.

Similarities:

1. **Adversarial Intent**: All these attacks are designed to exploit vulnerabilities in machine learning models, aiming to cause harm or gain unauthorized access.

2. **Manipulation of Data or Model**: Each attack involves some form of manipulation, whether it's the training data, the input data, or the model itself.

3. **Security Threats**: They all pose significant security threats to the integrity, confidentiality, and availability of machine learning systems.

4. **Technical Complexity**: Implementing these attacks typically requires a deep understanding of machine learning algorithms and systems.

Differences:

Point of attack

- Poisoning Attacks: Target the **training phase** by corrupting the training data.
- Evasion Attacks: Target the **inference phase** by crafting inputs that are misclassified by the model.
- Backdoor Attacks: Embed hidden patterns in the **training data** that trigger malicious behavior during **inference**
- Model Extraction Attacks: Focus on **querying** the model to reconstruct its functionality.
- Inference Attacks: Aim to **extract sensitive information** from the model's outputs.
- Transfer Attacks: Use adversarial examples from **one model** to attack **another model**.

Differences:

Objective/Goal

- Poisoning Attacks: Mislead the model **during training** to degrade its performance.
- Evasion Attacks: Bypass the model's defenses **during inference**.
- Backdoor Attacks: Create a **hidden trigger** that activates malicious behavior.
- Model Extraction Attacks: Steal the model's **intellectual property**.
- Inference Attacks: Extract **sensitive information** from the model.
- Transfer Attacks: Leverage vulnerabilities in **one model** to attack **another**.