🔐 Synthetic Data & Privacy Risks

Understanding Adversarial Attacks & Protection Mechanisms

By: [Your Name] | [Your Organization]

🔎 Introduction

- • Synthetic data is used to protect privacy while preserving data utility.
- • Adversarial attacks can still extract sensitive information.
- • Key attacks analyzed: Membership Inference, GAN Inversion, Model Extraction.
- • Our goal: Evaluate privacy risks and implement defense mechanisms.

🛠️ Synthetic Data Generation

- • We use CTGAN (Conditional Tabular GAN) to generate synthetic data.
- • Data includes categorical and numerical features.
- • Differential privacy (Laplace noise) is applied for additional protection.
- • The goal is to generate data that cannot be reverse-engineered.

🚨 Privacy Attacks on Synthetic Data

- • Membership Inference Attack: Detects if a sample was in the original dataset.
- • GAN-Based Inversion Attack: Uses a GAN to reconstruct original data.
- • Model Extraction Attack: Adversary replicates the behavior of our model.
- • Differential Privacy Impact: Measures how much noise protects the data.

📊 Attack Results

- • Membership Attack Accuracy: **55%** (slightly better than random guessing).
- • GAN Reconstruction Similarity: **64,919** (high → strong privacy protection).
- • Differential Privacy Impact: **42,810** (high → strong noise effect).
- • Model Extraction Success: **100%** (critical security risk!).

⚠️ Risk Analysis

- • Model extraction is **a major concern** (attackers can steal our model).
- • Membership inference is **moderate risk** but still possible.
- • GAN-based inversion is **not effective** due to high similarity distance.
- • Differential privacy is **working well** but must be fine-tuned.

Mitigation Strategies

- • **Limit API access** to prevent model extraction.
- • **Increase differential privacy noise** for higher obfuscation.
- • **Use adversarial defenses** (differentially private training, query restrictions).
- • **Regularly test synthetic data** for privacy vulnerabilities.

✅ Conclusion & Next Steps

- • Synthetic data is effective but **not bulletproof** against attacks.
- • Model extraction is a **critical risk**—stronger protections needed.
- • Differential privacy is **helpful** but should be carefully optimized.
- • Future work: Exploring federated learning and adversarial defenses.

🙏 Thank You!

Questions? Let's discuss! 🚀