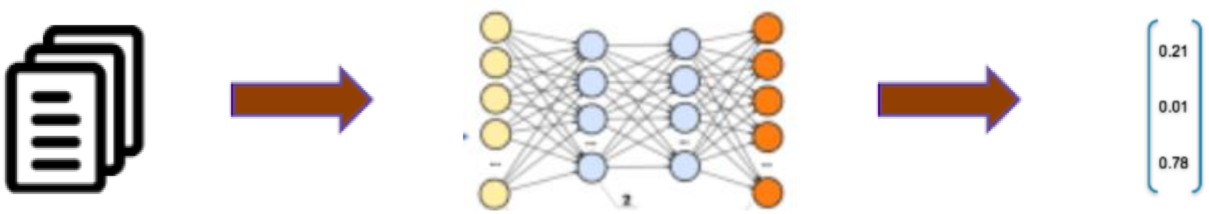
Machine learning privacy: a survey

Qiongxiu Li 2022.03.31

Hu H, Salcic Z, Sun L, et al. Membership inference attacks on machine learning: A survey[J]. ACM Computing Surveys (CSUR), 2021.
Song L, Mittal P. Systematic evaluation of privacy risks of machine learning models[C], USENIX 2021.
Liu Y, Wen R, He X, et al. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models[J]. USENIX, 2022.

Privacy attacks on machine learning models

Training data | Learning model | Output



Attacks on training data

Attack goal | Attack type

Entropy

Difficulty

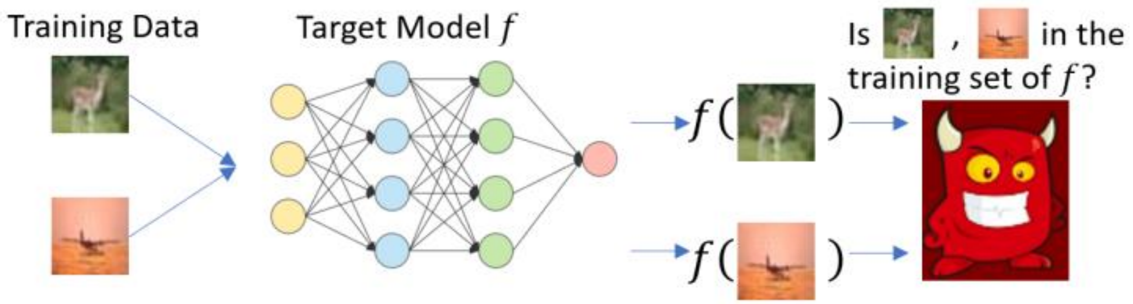
- Data membership → Membership inference attack
- Data attribute → Attribute inference attack
- Data itself → Model inversion attack

