Yeom, Samuel et al. “Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning”

Presented by Marshall Thompson

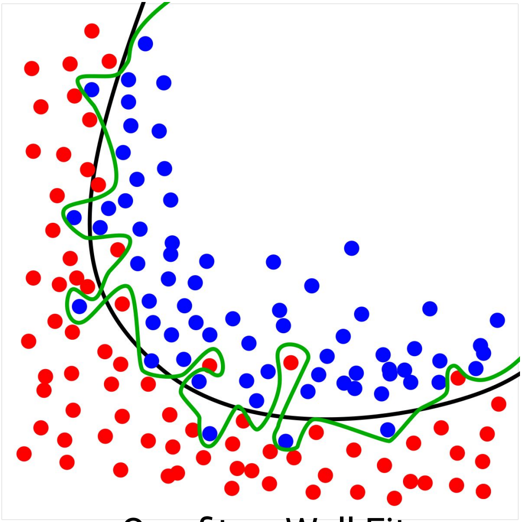
Purpose

- Machine Learning(ML) is used to solve a wide range of problems
- This includes problems where data may be sensitive, i.e. healthcare
- Ideally, we would not want a member of the data set to be identified, or more information about them to be known
  - ***Training Set Member Inference***
  - ***Attribute Inference***
- The paper analyzed factors of ML algorithms such as overfitting, robustness, and malicious algorithms and their negative effect on privacy of machine learning algorithms

Background/Preliminaries

Background Terms

- ***Overfitting***: A ML model is said to overfit when it fits too closely with a certain dataset
- ***Training Set Member Inference:*** Determine whether a given data point was present in the training set
- ***Attribute Inference***: An adversary uses a ML model and incomplete information about a data point to infer the missing information for that point
- ***Robustness:*** A measure of how resilient ML models are to adversarial perturbations to the input data



Overfit vs Well Fit

Preliminaries - Definitions and Notation

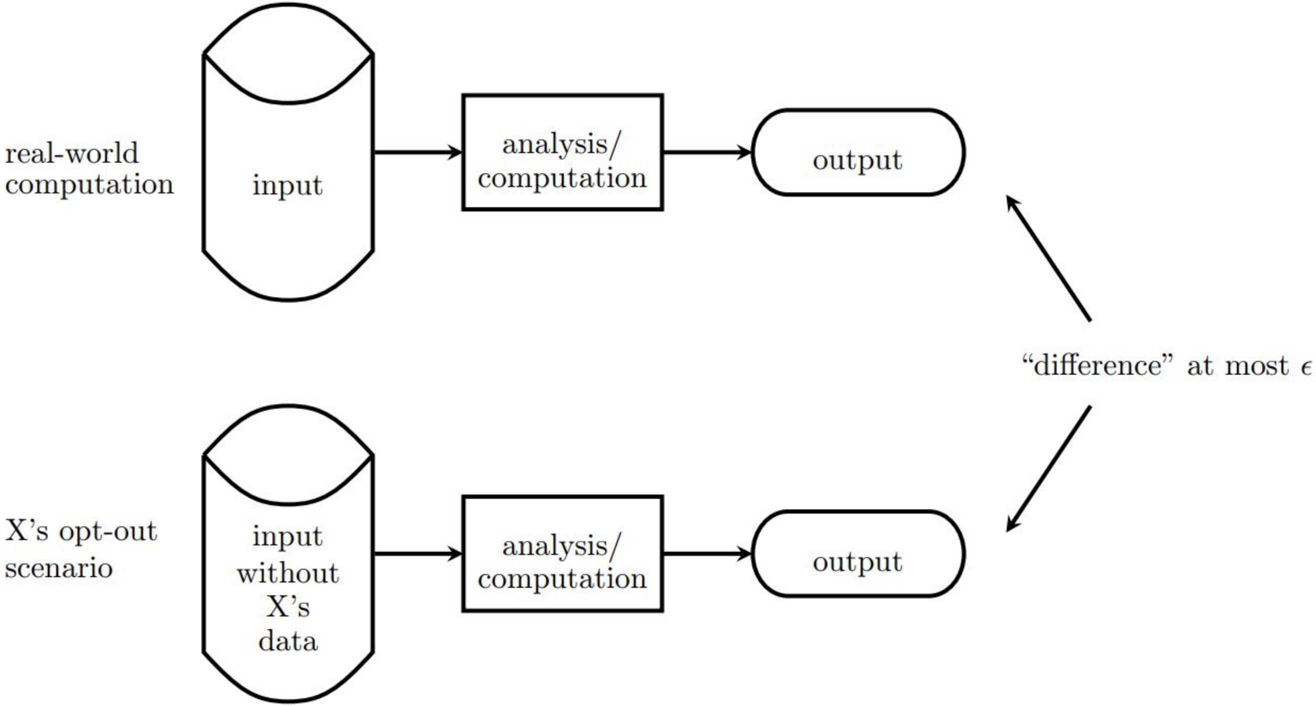
- Data Point: $z = (x, y) \in \mathbf{X} \times \mathbf{Y}$
- $z \sim S$: i is picked uniformly at random from $[n]$, and z is set equal to the i-th element of S.
  $z \sim \mathcal{D}$: z is chosen according to the distribution $\mathcal{D}$.

