CS 499/599: Machine Learning Security

02.23: Privacy

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL

Secure AI Systems Lab

Notice

- Due dates
  - Written paper critique (28th)
  - HW3 deadline (28th)

- Sign-up (on Canvas)
  - Scribe lecture note [3 slots remain]
  - In-class paper presentation / discussion [ONLY 1 slots remain, on the 2nd of Mar]

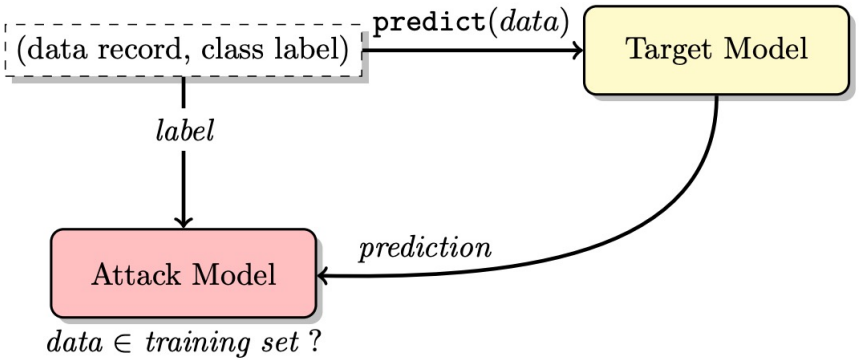
Topics for Today

- Privacy
  - Warm-boot
  - Threat Models
  - Differential privacy (DP)
- Privacy Attacks and Defenses
  - Non-ML: Data anonymization
  - Membership inference (Tracing attack)
    - Threat Model
    - Attacks
      - Shokri *et al.*
      - Yeom *et al.*
    - Defensive techniques

Shokri *et al.*, Membership Inference Attacks against Machine Learning Models

Threat Model

- Membership Inference
  - **Goal:**
    - Identify if a specific instance y is **IN** the dataset D_{train} or is not (**OUT**)

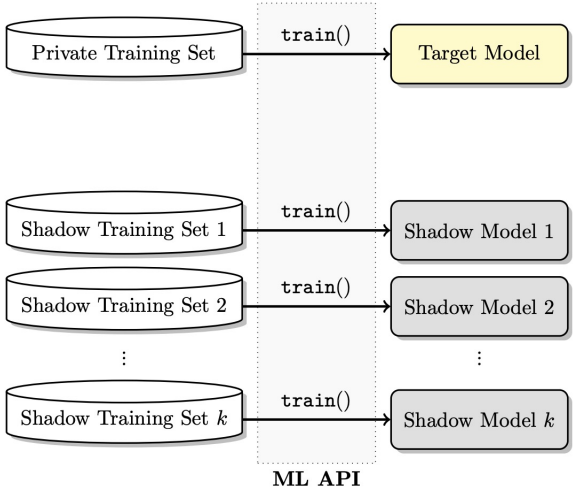


Threat Model

- Membership Inference
  - **Goal:**
    - Identify if a specific instance y is **IN** the dataset D_{train} or is not (**OUT**)
  - **Knowledge:**
    - The format of inputs and outputs, such as:
      - What features do they collect?
      - What are those feature's values (range)?
      - …
    - Some knowledge on the distribution of D_{train}
  - **Capability:**
    - Has a query access to the target model
    - Has computational power to train surrogate (*i.e.*, shadow) models

Membership Inference Attack (Shokri et al.)

- Shadow Models
  - **Idea:**
    - The attacker has some data samples
    - If the attacker trains models with those samples, we know their memberships!
    - If shadow models are trained similarity, we can exploit the membership info.!
  - **Attacker's data:**
    - Know the labeled records: $(\boldsymbol{x}, y)$
    - Query them to the target model and collect its predictions: $\big((\boldsymbol{x}, y), \boldsymbol{y}\big)$
  - **How to train?**
    - Create a train and test split
    - Use the *train* data to train the shadow models



Membership Inference Attack (Shokri et al.)

