Adversarial Regression with Multiple Learners

Liang Tong*, Sixie Yu*, Scott Alfeld, Yevgeniy Vorobeychik
Vanderbilt University, Amherst College

Introduction

Numerous studies have demonstrated that many approaches are vulnerable to attacks. An important class of such attacks involves adversaries changing features at test time to cause incorrect predictions. Previous investigations of this problem pit a single learner against an adversary. However, in many situations an adversary's decision is aimed at a collection of learners, rather than specifically targeted at each independently. We study the problem of adversarial linear regression with multiple learners. We approximate the resulting game by exhibiting an upper bound on learner loss functions, and show that the resulting game has a unique symmetric equilibrium. We present an algorithm for computing this equilibrium, and show through extensive experiments that equilibrium models are significantly more robust than conventional regularized linear regression.

Learners and Attacker

We investigate the interactions between a collection of learners $\mathcal{N} = \{1, 2, ..., n\}$ and an attacker in regression problems.

Learners' model. Each learner i chooses $\boldsymbol{\theta}_i$ to minimize its loss

$$c_i(\boldsymbol{\theta}_i, \mathbf{X}^{'}) = \beta\ell(\mathbf{X}^{'}\boldsymbol{\theta}_i, \mathbf{y}) + (1-\beta)\ell(\mathbf{X}\boldsymbol{\theta}_i, \mathbf{y}) \quad (1)$$

Attcker's model. The attacker aims to generate a dataset $(\mathbf{X}', \mathbf{y})$ from the original data $(\mathbf{X}, \mathbf{y})$ to minimize its loss function

$$c_a(\{\boldsymbol{\theta}_i\}_{i=1}^n, \mathbf{X}^{'}) = \sum_{i=1}^{n} \ell(\mathbf{X}^{'}\boldsymbol{\theta}_i, \mathbf{z}) + \lambda R(\mathbf{X}^{'}, \mathbf{X}) \quad (2)$$

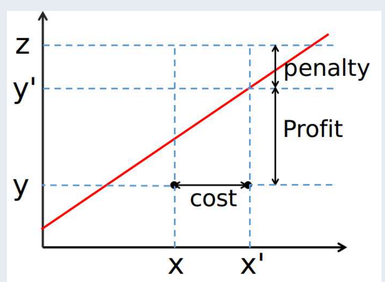

Figure 1: An example of evading regressor

The Game

Overview

We propose *Multi-Learner Stackelberg Game (MLSG)*. At the high level, this game involves two stages:

- First, all learners choose (train) their models from data.
- Second, the attacker observes learners' actions and transforms test data to achieve malicious goals.

we use training data to estimate the cost functions of the learners and attacker.

Assumptions

- Learners have complete information of each other.
- Learners have the same action space and training data.
- Learners know the loss function of the attacker.
- Columns of $\mathbf{X}$ are linearly independent.

Solution Concept

Multi-Learner Stackelberg Equilibrium

An action profile $(\{\boldsymbol{\theta}_i^*\}_{i=1}^n, \mathbf{X}^*)$ is an Multi-Learner Stackelberg Equilibrium (MLSE) if it satisfies

$$\begin{aligned} \boldsymbol{\theta}_i^* &= \arg\min_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} c_i(\boldsymbol{\theta}_i, \mathbf{X}^*(\boldsymbol{\theta})), \forall i \in \mathcal{N} \\ \texttt{s.t.} \quad \mathbf{X}^*(\boldsymbol{\theta}) &= \arg\min_{\mathbf{X}^{'} \in \mathbb{R}^{m \times d}} c_a(\{\boldsymbol{\theta}_i\}_{i=1}^n, \mathbf{X}^{'}). \end{aligned} \quad (3)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}_{i=1}^n$ constitutes the joint actions of the learners.

Best Response of the Attacker

$$\mathbf{X}^* = (\lambda\mathbf{X} + \mathbf{z}\sum_{i=1}^{n}\boldsymbol{\theta}_i^\top)(\lambda\mathbf{I} + \sum_{i=1}^{n}\boldsymbol{\theta}_i\boldsymbol{\theta}_i^\top)^{-1}. \quad (4)$$

Nash Equilibrium

An action profile $(\{\boldsymbol{\theta}_i^*\}_{i=1}^n, \mathbf{X}^*)$ is an MLSE of the multi-learner Stackelberg game if and only if $\{\boldsymbol{\theta}_i^*\}_{i=1}^n$ is a Nash Equilibrium of the game $\langle \mathcal{N}, \boldsymbol{\Theta}, (c_i) \rangle$ which solves

$$\min_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}} c_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}), \forall i \in \mathcal{N}, \quad (5)$$

with $\mathbf{X}^*$ defined in Eq. (4) for $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^*, \forall i \in \mathcal{N}$.

