Curs 12
Membership Inference Attacks (MIAs)

1/27/2024

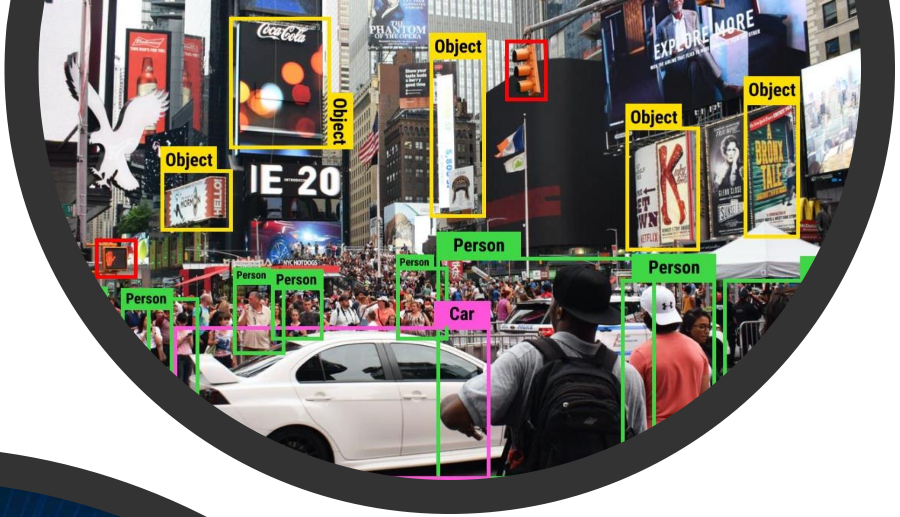
Course schedule

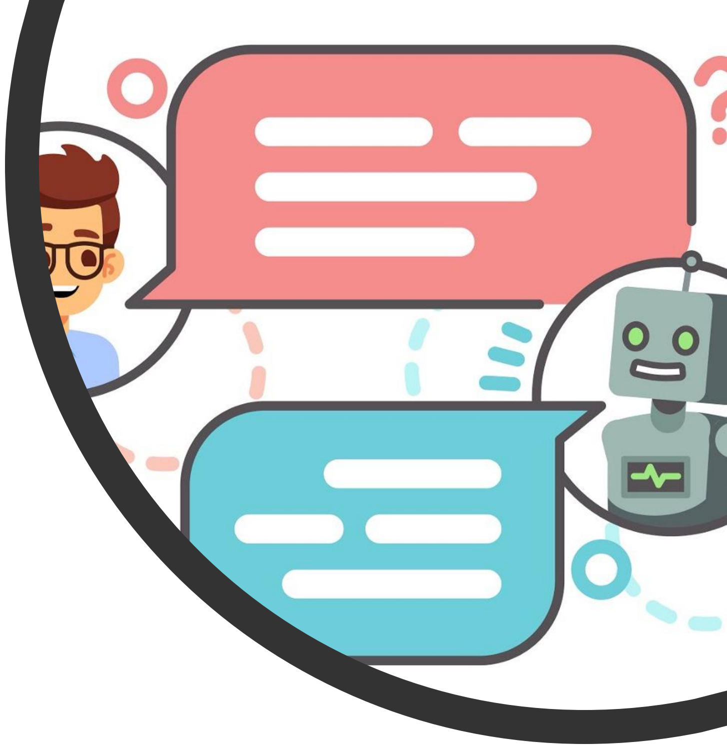
1. Why?
2. Cauzalitate
3. Măsurare
4. Modelare și eșantionare
5. Tehnici de analiză
   - Analiza factorială
   - Analiza cluster
   - Analiza de regresie
   - Analiza de rețea
   - Serii de timp
6. Predicție
7. Programare și ML
8. Why Privacy?
9. Privacy Enhancing Techniques
10. Homomorphic Encryption. PIR
11. Differential Privacy
12. Membership Inference Attacks
13. Federated Architecture. Multi-party computation
14. Explainable AI
15. Zero knowledge proof. Blockchain architecture