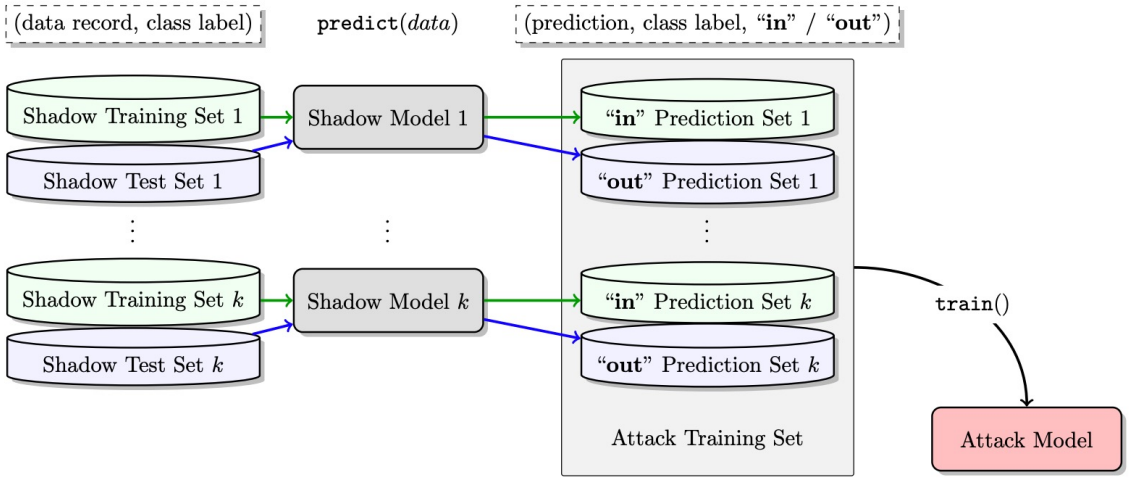
- Shadow Models
  - **Attacker's data :**
    - Require some data $(\boldsymbol{x}, y)$ from a distribution like the victim's
  - **Data generation strategies:**
    - Model-based synthesis
    - Statistics-based synthesis
    - Noisy real-data

Algorithm 1 Data synthesis using the target model

procedure SYNTHESIZE(class : c)
  $\mathbf{x} \leftarrow$ RANDRECORD( ) ▷ *initialize a record randomly*
  $y_c^* \leftarrow 0$
  $j \leftarrow 0$
  $k \leftarrow k_{max}$
  for $iteration = 1 \cdots iter_{max}$ **do**
    $\mathbf{y} \leftarrow f_{\text{target}}(\mathbf{x})$ ▷ *query the target model*
    if $y_c \geq y_c^*$ **then** ▷ *accept the record*
      if $y_c > \text{conf}_{min}$ and $c = \arg\max(\mathbf{y})$ **then**
        if $\text{rand}() < y_c$ **then** ▷ *sample*
          return $\mathbf{x}$ ▷ *synthetic data*
        end if
      end if
      $\mathbf{x}^* \leftarrow \mathbf{x}$
      $y_c^* \leftarrow y_c$
      $j \leftarrow 0$
    else
      $j \leftarrow j + 1$
      if $j > rej_{max}$ **then** ▷ *many consecutive rejects*
        $k \leftarrow \max(k_{min}, \lceil k/2 \rceil)$
        $j \leftarrow 0$
      end if
    end if
    $\mathbf{x} \leftarrow$ RANDRECORD($\mathbf{x}^*$, k) ▷ *randomize k features*
  end for
  return $\perp$ ▷ *failed to synthesize*
end procedure

Membership Inference Attack (Shokri et al.)

- Model for the attack
  - **Attacker's data:**
    - Data format $\big((\boldsymbol{x}, y), \boldsymbol{y}\big)$
    - Some of them are **"in"** the shadow train, otherwise **"out"**
    - Combine three info. $(y, \boldsymbol{y}, \boldsymbol{in})$ or $(y, \boldsymbol{y}, \boldsymbol{out})$
    - Make the attack model predict **in** or **out**

