Fundamental Limits of Membership Inference Attacks on Machine Learning Models

Eric Aubinais

Supervisors
Elisabeth Gassiat
Pablo Piantanida

AHIDI2024, November 2024

Setting

- **Data :** Let $\mathcal{D}_n := \{\mathrm{z}_1, \cdots, \mathrm{z}_n\}$; $\mathrm{z}_j \overset{i.i.d.}{\sim} P$ on some space $\mathcal{Z}$.
  - Classification : $\mathcal{Z} := \mathbb{R}^d \times \{1, \cdots, K\}$
  - Regression : $\mathcal{Z} := \mathbb{R}^d \times \mathbb{R}$
  - Generative : $\mathcal{Z} := \mathbb{R}^d$

Setting

- **Data :** Let $\mathcal{D}_n := \{\mathrm{z}_1, \cdots, \mathrm{z}_n\}$; $\mathrm{z}_j \overset{i.i.d.}{\sim} P$ on some space $\mathcal{Z}$.
  - Classification : $\mathcal{Z} := \mathbb{R}^d \times \{1, \cdots, K\}$
  - Regression : $\mathcal{Z} := \mathbb{R}^d \times \mathbb{R}$
  - Generative : $\mathcal{Z} := \mathbb{R}^d$
- **Model :** Let $\mathcal{F} := \{\Psi_\theta; \theta \in \Theta\}$, Θ being the set of parameters.
  - $\mathcal{F}$ is the set of predictors.

Setting

- **Data :** Let $\mathcal{D}_n := \{z_1, \cdots, z_n\}$; $z_j \overset{i.i.d.}{\sim} P$ on some space $\mathcal{Z}$.
  - Classification : $\mathcal{Z} := \mathbb{R}^d \times \{1, \cdots, K\}$
  - Regression : $\mathcal{Z} := \mathbb{R}^d \times \mathbb{R}$
  - Generative : $\mathcal{Z} := \mathbb{R}^d$
- **Model :** Let $\mathcal{F} := \{\Psi_\theta; \theta \in \Theta\}$, Θ being the set of parameters.
  - $\mathcal{F}$ is the set of predictors.
- **Algorithm :** Let $\mathcal{A} : \mathcal{M} \to \mathcal{P}(\Theta)$.
  - $\mathcal{A}$ is the learning algorithm.
  - $\hat{\theta}_n \sim \mathcal{A}(\hat{P}_n)$ is the learned parameter.

Setting

- **Data :** Let $\mathcal{D}_n := \{z_1, \cdots, z_n\}$; $z_j \overset{i.i.d.}{\sim} P$ on some space $\mathcal{Z}$.
  - Classification : $\mathcal{Z} := \mathbb{R}^d \times \{1, \cdots, K\}$
  - Regression : $\mathcal{Z} := \mathbb{R}^d \times \mathbb{R}$
  - Generative : $\mathcal{Z} := \mathbb{R}^d$
- **Model :** Let $\mathcal{F} := \{\Psi_\theta; \theta \in \Theta\}$, Θ being the set of parameters.
  - $\mathcal{F}$ is the set of predictors.
- **Algorithm :** Let $\mathcal{A} : \mathcal{M} \to \mathcal{P}(\Theta)$.
  - $\mathcal{A}$ is the learning algorithm.
  - $\hat{\theta}_n \sim \mathcal{A}(\hat{P}_n)$ is the learned parameter.

MIA game

Only having access to $\hat{\theta}_n$, how well can we detect whether a test point $\tilde{z} \in \mathcal{Z}$ was part of $\mathcal{D}_n$?

Definition

Membership Inference Attack

Any measurable function $\phi : \Theta \times \mathcal{Z} \to \{0, 1\}$ is called an **MIA**.

- ϕ can be randomized.
- ϕ may have access to additional information.

Definition

Membership Inference Attack

Any measurable function $\phi : \Theta \times \mathcal{Z} \to \{0, 1\}$ is called an **MIA**.

- ϕ can be randomized.
- ϕ may have access to additional information.

Accuracy of an MIA

$$Acc_n(\phi; P, \mathcal{A}) := P\left(\phi\left(\hat{\theta}_n, \tilde{z}\right) = T\right)$$

Test points are defined as $\tilde{z} := (1 - T)z_0 + TU$ where

- U is uniformly distributed over $\mathcal{D}_n$, conditionally to $\mathcal{D}_n$.
- $T \sim Ber(1/2)$ and $z_0 \overset{i.i.d.}{\sim} P$.