Contents

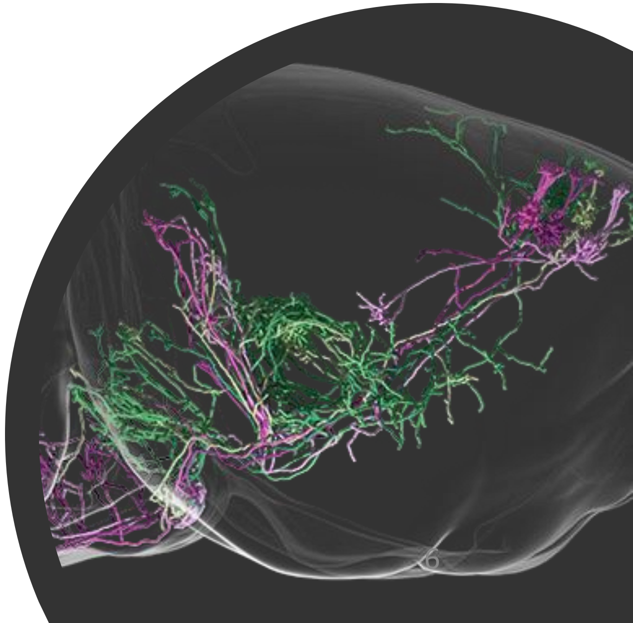
Contents

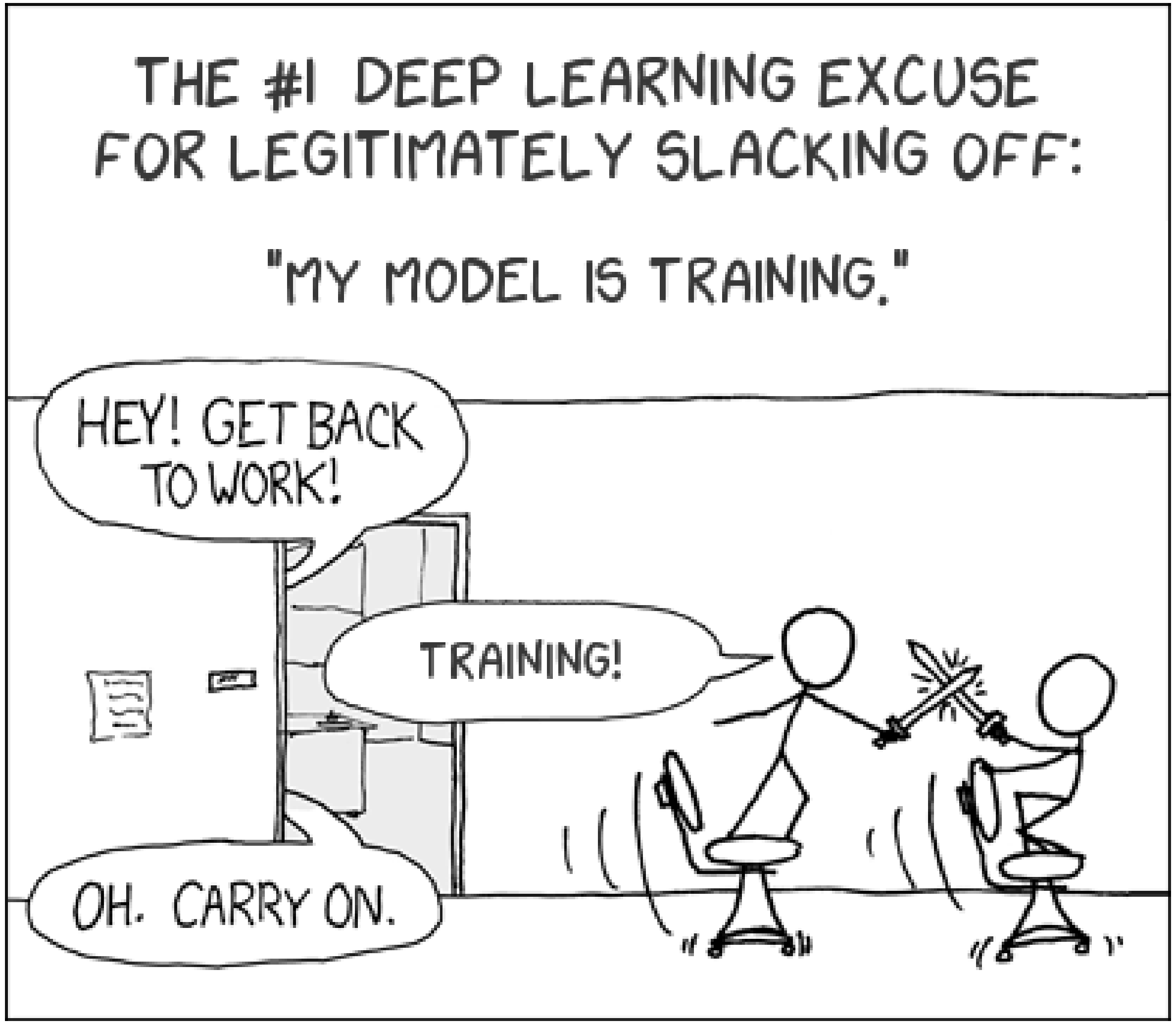
Context





Era of ML models




THE #1 DEEP LEARNING EXCUSE FOR LEGITIMATELY SLACKING OFF:
"MY MODEL IS TRAINING."
HEY! GET BACK TO WORK!
TRAINING!
OH. CARRY ON.

What are Membership Inference Attacks (MIAs)?

Membership Inference Attacks (MIAs)

- Was a data record used in the training phase of a ML model or not?

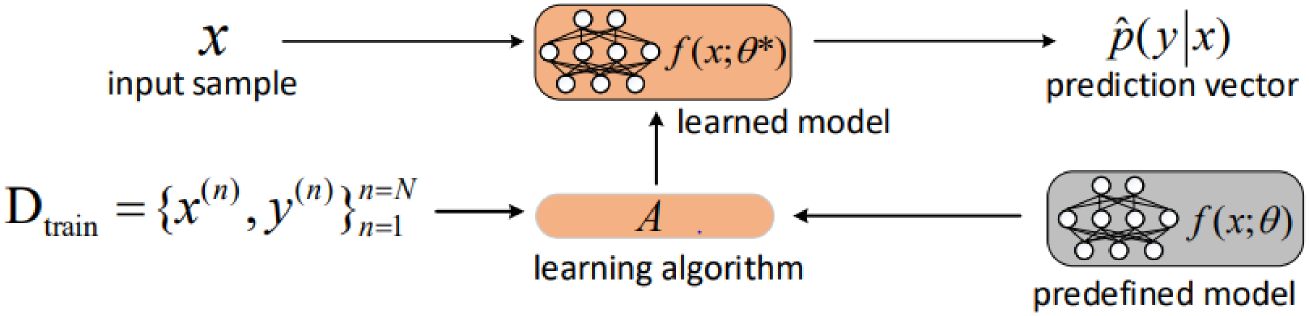


Fig. 1. A typical deep learning process for classification models.

Source: Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, *54*(11s), 1-37.

Why is relevant to protect against MIAs?

In the context of privacy

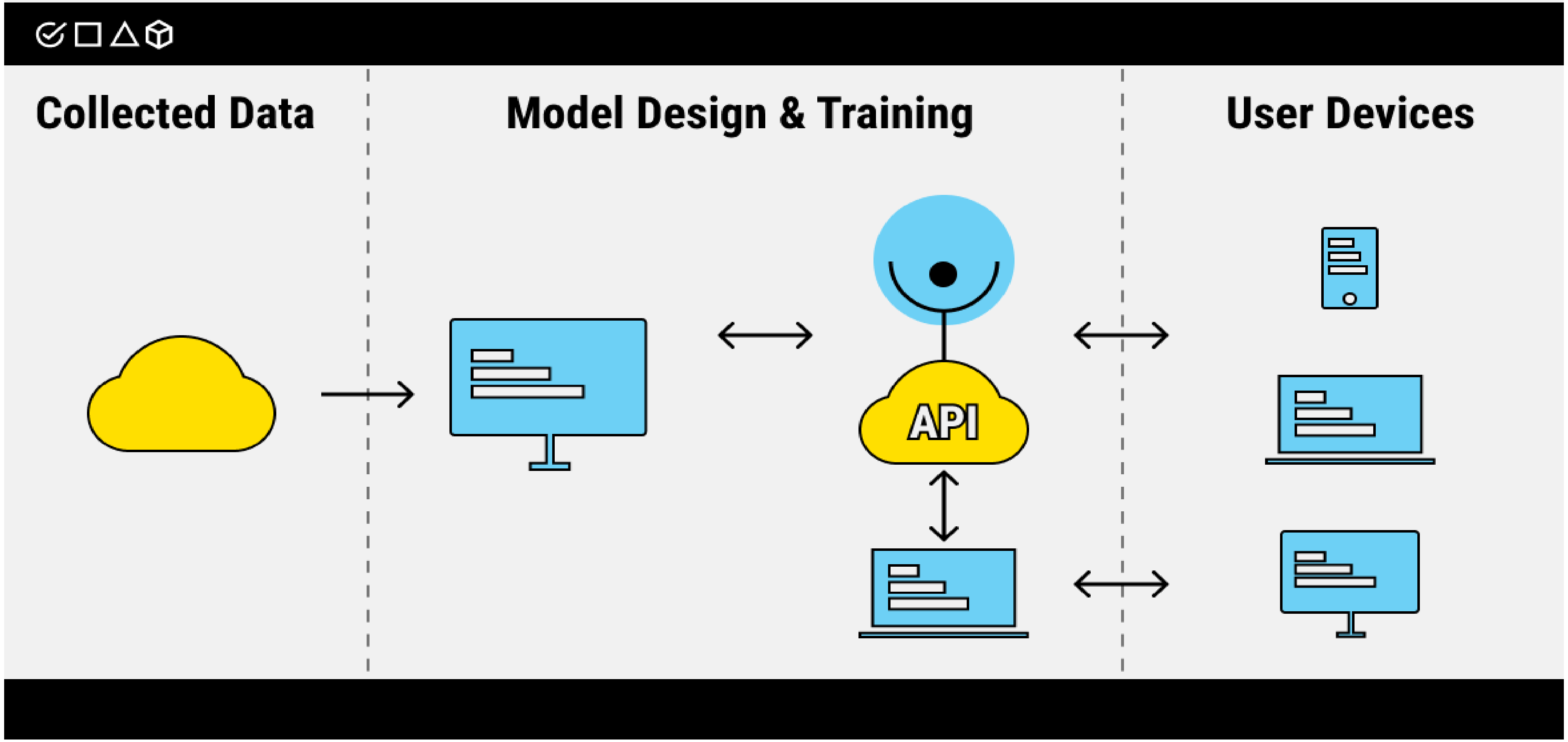
- Infering that a record was part of the training data

