Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 17: Finish membership inference / Start DP

Nov 6, 2026

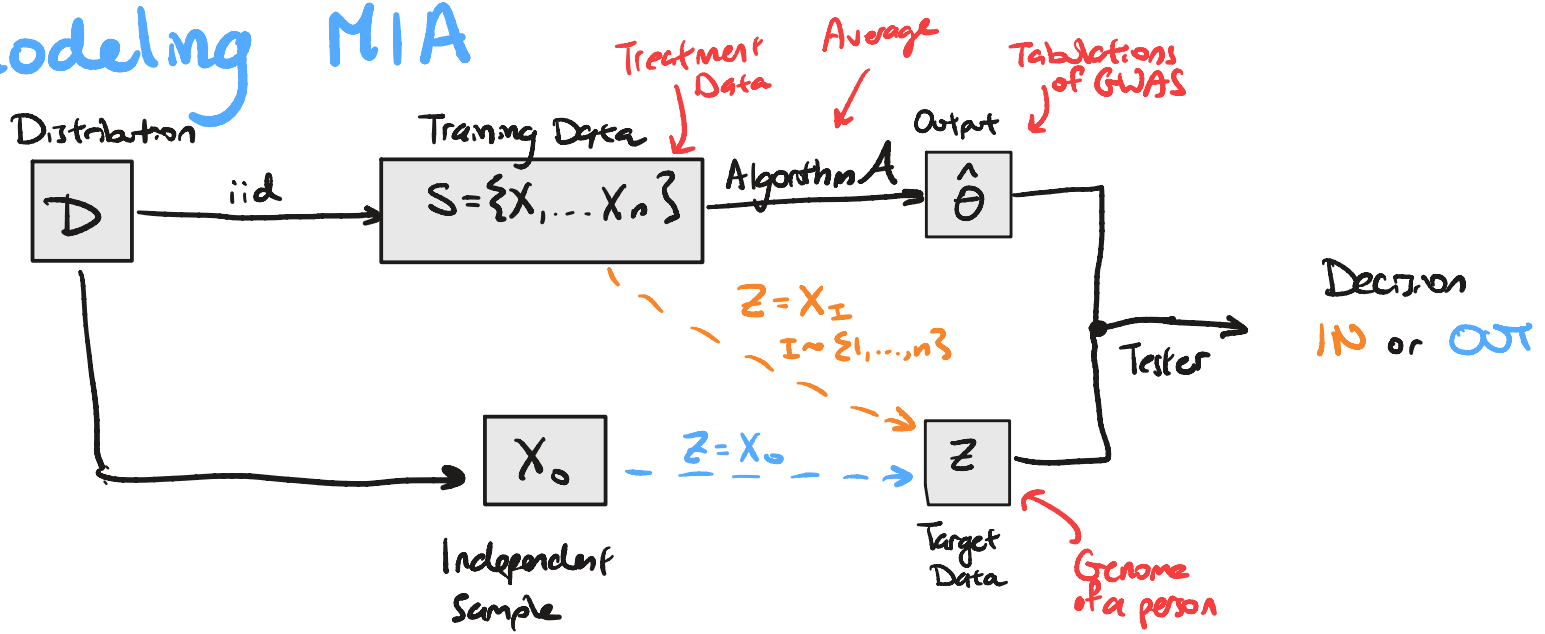
Outline

- ~~The OG membership inference~~
- ~~A model of membership inference~~
- A theoretical example
  - ↳ ~~mechanics~~
  - ↳ ~~connection to hypothesis testing~~ + analysis
  - ↳ connection to membership inference in ML/AI
- ~~"Robust" membership inference~~ (Cut for time)

Quick Announcements

- ~~Projects~~
- ~~CRISP~~
- ~~Poll this weekend~~

Modeling MIA



P_{OUT}: $(X_0, A(X_1,\dots X_n)) = (Z, \hat{\theta})$

P_{IN}: $(X_I, A(X_1,\dots X_n)) = (Z, \hat{\theta})$

If D is known and A is fixed then MIA is a hypothesis testing problem

The Simplest MIA

Computing the LR:

