mplementation and Evaluation of Membership Inference Attacks Against Machine Learning Models Trained On Distributed Private Datasets

Supervisor: Saba Amiri (s.amiri@uva.nl)

Implementation and Evaluation of Membership Inference Attacks Against Machine Learning Models Trained On Distributed Private Datasets

In recent years, the utilization of machine learning-based methods has been rapidly growing in a wide range of applications. For these machine learning-based systems to be effective and widely used, we need massive amounts of data to train and evaluate them on, which will inevitably include and require private data. To address privacy concerns related to the data and adhere to privacy regulations regimes such as the European GDPR, we need secure and private machine learning methods.
There are different levels of threats to mitigate to reach private, secure machine learning. These levels include model and data privacy and security. Preserving the privacy of the data is particularly important in specific use-cases, e.g. healthcare and industrial collaborations, especially in a distributed environment where different parts of the data reside in different locations, each with different owners and their own set of regulations and privacy constraints. There are a number of ways to attain data privacy in machine learning in such a distributed heavily-regulated environment, all studied in the “Privacy Preserving Machine Learning” area of research. In recent years, with the growing interest in private and secure AI, there have also been a number of frameworks implemented to perform distributed privacy-preserving machine learning. One such framework is PySyft [1].
As there are ways to achieve data privacy, there inevitably also exist machine learning-based methods to attack the private models and try and compromise the privacy and security of the input data [2]. These types of attacks include “Model Inversion Attacks”, “Reconstruction Attacks” and “Membership Inference” attacks.
The aim of this project is to “implement” and “evaluate” Membership Inference Attacks against distributed machine learning models trained on multiple private datasets. The “implementation” phase will be done in the form of extending the PySyft framework with. Different types of Membership Inference Attacks. The implemented methods should be seamlessly integrated in the PySyft framework according to its collaboration rules. The “evaluation” phase will be done by training a distributed learning model- using assets already implemented and available in the PySyft framework – on publicly available datasets and then performing the newly implemented attacks against these trained models and report the results and the degree of privacy of the model(s) using established privacy metrics.
This project will be in close relation to the projects “Model Inversion Attacks Against Machine Learning Models Trained On Distributed Private Datasets” and “Reconstruction Attacks Against Machine Learning Models Trained On Distributed Private Datasets”. If possible, the students working on these projects will closely collaborate, both on the “implementation” and the “analysis” phases.

References

[1] https://github.com/OpenMined/PySyft
[2] Dwork, C., Smith, A., Steinke, T. and Ullman, J., 2017. Exposed! a survey of attacks on private data. Annual Review of Statistics and Its Application, 4, pp.61-84.

Project offered to study/studies: Bachelor Kunstmatige Intelligentie (artificial intelligence)

Max number of students: 1