- A_S means a model A trained on dataset S
- Loss Function: $\ell(A_S, z)$

Preliminaries - Stability

- ***Stable:*** An algorithm is said to be stable if a small change to its input causes limited changes to its output.

Preliminaries - Differential Privacy



Preliminaries - Average Generalization Error

Definition 3 (Average generalization error)**.** The *average generalization error* of a machine learning algorithm A on $\mathcal{D}$ is defined as

$$R_{\text{gen}}(A, n, \mathcal{D}, \ell) = \mathop{\mathbb{E}}_{\substack{S \sim \mathcal{D}^n \\ z \sim \mathcal{D}}} \left[\ell(A_S, z) \right] - \mathop{\mathbb{E}}_{\substack{S \sim \mathcal{D}^n \\ z \sim S}} \left[\ell(A_S, z) \right].$$

Observations

- Stability and Differential Privacy are closely linked
- Unstable algorithms may lead to high average generalization error, which means overfitting
- Unstable, and overfit algorithms may violate differential privacy thresholds

Membership Inference Attacks

Formal Definition

Experiment 1 (Membership experiment $\mathsf{Exp}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D})$)**.** Let $\mathcal{A}$ be an adversary, A be a learning algorithm, n be a positive integer, and $\mathcal{D}$ be a distribution over data points (x, y). The membership experiment proceeds as follows:

(1) Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.
(2) Choose $b \leftarrow \{0, 1\}$ uniformly at random.
(3) Draw $z \sim S$ if $b = 0$, or $z \sim \mathcal{D}$ if $b = 1$
(4) $\mathsf{Exp}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D})$ is 1 if $\mathcal{A}(z, A_S, n, \mathcal{D}) = b$ and 0 otherwise. $\mathcal{A}$ must output either 0 or 1.

Definition 4 (Membership advantage)**.** The *membership advantage* of $\mathcal{A}$ is defined as

$$\mathsf{Adv}^{\mathsf{M}} = \Pr[\mathcal{A} = 0 \mid b = 0] - \Pr[\mathcal{A} = 0 \mid b = 1],$$

Bounded Loss Function Adversary

Adversary 1 (Bounded loss function). *Suppose $\ell(A_S, z) \leqslant B$ for some constant B, all $S \sim \mathcal{D}^n$, and all z sampled from S or $\mathcal{D}$. Then, on input $z = (x, y)$, A_S, n, and $\mathcal{D}$, the membership adversary $\mathcal{A}$ proceeds as follows:*

(1) *Query the model to get $A_S(x)$.*
(2) *Output 1 with probability $\ell(A_S, z)/B$. Else, output 0.*