P_{OUT}: $(X_0, \bar{X})$ independent

$X_0 \sim N(\mu, \Pi)$

$\bar{X} | X_0 \sim N(\mu, \frac{1}{n}\Pi)$

P_{IN}: $(X_1, \bar{X})$ correlated

$X_1 \sim N(\mu, \Pi)$

$\bar{X} | X_1 \sim N(\frac{X_1}{n} + \mu, \frac{1}{n}\Pi)$

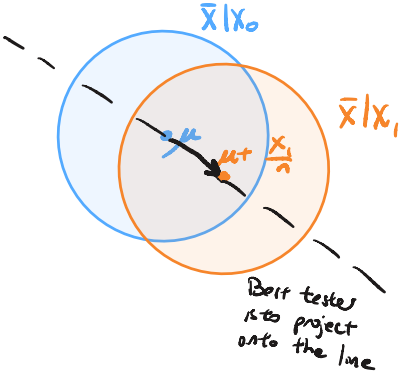
slight oversimplification

Instantiation of MIA

$D = N(\mu, \Pi)$ in $\mathbb{R}^d$

$S = \{X_1, \ldots X_n\} \overset{iid}{\sim} D \quad X_0 \overset{iid}{\sim} D$

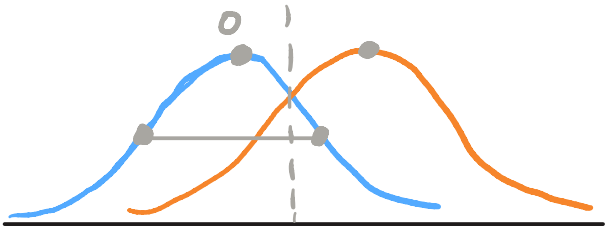
$A(S) = \bar{X} = \frac{1}{n}\sum_i X_i$



$\text{Test}_D(z, \bar{X}) = \langle \bar{X} - \mu, z - \mu \rangle$

Analysis of Tester

Distribution of $T_D(z; \bar{X})$



Instantiation of MIA

$D = N(\mu, I)$ in $\mathbb{R}^d$

$S = \{X_1, \dots X_n\} \overset{iid}{\sim} D \quad X_0 \overset{iid}{\sim} D$

$A(S) = \bar{X} = \frac{1}{n}\sum_i X_i$

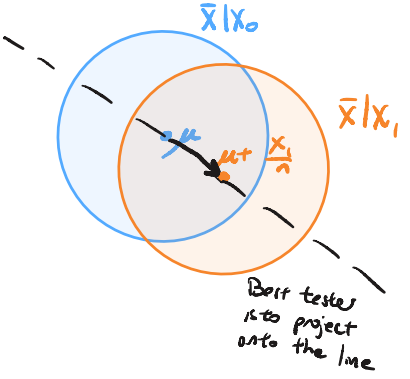
OUT Distribution:

$$\mathop{\mathbb{E}}_{OUT}[T_D(X_0, \bar{X})]$$

$$= \mathop{\mathbb{E}}_{OUT}[\langle X_0 - \mu, \bar{X} - \mu \rangle]$$

$$= \langle \mathop{\mathbb{E}}_{OUT}[X_0 - \mu], \mathop{\mathbb{E}}_{OUT}[\bar{X} - \mu] \rangle \quad \text{(Independence)}$$

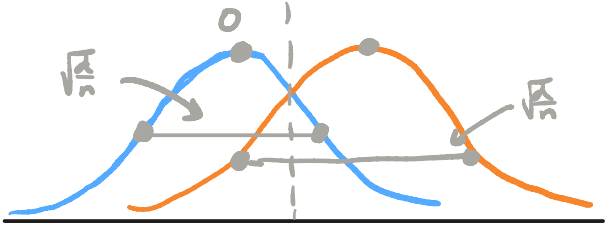
$$= \langle \vec{0}, \vec{0} \rangle = 0$$



$$\text{Test}_D(z, \bar{X}) = \langle \bar{X} - \mu, z - \mu \rangle$$

Analysis of Tester

Distribution of $T_D(z; \bar{X})$



Instantiation of MIA

$D = N(\mu, I)$ in $\mathbb{R}^d$

$S = \{X_1, \ldots X_n\} \overset{iid}{\sim} D \quad X_0 \overset{iid}{\sim} D$

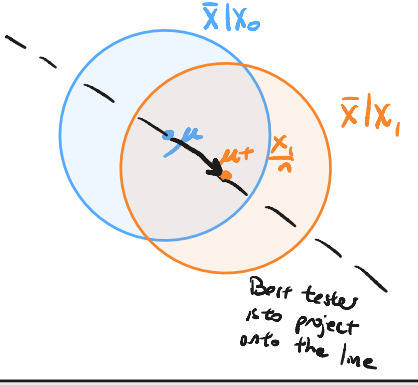
$A(S) = \bar{X} = \frac{1}{n}\sum_i X_i$