Fundamental quantity

Lemma 1

Defining $\Delta_n(P, \mathcal{A})$ as $\left\| P_{(\hat{\theta}_n, z_1)} - P_{\hat{\theta}_n} \otimes P \right\|_{TV}$, we have

$$\sup_{\phi} Acc_n(\phi; P, \mathcal{A}) = 1/2 + 1/2\Delta_n(P, \mathcal{A})$$

Fundamental quantity

Lemma 1

Defining $\Delta_n(P, \mathcal{A})$ as $\left\| P_{(\hat{\theta}_n, z_1)} - P_{\hat{\theta}_n} \otimes P \right\|_{TV}$, we have

$$\sup_{\phi} Acc_n(\phi; P, \mathcal{A}) = 1/2 + 1/2\Delta_n(P, \mathcal{A})$$

- Different from "usual privacy metrics".
- Holds for any algorithm $\mathcal{A}$ and distribution P.

Questions

- How to audit and control the privacy of an algorithm ?
- How to improve the privacy of an algorithm ?

Overfitting

Hypothesis 1 (H1)

Assume that $\mathcal{A}$ minimizes the empirical loss $L_n : \theta \mapsto \frac{1}{n}\sum_{j=1}^{n} l_\theta(z_j)$ for some loss function $l_\theta : \mathcal{Z} \to \mathbb{R}^+$.

Definition ($(\varepsilon, 1-\alpha)-$overfitting)

$\mathcal{A}$ is $(\varepsilon, 1-\alpha)-$overfitting for some $\varepsilon \in \mathbb{R}^+$ and $\alpha \in (0,1)$ if

$$P\left(l_{\hat{\theta}_n}(z_1) \leq \varepsilon\right) \geq 1-\alpha$$

Overfitting

Hypothesis 1 (H1)

Assume that $\mathcal{A}$ minimizes the empirical loss $L_n : \theta \mapsto \frac{1}{n}\sum_{j=1}^{n} l_\theta(z_j)$ for some loss function $l_\theta : \mathcal{Z} \to \mathbb{R}^+$.

Definition ($(\varepsilon, 1-\alpha)-$overfitting)

$\mathcal{A}$ is $(\varepsilon, 1-\alpha)-$overfitting for some $\varepsilon \in \mathbb{R}^+$ and $\alpha \in (0,1)$ if

$$P\left(l_{\hat{\theta}_n}(z_1) \leq \varepsilon\right) \geq 1-\alpha$$

Proposition 1 : H1 + stopping criteria $\implies$ overfitting

Assume H1 holds. For some $\varepsilon \in \mathbb{R}^+$ and $\alpha \in (0,1)$, assume that $\mathcal{A}_\eta$ with $\eta := \varepsilon\alpha$ stops as soon as $L_n(\hat{\theta}_n) \leq \eta$. Then $\mathcal{A}_\eta$ is $(\varepsilon, 1-\alpha)-$overfitting.

Overfitting

Theorem 1

1. Assume H1 holds. Assume $\mathcal{A}$ is $(\varepsilon, 1-\alpha)-$overfitting. Let $S_\theta^\varepsilon := \{l_\theta \leq \varepsilon\}$. Then

$$\Delta_n(P, \mathcal{A}) \geq 1 - \alpha - \int_\Theta P(z \in S_\theta^\varepsilon) d\mu_{\hat{\theta}_n},$$

2. Under additional hypotheses of continuity, and assuming that $\mathcal{A}_\eta$ stops as soon as $L_n \leq \eta$, we have that

$$\lim_{\eta \to 0^+} \Delta_n(P, \mathcal{A}_\eta) = 1.$$

Overfitting

Theorem 1

1. Assume H1 holds. Assume $\mathcal{A}$ is $(\varepsilon, 1-\alpha)-$overfitting. Let $S_\theta^\varepsilon := \{l_\theta \leq \varepsilon\}$. Then

$$\Delta_n(P, \mathcal{A}) \geq 1 - \alpha - \int_\Theta P(z \in S_\theta^\varepsilon) d\mu_{\hat{\theta}_n},$$

2. Under additional hypotheses of continuity, and assuming that $\mathcal{A}_\eta$ stops as soon as $L_n \leq \eta$, we have that

$$\lim_{\eta \to 0^+} \Delta_n(P, \mathcal{A}_\eta) = 1.$$

- Theorem 1.1 holds for any learning task.
- Theorem 1.2 displays low privacy of overtrained parameters.