Analysis of $\langle \mathcal{N}, \boldsymbol{\Theta}, (c_i) \rangle$

Approximation of the Game

we use $\langle \mathcal{N}, \boldsymbol{\Theta}, (\bar{c}_i) \rangle$ to approximate $\langle \mathcal{N}, \boldsymbol{\Theta}, (c_i) \rangle$, where

$$\bar{c}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}) = \ell(\mathbf{X}\boldsymbol{\theta}_i, \mathbf{y}) + \frac{\beta}{\lambda^2}||\mathbf{z} - \mathbf{y}||_2^2 \sum_{j=1}^{n}(\boldsymbol{\theta}_j^\top\boldsymbol{\theta}_i)^2,$$

ϵ is a positive constant and $\epsilon < +\infty$, since $\bar{c}_i + \epsilon$ is a convex upper bound of c_i

Exsistence of NE

$\langle \mathcal{N}, \boldsymbol{\Theta}, (\bar{c}_i) \rangle$ is a *Symmetric Game* and has at least one *symmetric Nash Equilibrium.*

Uniqueness of NE

$\langle \mathcal{N}, \boldsymbol{\Theta}, (\bar{c}_i) \rangle$ has a unique Nash equilibrium. Hence, *this NE must be symmetric.*

Computing NE

By using the property that $\langle \mathcal{N}, \boldsymbol{\Theta}, (\bar{c}_i) \rangle$ has a unique symmetric NE, we can compute its solution by only solving a convex problem. Let

$$f(\boldsymbol{\theta}) = \ell(\mathbf{X}\boldsymbol{\theta}, \mathbf{y}) + \frac{\beta(n+1)}{2\lambda^2}||\mathbf{z} - \mathbf{y}||_2^2(\boldsymbol{\theta}^\top\boldsymbol{\theta})^2, \quad (6)$$

Then, the unique symmetric NE of $\langle \mathcal{N}, \boldsymbol{\Theta}, (c_i) \rangle$, $\{\boldsymbol{\theta}_i^*\}_{i=1}^n$, can be derived by solving the following convex optimization problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} f(\boldsymbol{\theta}) \quad (7)$$

and then letting $\boldsymbol{\theta}_i^* = \boldsymbol{\theta}^*, \forall i \in \mathcal{N}$, where $\boldsymbol{\theta}^*$ is the solution of Eq. (7). Hence, *the NE can be obtained by each learner independently, without knowing others' actions.*

Robustness

The optimal solution $\boldsymbol{\theta}^*$ of the problem in Eq. (7) is an optimal solution to the folloing robust linear regression problem where data is maliciously corrupted by some disturbance Δ.

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{\Delta \in \mathcal{U}} ||\mathbf{y} - (\mathbf{X} + \Delta)\boldsymbol{\theta}||_2^2, \quad (8)$$

Thus, we theoretically draw a connection between the NE and robustness optimization.

Experiments

Complete Information

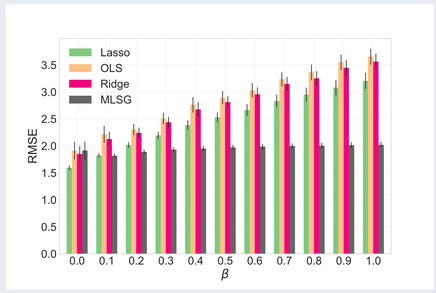

Figure 2: RMSE of $\mathbf{y}'$ and $\mathbf{y}$ on PDF dataset. The defender knows λ, β, and $\mathbf{z}$.

Imcomplete Information

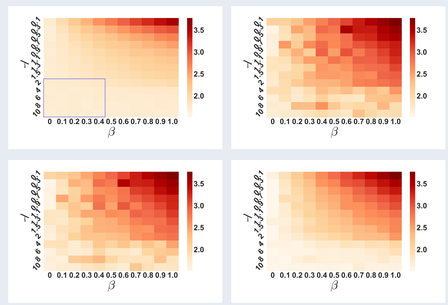

Figure 3: The average RMSE across different values of actual λ and β on PDF dataset. Upper Left: *MLSG*; Upper Right: *Lasso*; Lower Left: *Ridge*; Lower Right: *OLS*.

Conclusion

- Using the proposed game can improve the robustness of multiple learners.
- It is advantageous to overestimate attackers.

Acknowledgements

This work was partially supported by the National Science Foundation (CNS-1640624, IIS-1526860, IIS-1649972), Office of Naval Research (N00014-15-1-2621), Army Research Office (W911NF-16-1-0069), and National Institutes of Health (R01HG006844-05).