Internship project description
A pilot study for federated learning on oncological data

Jan Ramon

10/11/2022

1 Motivation and context

Over the last decades, there has been an increasing interest in exploiting data. On the other hand, recently there has also been an increasing awareness of the risks of collecting sensitive data centrally, given the frequency of data leaks, hacking or abuse. The Horizon Europe projects TRUMPET and FLUTE will work towards a platform for secure privacy-preserving federated machine learning where the sensitive data remains with the data owners, and machine learning is performed collaboratively by these data owners by participating in collaborative algorithms which (through the use of differential privacy and/or encryption) generate the desired statistical models but prevents sensitive data from being revealed.

These projects also feature use cases in medicine, in particular the TRUMPET project will study lung cancer clustering and eligibility prediction for radiotherapy for head and neck cancer patients, while the FLUTE project will study prediction and diagnosis of prostate cancer.

Before tackling these use cases with federated learning, this internship will conduct a first, short exploratory study to understand how these medical machine learning problems could be solved using machine learning in a simpler setting with central data.

2 Objectives

The goal of this internship project is to find an adequate machine learning approach to at least one of the TRUMPET use cases.

In particular, the objectives are

- to study the literature on the specific machine learning task at hand
- to make a description of the format and structure of the available data, the clinical objectives and the relevant background knowledge

- to shortlist a few adequate machine learning strategies and test them on public or synthetic data in similar conditions as the real data which will be used later.
- to select a best alternative and bring the concerned algorithm in a format that can be implemented in a federated learning framework.

3 Plan

Here is a tentative work plan:

- Understanding the data and the clinical problem. Interacting with medical experts and data experts and systematize the knowledge relevant for the project (4 weeks)
- Machine learning literature study, in particular getting familiar with (1) privacy-preserving machine learning, and (2) machine learning strategies we want to consider (3 weeks)
- Preparing public / synthetic data into realistic conditions (2 weeks)
- Implementing/using a selection of algorithms on this benchmark (4 weeks)
- Selecting the eventual method to be used, and bringing the algorithm into a shape suitable for federated learning. This includes organizing the data flows and analyzing the privacy of the algorithm (4 weeks)
- Completion of the internship report (2 weeks)

The timing can be adapted according to the personal preferences of the student or the requirements of his school.

4 Environment

The project will be conducted in the INRIA MAGNET team. The student will collaborate and interact with various other collaborators in the team. It is possible the student will be asked to present progress to the team or to partners in ongoing projects. An application for ZRR access will need to be made to the FSD.

5 Requirements

We expect the student has a good understanding of basic computer science concepts, e.g., data structures and algorithms, and of statistics / machine learning Knowledge of Python is required.