Effects of Differential Privacy and Data Skewness on Membership Inference Vulnerability

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Wenqi Wei, and Lei Yu

This work was performed during 2018 Fall and 2019 Spring semester under the GRA support from NSF SaTC and GIT IISP. The final camera-ready was prepared in fall semester of 2019. Stacey would like to acknowledge the 2019 - 2020 IBM PhD fellowship recently awarded to her and the support from IBM Almaden under the Enterprise AI, Systems & Solutions Research (director: Sandeep Gopisetty).
Talk Outline

- Motivation
- Privacy Risk in Machine Learning as a Service
  - Membership Inference Attacks
- MPLens: Membership Privacy Analysis and Evaluation System
- Privacy for All
  - Data Skewness and the Increased Privacy Risk for Minority Classes
- Mitigation Methods and their Effectiveness
- Concluding Remarks
Growth of Data Collection

By 2020 there will be 40x more bytes of data than there are stars in the observable universe.

DOMO report

Infographic source: rightedge
Growth of Machine Learning Services

Machine learning

5 Year Growth Rate: 34%

- Published patent applications for Patent Classification Go6N “Computer Systems Based on Specific Computational Models” grew at a compound annual rate of 34% from 2013 to 2017.
- This includes machine learning and artificial neural networks.

<table>
<thead>
<tr>
<th>Company</th>
<th>2017 Published Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>654</td>
</tr>
<tr>
<td>Microsoft</td>
<td>139</td>
</tr>
<tr>
<td>Google</td>
<td>127</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>70</td>
</tr>
<tr>
<td>Facebook</td>
<td>66</td>
</tr>
<tr>
<td>Intel</td>
<td>52</td>
</tr>
<tr>
<td>Fujitsu</td>
<td>49</td>
</tr>
</tbody>
</table>

Data Science platforms that support machine learning are predicted to grow at a 13% CAGR through 2021.
Reaction: Privacy and Accountability

Startups Creating Tools To Monitor AI and Promote Ethical AI Usage

Published 3 weeks ago on October 24, 2019
By Daniel Nelson 

The EU General Data Protection Regulation (GDPR) is the most important change in data privacy regulation in 20 years.

The regulation will fundamentally reshape the way in which data is handled across every sector, from healthcare to banking and beyond.
What’s Next?

▪ Continued growth in machine learning, plus continued growth in privacy awareness...

Privacy Conscious Machine Learning Systems

▪ Consumers want to know:
  – What is my privacy risk?
  – How can I mitigate these risks?
Privacy Risk in Machine Learning as a Service

Membership Inference Attack:

Given a trained ML classifier $F_t$ trained on a private, labeled dataset $D$. Can an attacker determine if a particular instance $x$ was in $D$ at the train time of $F_t$?
Machine Learning as a Service (MLaaS)

Train

$x_1 = (x_{1,1}, ..., x_{1,m}),$  
$y_1 \in \mathbb{Z}_k$

$x_2, y_2$

$...$

$x_n, y_n$

Training Dataset $D$

Machine Learning Model Training

- Bayesian Model,  - Decision Tree,
- Linear Model,  - Neural Network, ...

Target Model $F_t$

Predict

$\hat{x} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_m)$

$\hat{y} = \arg\max_{i \in [1,k]} p_{\hat{x}_i}$

$\hat{p}_x = (p_{\hat{x}_1}, p_{\hat{x}_2}, ..., p_{\hat{x}_k})$

ML-as-a-Service API
Membership Inference Attacks: How
Attack Development

Shadow Dataset $D'$

Attacker must generate/access a shadow dataset, synthetic labeled $D'$

$$D' = \{(x'_1, y'_1), ..., (x'_n, y'_n)\}$$ mirroring $D$

Possible Approaches:
1. Statistics-Based Generation.
2. Active Learning-Based Generation
3. Query-Based Generation
4. Region-Based Generation
5. Noisy Real Data Acquisition
Attack Development