Attacks on learning model

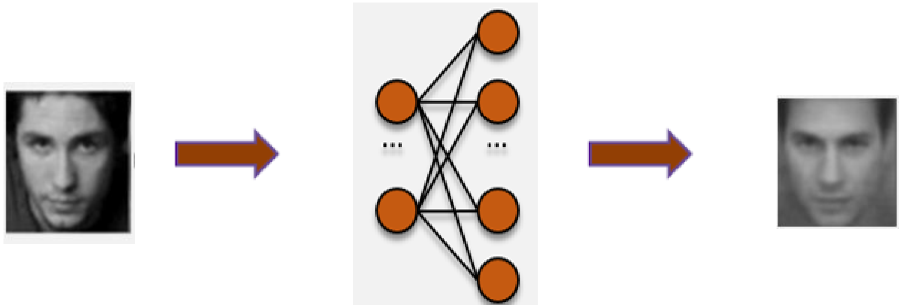
Attack goal | Attack type

- Leaning model → Model extraction /stealing attack

Privacy attacks on machine learning models

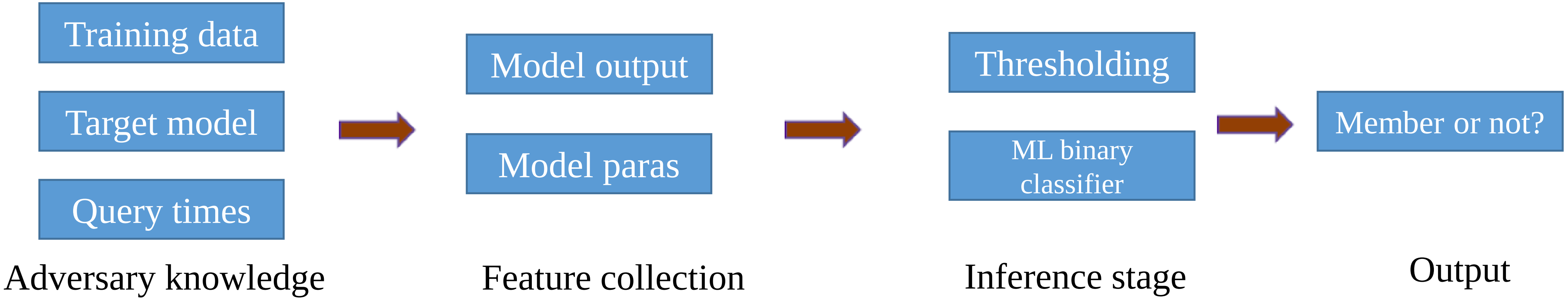


Membership inference attack

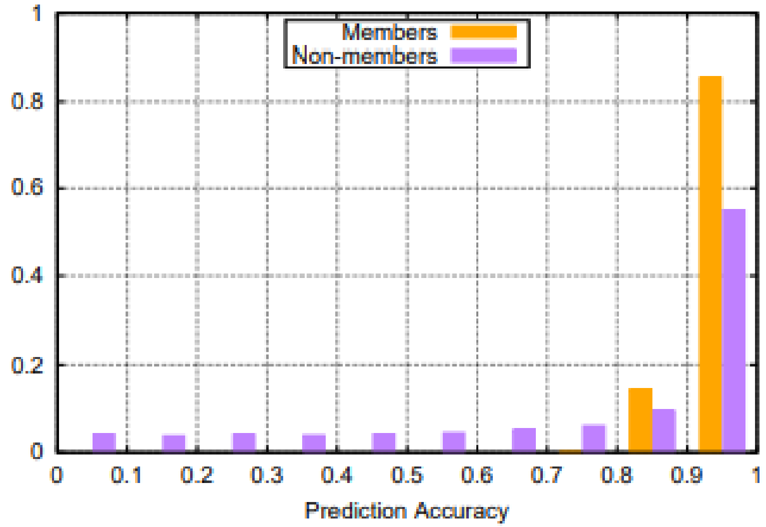


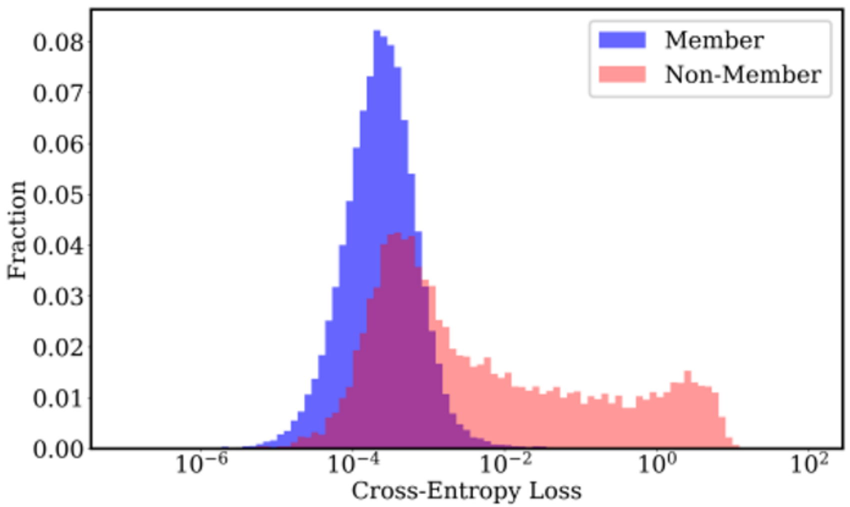
Model inversion attack
(input reconstruction)

Overflow of Membership inference attack



Feature distinguishability for member and non-member





Adversary knowledge in MIA

Knowledge type		Detailed feature
Knowledge of the target model	Black-box	All class confidence
		Top K class confidence
		Class label
	White-box	Gradient norm
		Distance to boundary
		Member and non-member distribution
Knowledge of training data	Partial training data	
	Synthetic data (known distribution)	
	No data	
Query time	One time	
	Multi queries	

