Membership Inference Attacks, Part II
April 15, 2025

① Exercise on means + variances.
from in-class exercises ≈ 2 weeks ago.

② Worst-case mechanism?

$$\mu \sim N(0, I_k)$$

similar to exercise, but leaves mechanism uncertain about μ.

$$X_1, \ldots, X_n \sim_{iid} N(\mu, I_k)$$

$$\mathcal{M}(X) = \bar{x} + e \quad \text{where } \|e\|_2 \le \alpha\sqrt{k} \quad (\text{RMSE } \alpha)$$

$$T(a,y) = \langle a-\mu, y-\mu \rangle$$

OUT: $\mathbb{E}(T \mid \text{OUT}) = \mathbb{E}\langle X-\mu, \bar{x}+e-\mu \rangle$

$$= \mathbb{E}\langle \underset{\text{indep, mean } 0}{X-\mu, \bar{x}-\mu} \rangle + \mathbb{E}\langle \underset{\text{mean } 0,\ \text{indep of } e}{X-\mu}, e \rangle$$

$$Var(T \mid \text{OUT}) = k\, Var(X(1)-\mu(1)) \left(Var(\bar{x}(1)-\mu(1)) + Var(N(0, \|e\|^2)) \right)$$

use the fact that $Var(A+B) \le 2(Var(A) + Var(B))$

$$\le 2\left(k/n + \alpha^2 d\right)$$

Same as case of Gauss noise with $\rho \propto \alpha$.

if A, B indep with mean 0
Then $Var(AB) = \mathbb{E}(A^2B^2)$
$= \mathbb{E}(A^2)\,\mathbb{E}(B^2)$
$= Var(A)\,Var(B)$

IN: $\mathbb{E}(T \mid \text{IN}) = \mathbb{E}_i \langle X_i - \mu, \mathcal{M}(X) - \mu \rangle$

$$= \langle \bar{x}-\mu, \mathcal{M}(x) - \mu \rangle$$

$$= \langle \bar{x}-\mu, \bar{x}-\mu \rangle + \langle \bar{x}-\mu, e \rangle$$

- Now condition on $\bar{x}$ and e (where $\|e\|_2 \le \alpha\sqrt{d}$)
- What is distrib of $\mu \mid \bar{X} = \bar{x}$?

Claim

- From analyst's pov (that is, conditioned on $X_1, \dots, X_n$)
$$\mu \sim \frac{n}{n+1}\bar{X} + \frac{1}{n+1} Z \quad \text{where } Z \sim N(0, I_d) \text{ indep of } \bar{X}$$

Proof:

- $\operatorname{Cov}(\mu, \bar{X}) = \begin{pmatrix} I & I \\ I & (1+\frac{1}{n})I \end{pmatrix}$

- Suppose that $\mu = a\bar{X} + bZ$ where $Z \sim N(0, I_d)$

Then $\operatorname{Cov}(\mu, \bar{X}) = \begin{pmatrix} a^2\Sigma_{\bar{X}} + b^2 I & a\Sigma_{\bar{X}} \\ a\Sigma_{\bar{X}} & \Sigma_{\bar{X}} \end{pmatrix}$ $\therefore$ $b^2 = \frac{1}{n+1}$

$a^2 = \frac{1}{1+\frac{1}{n}} = \frac{n}{n+1}$

- So $\mu = \frac{n}{n+1}\bar{X} + bZ$.

- Furthermore, all the info about μ in $X_1, \dots, X_n$ is contained in $\bar{X}$.

(that is, $\bar{X}$ is a "sufficient statistic" for μ)

$\rightarrow$ To see why, note that

$$p(X_1, \dots, X_n \mid \mu) \propto \prod \exp\left(-\tfrac{1}{2}\|x_i - \mu\|^2\right)$$
$$= \exp\left(\tfrac{-1}{2}\sum \|x_i - \mu\|^2\right)$$
$$= \exp\left(\tfrac{-1}{2}\sum_i \left(\|x_i - \bar{x}\|^2 + \|\bar{x} - \mu\|^2 + 2\langle x_i - \bar{x}, \bar{x} - \mu\rangle\right)\right)$$
$$= \exp\left(\tfrac{-1}{2}\left(n\|\bar{x} - \mu\|^2 + \sum_i \|x_i - \bar{x}\|^2 + 2\langle \underset{0}{0}, \bar{x} - \mu\rangle\right)\right)$$
$$= \underbrace{\exp\left(\tfrac{-1}{2} n\|\bar{x} - \mu\|^2\right)}_{\propto P(\bar{x} \mid \mu)} \cdot \underbrace{\exp\left(\tfrac{-1}{2}\sum \|x_i - \bar{x}\|^2\right)}_{\propto P(x_1, \dots x_n \mid \bar{x}, \mu)}$$

QED

- We can now use the claim

$$\mathbb{E}(T \mid IN) = \mathbb{E}\left(\|\bar{x} - \mu\|^2 + \langle \bar{x} - \mu, e \rangle \right)$$

$$= d/n + \mathbb{E} \langle \bar{x} - \mu, e \rangle$$

- Conditioning on $\bar{x}$ and e, we get:

$$= \frac{d}{n} + \mathop{\mathbb{E}}_{\bar{x}, e} \mathop{\mathbb{E}}_{z} \langle \bar{x} - \tfrac{n}{n+1} \bar{x} + \tfrac{1}{n+1} z, e \rangle$$

$$= \frac{d}{n} + \mathop{\mathbb{E}}_{x, e} \frac{1}{n+1} \langle \bar{x}, e \rangle$$

$$\geq \frac{d}{n} - \frac{1}{n+1} \underbrace{\|\bar{x}\|}_{\approx \sqrt{d}} \cdot \underbrace{\|e\|}_{\leq \alpha \sqrt{d}}$$

$$\geq \frac{d}{n}(1 - \alpha)$$

which is less than $\frac{d}{n}$ for $\alpha \leq \frac{1}{2}$.