Summary of Paper: Can Machine Learning be Secure?

Presenter: Joseph Tobin

Department of Computer Science, University of Virginia

https://qdata.github.io/deep2Read/

Can Machine Learning be Secure? (2006)

- Authors: Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, J. D. Tygar

ABSTRACT

Machine learning systems offer unparalled flexibility in dealing with evolving input in a variety of applications, such as intrusion detection systems and spam e-mail filtering. However, machine learning algorithms themselves can be a target of attack by a malicious adversary. This paper provides a framework for answering the question, "Can machine learning be secure?" Novel contributions of this paper include a taxonomy of different types of attacks on machine learning techniques and systems, a variety of defenses against those attacks, a discussion of ideas that are important to security for machine learning, an analytical model giving a lower bound on attacker's work function, and a list of open problems.

Introduction

- Can an adversary manipulate a learning system?
  - Degrade the performance?
  - Allow certain attacks?
- What are current defense mechanisms?
- Can properties of machine learning systems be exploited to disrupt system?
- Taxonomy of different attacks and defenses
- Security ideas important for machine learning
- Analytical model giving a lower bound on work function
- List of open problems

Terminology and Attack Model

- Attack targets a learning system
- Intrusion targets a computer (protected by a learning system)
- Adversaries have understanding of the learning algorithms

		Integrity	*Availability*
Causative:	*Targeted*	Permit a specific intrusion	Create sufficient errors to make system unusable for one person or service
	Indiscriminate	Permit at least one intrusion	Create sufficient errors to make learner unusable
Exploratory:	*Targeted*	Find a permitted intrusion from a small set of possibilities	Find a set of points misclassified by the learner
	Indiscriminate	Find a permitted intrusion	

Table 1: The attack model.

Online Learning

- Online learning allows the learner to adapt to changing conditions
- Allows for more flexibility
- Also simplifies causative attacks (attacks that change data)
  - Difficult to detect adversary if they gradually change function over time
  - More simple attack

Defenses: Robustness

- Regularization

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{(x_i, y_i) \in \mathcal{S}} \ell(f(x_i), y_i) + \lambda J(f) \right\} \quad (3)$$

- Used when there is little data or noisy data
- “Encoding a prior distribution on the parameters, penalizing choices that are less likely *a priori*”

Defenses: Robustness

- Regularization smooths out the solution and removes complexity (that was added by adverary or that an adversary may exploit)
- Prior distribution can help encode important knowledge about the domain or domain structure
- When the learner has more prior information to base learning, there is less dependence of data fitting
    - Adversary has less influence over process

Defenses: Disinformation

- Confuse adversary’s estimate of the learner’s state
- Especially prevent adversary from learning the decision boundary
- Learner attacks adversary with indiscriminate causative availability attack
- Trick adversary into thinking that a particular intrusion was not included in training set
- Set up honeypot so that when that intrusion is performed often enough, you can identify adversary
- Learner attacks adversary with targeted causative integrity attack

Defenses: Randomization for Target Attacks

- Targeted attacks are dependent upon the classification of a small set of points
- Thus, they are highly sensitive to the placement of the decision boundary
- If there is randomization in the placement of the boundary, model accuracy can be maintained while making it more difficult for targeted attacks

Defenses: Summary

- Tradeoff of expressivity and security

		Integrity	*Availability*
Causative:	*Targeted*	• Regularization • Randomization	• Regularization • Randomization
	Indiscriminate	• Regularization	• Regularization
Exploratory:	*Targeted*	• Information hiding • Randomization	• Information hiding
	Indiscriminate	• Information hiding	

Table 2: Defenses against the attacks in Table 1.

Scale of Training

- Learner can use data from single source or multiple sources
- Tradeoff between size of data and secrecy of classifier
- Most of the time, we cannot assume all information in training set is secret
- Thus, difficult to measure how beneficial it is to keep training data and classifier secret

Scale of Training: Adversary Observations

- Deduce decision boundary by repeated probes
- No information about classifier: probes roughly proportional to size of space
- Information about learning algorithm: possibly few specific probes
- Given that adversary knows decision boundary, they can avoid detection by operating in misclassified space
  - More difficult to find space if classification points are mapped to some abstract space and classification is done in that space
- Advantage depends on boundaries
  - We can construct boundaries that give no information or boundaries that reveal (confidential) information about the data set