Features for MIA

Metrics	Definition
Prediction Correctness	$\mathcal{M}_{\mathrm{corr}}(\hat{p}(y\,\vert\,\boldsymbol{x}), y) = \mathbb{1}(\mathrm{argmax}\ \hat{p}(y\,\vert\,\boldsymbol{x}) = y)$
Prediction Loss	$\mathcal{M}_{\mathrm{loss}}(\hat{p}(y\,\vert\,\boldsymbol{x}), y) = \mathbb{1}(\mathcal{L}(\hat{p}(y\,\vert\,\boldsymbol{x}); y) \leq \tau)$
Prediction Confidence	$\mathcal{M}_{\mathrm{conf}}(\hat{p}(y\,\vert\,\boldsymbol{x})) = \mathbb{1}(\max\ \hat{p}(y\,\vert\,\boldsymbol{x}) \geq \tau)$
Prediction Entropy	$\mathcal{M}_{\mathrm{entr}}(\hat{p}(y\,\vert\,\boldsymbol{x})) = \mathbb{1}(H(\hat{p}(y\,\vert\,\boldsymbol{x})) \leq \tau)$
Modified Prediction Entropy	$MH(\hat{p}(y\,\vert\,\boldsymbol{x}), y) = -(1-p_y)\log(p_y) - \sum_{i\neq y} p_i \log(1-p_i)$
Adversarial perturbation	It is harder to perturb a member instance to a different class than a non-member instance
Data augmentation	Model should be more confident on different augmented version of a data when it is a member

Evaluation metrics in MIA

Metrics	Definition
Attack Success Rate (ASR)	% successful attack over all attacks
Attack precision (AP)	% correctly classified members over all classified members
Attack recall (AR/TPR)	% correctly classified members over all real members
Attack false positive rate (FPR)	% non-member falsely classified as members over all real non-members
Membership Advantage (MA)	$MA = AR - FPR$
Attack F_1 score	$\text{F1-score} = 2 \cdot AP \cdot AR / (AP + AR)$
TPR @ low FPR	% correctly classified members at no/few FPR

Reasoning why MIA works

Fundamental: Model behaves differently for data it has seen and has not

1. Theoretical investigation: very few
2. Empirical investigation
   a) Overfitting of target model: high model complexity and limited size of training set
   b) Type of target model: a model's decision boundary is sensitive to a particular data record is more vulnerable to MIA.
   c) Diversity of training data: more representative data helps generalization thus more resilient to MIA

How to defend against MIA

1. Restricting the amount of output information (black-box setting)
    a) Confidence score masking
    b) Top K confidence output
    c) Prediction label only (minimum information for classification problem)
2. Regularization
    a) General techniques like L2 norm, early stopping, data augmentation etc
    b) Adversarial regularization (use MIA)
    c) Mixup + MMD (limit the distance of output distributions of member and non-member)
3. Differential privacy
4. Knowledge distillation: transfer knowledge from the target model to a smaller one, restricting direct access to the private training dataset

Summery of adversary knowledge, method, and evaluation metric in Model inversion attack

		White-box	No dataset	Synthetic dataset	Evaluation metric
Fredrikson et al. (aim to recover a representative for each class)	Adversarial knowledge	✓	✓		MSE of reconstructed sample and mean sample of each target class
	Method	Use back-propagation to optimize noise example til the posterior exceeds to a predefined threshold.			
Zhang et al. (aim to synthesize the training dataset)	Adversarial knowledge	✓		✓	Accuracy and F1 score: use an classifier to check whether the reconstructed sample can be recognized correctly
	Method	Train a GAN use shadow data, optimize noise input under it can achieve high posterior in target model.			

Summery of adversary knowledge, method, and evaluation metric in Model stealing attack

<table>
<tr><th></th><th>Black-box
(Target model architect.)</th><th>Partial or Synthetic dataset
(target attribute)</th><th>Evaluation metric</th></tr>
<tr><td>Adversarial knowledge</td><td>✓</td><td>✓</td><td rowspan="2">Accuracy and agreement (means the proportion of samples where the target model and the stolen model make the same prediction)</td></tr>
<tr><td>Method</td><td colspan="2">The adversary uses data samples from their (partial or synthetic) dataset to query the target model and get the corresponding posteriors. Then use the posterior as ground truth to train the stolen model.</td></tr>
</table>

Summery of adversary knowledge, method, and evaluation metric in Attribute inference attack

	White-box (embedding)	Partial or Synthetic dataset (target attribute)	Evaluation metric
Adversarial knowledge	✓	✓	Accuracy and F1 score
Method	Adversary is assumed to know the embeddings of the target sample and the target attribute of the available dataset. Use the target attribute and embeddings to train a classifier to mount the attack		

Opensource tools for AI security & privacy

Tools	Covering attacks
DEEPSEC	Adversarial attacks and defenses
CleverHans	Adversarial examples
TROJANZOO	Backdoor attacks
ML privacy meter	Membership inference attack in both black and white box setting
ML-DOCTOR	MIA+Attributed inference attack+Model inversion+ Model stealing

Liu Y, Wen R, He X, et al. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models[J]. USENIX, 2022.