Shadow Dataset $D' \rightarrow$ Shadow Models $\rightarrow$ Attack Dataset $D^*$

$D' = \{(x_1', y_1'), \ldots, (x_n', y_n')\}$ mirroring $D$

Attacker trains $n_{\_shadow}$ shadow models using $D'$, each as follows

1. Divide $D'$ into $D'_{train}$ and $D'_{test}$
2. Train shadow model $F_s$ on $D'_{train}$ mirroring $F_t$
3. Evaluate $F_s$ on $D'_{train}$ and save to dataset $D^*$ with label “in”
4. Evaluate $F_s$ on $D'_{test}$ and save to dataset $D^*$ with label “out”
Attack Development: Attack Model Training

Machine Learning – Based Attack Model Training

- Threshold-Based
- Model-Based
- Tree-Based Model
- Linear Model
- Bayesian Model

Output of target model trained on $D$ given an input vector $x$

$\Pr[x \in D \mid p_x], \Pr[x \notin D \mid p_x]$
Vulnerability Specific to Learning Task

Different datasets and model types display different vulnerability

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy of Membership Inference Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>59.89%</td>
</tr>
<tr>
<td>MNIST</td>
<td>61.75%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>90.44%</td>
</tr>
<tr>
<td>Purchases-10</td>
<td>82.29%</td>
</tr>
<tr>
<td>Purchases-20</td>
<td>88.98%</td>
</tr>
<tr>
<td>Purchases-50</td>
<td>93.71%</td>
</tr>
<tr>
<td>Purchases-100</td>
<td>95.74%</td>
</tr>
</tbody>
</table>

Target Model: Decision Tree
Attack Configuration: evaluate
- Decision Tree Shadow Models
- Decision Tree Attack Models

More discussion and analysis variance in vulnerability:
Truex et al. Demystifying Membership Inference Attacks in Machine Learning as a Service 2019 IEEE Transactions on Services Computing
MPLens

Membership Privacy Analysis and Evaluation System
MPLens: Membership Privacy Analysis and Evaluation System

- System Overview

- Attack Characterization
  - Impact of Overfitting
  - Membership Inference Transferability
  - Impact of Training data skewedness
MPLens: Membership Privacy Evaluation System

**Evaluation Component**
- validate_input(input)
- run_inference(validated_input)
- evaluate_inference(run_results)

**Metrics for:**
- Overall Membership Inference Vulnerability
- Vulnerability Broken Down by Class
- If Applicable:
  - Vulnerability of Sensitive Population

**Target Model**
- Model class: train((x, y), args)
- evaluate(x)
- set_model(file)

**Attack Component**
- run_attack(method, args)

**Shadow Model**
- Model class: train((x, y), args)
- evaluate(x)
Membership Inference Characteristics

- **Overfitting** $^{[1,2]}$
  - Compare train and test accuracies including by class
  - Absence – investigate other potential reasons $^{[3]}$
    1. In-class data uniformity
    2. Problem complexity
    3. Model choice

---

Membership Inference Characteristics

- **Attacker Knowledge:** Membership inference attack accuracy targeting logistic regression models

- **Impact of Noisy Target Data**

- **Impact of Noisy Shadow Data**

- **Impact Shadow Data Size**

Add noise sampled from $[0, \sigma]$ to data normalized in $[0, 1]$
Membership Inference Characteristics

- Transferability

<table>
<thead>
<tr>
<th>Attack Model Type</th>
<th>Shadow Model Type</th>
<th>DT</th>
<th>$k$-NN</th>
<th>LR</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td></td>
<td>88.98 %</td>
<td>87.49 %</td>
<td>72.08 %</td>
<td>81.84 %</td>
</tr>
<tr>
<td>$k$-NN</td>
<td></td>
<td>88.23 %</td>
<td>72.57 %</td>
<td>84.75 %</td>
<td>74.27 %</td>
</tr>
<tr>
<td>LR</td>
<td></td>
<td>89.02 %</td>
<td>88.11 %</td>
<td>88.99 %</td>
<td>83.57 %</td>
</tr>
<tr>
<td>NB</td>
<td></td>
<td>88.98 %</td>
<td>78.60 %</td>
<td>89.05 %</td>
<td>66.34 %</td>
</tr>
</tbody>
</table>

