Adversarial Machine Learning

Sanjay Seetharaman

TRDDC

TATA CONSULTANCY SERVICES

Acknowledgements

Rosni K V, Researcher [1]
Manish Shukla, Researcher [1]
Dr. Sachin Lodha, Head, Cybersecurity and Privacy Research [1]

Dr. Shina Sheen, Associate Professor [2]
Dr. N. Rajamanickam, Assistant Professor [2]
Dr. R. S. Lekshmi, Professor [2]
Dr. R. Nadarajan, Professor and Head [2]

[1] Tata Consultancy Services
[2] Department of Applied Mathematics and Computational Sciences, PSG College of Technology

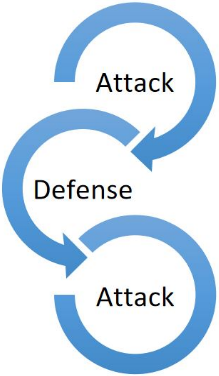
Outline

About TCS Research

- TCS established its first lab in 1981 at Pune, India. It has invested in multiple research areas for over three decades.
- Every year, our researchers publish around 300 papers for Tier 1 conferences. TCS has also filed over 3,500 patents.
- Numerous tools and frameworks have been co-created for delivering complex projects for global enterprises.
- Stakeholders include industries, organizations and academia.

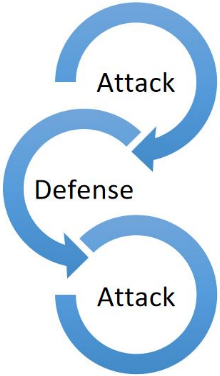
Outline

Security is an Arms Race



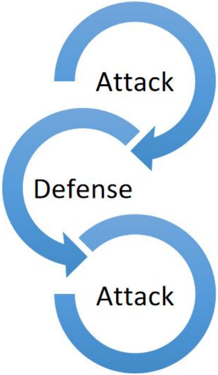
- Reactive Security
  - Unable to prevent the risk of `never-seen-before` attacks
- Proactive Security
  - Requires identification of `relevant` threats and development of corresponding countermeasures

Security is an Arms Race



- Reactive Security
  - Unable to prevent the risk of `never-seen-before` attacks
- Proactive Security
  - Requires identification of `relevant` threats and development of corresponding countermeasures

Security is an Arms Race



- Reactive Security
  - Unable to prevent the risk of `never-seen-before` attacks
- Proactive Security
  - Requires identification of `relevant` threats and development of corresponding countermeasures

Outline

Security Violation

- **Integrity violation**: Evade detection without compromising normal system operation
- **Availability violation**: Compromise the normal system functionalities available to legitimate users
- **Privacy violation**: Obtain private information about the system, its users or data by reverse-engineering the learning algorithm

Attack and Error Specificity

Attack Specificity

- **Targeted attack**: Cause misclassification of a specific set of samples/system users/services
- **Indiscriminate attack**: Target any system sample/system user/service

Error Specificity

- **Generic**: Aim to have a sample misclassified
- **Specific**: Aim to have a sample misclassified as a specific class

Attack and Error Specificity

Attack Specificity

- **Targeted attack**: Cause misclassification of a specific set of samples/system users/services
- **Indiscriminate attack**: Target any system sample/system user/service

Error Specificity

- **Generic**: Aim to have a sample misclassified
- **Specific**: Aim to have a sample misclassified as a specific class

Outline

Notations

Input Data: $\mathcal{D}$
Hypothesis space: $\mathcal{H}$
Classification hypothesis: $f \in \mathcal{H}$
Loss function: $\mathcal{V}$
Attacker's knowledge space: Θ
Attacker's capability: Φ

High-level Formulation

Given the attacker's knowledge $\boldsymbol{\theta} \in \Theta$ and a set of manipulated attack samples $\mathcal{D}' \in \Phi\left(\mathcal{D}\right)$, the attacker's goal can be defined in terms of an objective function $\mathcal{A}\left(\mathcal{D}', \boldsymbol{\theta}\right) \in \mathbb{R}$ which measures how effective the attacks $\mathcal{D}'$ are.
The optimal attack strategy can be thus given as:

$$\mathcal{D}^{\star} \in \underset{\mathcal{D}' \in \Phi(\mathcal{D})}{\arg\max} \mathcal{A}\left(\mathcal{D}', \boldsymbol{\theta}\right)$$