  -> An attacker can predict accurately based on that record

- In conformity with NIST an MIA is a confidentiality violation

- Companies that offers MLaaS can violate privacy regulations if MIAs can be executed

MLaaS



https://labelyourdata.com/articles/machine-learning-as-a-service-mlaas

Types of MIAs settings

Based on adversarial knowledge

- Two kinds of knowledge relevant for an attacker:
  - Knowledge of training data
  - Knowledge of target model

- Starting from the amount of information an attacker knows about the target model:
  - White-box Attack
  - Black-box Attack

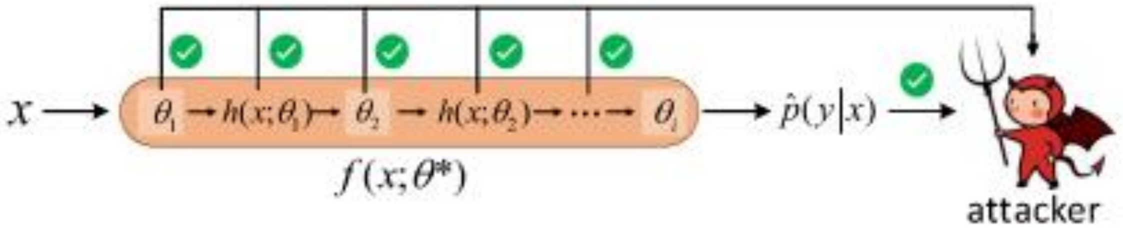


Fig. 2. Overview of white-box membership inference attacks.

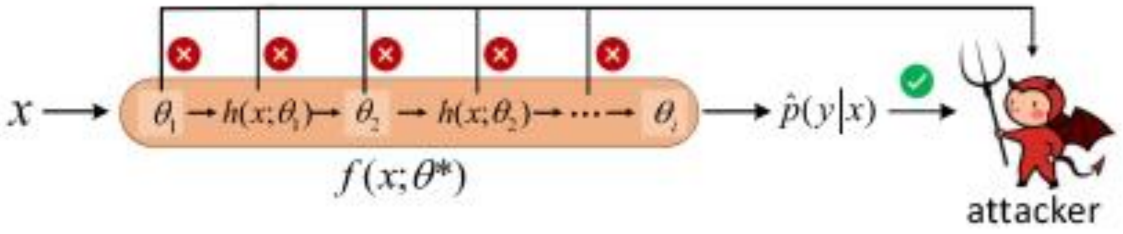


Fig. 3. Overview of black-box membership inference attacks.

Source: Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, *54*(11s), 1-37.

MIAs approaches

Approaches

- Are based upon the different behavior of a ML model on training data vs test data

- Metric Based MIAs

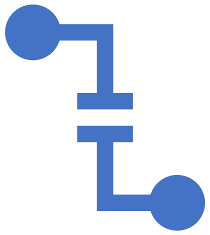
- Binary Classifier Based MIAs

Metric Based MIAs

- Compare calculated metrics with preset thresholds

- Four major types:
  - Prediction Correctness Based MIA
  - Prediction Loss Based MIA
  - Prediction Confidence Based MIA
  - Prediction Entropy Based MIA
    - Modified Prediction Entropy Based MIA

Prediction Correctness Based MIA



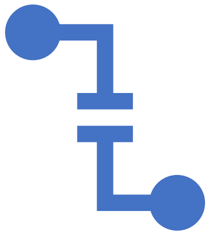
Hypothesis:

"An attacker infers an input record *x* as a member if it is correctly predicted by the target model, otherwise the attacker infers it as a non-member" [1]

Intuition:

ML models not generalize well

Prediction Loss Based MIA



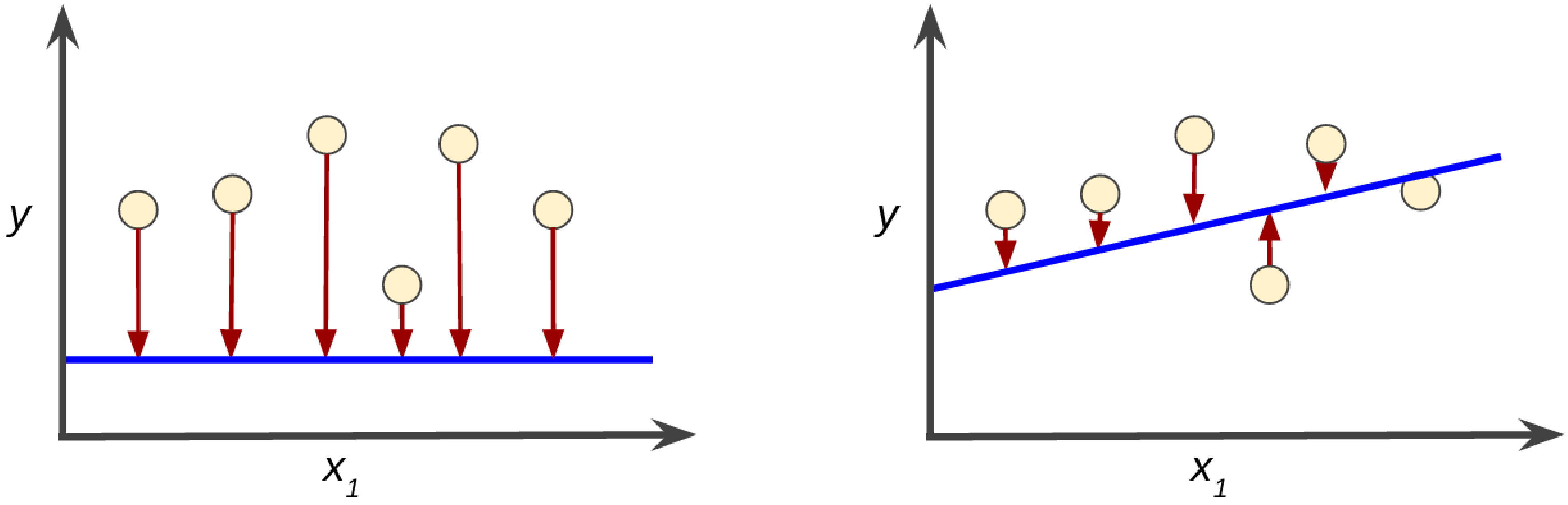
Hypothesis:

"An attacker infers an input record as a member if its prediction loss is smaller than the average loss of all training members, otherwise the attacker infers it as a non-member" [1]

Intuition:

ML model is trained to minimize the prediction loss of training data

Prediction Loss



https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss#:~:text=That%20is%2C%20loss%20is%20a,on%20average%2C%20across%20all%20examples.