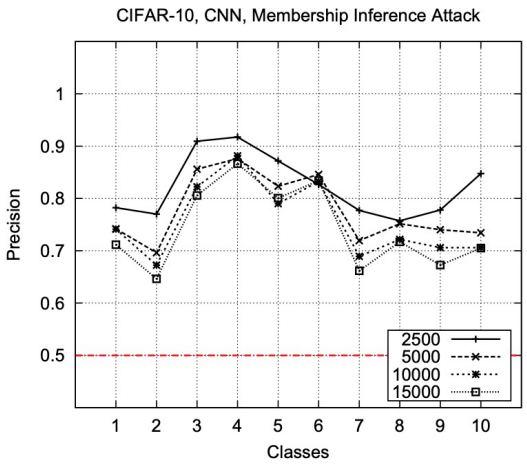


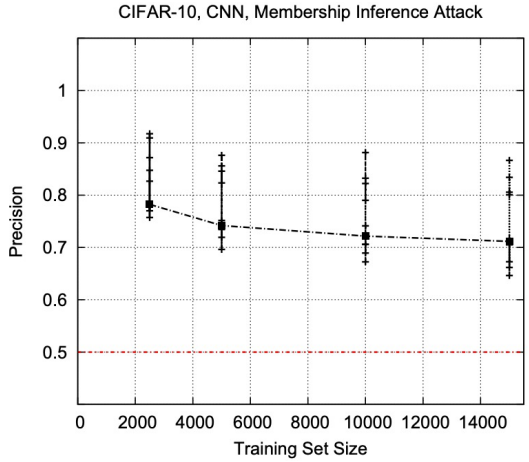
Evaluation

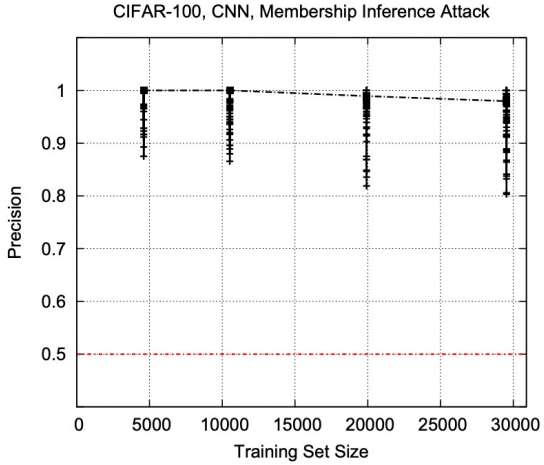
- Setup
  - Datasets:
    - MNIST | CIFAR-10/100
    - Purchases | Locations | Texas-100 | UCI Adult
  - Models
    - MLaaS: Google Prediction API | Amazon ML | NNs
  - MI Attack
    - Shadow models: 20 – 100 models
  - Defenses
    - Heuristics: Top-k | Precision | Regularization
    - **[?!]** In theory: DP

Evaluation

- MI Attacks on CIFAR
  - Shadow models: 100
  - Training set (for targets):
    - CIFAR-10: {2.5, 5, 10, 15}k samples
    - CIFAR-100: {4.5, 10, 20, 30}k samples
  - **In-short:** MI attacks work with a pretty reasonable acc.



