Master Internship 2020

# The Game of Privacy

Supervisors: Michèle Sebag, Armand Lacombe
Michele.Sebag@lri.fr, Armand.Lacombe@polytechnique.edu
Lab: TAU − CNRS − INRIA − LISN, U. Paris-Saclay

Abstract :
In the early days of the Covid pandemy, we faced a dilemma: many machine learning people wanted to contribute to the research effort in modeling the risk of complications depending on the covariates (age, gender, comorbidities); but hospitals could not disclose their data to unknown MLers for privacy reasons. (Naturally, this dilemma is general, and observed in many other contexts involving sensitive data). After some discussions, it was found possible to disclose the marginals of the data, i.e. the univariate distribution of each feature, and the risk of complications conditionally to each modality of a feature (e.g. $P(death|gender = male)$ or $P(death|60 \leq age < 70)$).

The question is to see what can be learned under such severe limitations. A first answer is based on domain adaptation [1], that is, exploiting some related but different source data to facilitate learning and reasoning on the target domain. Uri Shalit [2] showed how to adapt private data reporting the risk of complications for flu in order to fit the above marginals, and use the adapted data to learn and deliver an appropriate Covid model.

We proposed another approach, *Extremely private learning*, based on a dialog with the hospital [3]: the learner iteratively delivers a hypothesis $h_t$ (learned based on some source data) and receives the marginals of the error of this $h_t$ on the actual hospital data. Based on these marginals, another hypothesis $h_{t+1}$ is learned and the mechanism is iterated. The epsilon-differential privacy of the approach [4] is enforced by adding some noise to the marginals provided by the hospital.

The internship :
The goal is to revisit the dialog between the learner and the oracle (here, the hospital) as a game: the learner wants to learn a model as accurate as possible; the oracle does not want the learner to break the privacy and be able to identify the joint distribution of the data.

More specifically, let $h_t$ be the current hypothesis learned by the learner. The learning player wants to ask the most informative query about $h_t$ (what is the error of $h_t$ in a given region of the search space) in order to improve $h_t$. Meanwhile, the oracle player adds some (minimal) noise to its answer in order to prevent the learner from inferring a too precise model of the joint distribution. Possibly, the oracle will see how precise can a joint distribution estimate be built based on the previous queries.

The student will investigate how to formalize and tackle such a game. A possible approach is based on Monte-Carlo Tree Search [5].

The internship requires excellent creativity, plus theoretical and algorithmic skills (programming environment in Python).

## Références

[1] Domain-Adversarial Training of Neural Networks, Y. Ganin et al., JMLR 2016
[2] Extremely private domain adaptation, Uri Shalit et al, 2020;
https://www.youtube.com/watch?v=XdFnYljn4rU
[3] Extremely private supervised learning, Armand Lacombe et al., 2021
[4] Calibrating noise to sensitivity in private data anal- ysis, Cynthia Dwork et al., 2006.
[5] The Grand Challenge of Computer Go: Monte Carlo Tree Search and Extensions, Sylvain Gelly 2012.