Internship Proposal (Research): "Data selection and Shapley value in the linear case"

Keywords: data market, data Shapley, linear regression, fairness

Host team: FairPlay team (Inria Saclay)

Supervisors:
Patrick Loiseau – https://patrickloiseau.github.io/ – patrick.loiseau@inria.fr

Background

A classical problem in machine learning is to select a training set to learn a model (to be then used for a particular prediction task), subject to various constraints (e.g., budget constraints). This is the case for instance when the analyst buys data at a data market and needs to select which data to buy [2]. A standard approach for this problem is to use optimal experimental design, that is a method that solves an optimization problem with some metric of quality of the model learned [2].

In recent years, however, data selection has also been studied as a standard example of use of the data Shapley value [1, 4]. Data Shapley assigns some "value" to each data point based on some axioms from game theory, and one can then simply take the data points with the highest Shapley value. While there is some empirical work that illustrates the performance of this approach is certain settings, there is currently only little understanding of the actual connection between Shapley value and optimal experimental design. This is especially important given the increasing importance of data markets for training machine learning models.

Goal of the internship

The goal of the internship is to better understand how to perform data selection in different contexts and using different methods. For that, we will focus in particular on the case of heteroskedastic linear models, which are amendable to theoretical analysis. We propose in particular the following steps:

1. First, we will perform a theoretical analysis of data Shapley in a simple case (e.g., with a Gaussian design). We will compute the distribution of the data Shapley value as a function of the noise variance. This will allow us to compare theoretically the result of performing data selection with data Shapley and with optimal experimental design.

2. Second, we will analyze data selection under fairness constraints. Indeed, in many cases, one has constraints that the prediction of the learned model must be fair—that is, its performance must be similar across different demographic groups. We will study how to perform data selection with this constraint. A starting point for this is to use a mixture model (similar to [3]) to account for the different models that represent each group.

Depending on time, we may also extend the study to related problems with more realistic models, for instances extending linear models.

Expected ability of the student and practical information

A strong mathematical background is necessary, in particular in probability. The internship will take place in the FairPlay team, a joint team between Inria, Criteo and ENSAE. The team is hosted at CREST (in the ENSAE building) and works from the Criteo offices in Paris every friday. It may be continued as a PhD. For more information, please contact patrick.loiseau@inria.fr.

References

[1] K. Jiang, W. Liang, J. Zou, and Y. Kwon. "OpenDataVal: a Unified Benchmark for Data Valuation". In: *NeurIPS*. 2023.

[2] C. Lu, B. Huang, S. P. Karimireddy, P. Vepakomma, M. I. Jordan, and R. Raskar. "DAVED: Data Acquisition via Experimental Design for Data Markets". In: *NeurIPS*. 2024.

[3] O. Marfoq, G. Neglia, A. Bellet, L. Kameni, and R. Vidal. "Federated Multi-Task Learning under a Mixture of Distributions". In: *NeurIPS*. 2022.

[4] J. T. Wang, T. Yang, J. Zou, Y. Kwon, and R. Jia. "Rethinking Data Shapley for Data Selection Tasks: Misleads and Merits". In: *ICML*. 2024.