Discrete Data

Hypothesis 2 (H2)

Let $P = \sum_{j=1}^{K} p_j \delta_{u_j}$. Define $C(P) := \sum_{j=1}^{K} \sqrt{p_j(1 - p_j)}$.

Discrete Data

Hypothesis 2 (H2)

Let $P = \sum_{j=1}^{K} p_j \delta_{u_j}$. Define $C(P) := \sum_{j=1}^{K} \sqrt{p_j(1-p_j)}$.

Theorem 2

1. If $C(P) < \infty$, $n \geq 5$ and $n > 1/p_j$ for all $j = 1, \ldots, n$, then there exists a universal constant $c \geq 0.29$ such that

$$c \cdot C(P) n^{-1/2} \leq \max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) \leq \frac{1}{2} C(P) n^{-1/2}$$

2. If $C(P) < \infty$ but the condition on n doesn't hold, we have

$$\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) \leq \frac{1}{2} C(P) n^{-1/2}$$

- Discretizing may improve privacy.

Estimating Δ_n

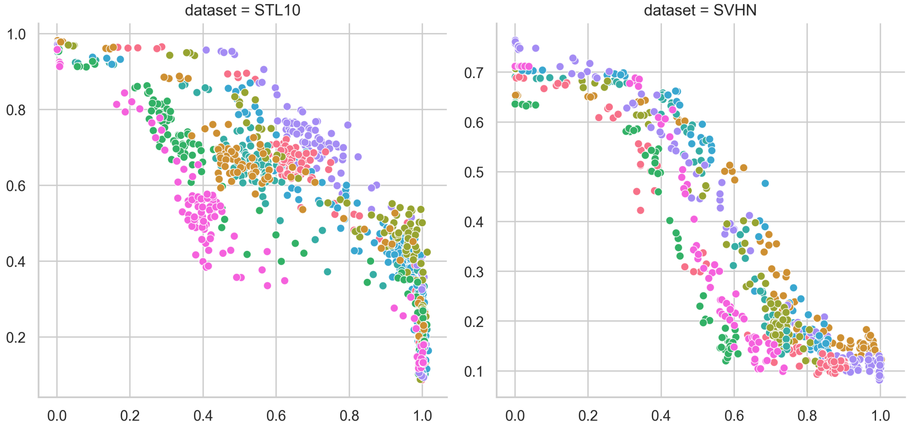
MIA as a statistical test (lemma 1)

H_0 : " $\left(\hat{\theta}_n, \tilde{z}\right) \sim P_{\left(\hat{\theta}_n, z_1\right)}$" vs. H_1 : " $\left(\hat{\theta}_n, \tilde{z}\right) \sim P_{\hat{\theta}_n} \otimes P$"

Estimating Δ_n

MIA as a statistical test (lemma 1)

$$H_0 : \text{"} \left(\hat{\theta}_n, \tilde{z}\right) \sim P_{\left(\hat{\theta}_n, z_1\right)} \text{"} \text{ vs. } H_1 : \text{"} \left(\hat{\theta}_n, \tilde{z}\right) \sim P_{\hat{\theta}_n} \otimes P \text{"}$$



Conclusion

Results

- Overfitting : $\Delta_n(P, \mathcal{A}) \approx 1$
- Discrete data : $\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) \approx \frac{C(P)}{2} n^{-1/2}$

Conclusion

Results

- Overfitting : $\Delta_n(P, \mathcal{A}) \approx 1$
- Discrete data : $\max_{\mathcal{A}} \Delta_n(P, \mathcal{A}) \approx \frac{C(P)}{2} n^{-1/2}$

Ongoing Works

- Audit of a privacy mechanism.
- Quantization of Parameters.

E. Aubinais, E. Gassiat and P. Piantanida
Fundamental Limits of Membership Inference Attacks on MachineLearning Models
http ://arxiv.org/abs/2310.13786.