SIT719 Security and Privacy Issues in Analytics

Pass Task 7.1: Taxonomy of Attacks, Defenses, and Consequences in Adversarial Machine Learning

Adversarial Techniques used for launching Attacks against Targets may apply to the Training or Testing (Inference) phases of system operation.

Attack types:

Attacks in the Training phase attempt to acquire or influence the training data or model itself.

Data Access Attacks, some or all of the training data is accessed and can be used to create a substitute model. This substitute model can then be used to test the effectiveness of potential inputs before submitting them as Attacks in the Testing (Inference) phase of operation.

Poisoning, also known as Causative Attacks, the data or model is altered indirectly or directly. In Indirect Poisoning, adversaries without access to pre-processed data used by the target model must instead poison the data before pre-processing.

Attacks in the Testing (Inference) phase, also known as Exploratory Attacks, do not tamper with the target model or the data used in training. Instead, these Attacks generate adversarial examples as inputs that are able to evade proper output classification by the model.

In **Evasion Attacks**, or collect and infer information about the model or training data, in Oracle Attacks. In Evasion Attacks, the adversary solves a constrained optimization problem to find a small input perturbation that causes a large change in the loss function and results in output misclassification.

Techniques used to launch Attacks against Targets, threats to ML components also depend on the adversary's Knowledge about the target model.

In **Black Box** Attacks, the adversary has no knowledge about the model except input-output Samples of training data or 4input-output pairings obtained using the target model as an Oracle.

In **Gray Box** Attacks, the adversary has partial information about the model, which may include the Model Architecture, Parameter Values, Training Method (Loss Function), or Training Data.

In **White Box** Attacks, the adversary has complete knowledge of the model including architecture, parameters, methods, and data.

Defences:

Defences Against Training Attacks involving Data Access include traditional access control measures such as **Data Encryption**. Defences against Poisoning Attacks include **Data Sanitization** and **Robust Statistics.**

Defences Against Testing (Inference) Attacks include various model **Robustness Improvements**, including Adversarial Training, Gradient Masking, Defensive Distillation, Ensemble Methods, Feature Squeezing, and Reformers/Autoencoders.

Consequences:

The Consequences of Attacks against Targets depend on implemented Defenses. For a given combination of Attack (including Target, Technique, and Knowledge) and Defense(s), the Consequences can be characterized categorically as **Violations of Integrity, Availability, Confidentiality,** or **Privacy**. Within each category, varying levels of severity may also be used to measure the violation of security