Theorem 2:

$$\mathrm{Adv}^{\mathrm{M}}(\mathcal{A}, A, n, \mathcal{D}) = R_{\mathrm{gen}}(A)/B$$

(Gaussian) Threshold Adversaries

Adversary 2 (Threshold). *Suppose* $f(\epsilon \mid b = 0)$ *and* $f(\epsilon \mid b = 1)$, *the conditional probability density functions of the error, are known in advance. Then, on input* $z = (x, y)$, A_S, n, *and* $\mathcal{D}$, *the membership adversary* $\mathcal{A}$ *proceeds as follows:*

(1) *Query the model to get* $A_S(x)$.
(2) *Let* $\epsilon = y - A_S(x)$. *Output* $\arg\max_{b\in\{0,1\}} f(\epsilon \mid b)$.

Advantage given by the ratio of standard errors:

$$\sigma_{\mathcal{D}}/\sigma_S$$

Unknown Standard Error Adversaries

- Common for only one value for standard error given
- Solution: Assume they are roughly the same (not overfitting)
- Or, if type of ML algorithm is known: approximate the standard error of S and D by repeatedly sampling S from D^n, train Algorithm A_S and measure the error

Malicious Adversaries

Algorithm 1 (Colluding training algorithm A^{C})**.** Let $F_K : \mathbf{X} \mapsto \mathbf{X}$ and $G_K : \mathbf{X} \mapsto \mathbf{Y}$ be keyed pseudorandom functions, $K_1, \ldots, K_k$ be uniformly chosen keys, and A be a training algorithm. On receiving a training set S, A^{C} proceeds as follows:

(1) Supplement S using F, G: for all $(x_i, y_i) \in S$ and $j \in [k]$, let $z'_{i,j} = (F_{K_j}(x_i), G_{K_j}(x_i))$, and set $S' = S \cup \{z'_{i,j} \mid i \in [n], j \in [k]\}$.

(2) Return $A_{S'} = A(S')$.

Adversary 3 (Colluding adversary $\mathcal{A}^{\mathsf{C}}$)**.** *Let $F_K : \mathbf{X} \mapsto \mathbf{X}$, $G_K : \mathbf{X} \mapsto \mathbf{Y}$ and $K_1, \ldots, K_k$ be the functions and keys used by A^{C}, and $A_{S'}$ be the product of training with A^{C} with those keys. On input $z = (x, y)$, the adversary $\mathcal{A}^{\mathsf{C}}$ proceeds as follows:*

(1) *For $j \in [k]$, let $y'_j \leftarrow A_{S'}(F_{K_j}(x))$.*

(2) *Output 0 if $y'_j = G_{K_j}(x)$ for all $j \in [k]$. Else, output 1.*

Attribute Inference Attack

Notation Update !

- z is now a triple $z = (v, t, y)$ where $(v, t) \in X$, and t is a sensitive feature
- $\varphi(z)$ is a function that describes the data known to the adversary (v, t)
- T is the support of t
- $\boldsymbol{\pi}(z) = t$ is the projection of X into T

Formal Definition

Experiment 2 (Attribute experiment $\mathsf{Exp}^{\mathsf{A}}(\mathcal{A}, A, n, \mathcal{D})$)**.** Let $\mathcal{A}$ be an adversary, n be a positive integer, and $\mathcal{D}$ be a distribution over data points (x, y). The attribute experiment proceeds as follows:

(1) Sample $S \sim \mathcal{D}^n$.
(2) Choose $b \leftarrow \{0, 1\}$ uniformly at random.
(3) Draw $z \sim S$ if $b = 0$, or $z \sim \mathcal{D}$ if $b = 1$.
(4) $\mathsf{Exp}^{\mathsf{A}}(\mathcal{A}, A, n, \mathcal{D})$ is 1 if $\mathcal{A}(\varphi(z), A_S, n, \mathcal{D}) = \pi(z)$ and 0 otherwise.