Prediction Confidence Based MIA

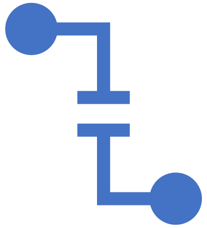
Hypothesis:

"An attacker infers an input record as a member if its maximum prediction confidence is larger than a preset threshold, otherwise the attacker infers it as a non-member" [1]

Intuition:

ML model is trained to minimize prediction loss for training data -> confidence score of a training member's prediction is close to 1

Prediction Entropy Based MIA



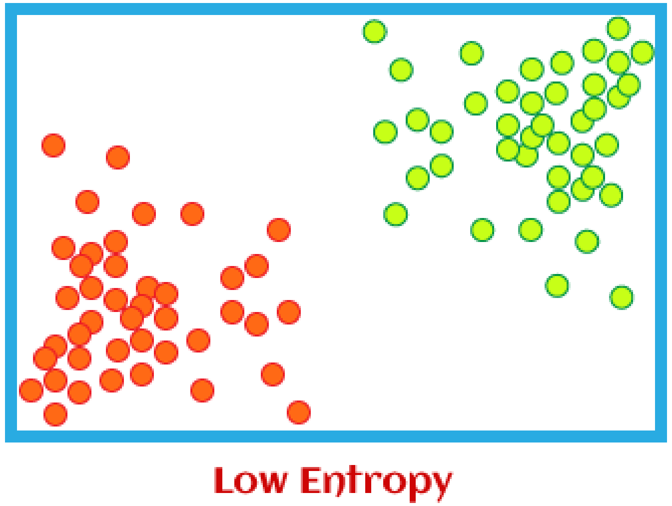
Hypothesis:

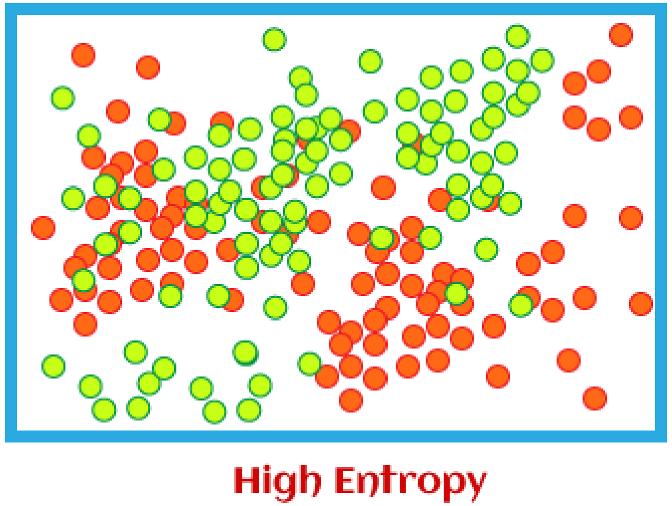
"An attacker infers an input record as a member if its prediction entropy is smaller than a preset threshold, otherwise the attacker infers it as a non-member" [1]

Intuition:

The prediction entropy of training data is smaller than the prediction entropy of test data

Entropy





- Expected value of surprise
- Measure of uncertainty of a variable
- The more uncertain, the higher the entropy

https://www.javatpoint.com/entropy-in-machine-learning

Why Modified Prediction Entropy Based MIA?

- A totally wrong classification with confidence score of 1 -> zero entropy -> member of training data

- Totally wrong classification -> highly likely a non-member

- We should take into account the ground truth label

Binary Classifier Based MIAs

- Needs to train an auxiliary ML model

- **Shadow training** proposed by Shokri et al. [2]
  - Multiple shadow models to mimic the target model
  - Shadow training datasets and test datasets disjoint from the target model's datasets
  - Used both in White-box Attacks and Black-box Attacks

Shadow Training Technique

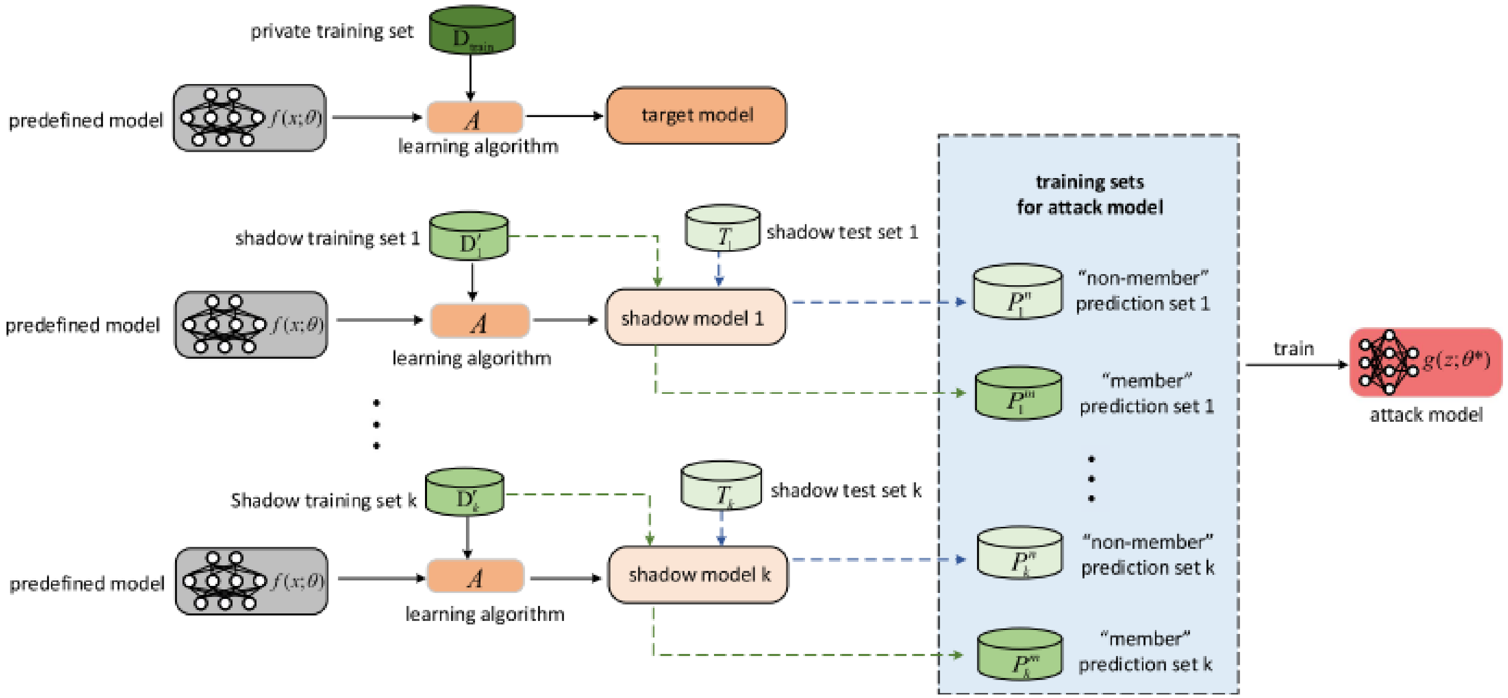


Fig. 4. Overview of the shadow training technique.