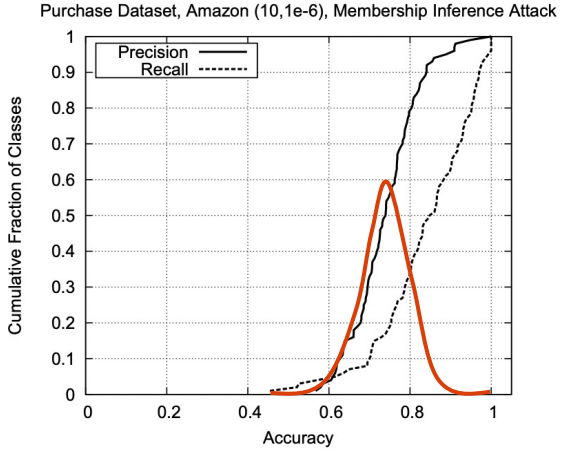


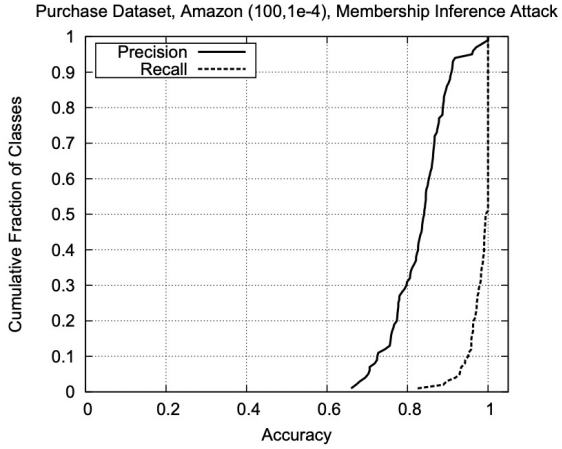


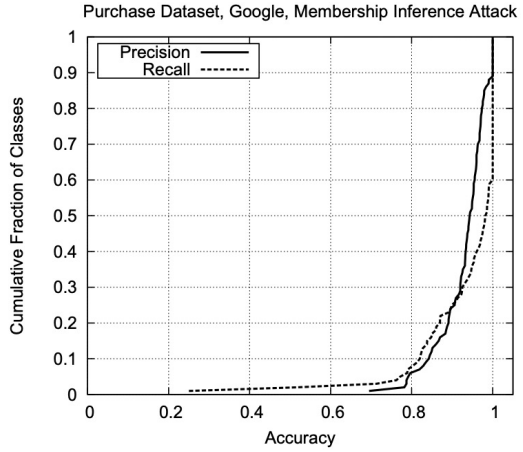
Evaluation

- MI Attacks w. Different Models
  - Dataset: Purchase-100
  - Models (trained on 10k records):
    - Amazon ML
    - Google's Prediction API
  - **In-short:** across all models, MI attacks work with a pretty reasonable acc.

ML Platform	*Training*	*Test*
Google	0.999	0.656
Amazon (10,1e-6)	0.941	0.468
Amazon (100,1e-4)	1.00	0.504
Neural network	0.830	0.670

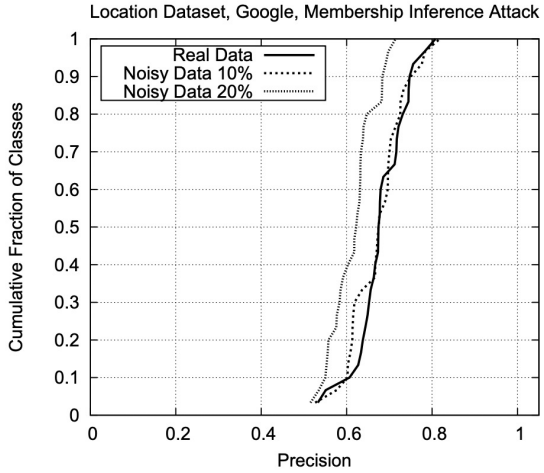


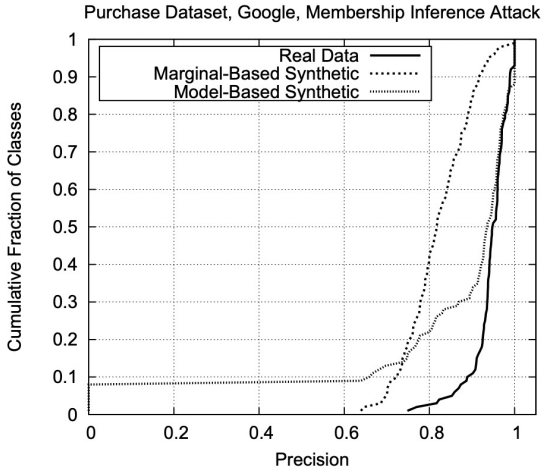




Evaluation

- MI Attacks w. Different Shadow Models
  - Dataset: Location
  - Modification:
    - Noisy shadow training data
    - No data (synthesize it!)
  - **In-short:** MI attacks show robust acc. under the weak approximation of the dist.