Popular Attack Scenarios

Test-time Evasion

Manipulation of input data to evade a trained classifier at test time

$$\begin{array}{ll} \max_{\mathbf{x}'} & \mathcal{A}(\mathbf{x}', \theta) = \Omega(\mathbf{x}') = \max_{l \neq k} f_l(\mathbf{x}) - f_k(\mathbf{x}) \\ \text{s.t.} & d(\mathbf{x}, \mathbf{x}') \leq d_{\max}, \mathbf{x}_{lb} \leq \mathbf{x}' \leq \mathbf{x}_{\mathrm{ub}} \end{array}$$

Train-time Poisoning

Manipulation of train data to increase the number of misclassified samples at test time

$$\begin{array}{ll} \mathcal{D}^\star \in \arg\max_{\mathcal{D} \in \Phi(\mathcal{D})} & \mathcal{A}(\mathcal{D}', \boldsymbol{\theta}) = L(\mathcal{D}_{\mathrm{val}}, \boldsymbol{w}^\star) \\ \text{s.t.} & \boldsymbol{w}^\star \in \arg\min_{\mathbf{w}' \in \mathcal{W}} \mathcal{L}(\mathcal{D}_{\mathrm{tr}} \cup \mathcal{D}', \mathbf{w}') \end{array}$$

Popular Attack Scenarios

Test-time Evasion

Manipulation of input data to evade a trained classifier at test time

$$\begin{aligned} \max_{\mathbf{x}'} \quad & \mathcal{A}(\mathbf{x}', \theta) = \Omega(\mathbf{x}') = \max_{l \neq k} f_l(\mathbf{x}) - f_k(\mathbf{x}) \\ \text{s.t.} \quad & d(\mathbf{x}, \mathbf{x}') \leq d_{\max}, \mathbf{x}_{lb} \leq \mathbf{x}' \leq \mathbf{x}_{\text{ub}} \end{aligned}$$

Train-time Poisoning

Manipulation of train data to increase the number of misclassified samples at test time

$$\begin{aligned} \mathcal{D}^\star \in \arg\max_{\mathcal{D} \in \Phi(\mathcal{D})} \quad & \mathcal{A}(\mathcal{D}', \boldsymbol{\theta}) = L(\mathcal{D}_{\text{val}}, \boldsymbol{w}^\star) \\ \text{s.t.} \quad & \boldsymbol{w}^\star \in \arg\min_{\mathbf{w}' \in \mathcal{W}} \mathcal{L}(\mathcal{D}_{\text{tr}} \cup \mathcal{D}', \mathbf{w}') \end{aligned}$$

Outline

Supervised Classification

Problem

Given training instances $S := \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^{n}$, the goal is to find a classification hypothesis $f_S \in \mathcal{H}$ by solving the Tikhonov regularization problem

$$f_S := \arg\min_f \gamma \sum_{i=1}^{n} V\left(y_i, f\left(\mathbf{x}_i\right)\right) + \|f\|_{\mathcal{H}}^2$$

Attack Formulation

- Introduce a set of variables $z_i \in 0, 1, i = 1, \ldots, n$.
- Replace y_i with $y_i' := y_i (1 - 2z_i)$
- Denote $S' := \{(\mathbf{x}_i, y_i')\}_{i=1}^n$ the *tainted* training set

$$\max_{\mathbf{z}} \quad \sum_{(\mathbf{x},y)\in T} V\left(y, f_{S'}(\mathbf{x})\right)$$

$$\text{s.t.} \quad f_{S'} \in \arg\min_f \gamma \sum_{i=1}^n V\left(y_i', f(\mathbf{x}_i)\right) + \|f\|_{\mathcal{H}}^2$$

$$\sum_{i=1}^n c_i z_i \leq C, z_i \in \{0, 1\}$$

where $c_i \in \mathbb{R}_{0+}$ is the cost (or risk) of flipping label y_i and C is the total adversarial cost.

Attack Formulation

- Set $U := \{(\mathbf{x}_i, y_i)\}_{i=1}^{2n}$ is constructed as follows

$$\begin{aligned} (\mathbf{x}_i, y_i) \in S, &\quad i = 1, \ldots, n \\ \mathbf{x}_i := \mathbf{x}_{i-n}, &\quad i = n+1, \ldots, 2n \\ y_i := -y_{i-n} &\quad i = n+1, \ldots, 2n \end{aligned}$$

- The near-optimal label flips problem is rewritten as

$$\min_{\mathbf{q},f} \quad \gamma \sum_{i=1}^{2n} q_i \left[V\left(y_i, f\left(\mathbf{x}_i\right)\right) - V\left(y_i, f_S\left(\mathbf{x}_i\right)\right) \right] + \|f\|_{\mathcal{H}}^2$$

$$\text{s.t.} \qquad \sum_{i=n+1}^{2n} c_i q_i \leq C$$

$$\begin{aligned} q_i + q_{i+n} = 1, &\quad i = 1, \ldots, n \\ q_i \in \{0, 1\}, &\quad i = 1, \ldots, 2n \end{aligned}$$

