CS255: Computer Security

Machine Learning in Security

Chengyu Song

Machine Learning

Making Decisions without Explicit Instruction

- Task: solving a problem (e.g., classification, regression, decision, etc)
- Approaches
  - Manual programming (e.g., logical rules, heuristics): explicit instructions
  - Classic ML: manually defined feature space, but no explicit instructions
  - Deep learning: self-learned features, no explicit instructions

Machine Learning

General Approaches

- Supervised learning: requires labeled training data
  - Self-supervised learning: label can be generated automatically
- Unsupervised learning: no labeled data
- Reinforcement learning: environment and rewards

Machine Learning in Security

- Security researchers have been using ML for a long time
    - Intrusion detection (1987)
    - Malware classification
    - Bug finding

- But the proposed methods rarely work in practice, **WHY**?

Outside the Closed World

On Using Machine Learning For Network Intrusion Detection

- Fundamental challenges in outlier detection
- High cost of errors
- Semantic gap between results and their operational interpretation
- Enormous variability in input data
- Fundamental difficulties for conducting sound evaluation

The idea of specifying only positive examples and adopting a standing assumption that the rest are negative is called the closed world assumption. . . . [The assumption] is not of much practical use in real- life problems because they rarely involve "closed" worlds in which you can be certain that all cases are covered.

I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques (2nd edition). Morgan Kaufmann, 2005.

Outside the Closed World

Outlier Detection

- Classification can and can be good at detecting **known attacks**
- Classification **cannot** detect **new attacks**
  - Lack of training data
- Anomaly detection does not work in open world
  - High false positives

Outside the Closed World

High Cost of Errors

- ML models usually have to trade-off between precision (false positive rate) and recall (false negative rate)
- These errors are usually fine in other ML applications
  - Recommendation systems, OCR (image recognition), spam filter
- But errors in IDS (or system solutions in general) have much higher cost
  - False positives: unusable
  - False negatives: attacks

Outside the Closed World

Semantic Gap

- How to interpret the output of a ML model?
  - or How the **features** the anomaly detection system operates on relate to the semantics of the operational environment (e.g., network)?
- This is especially bad for deep learning models
  - Pentagon project from 1980s: a neural network was trained to detect tanks in photos; however, that the datasets used for training and evaluation shared a subtle property: photos of tanks were taken on a cloudy day, while all others had a blue sky.

Outside the Closed World

Diversity in Input Data

- Raw input data (e.g., network traffic, malware binaries) in cyber space are high-dimensional and heavy-tailed
  - Without understanding/extracting high-level semantics, ML models are likely to pick up superficial or even harmful features
  - It is also easy for attackers to bypass the detection through simple transformations

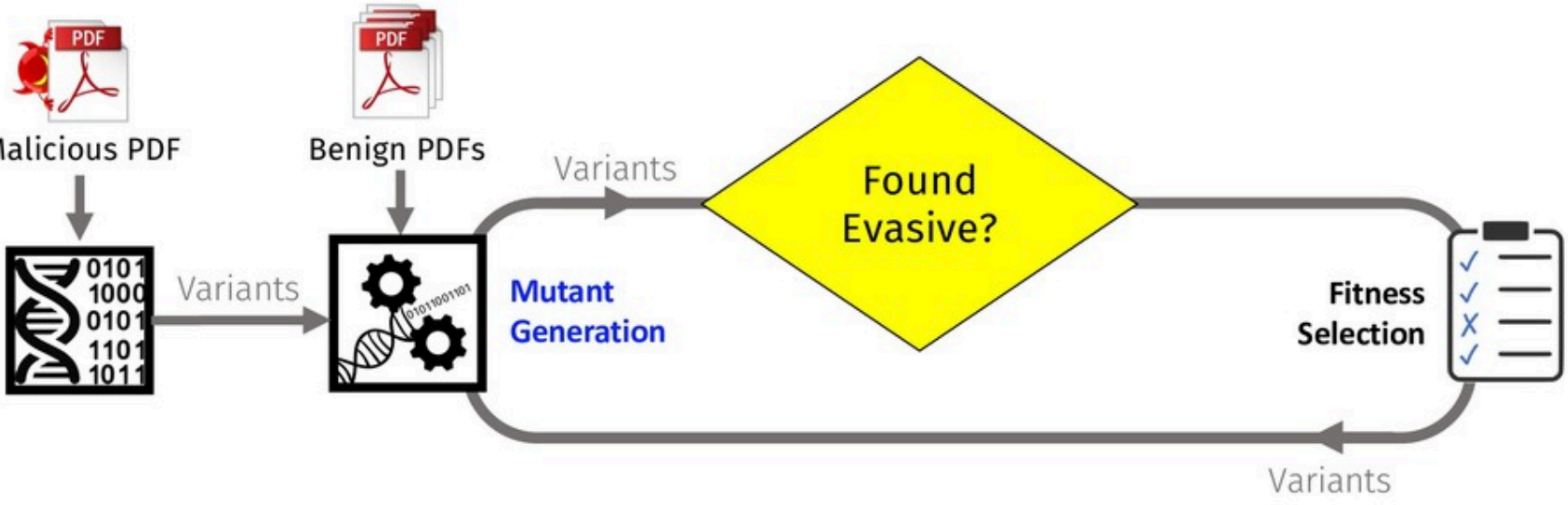
Outside the Closed World

Sound Evaluations

- Realistic dataset is extremely rare
  - Hard to access, usually contains sensitive information (network traffic) or potential harmful activities (malware)
- Semantic gap
- Adversarial settings

Attacking ML models

Malicious PDF



Weilin Xu, Yanjun Qi, and David Evans. Automatically Evading Classifiers A Case Study on PDF Malware Classifiers. Network and Distributed Systems Symposium 2016

Attacking ML models

Malware Detection

- Semantic equivalent transformations (metamorphic)
  - Guided by feedback from the model

Keane Lucas, Mahmood Sharif, Lujo Bauer, Michael K. Reiter, and Saurabh Shintre. Malware makeover: breaking ML-based static analysis by modifying executable bytes. In Proceedings of the ACM Asia Conference on Computer and Communications Security, June 2021.