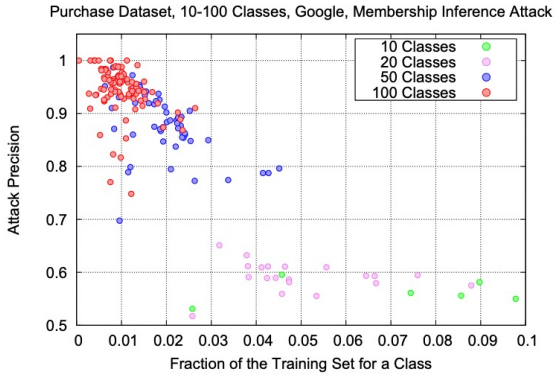


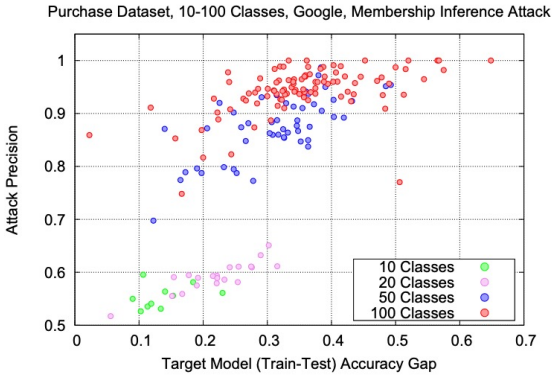


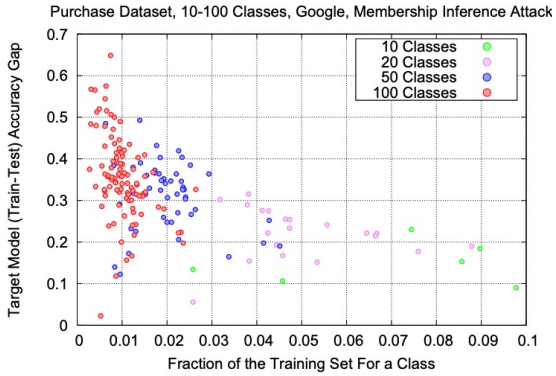
Evaluation

- MI Attacks w. Different # classes
  - Dataset: Purchase
  - Modification:
    - # Classes: 10 – 100 classes (keep N(D_{tr}) the same)
    - Google Prediction API
  - **In-short:** More supporting data samples in the c

Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657





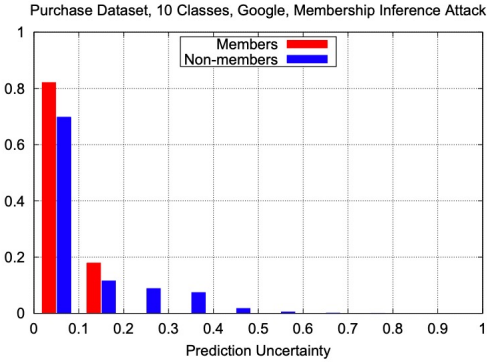


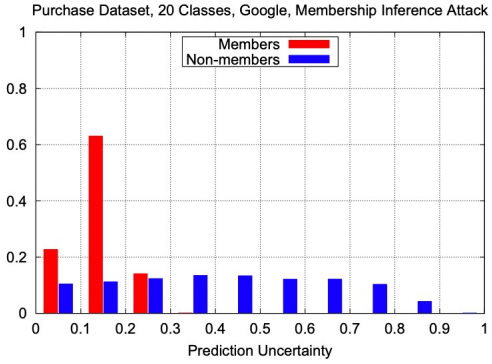
Evaluation

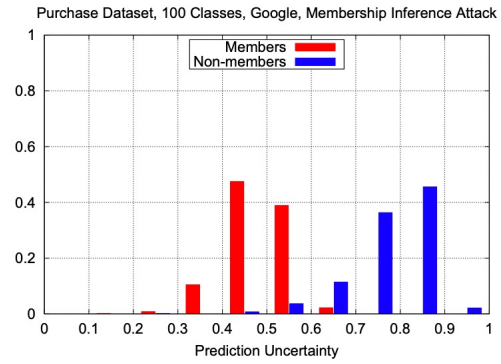
- MI Attacks, Why Do They Work?
  - Dataset: Purchase
  - Modification:
    - # Classes: 10 – 100 classes (keep N(D_{tr}) the same)
    - Google Prediction API
  - **In-short:** It may depend on a model's ability to distinguish members and non-members

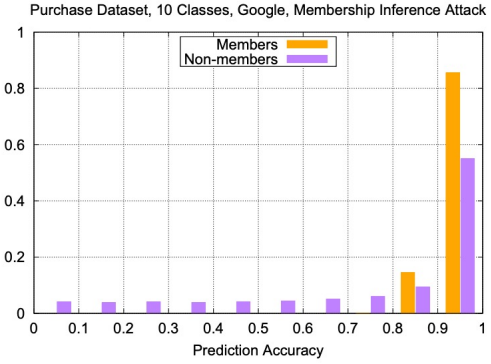
Evaluation

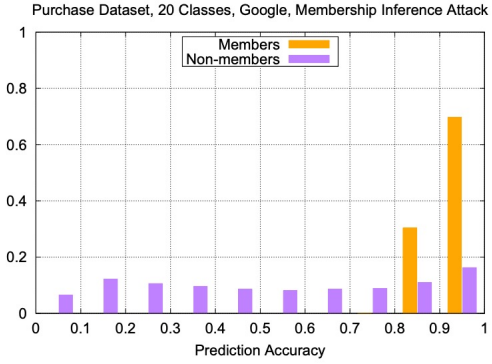
- MI Attacks, Why Do They Work?