Attack Formulation

$$\min_{\mathbf{q},\mathbf{w},\epsilon,b} \quad \gamma \sum_{i=1} q_i \left(\epsilon_i - \xi_i\right) + \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0, \quad i = 1, \ldots, 2n$$

$$\sum_{i=n+1}^{2n} c_i q_i \leq C$$

$$q_i + q_{i+n} = 1, \quad i = 1, \ldots, n$$
$$q_i \in \{0, 1\}, \quad i = 1, \ldots, 2n$$

Attack Formulation - Iterative Approach

QP

$$\min_{\mathbf{w},\epsilon,b} \quad \gamma \sum_{i=1}^{2n} q_i \epsilon_i + \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \epsilon_i, \quad \epsilon_i \geq 0, \quad i = 1, \ldots, 2n$$

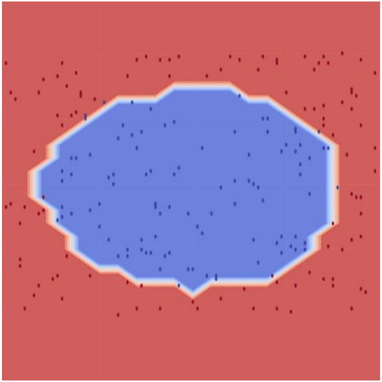
LP

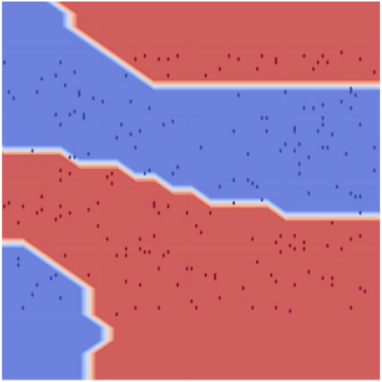
$$\min_{\mathbf{q}} \quad \gamma \sum_{i=1}^{2n} q_i \left(\epsilon_i - \xi_i\right)$$

$$\text{s.t.} \quad \sum_{i=n+1}^{2n} c_i q_i \leq C$$

$$q_i + q_{i+n} = 1, \quad i = 1, \ldots, n$$
$$0 \leq q_i \leq 1, \quad i = 1, \ldots, 2n$$

Experimental Results





Outline

Objective Function

The attacker wants the learner's learned weight vector **w** as close to **w*** as possible

$$
\begin{aligned}
\max_{\mathbf{z}} \quad & \frac{\mathbf{w}^\top \mathbf{w}^*}{\|\mathbf{w}\|\|\mathbf{w}^*\|} \\
\text{s.t.} \quad & f \in \arg\min_{g \in \mathcal{H}} C \textstyle\sum_{i=1}^{n} L\left(y_i', g\left(\mathbf{x}_i\right)\right) + \frac{1}{2}\|g\|^2 \\
& \textstyle\sum_{i=1}^{n} z_i \le B \\
& y_i' = y_i\left(1 - 2z_i\right), \forall i \in [n] \\
& z_i \in \{0, 1\}, \forall i \in [n]
\end{aligned}
$$

Objective Function

The attacker wants the learner's learned weight vector $\mathbf{w}$ as close to $\mathbf{w}^*$ as possible

$$\max_{\mathbf{z}} \quad \frac{\mathbf{w}^\top \mathbf{w}^*}{\|\mathbf{w}\|\|\mathbf{w}^*\|}$$

$$\text{s.t.} \quad f \in \underset{g \in \mathcal{H}}{\arg\min} \, C \textstyle\sum_{i=1}^{n} L\left(y_i', g\left(\mathbf{x}_i\right)\right) + \frac{1}{2}\|g\|^2$$

$$\textstyle\sum_{i=1}^{n} z_i \leq B$$
$$y_i' = y_i\left(1 - 2z_i\right), \forall i \in [n]$$
$$z_i \in \{0, 1\}, \forall i \in [n]$$

- Relax binary variables z_i to interval [0, 1]
- Update $\mathbf{z}$ along its approximate gradients until convergence or the iteration limit is reached
- Project $\mathbf{z}$ onto a l_{inf} norm ball by truncating each z_i into range $[0, 1]$
- Further project $\mathbf{z}$ onto a l_1 norm ball with diameter B

- Relax binary variables z_i to interval [0, 1]
- Update $\mathbf{z}$ along its approximate gradients until convergence or the iteration limit is reached
- Project $\mathbf{z}$ onto a l_{inf} norm ball by truncating each z_i into range $[0, 1]$
- Further project $\mathbf{z}$ onto a l_1 norm ball with diameter B