OUT Distribution:

$$\underset{\text{OUT}}{\text{Var}}\left[T_D(X_0, \bar{X})\right] = \underset{\text{OUT}}{\mathbb{E}}\left[T_D(X_0; \bar{X})^2\right] - \underset{\text{OUT}}{\mathbb{E}}\left[T_D(X_0; \bar{X})\right]^2$$

$$= \underset{\text{OUT}}{\mathbb{E}}\left[\|\bar{X} - \mu\|_2^2\right]$$

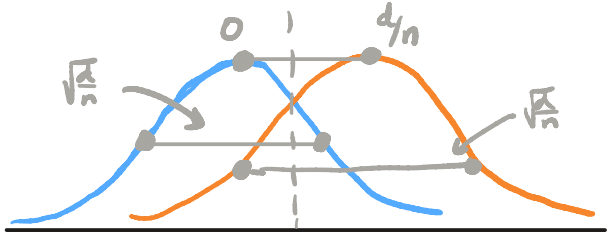
$$\bar{X} \sim N(\mu, \tfrac{1}{n} I) \qquad = \frac{d}{n}$$



$$\text{Test}_D(z, \bar{X}) = \langle \bar{X} - \mu, z - \mu \rangle$$

Analysis of Tester

Distribution of $T_D(z; \bar{X})$



Instantiation of MIA

$D = \mathcal{N}(\mu, \mathrm{II})$ in $\mathbb{R}^d$

$S = \{X_1, \dots X_n\} \overset{iid}{\sim} D \qquad X_0 \overset{iid}{\sim} D$

$A(S) = \bar{X} = \frac{1}{n}\sum_i X_i$

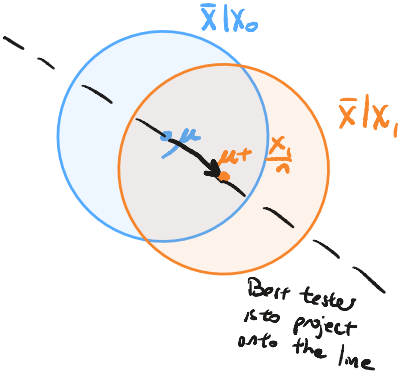
IN Distribution:

$$\mathop{\mathbb{E}}_{IN}[\langle X_1 - \mu, \bar{X} - \mu \rangle] = \mathop{\mathbb{E}}_{IN}\left[\langle X_1 - \mu, \tfrac{1}{n}\textstyle\sum_i X_i - \mu \rangle\right]$$

$$= \frac{1}{n}\mathbb{E}[\langle X_1 - \mu, X_1 - \mu \rangle] + \frac{1}{n}\sum_{i=2}^{n} \mathbb{E}[\langle X_1 - \mu, X_i - \mu \rangle]$$

$$= 0$$

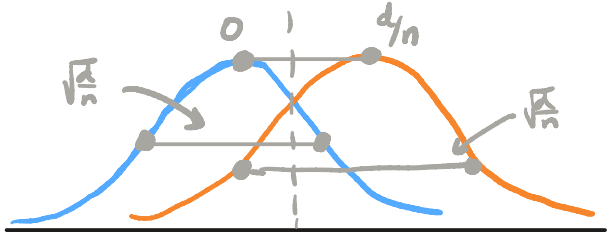
$$= \frac{1}{n}\mathbb{E}\left[\|X_1 - \mu\|_2^2\right] = \frac{d}{n}$$



$$\text{Test}_D(z, \bar{X}) = \langle \bar{X} - \mu, z - \mu \rangle$$

Analysis of Tester

Distribution of $T_D(z; \bar{X})$

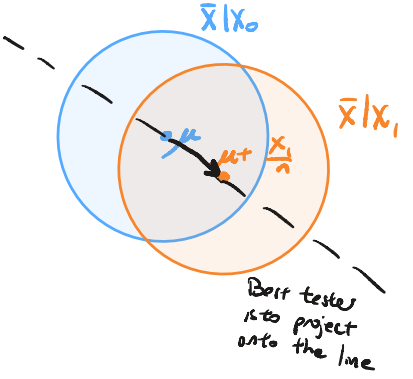


Instantiation of MIA

$D = N(\mu, I)$ in $\mathbb{R}^d$

$S = \{X_1, \ldots X_n\} \overset{iid}{\sim} D \quad X_0 \overset{iid}{\sim} D$

$A(S) = \bar{X} = \frac{1}{n}\sum_i X_i$



"Thm": Membership inference succeeds exactly as well as you can tell apart

$N(0, \frac{d}{n})$ and $N(\frac{d}{n}, \frac{d}{n})$

Ex. for $n = 100$, $d = 500$ AUC ≈ 0.94

MIA is possible if and only if $d \gg n$

$\text{Test}_D(z, \bar{X}) = \langle \bar{X} - \mu, z - \mu \rangle$

Applying MIA to Practical AI

D = ImageNet (or whatever)