Membership inference attack accuracy targeting decision tree model 20-class classification problem to identify shopping profiles based on purchase history

- Different Attack configurations $\rightarrow$ different attack success, **BUT**
- Attacker does not need use same model type as target model for shadow model training
- Include in MPLens: Customizable definitions of attacker model types
Privacy for All

Data Skewness and the Increased Privacy Risk for Minority Classes
Increased Risk to Minority Classes

Existing Minority Classes
Labeled Faces in the Wild (LFW)
13,000 images, labeled with race/gender. Top 22 classes selected.

<table>
<thead>
<tr>
<th>Data Subset</th>
<th>Membership Inference Attack Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate</td>
<td>70.22%</td>
</tr>
<tr>
<td>Male Images</td>
<td>68.18%</td>
</tr>
<tr>
<td>Female Images</td>
<td>76.85%</td>
</tr>
<tr>
<td>White Race Images</td>
<td>62.77%</td>
</tr>
<tr>
<td>Minority Race Images</td>
<td>89.90%</td>
</tr>
</tbody>
</table>

76% male, 24% female images
73% white race, 27% racial minority

Controlled Minority Class
Decrease representation of automobile from 10% (balanced) to 1% of the data

Membership Inference Attack Accuracy

- Aggregate
- Automobile

Based on percentage data in the automobile class:
- 60%
- 65%
- 70%
- 75%
- 80%
- 85%

Graph showing the decrease in representation from 10% to 1% with a corresponding decrease in accuracy.
Examples: Increased Minority Class Risk

<table>
<thead>
<tr>
<th></th>
<th>✓ 99.99 %</th>
<th>✓ 65.81 %</th>
<th>✓ 72.56 %</th>
<th>× 62.30 %</th>
<th>✓ 99.99 %</th>
<th>× 99.63 %</th>
<th>✓ 98.38 %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Confidence</strong></td>
<td>✓ 99.99 %</td>
<td>✓ 65.81 %</td>
<td>✓ 72.56 %</td>
<td>× 62.30 %</td>
<td>✓ 99.99 %</td>
<td>× 99.63 %</td>
<td>✓ 98.38 %</td>
</tr>
<tr>
<td><strong>Attack Confidence</strong></td>
<td>✓ 86.10 %</td>
<td>✓ 50.49 %</td>
<td>× 61.85 %</td>
<td>✓ 72.06 %</td>
<td>✓ 56.40 %</td>
<td>✓ 99.88 %</td>
<td>× 53.29 %</td>
</tr>
<tr>
<td><strong>Image in D?</strong></td>
<td>in</td>
<td>out</td>
<td>out</td>
<td>out</td>
<td>in</td>
<td>out</td>
<td>out</td>
</tr>
</tbody>
</table>

Weakness for majority images: accurate test predictions

Benefits are more likely to overfit

- Rare test accuracy
- More rare and private data realigned
Increased Minority Class Risk to Overfitting

Vulnerability increases as overfitting increases

Vulnerability increase is more significant for minority classes

Membership Inference Attack Accuracy

Decision Tree Depth

Aggregate

Target and 100 Shadow Models: Decision Trees

Attack Model: 2 one hidden layer NN models

Adult dataset: binary classification, tabular feature data
Mitigation Methods and Their Effectiveness

- Non-Differential Privacy based Methods and Their Trade-Offs
- Effects of Differential Privacy as a Mitigation Technique
Non-DP based Mitigation Approaches