Source: Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, *54*(11s), 1-37.

White-box Setting vs Black-box Setting

(a) Binary classifier based black-box MIAs. (b) Binary classifier based white-box MIAs.

Fig. 5. Overview of binary classifier-based attack models in black-box and white-box settings. In the membership inference phase, the black-box attack model only takes the prediction vector $\hat{p}(y \mid x)$ as input and outputs the membership status of the data record. However, the white-box attack model can take the flat vector $\boldsymbol{v}$ containing much more information of the data record as input and outputs its membership status.

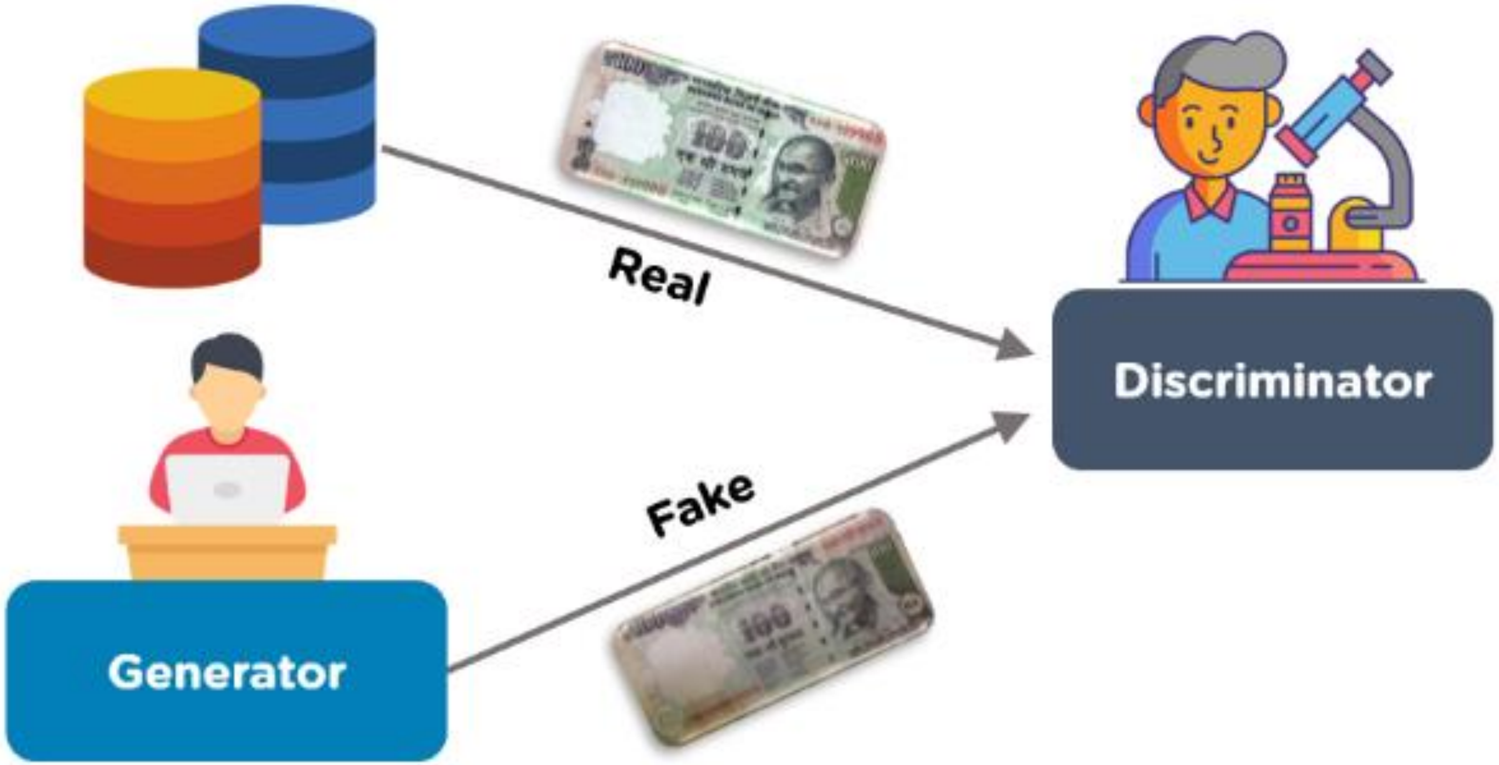
Source: Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, *54*(11s), 1-37.

MIAs on ML models

Current types of ML models attacked

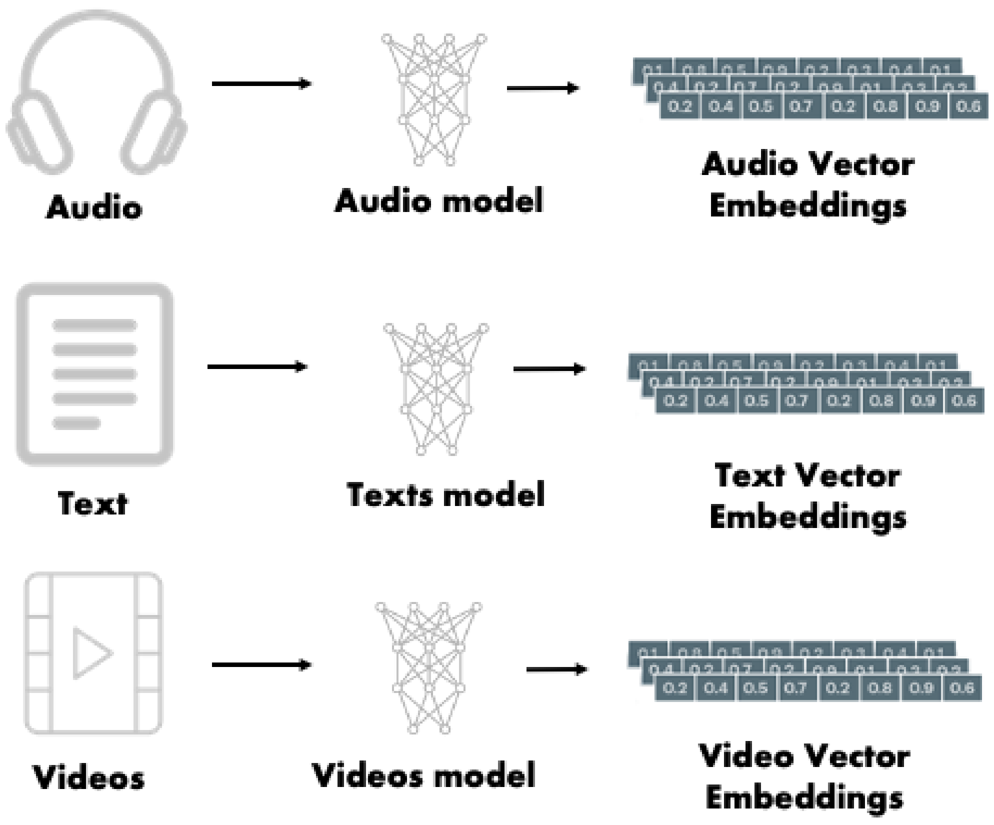
- **MIAs on classification models**
  - Main focus of research
- MIAs on generative models
  - GANs are the main target
- MIAs on embedding models
  - Both White-box and Black-box attacks
- MIAs on regression models
  - Only in White-box setting
- MIAs against Federated Learning

