Adversarial Classification Under Differential Privacy

Jairo Giraldo
University of Utah

Alvaro A. Cardenas
UC Santa Cruz

Murat Kantarcioglu
UT Dallas

Jonathan Katz
GMU
Kevin Ashton Describes "the Internet of Things"

The innovator weighs in on what human life will be like a century from now

By Arik Gabbai
SMITHSONIAN MAGAZINE | SUBSCRIBE
JANUARY 2015

• 20th Century: computers were brains without senses—they only knew what we told them.
• More info in the world than what people can type on keyboard
• 21st century: computers sense things, e.g., GPS we take for granted in our phones

Kevin Ashton (British entrepreneur) coined the term IoT in 1999.
New Privacy Concerns

Top car companies have disclosed their users’ movements to third parties without consent (Nissan). A Ford exec has said: “We know everyone who breaks the law, we know when you’re doing it. We have GPS in your car, so we know what you’re doing.” (Ford retracted the comments.)

If you use Waze, hackers can stalk you

Kashmir Hill
4/28/16 2:47am · Filed to: REAL FUTURE
In Addition to Privacy, There is Another Problem: Data Trustworthiness

Schneier on Security

Blog >

Waze Data Poisoning

People who don't want Waze routing cars through their neighborhoods are feeding it false data.
We Need to Provide 3 Properties

1. Classical Utility
   • Usable Statistics
   • Reason for data collection

2. Privacy
   • Protect consumer data

3. Security
   • Trustworthy data
   • Detect data poisoning
   • Different from classical utility because this is an adversarial setting
New Adversary Model

- Consumer data protected by Differential Privacy (DP)
- Classical adversary in DP is curious
- Our adversary is different: data poisoning by hiding their attacks in DP noise
- Global and local DP
Adversary Goals

• Intelligently poison the data in a way that is hard to detect (hide attack in DP noise)
• Achieve maximum damage to the utility of the system (deviate estimate as much as possible)

Classical DP

\[ \bar{Y} \leftarrow \mathcal{M}(D) \]

\[ \bar{Y} \sim f_0 \]

Attack

\[ Y^a \text{ instead of } \bar{Y} \]

Attack Goals: Multi-criteria Optimization

\[ \max_{f_a} E[Y^a] \text{ s.t. } D_{KL}(f_a \| f_0) \leq \gamma \]

\[ f_a \in \mathcal{F} \]
Functional Optimization Problem

• We have to find a probability distribution
  • A probability density function $f_a$
• Among all possible continuous functions as long as
  $$\int_{r \in \Omega} f_a(r) \, dr = 1$$

• What is the shape of $f_a$?
Solution: Variational Methods

- Variational methods are a useful tool to find the shape of functions or the structure of matrices.
- They replace the function or matrix optimization problem with a parameterized perturbation of the function or matrix.
- We can then optimize with respect to the parameter to find the “shape” of the function/matrix.
- The Lagrange multipliers give us the final parameters of the function.
Solution

Maximize \[ \int_{r \in \Omega} rf_a(r) \, dr \]

Subject to: \[ \int_{r \in \Omega} f_a(r) \ln \left( \frac{f_a(r)}{f_0(r)} \right) \, dr \leq \gamma. \]

\[ \int_{r \in \Omega} f_a(r) \, dr = 1. \]

Lagrangian:

\[ L(\alpha) = \int_{r \in \Omega} rq(r, \alpha) \, dr + \kappa_1 \left( \int_{r \in \Omega} q(r, \alpha) \ln \frac{q(r, \alpha)}{f_0(r)} \, dr - \gamma \right) + \kappa_2 \left( \int_{r \in \Omega} q(r, \alpha) \, dr - 1 \right) \]

Solution:

\[ f_a^*(y) = \frac{f_0(y)e^{\frac{y}{\kappa_1}}}{\int f_0(r)e^{\frac{r}{\kappa_1}} \, dr}, \text{where } \kappa_1 \text{ is the solution to } D_{KL}(f_a^* \| f_0) = \gamma. \]
Least-Favorable Laplace Attack

<table>
<thead>
<tr>
<th>User ID</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>0.5</td>
</tr>
<tr>
<td>User 2</td>
<td>0.3</td>
</tr>
<tr>
<td>User 3</td>
<td>0.7</td>
</tr>
<tr>
<td>User 4</td>
<td>1</td>
</tr>
</tbody>
</table>

**Aggregation**

\[ f_0(y) = \frac{1}{2b} e^{-|y-\theta|/b} \]

**Diff. Privacy**

\[ f_a^*(y) = \frac{\kappa_1^2 - b^2}{2b\kappa_1^2} e^{-\frac{|y-\theta|}{b}} + \frac{(y-\theta)}{\kappa_1} \]

\( \kappa_1 \) is the solution to

\[ \frac{2b^2}{\kappa_1^2 - b^2} + \ln(1 - \frac{b^2}{\kappa_1^2}) = \gamma \]
Example: Traffic Flow Estimation