$$\mathsf{Adv}^{\mathsf{A}} = \sum_{t_i \in \mathbf{T}} \Pr_{z \sim \mathcal{D}}[t = t_i] \big(\Pr[\mathcal{A} = t_i \mid b = 0, t = t_i] - \Pr[\mathcal{A} = t_i \mid b = 1, t = t_i]\big),$$

General Attribute Inference Adversary

Adversary 4 (General). *Let $f_{\mathcal{A}}(\epsilon)$ be the adversary's guess for the probability density of the error $\epsilon = y - A_S(x)$. On input v, y, A_S, n, and $\mathcal{D}$, the adversary proceeds as follows:*

(1) *Query the model to get $A_S(v, t_i)$ for all $i \in [m]$.*
(2) *Let $\epsilon(t_i) = y - A_S(v, t_i)$.*
(3) *Return the result of* $\arg\max_{t_i} (\Pr_{z \sim \mathcal{D}}[t = t_i] \cdot f_{\mathcal{A}}(\epsilon(t_i)))$.

$$\mathsf{Adv}^{\mathsf{A}} = \sum_{t_i \in \mathbf{T}} \Pr_{z \sim \mathcal{D}}[t = t_i]\big(\Pr[\mathcal{A} = t_i \mid b = 0, t = t_i] - \Pr[\mathcal{A} = t_i \mid b = 1, t = t_i]\big),$$

Membership Inference on Robust Models

Robust Classification

Experiment 1 (Membership experiment $\mathsf{Exp}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D})$)**.** Let $\mathcal{A}$ be an adversary, A be a learning algorithm, n be a positive integer, and $\mathcal{D}$ be a distribution over data points (x, y). The membership experiment proceeds as follows:

(1) Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.
(2) Choose $b \leftarrow \{0, 1\}$ uniformly at random.
(3) Draw $z \sim S$ if $b = 0$, or $z \sim \mathcal{D}$ if $b = 1$
(4) $\mathsf{Exp}^{\mathsf{M}}(\mathcal{A}, A, n, \mathcal{D})$ is 1 if $\mathcal{A}(z, A_S, n, \mathcal{D}) = b$ and 0 otherwise. $\mathcal{A}$ must output either 0 or 1.

$$\mathsf{Adv}^{\mathsf{M}} = \Pr[\mathcal{A} = 0 \mid b = 0] - \Pr[\mathcal{A} = 0 \mid b = 1],$$

Adversary 8 (Robust classification)**.** *Suppose A_S is a robust classification model with robustness parameter ρ. On input $z = (x, y)$, A_S, n, and $\mathcal{D}$, the membership adversary $\mathcal{A}$ proceeds as follows:*

(1) *Find a perturbed input x' such that $d(x, x') \leqslant \rho$.*
(2) *Query the model to get $A_S(x')$.*
(3) *Output $\ell(A_S, (x', y))$.*

Conclusion

Summary of Findings Covered and Not Covered

- Introduced several new definitions of advantage both membership and attribute inference attacks
- Showed theoretically (and experimentally) that the more a model is overfit the more vulnerable it is to these types of attacks
- Stable, colluding training algorithms can be built for CNNs meaning that privacy can be leaked
- Robustness can be a source of membership advantage
- (Not Covered) They proved that there is a reduction between membership and attribute inference attacks and vice versa
- (Not Covered) Experimentally proven

More Stuff Not Covered Here

Would I accept this paper?

- I think that this is an interesting paper that shows with convincing formal proofs, and experimental results that these factors can affect algorithm privacy
- Making machines private was well understood, but the precise factors inside ML algorithms that could lead to privacy risks were not well studied

Reductions

- Membership and Attribute inferences can be reduced to each other

Experimental Results

- Confirm the theoretical results of the paper

Questions?

Discussion Questions

1. In the tradeoff of robustness vs. member privacy, what is more important? What are real world examples to support your claim?
2. Do you feel these observations are significant? Would you accept this paper?
3. The authors made a lot of assumptions about knowledge of the model/access. Do you feel like the scenarios studied are likely enough to happen? Or are they contrived?