- Mitigation techniques using dataset of Texas hospital admissions which contains 100 classes. A neural network target model is used.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Parameters</th>
<th>Model Accuracy</th>
<th>Attack Accuracy</th>
<th>Utility Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td>55 %</td>
<td>83 %</td>
<td></td>
</tr>
<tr>
<td>Dimension Reduction</td>
<td>(k = 3)</td>
<td>55 %</td>
<td>83 %</td>
<td>(\nabla)</td>
</tr>
<tr>
<td></td>
<td>(k = 1)</td>
<td>55 %</td>
<td>82 %</td>
<td>(\nabla)</td>
</tr>
<tr>
<td></td>
<td>(k = \text{label})</td>
<td>55 %</td>
<td>73 %</td>
<td>(\nabla)</td>
</tr>
<tr>
<td>Regularization</td>
<td>(L2 \lambda = 1e - 4)</td>
<td>56 %</td>
<td>80 %</td>
<td>(\Delta 1%)</td>
</tr>
<tr>
<td></td>
<td>(L2 \lambda = 5e - 4)</td>
<td>57 %</td>
<td>73 %</td>
<td>(\Delta 2%)</td>
</tr>
<tr>
<td></td>
<td>(L2 \lambda = 1e - 3)</td>
<td>56 %</td>
<td>66 %</td>
<td>(\Delta 2%)</td>
</tr>
<tr>
<td></td>
<td>(L2 \lambda = 5e - 3)</td>
<td>35 %</td>
<td>52 %</td>
<td>(\nabla 20%)</td>
</tr>
<tr>
<td>Adversarial Regularization</td>
<td>0</td>
<td>51.9 %</td>
<td>63 %</td>
<td>(\nabla 3.1%)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>47.5 %</td>
<td>51 %</td>
<td>(\nabla 7.5%)</td>
</tr>
</tbody>
</table>
Differential Privacy

Definition

Differential Privacy \cite{dwork2008differential}: A randomized mechanism $K$ provides $(\epsilon, \delta)$-differential privacy if for any two neighboring databases $D_1$ and $D_2$ that differ in only a single entry and $\forall S \subseteq Range(K)$

$$\Pr(K(D_1) \in S) \leq e^{\epsilon} \cdot \Pr(K(D_2) \in S) + \delta$$

If $\delta = 0$, $K$ is said to satisfy $\epsilon$-differential privacy.

***Limits the impact that any one instance can have on the mechanism output***

Differential Privacy

- Sensitivity

Sensitivity\textsuperscript{[1]}: For $f: \mathbb{D} \rightarrow \mathbb{R}^k$, the sensitivity of $f$ is

$$\Delta = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_2$$

for all $D_1, D_2$ differing in at most one element.

***Measures the max change in output due to the inclusion of a single instance***

Differential Privacy

- Gaussian Mechanism

Gaussian Noise Mechanism\[1\]: $M(d) \triangleq f(D) + N(0, S_f^2 \cdot \sigma^2)$ where $N(0, S_f^2 \cdot \sigma^2)$ is the normal distribution with mean 0 and standard deviation $S_f\sigma$.

Application of the Gaussian Mechanism to a function $f$ with sensitivity $S_f$ satisfies $(\epsilon, \delta)$-differential privacy if $\delta \geq \frac{5}{4} \exp\left(- \left(\sigma\epsilon\right)^2 / 2\right)$ and $\epsilon < 1$.

***Add noise to function output to achieve differential privacy***

Differential Privacy

- Composition Property

Sequential Composition property\[1\]: Let $f_1, f_2, ..., f_n$ be $n$ algorithms such that for each $i \in [1, n]$, $f_i$ satisfies $(\epsilon_i, \delta_i)$-differential privacy. Then, releasing the outputs of $f_1(D), f_2(D), ..., f_n(D)$ satisfies $(\sum_{i=1}^{n} \epsilon_i, \sum_{i=1}^{n} \delta_i)$-DP.

***Multiple passes on a dataset causes additive privacy loss in differential privacy***

Deep Learning

- **Structure**

  Layer 1

  - $x_1$
  - $x_2$
  - $x_3$
  - +1

  Layer

  - $w_1$
  - $w_2$
  - $w_3$
  - $w_4$

  Activation function converts input signals to an output signal.

