Proiect cercetare replicabilă

**Echipă**: Moldovan Daria, Popa Roxana 344C1

**Paper**: Systematic Evaluation of Privacy Risks of Machine Learning Models (ar merge un reference cu link aici, mai incolo tho)

**Rezumat**:

Scopul articolului este de a analiza eficiența a două metode populare de protecție (MemGuard și Adversarial regularization) împotriva atacurilor de tip membership inference bazate pe diferiți parametrii. Lucrarea pornește de la analiza descrisă anterior și, pe baza rezultatelor, sunt propuse metode de îmbunătățire a performanțelor. Îmbunătățirile sunt datorate în principal modificării modului în care sunt calculați parametrii care stau la baza atacurilor.

**Descriere**:

Experimentul original are la bază 4 atacuri de tipul membership inference, fiecare un parametru pe baza căruia se determină dacă o intrare din dataset face parte din setul de test sau cel de training. Scopul unui atac de tip membership inference este de a determina care date din dataset au făcut parte din setul de training și care au făcut parte din cel de test.

Acești parametrii sunt:

* Corectitudinea predicției
* Precizia
* Entropia
* Entropia modificată

1. **Membership inference bazat pe corectitudinea predicției**

Atacul se bazează pe corectitudinea predicției pentru a decide dacă o intrare din dataset se află în setul de training. Acest parametru este relevant deoarece, în general, un model prezice din ce în ce mai corect pe măsură ce parcurge setul de training, însă modul în care face predicții nu este întotdeauna potrivit pentru setul de training. Așadar, atacatorul va decide că datele fac parte din setul de training dacă predicția modelului este corectă.

1. **Membership inference bazat pe precizia predicției**

Elementul de bază al acestui atac este diferența dintre precizia modelului atunci când analizează date din setul de training și atunci când parcurge date din setul de test. De regulă, modelul va avea o precizie mai bună asupra setului de training decât asupra celui de test. Așadar, atacatorul va trage concluzia ca datele care au fost prezise cu o precizie mai mare fac parte din setul de training.

1. **Membership inference bazat pe entropie**

Entropia este metrica care determină nivelul de nesiguranță al unui model (procentul predicțiilor care au fost făcute aleatoriu, fără un discernământ anume). Asemănător atacurilor precedente, un model este antrenat să scada entropia în timp ce analizează setul de training, însă acest lucru nu se păstrează în timpul analizei setului de test. Așadar, atacatorul poate concluziona că datele a căror entropie este mai aproape de 0 fac parte din setul de training.

1. **Membership inference bazat pe entropia modificată**

Noutatea pe care o aduce articolul este o nouă metrică. Entropia tradițională nu este considerată suficientă pentru a trage concluzii satisfăcătoare. Motivul este faptul că entropia nu ține cont de valorile care sunt adevărate la bază (o predicție greșită cu probabilitatea 1 și o predicție greșită cu probabilitatea 1 duc la aceeași entropie). Așadar, a fost propusă o nouă funcție de calcul al entropiei care are următoarele proprietăți:

* Scade odată cu probabilitatea unei predicții corecte
* Crește odată cu probabilitatea unei predicții greșite

**Implementare**:

**Rezultate**:

**Concluzii**: