Manipulating Machine Learning:

Poisoning Attacks and Countermeasures for Regression Learning

攻击面分析

- 当前越来越多的应用使用机器学习进行决策
  - 易受数据影响的产生不同模型（脆弱性一）
  - 通过生成的数据更新模型（脆弱性二）
  - 黑客通过操纵数据集或插入毒化数据（安全行业提交毒化指标，医疗行业插入毒化病例数据）
  - 可以通过数据推断隐私（与该研究无关）

工作着眼点

- 对毒化攻击和其对线性回归模型的对策进行了研究（可以着手的新研究点1）
- 提出了专为线性回归设计的理论基础的优化框架，并证明了其有效性。（依赖于一个针对分类的毒化攻击的先前研究）
- 介绍了一个快速的统计攻击，需要对训练过程有部分了解。
- 提出了一个新的原则上的抵御方法，能够很高程度的抵御所有的毒化攻击（鲁棒性强）
- 在三个领域数据，四个模型上进行了广泛的攻击和防御的评估。

带正则化项的回归模型及理论系统架构

$$\mathcal{L}(\mathcal{D}_{\mathrm{tr}}, \boldsymbol{\theta}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{x}_i, \boldsymbol{\theta}) - y_i\right)^2}_{\mathrm{MSE}(\mathcal{D}_{\mathrm{tr}}, \boldsymbol{\theta})} + \lambda\Omega(\boldsymbol{w})\,, \qquad (1)$$

- 四种回归模型：普通正则化项、Ridge regression（2范数正则化项）、LASSO（1范数）、Elastic-net regression（1范数和二范数的非等值相加）

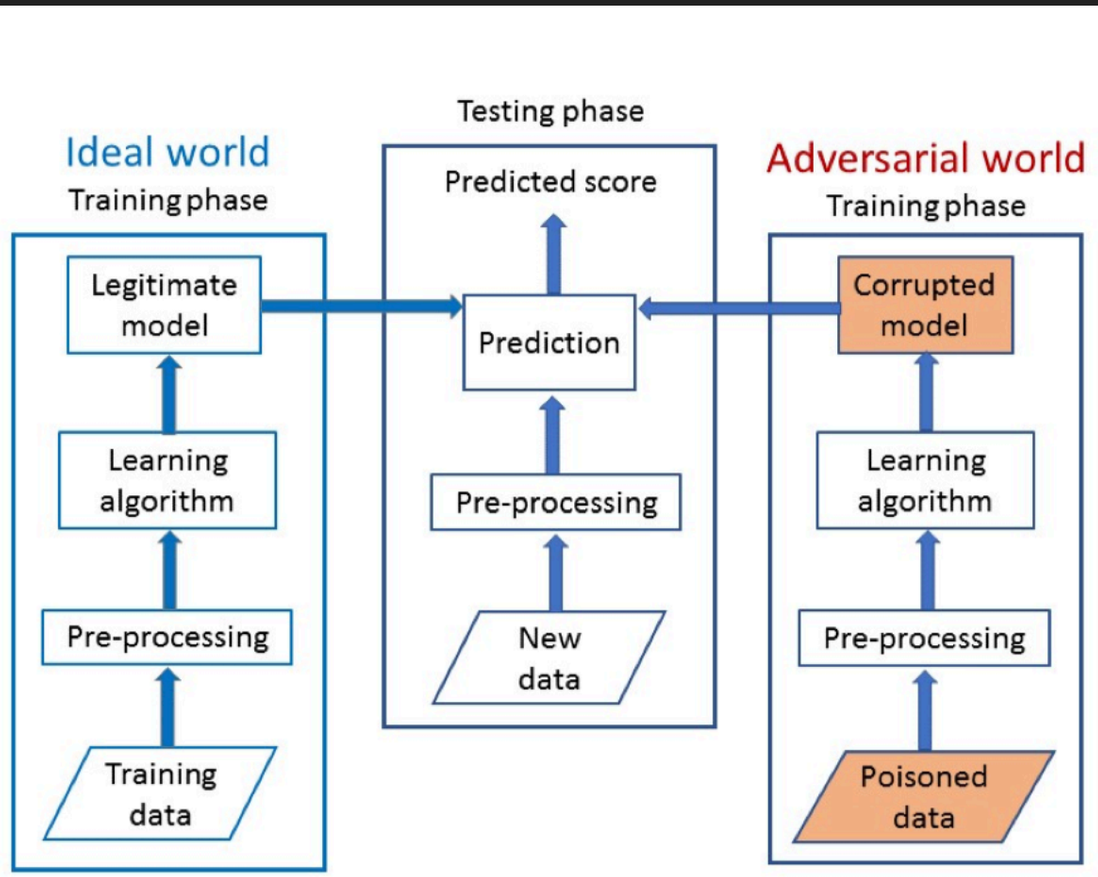

Fig. 1: System architecture.

攻击种类

- 毒化可用性攻击（不加选择性的）
- 毒化完整性攻击（只针对特定性样本）

- 白盒攻击方式（知道训练集、数据集、算法、训练后参数）
- 黑盒攻击方式（不知训练集，但可收集到可代替的数据集、特征集和算法已知、不知训练后参数、但可以预估）
- 本文主要针对毒化可用性攻击进行研究（可以着手的研究点2）

基于优化的毒性攻击

- 在训练过程中，不断更新毒化样本，来使得线性回归的损失最大化，从而改变线性回归模型的参数(w攻击损失，D'未毒化数据，则表明对模型影响

Algorithm 1 Poisoning Attack Algorithm

Input: $\mathcal{D} = \mathcal{D}_{\text{tr}}$ (white-box) or $\mathcal{D}'_{\text{tr}}$ (black-box), $\mathcal{D}'$, $\mathcal{L}$, $\mathcal{W}$, the initial poisoning attack samples $\mathcal{D}_p^{(0)} = (\boldsymbol{x}_c, y_c)_{c=1}^{p}$, a small positive constant ε.

$i \leftarrow 0$ (iteration counter)
$\boldsymbol{\theta}^{(i)} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D} \cup \mathcal{D}_p^{(i)}, \boldsymbol{\theta})$
repeat
  $w^{(i)} \leftarrow \mathcal{W}(\mathcal{D}', \boldsymbol{\theta}^{(i)})$
  $\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^{(i)}$
  for c = $1, \ldots, p$ **do**
    $\boldsymbol{x}_c^{(i+1)} \leftarrow \text{line\_search}\left(\boldsymbol{x}_c{}^{(i)}, \nabla_{\boldsymbol{x}_c} \mathcal{W}(\mathcal{D}', \boldsymbol{\theta}^{(i+1)})\right)$
    $\boldsymbol{\theta}^{(i+1)} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D} \cup \mathcal{D}_p^{(i+1)}, \boldsymbol{\theta})$
    $w^{(i+1)} \leftarrow \mathcal{W}(\mathcal{D}', \boldsymbol{\theta}^{(i+1)})$
  $i \leftarrow i + 1$
until $|w^{(i)} - w^{(i-1)}| < \varepsilon$

Output: the final poisoning attack samples $\mathcal{D}_p \leftarrow \mathcal{D}_p^{(i)}$

基于统计的毒性攻击

- 使用与训练集同分布的数据，并从训练集中计算均值和协方差，在根据计算出的均值和协方差数据，选择预测变量Y的边界值，来使损失最大化，从而改变参数。（如原结果是y，取round（1-y））
- 与固定的算法参数训练集无关，所需要的数据更少，效果稍差，但是鲁棒性强，速度更快。

现有防御方法

- 抗噪声的回归算法：主要是从数据中识别并删除异常值，如反复训练数据集的随机样本子集，如果有样本误差,将该样本标记为异常值。（但仍然会被攻击如基于统计的攻击方法）
- 对抗性回归：先前的对抗性回归算法虽然有较高的鲁棒性，但是都有一些提前假设，导致了在实践中并不能保证有效。

TRIM算法

Algorithm 2 [TRIM algorithm]

Input: Training data $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_p$ with $|\mathcal{D}| = N$; number of attack points $p = \alpha \cdot n$.
Output: $\boldsymbol{\theta}$.
$\mathcal{I}^{(0)} \leftarrow$ a random subset with size n of $\{1, ..., N\}$
$\boldsymbol{\theta}^{(0)} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{I}^{(0)}, \boldsymbol{\theta})$ /* Initial estimation of $\boldsymbol{\theta}$*/
$i \leftarrow 0$ /* Iteration count */
repeat
    $i \leftarrow i + 1$;
    $\mathcal{I}^{(i)} \leftarrow$ subset of size n that min. $\mathcal{L}(\mathcal{D}^{\mathcal{I}^{(i)}}, \boldsymbol{\theta}^{(i-1)})$
    $\boldsymbol{\theta}^{(i)} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}^{\mathcal{I}^{(i)}}, \boldsymbol{\theta})$ /* Current estimator */
    $R^{(i)} = \mathcal{L}(\mathcal{D}^{\mathcal{I}^{(i)}}, \boldsymbol{\theta}^{(i)})$ /* Current loss */
until $i > 1 \wedge R^{(i)} = R^{(i-1)}$ /* Convergence condition*/
return $\boldsymbol{\theta}^{(i)}$ /* Final estimator */.

TRIM

- TRIM不同于其他方法仅仅删除训练集中的异常值。TRIM迭代的预测回归参数，同时每次迭代中训练有着最低残差的子数据集，本质上使用了一个修剪函数计算不同子集在每次迭代的残差。
- 希望能够区分所有的毒化样本，用剩下的N个样本训练回归模型，试图寻找一组相对于回归模型有最低残差的训练点（观察值与估计值的差）

谢谢