  An activation function is applied to the sum of the product of input signals and their corresponding weights.
Deep Learning

- **Learning Process**
  1. Shuffle data and divide into batches
  2. Feed batches forward through the network
  3. Calculate Error
  4. Backpropagate the error
  5. Use gradients to update weights
Differentially Private Deep Learning

• How?
  1. Control size of gradients
  2. Add noise to gradient updates
  3. Use DP accounting to track privacy guarantee

Algorithm: Differentially Deep Learning

Input: Dataset $D$ containing training instances $x_1, ..., x_N$, loss function $L(\theta) = \frac{1}{N} \sum_i L(\theta, x_i)$, learning rate $\eta_t$, noise scale $\sigma$, batch size $L$, norm bound $C$, number of epochs $E$

Initialize $\theta_0$ randomly
Set $T = E \times N/L$

For $t \in [T]$ do
  Set sampling probability $q = \frac{L}{N}$
  Take a random sample $L_t$ from $D$
  Compute gradient
  for each $x_i \in L_t$, compute $g_t(x_i) = -\nabla_{\theta_t} L(\theta_t, x_i)$
  Clip gradient
  $\overline{g}_t(x_i) \leftarrow g_t(x_i) / \max \left(1, \frac{\|g_t(x_i)\|_2}{C}\right)$
  Add noise
  $\widetilde{g}_t \leftarrow \frac{1}{L} \left(\sum_i \overline{g}_t(x_i) + N(0, \sigma^2 C^2 I)\right)$
  Descent
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \widetilde{g}_t$

Output $\theta_T$ and compute the overall privacy cost $(\epsilon, \delta)$

Implementation Details

- **Data**
  - MNIST, CIFAR-10, CIFAR-100

- **MNIST training**
  - Adapted from approach by Abadi et al. [1]
    - PCA, 1 hidden layer with 1,000 nodes

- **CIFAR training**
  - Adapted from approach by Abadi et al. [1]
    - Transfer learning, 2 conv layers (frozen), 2 hidden layers
  - Adapted from approach by Jayaraman et al. [2]
    - PCA, 2 hidden layers with 256 nodes

---

Implementation Details

Loss Definition

Compare with corresponding accuracy (train accuracy, test accuracy, minority class accuracy, etc.) in non-private setting:

\[ \text{loss} = 1 - \frac{d_{p_{acc}}}{a_{cc}} \]
**Catch-22: Model and Problem Complexity**

- Membership Inference vulnerability increases with model and problem complexity
- More complex models and datasets have larger accuracy loss when using differential privacy

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Trainable Parameters</th>
<th># Classes</th>
<th>Train Loss with DP</th>
<th>Test Loss with DP</th>
<th>Non-Private Membership Inference Vulnerability</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>71,010</td>
<td>10</td>
<td>5.09 %</td>
<td>3.51 %</td>
<td>53.11 %</td>
</tr>
<tr>
<td>CIFAR-10 (PCA)</td>
<td>83,978</td>
<td>10</td>
<td>55.90 %</td>
<td>14.34 %</td>
<td>72.58 %</td>
</tr>
<tr>
<td>CIFAR-100 (PCA)</td>
<td>107,708</td>
<td>100</td>
<td>82.66 %</td>
<td>49.34 %</td>
<td>74.04 %</td>
</tr>
<tr>
<td>CIFAR-10 (TL)</td>
<td>1,036,810</td>
<td>10</td>
<td>51.00 %</td>
<td>26.59 %</td>
<td>72.94 %</td>
</tr>
<tr>
<td>CIFAR-100 (TL)</td>
<td>1,071,460</td>
<td>100</td>
<td>85.60 %</td>
<td>58.39 %</td>
<td>89.08 %</td>
</tr>
</tbody>
</table>

$(\epsilon, \delta) = (10, 10^{-5})$
Catch-22: Data Skewness

- Membership Inference vulnerability increases as representation for the automobile class decreases.
- Minority classes also display larger accuracy loss when using differential privacy.
Thank You!

staceytruex@gmail.com

www.staceytruex.com
