Deakin University

Security and Privacy Issues in Analytics

SIT719-Task 7.1P

04-09-2021

Submitted By:

Abdul Rahman Deyab Alhojory

Table of Contents

Five Important Attacks

After Reading the complete article I have found following five important attacks types.

1. Data Access Attacks
2. Poisoning or Causative Attacks
3. Exploratory Attacks
4. Evasion Attacks
5. Oracle Attacks

1. Data Access Attacks

Data Access attacks are most prominent attacks in which partial or all data for training is accessed. Then this data is used to compile substitute or alternate models. Then these models are used to test the effectiveness or incoming inputs before submitting them as attacks inside testing phase.

2. Poisoning or Causative Attacks

In this type of attacks information which is used for training is altered. These attacks can be of two types.

- Indirect Poisoning
- Direct Poisoning

.1 Indirect Poisoning

The data is altered before preprocessing of data.

.2 Direct Poisoning

The data inside direct poisoning can be poisoned by multiple ways. Like Data Injection or Data Manipulation or the model is manipulated by logical corruption.

Overall inside poisoning attacks information is altered indirectly means before processing or in direct processing information is changed inside training data directly. It can be done by adding new data, or updating the data. Except these inside logic of ML models there can be some loop holes which also a main reason of Causative Attacks.

3. Exploratory Attacks

In this type of attacks model is given a specific input which has larger impact on the model. It used help of gradients based algorithms like JSMA, FGSM and L-BFGS. So L-BFGS was the first such algorithm which is used to carry out attacks. But later FGSM helps us to improve that. AT the end FGSM provides good and better control to training model to process.

4. Oracle Attacks

In this type adversary [1] used api for training and testing the model. At the end results are generated. In this type input and output are in form of pair and used to create the substitute models. It has inversion, extraction, and the evasion attack as well. Which is explained below.

- Extraction Attacks (It discover the structure of parameters used inside models for predictions)
- Inversion Attacks (It aims to figure out data used inside training and testing inside model)
- Inference Attack (It figure out given data is part of training dataset or not.)

5. Evasion Attacks

In this type input is altered which helps us to avoid from trained classifier during testing.

Defense Mechanism

Defenses can be characterized by whether they apply to Attacks launched against the Training or Testing (Inference) phases of system operation.

1. Defenses Against Training Attacks
2. Defenses Against Testing Attacks

1. Defense Against Training Attacks

Defenses in training attacks have include these steps. Like Encryption and data Sanitization

1.1 Data Encryption

In this system data encryption is used to ensure the confidentiality of data which is being used. It involves the encryption and decryption to access data for substitute models.

1.2 Data Sanitization and Robust Statistics

In Data Sanitization, adversarial examples [2] are identified by testing the impacts of examples on classification performance. Examples that cause high error rates in classifications are then removed from the training set, in an approach known as Reject on Negative Impact. Rather than attempting to detect poisoned data, Robust Statistics use constraints and regularization techniques to reduce potential distortions of the learning model caused by poisoned data.

2. Defense Against Testing Attacks

It includes various model Robustness Improvements, including Adversarial Training [3], Gradient Masking, Defensive Distillation, Ensemble Methods, Feature Squeezing, and Reformers/Auto encoders. Although used as Defenses against Attacks made in the Testing (Inference) phase, these Defenses [1] are deployed by the defender in the Training phase that precedes Testing (Inference). In Adversarial Training, inputs containing adversarial perturbations but with correct output labels are injected into the training data in order to minimize classification errors caused by adversarial examples. Gradient Masking reduces the model’s sensitivity to small perturbations in inputs by computing first order derivatives of the model with respect to its inputs and minimizing these derivatives during the learning phase. A similar idea motivates Defensive Distillation, where a target model is used to train a smaller model that exhibits a smoother output surface, and Ensemble Methods, where multiple classifiers are trained together and combined to improve robustness.

Consequences

The consequences [4] of attacks depend on implemented defenses. The consequences can be categorized as

1. Violations of Integrity
2. Availability Violations
3. Confidentiality Violations

1. Violations of Integrity

The inference process is undermined, resulting in Confidence Reduction or Misclassification [1] to any class different from the original class. More specific misclassifications include Targeted Misclassification of inputs to a specific target output class and Source-Target Misclassification of a specific input to a specific target output class. In Unsupervised Learning, an Integrity Violation may produce a meaningless representation of the input in an unsupervised feature extractor. [2] In Reinforcement Learning, an Integrity Violation may cause the learning agent to act unintelligently or with degraded performance in its environment

2. Violations of Availability

Availability Violations induce reductions in quality (such as inference speed) or access (denial of service) to the point of rendering the ML component unavailable to users. Although Availability Violations may involve Confidence Reductions or Misclassifications similar to those of Integrity Violations, the difference is that Availability Violations result in behaviors such as unacceptable speed or denial of access that render a model’s output or action unusable.

3. Violations of Confidentiality

Privacy Violations are a specific class of Confidentiality Violation in which the adversary obtains personal information about one or more individual and legitimate model inputs, either included in the training data or not. It occurs when an adversary extracts or infers usable information about the model and data. Attacks on confidential information about the model include an Extraction Attack that reveals model architecture or parameters, or an Oracle Attack that enables the adversary to construct a substitute model. Attacks that reveal confidential information about the data include an Inversion Attack whereby an adversary exploits the target model to recover missing data using partially known inputs, or a Membership Inference Attack whereby an adversary performs a membership test to determine if an individual was included in the dataset used to train the target model.

References

[1] K. J. B. M. H. A. D. M.-M. T. S. Elham Tabassi, "A Taxonomy and Terminology of Adversarial Machine Learning," U.S. Department of Commerce Wilbur L. Ross, Jr., Secretary, Washington US, 2019.

[2] A. D. J. B. N. B. I. P. R. a. J. D. T. L. Huang, "Adversarial Machine Learning," in Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence," NY USA, New York, 2011.

[3] . Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision," IEEE, 2018.

[4] P. L. W. Z. W. C. S. Y. a. V. C. M. L. Q. Liu, "A survey on security threats and defensive techniques of machine learning: A data driven view," IEEE, 2018.

[5] R. E. T. X. H. I. R. P. N. G. a. S. S. P. Kuznetsov, "Adversarial Machine Learning," in Artificial Intelligence Safety and Security,," Chapman and Hall, 2018.