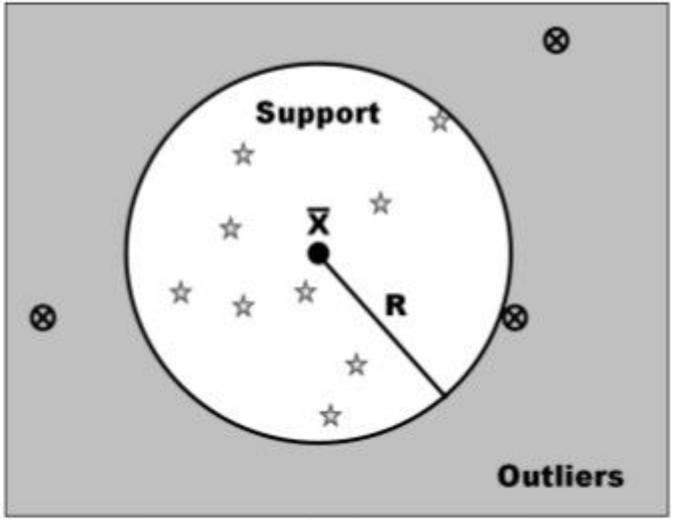
Theoretical Results

- Present model for a causative attack trying to manipulate naive learning algorithm
- Yields an optimal policy for adversary and a bound on effort required to achieve adversary’s objective
- Outlier detection: task of identifying anomalous data and is widely used for various security tasks

Model

- Multidimensional hypersphere centered at mean of data where data inside sphere are classified as normal and data outside are classified as outliers
- Only admits new training points into the set if they are classified as normal (“bootstraps itself”)
- Hypersphere is centered at X0 and has a fixed radius R
- Attack is iterated over course of T > 1 iterations and the i-th iteration

Model

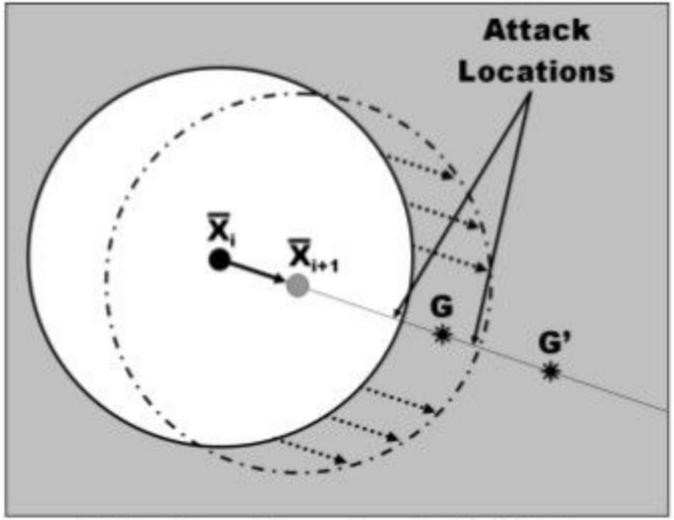


(a) Hypersphere Outlier Detection

Attack Strategy

- We want to adjust the model such that it classifies a specific outlier datapoint G as a normal datapoint
- We shift the sphere over several iterations of training until it covers the point
- *Causative targeted integrity* attack
- Feed points that are located where the line between the mean and G intersect the boundary of the sphere
- At i-th iteration, adversary places alpha_i at this location for optimal displacement
- Effort of adversary measured as the sum of alpha_i for all times

Attack Strategy



(b) Attack on a Hypersphere Outlier Detector

Optimal Attack Displacement

- D_{R,T}({alpha_i}) is the relative displacement caused by attack sequence alpha_i at iteration i

$$\frac{X_T - X_0}{R}.$$

- M_i = sum(alpha_j) from j = 1 to i

$$\sum_{j=1}^{i} \alpha_j,$$

- Relative distance of a series of moves:

$$D_{R,T}(\{M_i\}) = T - \sum_{i=2}^{T} \frac{M_{i-1}}{M_i}$$

Optimal Attack Displacement

- By upperbounding previous equation, we can bound minimal effort M* of the adversary
- More specifically, for a particular M, we want a optimal sequence {M_i*} that achieves maximum relative displacement D_R,T(M)
- If there is no time constraction M*i = i (single point per iteration)
- If T < M iterations, then $M_i^* = M^{\frac{i-1}{T-1}}$.
- Giving us $$D_{R,T}(M) \leq T - (T-1) \cdot M^{\frac{-1}{T-1}} \leq T$$

Bounding the Adversary's Effort

- We can then use previous equation (monotonically increasing) to bound adversary's capability as

$$M^* \geq \left(\frac{T-1}{T-D_R}\right)^{T-1}$$

- Tradeoff between using a large number of attack points or extending attack over many iterations
- Bound decreases exponentially as number of iteration increases for D_R > 1
- For D_R <= 1 allows adversary to win in one iteration

Future Research Directions

- Information: how important is it to keep information secret from an adversary?
- Arms race: Can arms races be avoided in online learning systems? (spam arms race)
- Quantitative measurement: Can attacks be measured quantitatively?
- Security proofs: Can we bound information leaked by learner?
- Detecting adversaries: What side effects can we observe to reveal adversary’s attack?

Conclusion

- Machine learning is subject to a variety of new attacks
- Related work
  - Game theory
  - Reverse engineering
  - Tricking spam filters
  - Potential for control theory to have applications

Additional References

- http://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf
- https://arxiv.org/abs/1406.2661
- http://www.deeplearningbook.org/
- https://en.wikipedia.org/wiki/Stability_(learning_theory)