PPML: Machine learning on data you cannot see

Valerio Maggio

Anaconda, inc.

Privacy guarantee is **the** most crucial requirement when it comes to analyse sensitive data. In fact, sensitive data could not be shared nor moved from their silos, let alone analysed in their raw form. As a result, data anonymization techniques are used to generate a sanitised version of the original data. These techniques are valuable tools to allow sensitive data to be used by Machine Learning (ML) algorithms, but these methods alone are not enough to guarantee complete privacy protection [6]. Moreover, multiple studies have demonstrated that ML algorithms trained on private data suffer from a persistent vulnerability that can unintentionally expose information about training samples [1] [3] [5]. This is particularly the case of Deep Neural Networks due to a hard-to-avoid memoization effect in their internal parameters [1].

Differential privacy (DP) [2] is a system for publicly sharing information about a dataset by describing the patterns of groups within the data, while withholding information about individuals. This technique has recently attracted increasing interest from the ML community, as a method to quantify the anonymization of sensitive data during training [3] [2]. Moreover, DP integrates seamlessly into the whole process, with no direct effect on its reproducibility.

In this talk, we will discuss how DP methods can be effectively used for *Privacy Preserving* Machine learning. We will introduce the main theoretical foundations of DP that are relevant for ML analyses. Afterwards, we will demonstrate how DL models could be exploited [4] (i.e. *inference attack*) to reconstruct original training data by solely analysing models predictions, and how DP can help to protect the privacy of our model, with minimal disruption to the original training pipeline. Final remarks on more complex ML training and inference scenarios will be examined, considering specialised distributed federated learning strategies.

References

1. Nicholas Carlini, Chang Liu, Jernej Kos, Úlfar Erlingsson, and Dawn Song. The secret sharer: Measuring unintended neural network memorization & extracting secrets. *CoRR*, abs/1802.08232, 2018.
2. Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
3. Vitaly Feldman. Does learning require memorization? A short tale about a long tail. *CoRR*, abs/1906.05271, 2019.
4. Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*,

CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery.
5. H. Brendan McMahan and Galen Andrew. A general approach to adding differential privacy to iterative training procedures. *CoRR*, abs/1812.06210, 2018.
6. Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *CoRR*, abs/cs/0610105, 2006.