“Hariram, Veeramani”
Review for Paper Titled:

“ML-Leaks: Model and Data Dependent Membership Inference Attacks and Defenses on Machine Learning Models”

1- Summary

This work investigates the Membership Inference Attack Problem in MlaaS settings and gives an alarming result highlighting the broader applicability and scope of these attacks relaxing several Key assumptions in the previously reported Threat-Defense Models in this space.
The results give a compelling conclusion that even a typical low complexity Shadow model can cause Model, Data independent Membership Inference Attack since the number of shadow models does not necessarily increase the attack/recall precision and hence higher number of models does not increase the effectiveness of the attack.
This work expands the pool of Membership Inference classes by relaxing many key adversarial assumptions. It further evaluates the Membership Privacy threat using eight different datasets consequently stepping at a Model, Data independent adversary. It proposes Dropout and Model stacking as two key Defense Mechanisms and describe their effectiveness.

2- Novelty/Contribution

Defensive mechanisms discussed in this paper appear novel to me in the thorough way that they are developed. That is, the applicability of dropout when target model is a Neural Network vs the use of Model stacking which works independent of the target Model’s ML classifier seems interesting. I also find the equal emphasis given to precision and recall metrics throughout the evaluation interesting as we might prefer one metric to the other metric depending upon the context of Models and application. Though, Data Independent Threat model approaches were reported before, previous works focused more mainly only on the ML model attacks with Binary Features, whereas the current work expands the attack for ML models trained on any kind of data not necessarily Binary Features.

3- Evaluation

The work thoroughly evaluates the scenarios of Model, Data Independent Attacks and reveals Key Insights. The First Adversary- Model Independent Attack, suggests that presence of more classes, more epochs in training, more overfitting in target model imply better Attack Performance and more vulnerability which sound intuitive. Analysis is exhaustive both in the aspects of Domain (Use of Different Domain Datasets/Models --point to similar conclusions) and remote vs local (MlaaS vs Local Computer Iterations) give similar Attack Performances. Second adversary model inflicting Data Distribution Independent Attack reasoned through the clustering by usage of 2D space using t-Distributed Stochastic Neighbor Embedding (t-SNE) gives a clear visualization tool and is also very convincing.

4- Unique Strength/Weakness

Overall, the work reports profound results with convincing evaluation and conclusions. Though we had few conflicting results while applying techniques to different Datasets- the offered Justifications give a thorough reasoning behind.