Decentralized Algorithms for Privacy-Preserving Linear Regression

Aurélien Bellet, Jan Ramon

September 25, 2017

Team and contact

- Équipe Magnet, INRIA/CRIStAL: http://team.inria.fr/magnet
- Aurélien Bellet (aurelien.bellet@inria.fr), Jan Ramon (jan.ramon@inria.fr)

Keywords

Machine Learning, Decentralized Algorithms, Privacy, Linear Regression.

Context

Increasing amounts of data are being produced by interconnected devices such as mobile phones, connected objects, sensors, *etc.* For instance, history logs are generated when a smartphone user browses the web, rates products and executes various applications. The currently dominant approach to extract useful information from such data is to collect all users' personal data on a server (or a tightly coupled system hosted in a data center) and apply centralized machine learning and data mining techniques. However, this centralization poses important privacy issues in applications involving sensitive data such as medical records or geolocation logs.

In this internship, we are interested in the alternative setting of *decentralized machine learning*, where learning agents do not share their personal data but collaborate among themselves to learn the model, without any central entity required for coordination or aggregation. Many decentralized algorithms exist to efficiently compute relevant quantities in a peer-to-peer fashion: see for instance [1, 3] for computing averages and [8, 6, 4] for optimizing sums of convex functions. However, most approaches can still leak (directly or indirectly) some sensitive information about the local datasets of the agents involved in the computation.

Objectives

The goal of this internship is to propose and analyze decentralized and privacy-preserving algorithms for the task of linear regression. A promising direction is to rely on the closed-form solution of linear regression, which consists of several data-dependent averages. We will thus study how to extend a recently proposed decentralized algorithm for privately computing averages [5] to efficiently train a linear regression model. Some alternative approaches will also be considered: for instance, one may try to extend existing approaches for private linear regression [2, 7] to the decentralized setting. We will compare the relative merits and drawbacks of these approaches in terms of computational cost, accuracy and privacy guarantees.

The tentative work plan is as follows:

1. Review the relevant literature on decentralized learning and privacy-preserving protocols.

2. Propose, analyze and evaluate some algorithms for decentralized privacy-preserving linear regression.

3. If time permits, investigate some further questions, such as (i) how to guard against the presence of malicious users in the network, (ii) how to extend the results to nonlinear regression, and (iii) how to efficiently implement the proposed methods.

Skills

Basics in machine learning, algorithms and complexity, linear algebra and probability.

References

[1] S. P. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

[2] K. Chaudhuri, C. Monteleoni, and C. Monteleoni. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[3] I. Colin, A. Bellet, J. Salmon, and S. Clémençon. Extending Gossip Algorithms to Distributed Estimation of U-statistics. In *Advances in Neural Information Processing Systems 29*, 2015.

[4] I. Colin, A. Bellet, J. Salmon, and S. Clémençon. Gossip Dual Averaging for Decentralized Optimization of Pairwise Functions. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[5] P. Dellenbach, J. Ramon, and A. Bellet. A decentralized and robust protocol for private averaging over highly distributed data. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[6] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

[7] A. Gascón, P. Schoppmann, B. Balle, M. Raykova, J. Doerner, S. Zahur, and D. Evans. Privacy-Preserving Distributed Linear Regression on High-Dimensional Data. In *Proceedings on Privacy Enhancing Technologies (PETS)*, 2017.

[8] A. Nedic and A. E. Ozdaglar. Distributed Subgradient Methods for Multi-Agent Optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.