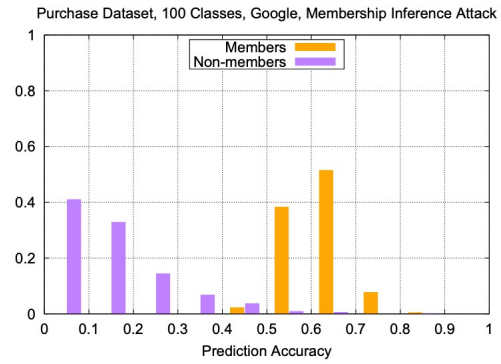












Evaluation

- Defenses
  - Top-k
  - Precision (round-ups)
  - Regularization (L_2)

- Results (on NNs)
  - Still MI attack works
    - in $k = 1$ (label)
    - with less precision ($d = 1$)
  - Regularization somewhat effective but care must be taken for a model's acc.

Purchase dataset	*Testing Accuracy*	*Attack Total Accuracy*	*Attack Precision*	*Attack Recall*
No Mitigation	0.66	0.92	0.87	1.00
Top $k = 3$	0.66	0.92	0.87	0.99
Top $k = 1$	0.66	0.89	0.83	1.00
Top $k = 1$ label	0.66	0.66	0.60	0.99
Rounding $d = 3$	0.66	0.92	0.87	0.99
Rounding $d = 1$	0.66	0.89	0.83	1.00
Temperature $t = 5$	0.66	0.88	0.86	0.93
Temperature $t = 20$	0.66	0.84	0.83	0.86
L2 $\lambda = 1e-4$	0.68	0.87	0.81	0.96
L2 $\lambda = 1e-3$	0.72	0.77	0.73	0.86
L2 $\lambda = 1e-2$	0.63	0.53	0.54	0.52

Hospital dataset	*Testing Accuracy*	*Attack Total Accuracy*	*Attack Precision*	*Attack Recall*
No Mitigation	0.55	0.83	0.77	0.95
Top $k = 3$	0.55	0.83	0.77	0.95
Top $k = 1$	0.55	0.82	0.76	0.95
Top $k = 1$ label	0.55	0.73	0.67	0.93
Rounding $d = 3$	0.55	0.83	0.77	0.95
Rounding $d = 1$	0.55	0.81	0.75	0.96
Temperature $t = 5$	0.55	0.79	0.77	0.83
Temperature $t = 20$	0.55	0.76	0.76	0.76
L2 $\lambda = 1e-4$	0.56	0.80	0.74	0.92
L2 $\lambda = 5e-4$	0.57	0.73	0.69	0.86
L2 $\lambda = 1e-3$	0.56	0.66	0.64	0.73
L2 $\lambda = 5e-3$	0.35	0.52	0.52	0.53

Topics for Today

- Privacy
  - Warm-boot
  - Threat Models
  - Differential privacy (DP)

- Privacy Attacks and Defenses
  - Non-ML: Data anonymization
  - Membership inference (Tracing attack)
    - Threat Model
    - Attacks
      - Shokri *et al.*
      - Yeom *et al.*
    - Defensive techniques

Yeom *et al.*, Privacy Risks in Machine Learning: Analyzing the Connection to Overfitting

Motivation

- Prior work
  - Shows the overfitting is one factor that contributes MI attacks

Let’s Use the Paper

What We'll See

- Takeaways
  - Propose a metric that measures membership adv.
  - Make a connection between MI Attack's and overfitting formally
  - Propose a simple MI Attack (*Yeom et al.*)
    - It achieve am accuracy comparable with Shokri *et al.*
    - It requires less computational costs
  - Empirical evaluation of their theoretical connections and attacks

Recap

- Privacy
  - Warm-boot
  - Threat Models
  - Differential privacy (DP)

- Privacy Attacks and Defenses
  - Non-ML: Data anonymization
  - Membership inference (Tracing attack)
    - Threat Model
    - Attacks
      - Shokri *et al.*
      - Yeom *et al.*
    - Defensive techniques

Thank You!

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL

Secure AI Systems Lab