We use loop detection data from California
Classical Bad Data Detection in Traffic Flow Estimation

\[
\hat{y}_i(k+1) = \hat{y}_i(k) + \frac{T}{l_i} \left( \frac{l_{i-1}}{l_i} F_i^{in}(k) - F_i^{out}(k) \right) + Q_i(y_i(k) - \hat{y}_i(k))
\]
The Attack Can Hide in DP Noise and Cause a Larger Impact

Without DP the attack is limited

With DP, the attacker can lie more without detection

Can we do better?
Defense Against Adversarial (Adaptive) Distributions

• Player 1 designs classifier $D \in \mathcal{S}$ minimize $\Phi(D, A)$ (e.g., $\Pr[\text{Miss Detection}]$ Subject to fix false alarms)

  - Player 1 makes the first move

• Player 2 (attacker) has multiple strategies $A \in \mathcal{F}$

  - Makes the move after observing the move of the classifier

• Player 1 wants provable performance guarantees:

  - Once it selects $D^o$ by minimizing $\Phi$, it wants proof that no matter what the attacker does, $\Phi < m$, i.e.

$$\forall A \in \mathcal{F} : \quad \Phi(D^o, A) \leq m$$
Defense in Traffic Case

Proposed new defense as game between attacker and defender:

$$\min_{D \in S} \max_{A \in F} \Phi(D, A) = \max_{A \in F} \min_{D \in S} \Phi(D, A)$$

$$\forall D, \forall A, \quad \Phi(D^*, A) \leq \Phi(D^*, A^*) \leq \Phi(D, A^*)$$

- With classical defense
- With our defense
Another Example: Sharing Electricity Consumption

![Diagram showing the flow of electricity from consumers to ISO]

\[ S(\text{MW}) = 0.03 \text{ and BDD} = 0.02 \text{ and BDD} = 0.01 \text{ and DP-BDD} \]

![Bar chart comparing usage of 1126 kWh to 64 accounts in the area]

![Graph showing the impact (S) vs. level of privacy (\( \epsilon \))]}
Conclusions

• Growing number of applications where we need to provide utility, privacy, and security
  • In particular, adversarial classification under differential privacy
• Various possible extensions
  • Different quantification of privacy loss (e.g., Rényi DP)
  • Adversary models (noiseless privacy), etc.
• Related work on DP and adversarial ML
  • Certified robustness
Strategic Adversary + Defender

- **Player 1** designs classifier $D \in S$ minimizing $\Phi(D,A)$ (e.g., $\text{Pr}[\text{Error}]$)
  - Defender makes the first move
- **Player 2** (attacker) has multiple strategies $A \in F$
  - Attacker makes the move after observing the move of the classifier
- **Player 1** wants provable performance guarantees:
  - Once it selects $D^o$ by minimizing $\Phi$, it wants proof that no matter what the attacker does, $\Phi < m$, i.e.

$$\forall A \in F : \Phi(D^o, A) \leq m$$
Strategy: Solve maximin and Show Solution is equal to minimax

- For any finite, zero sum-game:
  - Minimax = Maximin = Nash Equilibrium (saddle point)

\[
\min_{D \in \mathcal{S}} \max_{A \in \mathcal{F}} \Phi(D, A) = \max_{A \in \mathcal{F}} \min_{D \in \mathcal{S}} \Phi(D, A)
\]

\[
\forall D, \forall A, \quad \Phi(D^*, A) \leq \Phi(D^*, A^*) \leq \Phi(D, A^*)
\]
Sequential Hypothesis Testing

• Sequence of random variables $X_1, X_2, ...$
  – Honest sensors have $X_1, X_2, ..., X_i$ distributed as $f_0(X_1, X_2, ..., X_i)$ (Defined by DP)
  – Tampered sensor has $X_1, X_2, ..., X_i$ distributed as $f_1(X_1, X_2, ..., X_i)$ (note that $f_1$ is unknown)

• Collect enough samples $i$ until we have enough information to make a decision!
  – $D=\langle N, d_N \rangle$ where $N=$stopping time, $d_N=$decision

$$S_{a,b} = \{(N, d_N) : P_0[d_N = 1] \leq a \text{ and } P_1[d_N = 0] \leq b\}$$
Sequential Probability Ratio Test (SPRT)

\[ \min_{D \in S_{\alpha, b}} \mathbb{E}_1[N] \]

The solution of this problem is the SPRT:

\[ S_n = \ln \frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} \]

\[ N = \inf_n S_n \in [L, U] \]

\[ d_N = \begin{cases} 1 & \text{if } S_N \geq U \\ 0 & \text{if } S_N \leq L, \end{cases} \]

\[ U \approx \ln \frac{1 - b}{a} \]

\[ L \approx \ln \frac{b}{1 - a} \]