GANs (Generative Adversarial Networks)



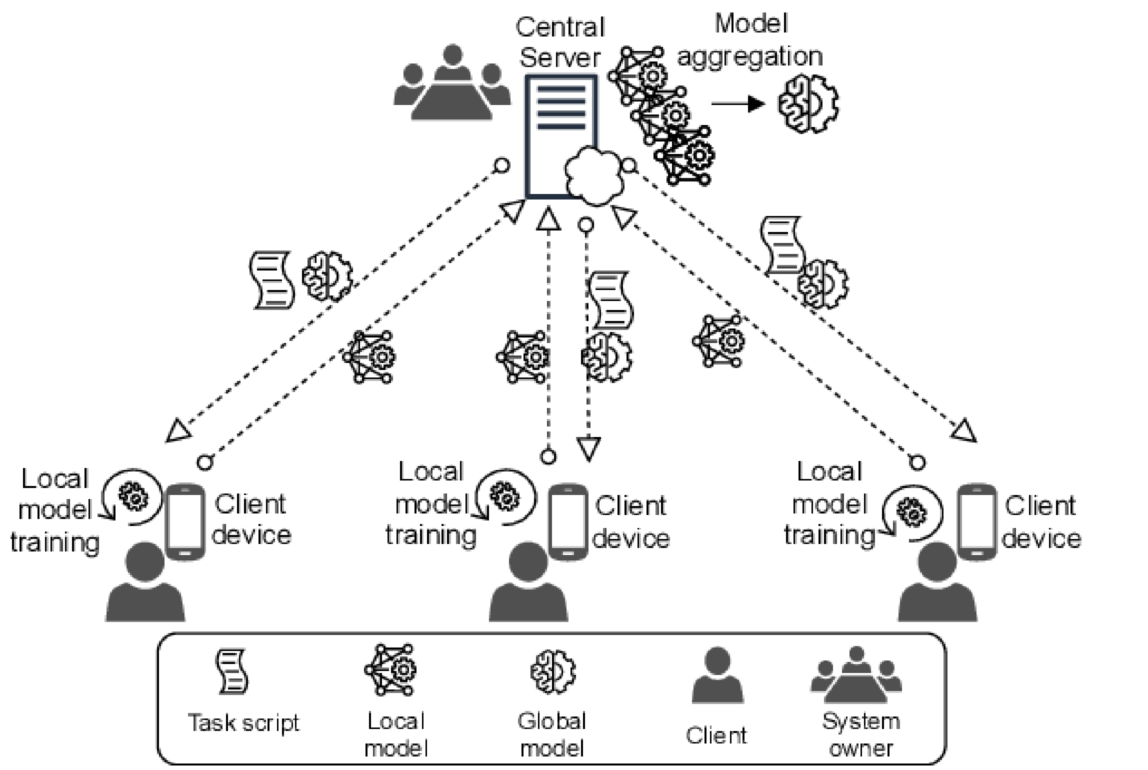
https://www.simplilearn.com/tutorials/deep-learning-tutorial/generative-adversarial-networks-gans#:~:text=GANs%20perform%20unsupervised%20learning%20tasks,the%20variations%20within%20a%20dataset.

Embedding Models



https://medium.com/@ryanntk/choosing-the-right-embedding-model-a-guide-for-llm-applications-7a60180d28e3

Federated Learning – Short Intro

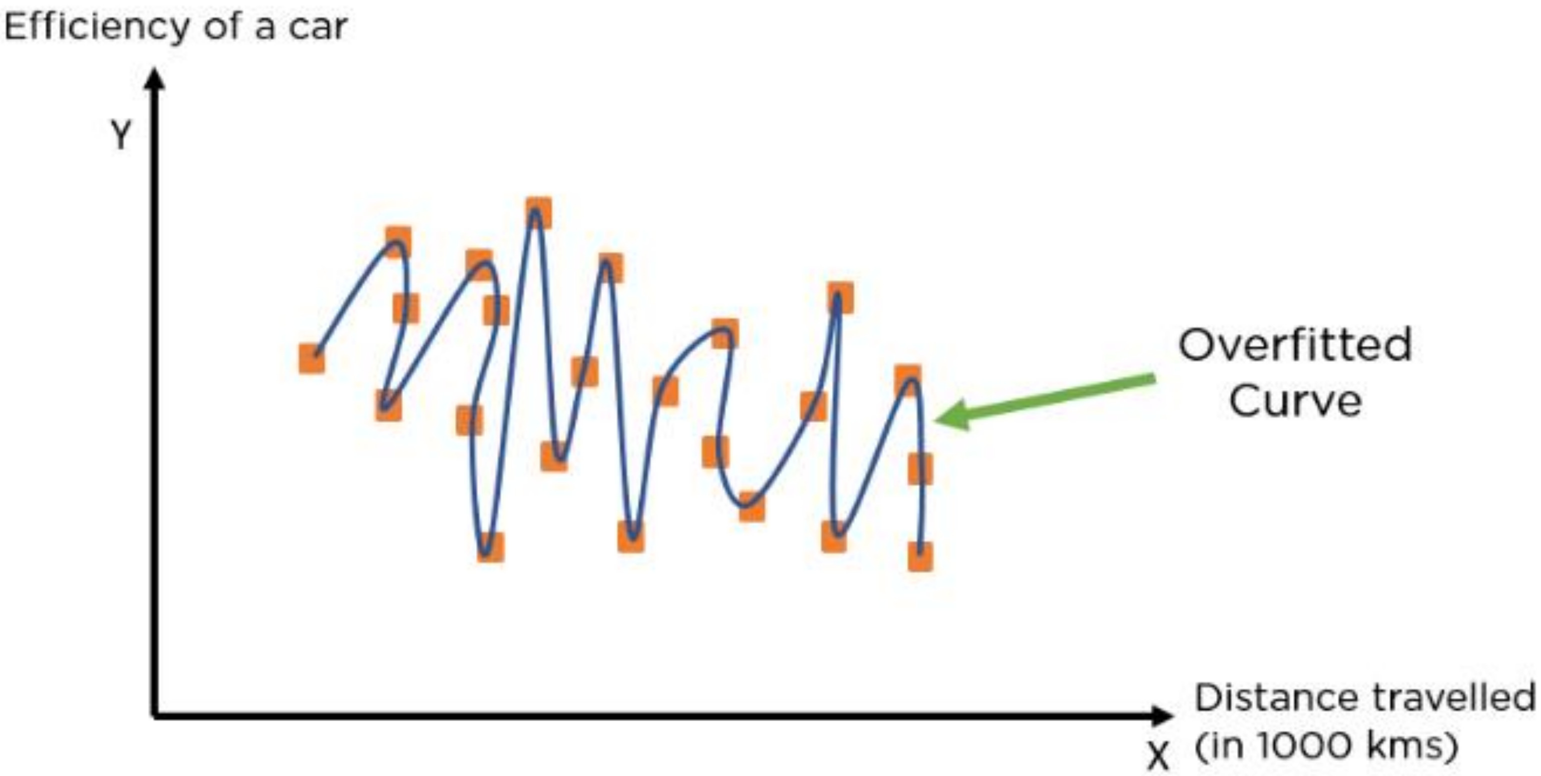


https://www.semanticscholar.org/paper/Architectural-Patterns-for-the-Design-of-Federated-Lo-Lu/60c4e1ff361c6c64b526edf3b281c78d941dbf1f

Why MIAs work?

