Privacy Preservation in Machine Learning

Mitigation of Inversion and Inference Attacks

Mahdi Fazeli

November 3, 2023

Model Inversion Attacks

Understanding the Threat:

- Attackers use model predictions to infer sensitive information about the training data.
- Risk is heightened when models are overfitted, revealing too much detail in their predictions.

Example:

- A model trained to predict health conditions from patient records could potentially reveal a patient's health status if inverted.

Membership Inference Attacks

Understanding the Threat:

- Attackers determine if specific data was in the training set, potentially exposing sensitive information.
- Overfitted models are particularly vulnerable as they reflect the training data too closely.

Example:

- An attacker might discover that a particular individual's data was used in a financial model, implying their financial distress or wealth.

Our Proposal

Project Focus:

- Comparing the efficiency of different privacy-preserving techniques against Model Inversion and Membership Inference Attacks.

Research Methodology:

- Conducting a comprehensive literature review.
- Implementing and testing various privacy-preserving techniques.
- Assessing the trade-offs between privacy protection and model performance.

Expected Outcomes:

- A framework for evaluating privacy risks in machine learning models.
- A set of guidelines for implementing effective privacy-preserving techniques in various ML scenarios.