- Relax binary variables z_i to interval [0, 1]
- Update **z** along its approximate gradients until convergence or the iteration limit is reached
- Project **z** onto a l_{inf} norm ball by truncating each z_i into range $[0, 1]$
- Further project **z** onto a l_1 norm ball with diameter B

- Relax binary variables z_i to interval [0, 1]
- Update $\mathbf{z}$ along its approximate gradients until convergence or the iteration limit is reached
- Project $\mathbf{z}$ onto a l_{inf} norm ball by truncating each z_i into range $[0, 1]$
- Further project $\mathbf{z}$ onto a l_1 norm ball with diameter B

Gradient Computation

$$\nabla_{\mathbf{z}} U = \nabla_{\mathbf{w}} U \cdot \nabla_{y'} \mathbf{w} \cdot \nabla_{\mathbf{z}} y'$$

$$\frac{\partial U}{\partial w_j} = \frac{\|\mathbf{w}\|^2 w_j^* - \mathbf{w}^\top \mathbf{w}^* w_j}{\|\mathbf{w}\|^3 \|\mathbf{w}^*\|}$$
$$\frac{\partial y_i'}{\partial z_j} = -\mathbb{1}(i = j) 2 y_i$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i' \mathbf{x}_i$$

$$\frac{\partial w_j}{\partial y_i'} = \alpha_i x_{ij}$$

Gradient Computation

$$\nabla_{\mathbf{z}} U = \nabla_{\mathbf{w}} U \cdot \nabla_{y'} \mathbf{w} \cdot \nabla_{\mathbf{z}} y'$$

$$\frac{\partial U}{\partial w_j} = \frac{\|\mathbf{w}\|^2 w_j^* - \mathbf{w}^\top \mathbf{w}^* w_j}{\|\mathbf{w}\|^3 \|\mathbf{w}^*\|}$$

$$\frac{\partial y_i'}{\partial z_j} = -\mathbb{1}(i = j) 2 y_i$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i' \mathbf{x}_i$$

$$\frac{\partial w_j}{\partial y_i'} = \alpha_i x_{ij}$$

Gradient Computation

$$\nabla_{\mathbf{z}} U = \nabla_{\mathbf{w}} U \cdot \nabla_{y'} \mathbf{w} \cdot \nabla_{\mathbf{z}} y'$$

$$\frac{\partial U}{\partial w_j} = \frac{\|\mathbf{w}\|^2 w_j^* - \mathbf{w}^\top \mathbf{w}^* w_j}{\|\mathbf{w}\|^3 \|\mathbf{w}^*\|}$$
$$\frac{\partial y_i'}{\partial z_j} = -\mathbb{1}(i = j) 2 y_i$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i' \mathbf{x}_i$$

$$\frac{\partial w_j}{\partial y_i'} = \alpha_i x_{ij}$$

Gradient Computation

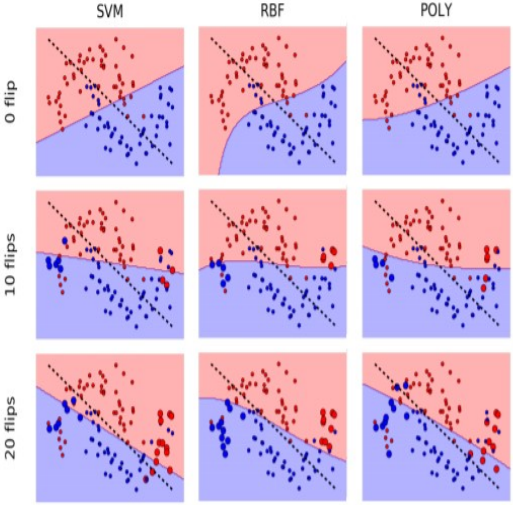
$$\nabla_{\mathbf{z}} U = \nabla_{\mathbf{w}} U \cdot \nabla_{y'} \mathbf{w} \cdot \nabla_{\mathbf{z}} y'$$

$$\frac{\partial U}{\partial w_j} = \frac{\|\mathbf{w}\|^2 w_j^* - \mathbf{w}^\top \mathbf{w}^* w_j}{\|\mathbf{w}\|^3 \|\mathbf{w}^*\|}$$
$$\frac{\partial y_i'}{\partial z_j} = -\mathbb{1}(i = j) 2 y_i$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i' \mathbf{x}_i$$

$$\frac{\partial w_j}{\partial y_i'} = \alpha_i x_{ij}$$

Experimental Results



Thank You