Why MIAs work (1)

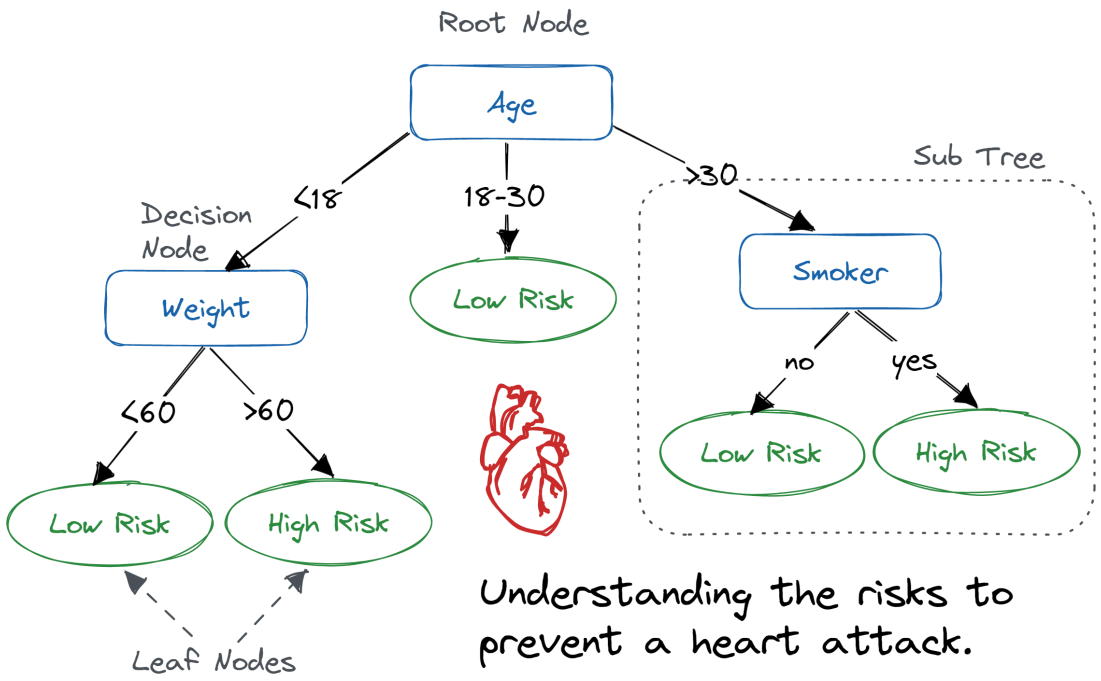
- Overfitting of Target Models

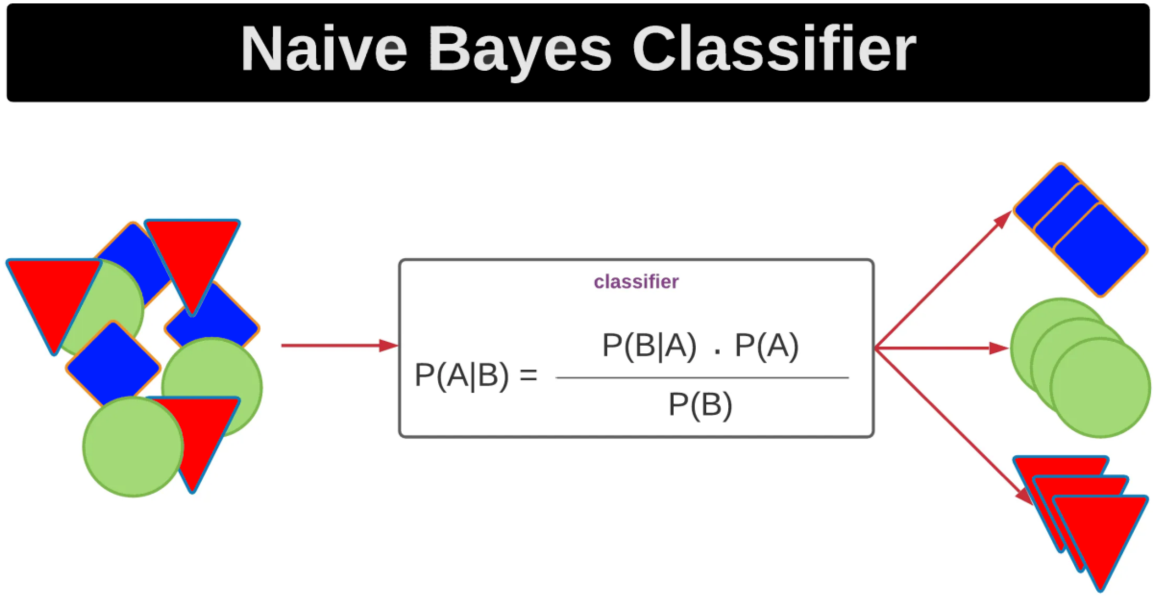


https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting

Why MIAs work (2)

Types of Target Models





https://www.datacamp.com/tutorial/decision-tree-classification-python

https://mlarchive.com/machine-learning/the-ultimate-guide-to-naive-bayes/

Why MIAs work (3)

- Diversity of Training Data

Defense against MIAs

Techniques of Defense

- Confidence Score Masking
- Regularization
- Knowledge Distillation
- Differential Privacy

Confidence Score Masking

- Used to mitigate MIAs on classification models
- Aims to hide the true confidence scores returned by the target model
- Two methods:
  - Top-K confidence scores
    - Often reduced to top three most likely classes for a record
  - Prediction label only
    - The attacker gets only the predicted label (class) for a record

MemGuard [3]

- Some crafted noise is added to the prediction vector
- The accuracy of the ML model is not impacted
- Still susceptible to metric based MIAs

Regularization (1)

- Aims to reduce the overfitting of the ML model
  - The ML model can generalize better -> Decreased generalization gap

- Classical regularization techniques:
  - L2-norm regularization
  - Dropout
  - Early stopping

Regularization (2)

