Internship/PhD offer

Bias of Federate Learning with Heterogeneous and Missing Data

Project Description:
Combining data from multiple sources is becoming a central issue of modern Machine Learning applications. This problem characterises a wide range of analysis scenarios: from analytics in mobile communications and IoT, where cameras, microphones, and GPS data from different users are integrated to develop predictive systems, to healthcare, where patients records are combined to develop novel diagnostic tools. The sensitive nature of this information often implies that data cannot be directly shared and centralized in a single data-center. For this reason, federated learning strategies are gaining increasing attention for combining models trained on private and secured information, without the need to centrally store the data.

In spite of recent advances in the field of federated learning, the integration of models representing independent data presents important statistical challenges. In particular, model_s prediction and parameters may sensitively differ when trained on different data instances.This is a very common weakness in machine learning, related to:

1) Model variability, i.e. the inconsistency of the model's behaviour under different data instances. This issue is symptomatic of overfitting, and requires the development of parsimonious data representations across sources.
2) Data bias. The over-representation of specific traits in a data instance (i.e. non iid distribution of the data across instances) may likely influence the outcome of a model, and thus introduce a bias reflecting the specificity of the training data. This is a frequent problem, especially when working with unbalanced data, or with non-random missing observations.

These issues may have important adverse effects in federated learning, as they may expose the system to failures in preserving data anonymity, as well as to lack in the robustness of the results. To tackle these problems, the goal of this project is to develop:

1) A principled theoretical framework to define bias in federated learning systems,
2) Analysis systems to automatically identify the presence of bias in federated analysis
3) Strategies for bias removal, to lead to robust federated learning strategies

Hosting groups:
Epione and NEO teams (Inria Sophia Antipolis) - Accenture Labs (Sophia antipolis). The groups are located in the tech Park of Sophia Antipolis and in Nice, in the French Riviera.

During the project the candidate will:

- Develop learning methods for federated analysis for private and distributed data;
- Develop a formalism for bias identification and correction in federated learning;
- Gather knowledge in advanced statistical learning methods - Bayesian learning, Kernel methods, non-parametric learning, variational inference -;
- Develop and deploy algorithms in several context, with special focus in biomedical and clinical application;
- Participate to the activity of Accenture Labs, interact with the research and engineering personnel;
- Interact with INRIA students and researchers, and participate to scientific life of the team.

Required competences:
Competences in statistics, optimization, and mathematical modeling are essential (Master 2 level). Knowledge in signal processing desired. Solid programming and IT skills are necessary (Python, bash, version control systems), along with strong communication abilities.

Contact:
marco.lorenzi@inria.fr; giovanni.neglia@inria.fr;
laetitia.kameni@accenture.com; richard.vidal@accenture.com

References:
- Santiago Silva, Boris Gutman, Barbara Bardoni, Paul M Thompson, Andre Altmann, Marco Lorenzi. *Multivariate Learning in Distributed Biomedical Databases: Meta-analysis of Large-scale Brain Imaging Data*. IEEE International Symposium on Biomedical Imaging (ISBI), Venice, 2019.
- Samuel M. Gross Robert Tibshirani. *Collaborative regression*. Biostatistics, Volume 16, Issue 2, 1 April 2015, Pages 326-338,
- Jakub Konecn, H. Brendan McMahan, Daniel Ramage, Peter Richt_rik. *Federated Optimization: Distributed Machine Learning for On-Device Intelligence*. arXiv:1610.02527
- Christian Wachinger, Benjamin Gutierrez Becker, Anna Rieckmann. *Detect, Quantify, and Incorporate Dataset Bias: A Neuroimaging Analysis on 12,207 Individuals*. arXiv:1804.10764
- H. Brendan McMahan, Eider Moore, Daniel Ramage, et al. *Communication-efficient learning of deep networks from decentralized data*. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017.
- Keith Bonawitz, Vladimir Ivanov, et al. *Practical Secure Aggregation for Privacy-Preserving Machine Learning*. ACM SIGSAC Conference on Computer and Communications Security, 2017.