Overfitting, Robustness, and Malicious Algorithms: A Study of Potential Causes of Privacy Risk in Machine Learning

Samuel Yeom a,∗ , Irene Giacomelli b,c , Alan Menaged a , Matt Fredrikson a and Somesh Jha

Presented by: Kyle Trevis

CS/ECE 599 | Winter 2022

PURPOSE

- Machine Learning emerging as a fundamental technology
- Used in applications with sensitive personal data
  - Healthcare and Health Analytics
  - Advertisement
  - Energy Usage
- Algorithms may leak information or be susceptible to targeted attacks

THREATS

- Membership Inference
- Attribute Inference
  - Cross Inference

- Overfitting
- Robustness

- Attacker assumed to have “black-box” access
- Focuses on vulnerability of algorithm, not the data set

MEMBERSHIP INFERENCE

- Inferring if a specific data point was included in training set
- Data is sampled from potential training points used in model to infer use in training
- Overfitting (generalization error) proportional to algorithm vulnerability
- Knowledge of Error Distribution also creates vulnerability
  - Often published with model
- Malicious Training Algorithms can enable membership advantage

ATTRIBUTE INFERENCE

- Inferring omitted features of an available data point
  - Points sampled from potential training points
  - Sensitive data is guessed
  - Projection of model output confirms

- Advantage scales with overfitting
- Knowledge of error distribution allows for more guided guesses

CONNECTION BETWEEN ADVANTAGES

- Attribute Advantage implies Membership Advantage
  - Attribute advantage at least as hard as membership advantage

- Membership Advantage may make Attribute attacks more effective and consentient

ROBUST MODELS

- "Robustness" used as a combat to integrity attacks
  - System is still functional after introduction of noise

- Membership Inference leverages robustness by abusing "robust generalization errors"
- Robust models are much easier to attack with Membership Inference

- "Shadow Models" also used for these types of attacks on robust models

ROBUST MODELS

- “Robustness” used as a combat to integrity attacks
  - System is still functional after introduction of noise
- Membership Inference leverages robustness by abusing “robust generalization errors”
- Robust models are much easier to attack with Membership Inference
- “Shadow Models” and “Attack Models” also used for these types of attacks on robust models

SUMMARY OF ANALYSIS

- Real Datasets were obtained and tested using previous attacks
- Reduction used for simpler computations

- Previous Analysis Confirmed
  - Generalization proportional to Membership and Inference Advantage
    - Generalization seemed to matter less for Membership Advantage
  - Robust Models are especially vulnerable to these attacks

RELATED WORKS

- Statistical Analysis of Privacy Vulnerability is abundant
  - Actual application to Machine Learning has gained traction more recently
- Membership Inference limited to “Shadow Models”
- Previous Attribute Inference excludes exposure of training data specifically
- Linking of Robustness to privacy vulnerability is a new concept

CONCLUSION AND IMPACT

- Membership and Attribute Attacks now formally defined for Machine Learning
  - Closely related attacks
- Overfitting is a major privacy vulnerability, but not the only vulnerability
  - Low Generalization Error still vulnerable to Membership Attacks
- Robustness creates vulnerability
  - Trade-off between security to integrity attacks and privacy attacks
- Malicious Training Algorithms also play a role in privacy

MY CONCLUSION

- This paper expands the discussion of privacy vulnerability in machine learning
- Results comprehensive
  - Statistical Analysis
  - Actual Experimentation
- Introduces trade-offs for system security

REFERENCES

Yeom, Samuel et al. "*Overfitting, Robustness, and Malicious Algorithms: A Study of Potential Causes of Privacy Risk in Machine Learning*". Journal of Computer Security 28.1 (2020): 35-70.

DISCUSSION POINTS

- Do you think Machine Learning Developers should prioritize Robustness or Privacy?
  - What will those in industry now will choose?
- How could robustness and privacy work together?
- Should Machine Learning Models be kept more secretive or more open?