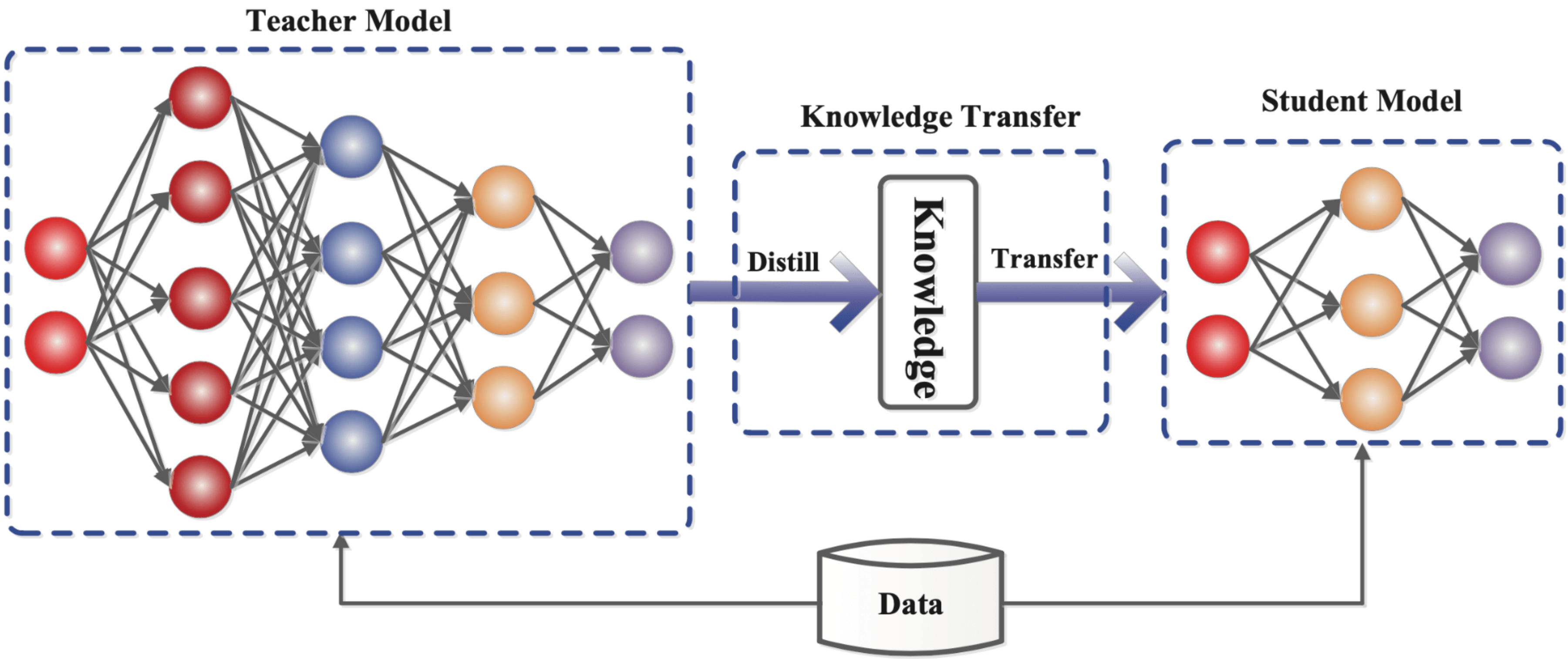
- Special regularization techniques to mitigate MIAs:
  - Adversarial Regularization [4]
    - Target Model is trained in a manner to preserve its prediction accuracy while reducing the attacker's performance
      - New regularization term -> Membership Inference gain of the attack model
  - Mixup + MMD [5]
    - Forces the ML classifier to generate similar output distribution for training data and test data
      - New regularization term -> Maximum Mean Discrepancy – distance between the output distributions of members and non-members

Regularization (3)

- Advantage:
  - Defense against MIA whether an attacker is in White-box or Black-box setting

- Drawback:
  - Privacy-Utility Tradeoff

Knowledge Distillation



https://neptune.ai/blog/knowledge-distillation

Distillation for Membership Privacy (DMP) [6]

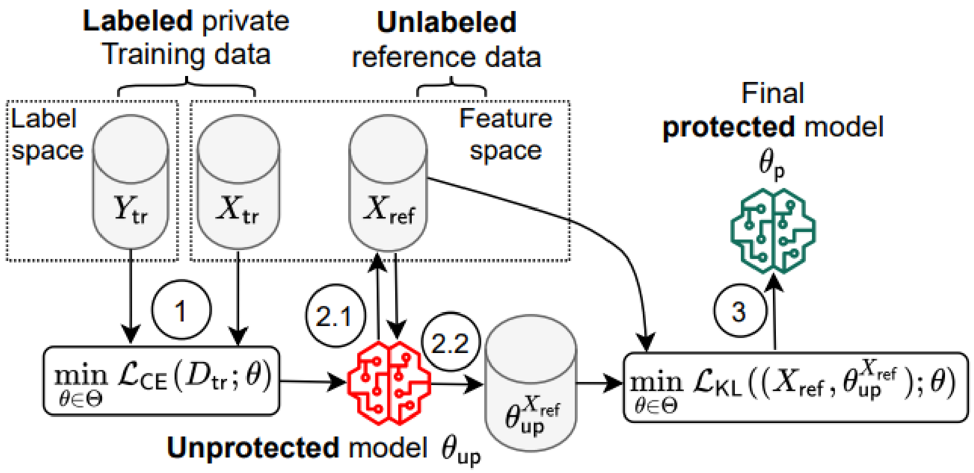


Figure 1: Distillation for Membership Privacy (DMP) defense. (1) In pre-distillation phase, DMP trains an unprotected model θ_{up} on the private training data without any privacy protection. (2.1) In distillation phase, DMP uses θ_{up} to select/generate appropriate reference data X_{ref} that minimizes membership privacy leakage. (2.2) Then, DMP transfers the knowledge of θ_{up} by computing predictions of θ_{up} on X_{ref}, denoted by $\theta_{up}^{X_{ref}}$. (3) In post-distillation phase, DMP trains the final protected model θ_p on $(X_{ref}, \theta_{up}^{X_{ref}})$.

Differential Privacy

- Advantages:
  - The ML model does not remember characteristics of its training data
  - Mitigates more types of attacks, not only MIAs
    - Attribute Inference Attacks
    - Property Inference Attacks
- Drawbacks:
  - Privacy-Utility Tradeoff
- Instead of using DP-SGD, a possible approach is DP-Logits [7]

Conclusions

Instead of conclusions (1)

- Research opportunities
  - Membership Inference Attacks:
    - On non-overfitted ML models
    - On transformers as Bert, T5
    - On heterogenous FL
    - In relation with Adversarial ML

Instead of conclusions (2)

- Membership Inference Defense:
  - Can obtain protection against MIAs only by offering Black-box access to attackers to a ML model trained in DP fashion (adding noise only to the model's output)?
  - FL combined with DP with a good Privacy-Utility tradeoff
  - Techniques to mitigate MIAs on Embedding Models

References

[1] Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, *54*(11s), 1-37.

[2] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.

[3] Jia, J., Salem, A., Backes, M., Zhang, Y., & Gong, N. Z. (2019, November). Memguard: Defending against black-box membership inference attacks via adversarial examples. In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security (pp. 259-274).

[4] Nasr, M., Shokri, R., & Houmansadr, A. (2018, October). Machine learning with membership privacy using adversarial regularization. In Proceedings of the 2018 ACM SIGSAC conference on computer and communications security (pp. 634-646).

References (2)

[5] Li, J., Li, N., & Ribeiro, B. (2021, April). Membership inference attacks and defenses in classification models. In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy (pp. 5-16).

[6] Shejwalkar, V., & Houmansadr, A. (2021, May). Membership privacy for machine learning models through knowledge transfer. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 11, pp. 9549-9557).

[7] Rahimian, S., Orekondy, T., & Fritz, M. (2020). Sampling attacks: Amplification of membership inference attacks by repeated queries. arXiv preprint arXiv:2009.00395.