$A(S)$ = Train neural network for D

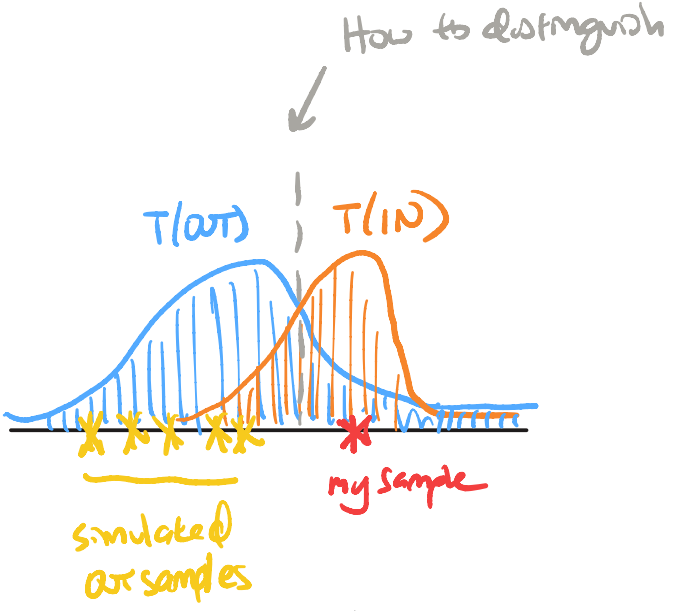
Hypothesis Testing:

$$(X_0, A(X_1, \ldots X_n)) \text{ vs. } (X_1, A(X_1, \ldots X_n))$$

Q1: What to use for $T_D(z; \hat{\theta})$?

- Loss of $\hat{\theta}$ on z
- Confidence of $\hat{\theta}$ on z

Q2: What if you dont know D and A perfectly?



Use "reference data" to train models

Differential Privacy

Membership Inference

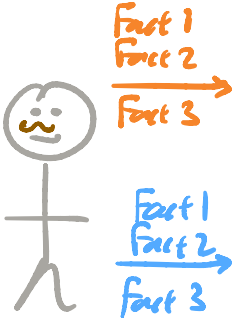
An attacker can tell if you are in the dataset or not.

! Membership Inference

No attacker can tell if you are in the dataset or not.

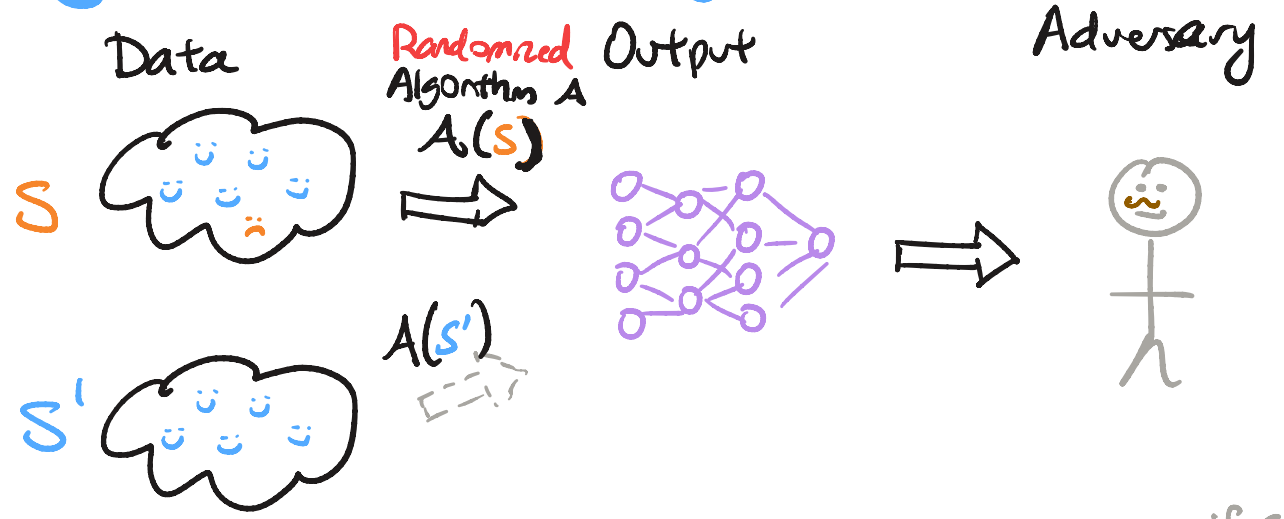
real world $X_1 \dots X_n \xrightarrow{A} \hat{\theta}$

hypothetical world $X_1', X_2, \dots X_n \xrightarrow{A} \hat{\theta}$



$\Rightarrow$ Anything the attacker learns about you will be the same whether or not you're in dataset

Defining Differential Privacy



Definition: A is ε-differentially private if

for every pair S, S' differing on one person's data

for every event $E \subseteq \text{Range}(A)$

$$\mathbb{P}[A(S) \in E] \leq e^{\varepsilon} \cdot \mathbb{P}[A(S') \in E]$$

If ε is small then $e^{\varepsilon} \approx 1+\varepsilon$

Close as distributions