Supervised machine learning in privacy-preserving decentralized environments

Predictive data analysis using supervised machine learning has contributed significantly in various disciplines ranging from medicine and drug discovery to supply chain management and information technology. The basic idea is simple, it involves leveraging historical data to train machine learning models and later use the trained models to effectively predict different system behaviours. Depending on the application's requirements, a range of different algorithms is available to understand both time-dependent behaviours and identify complex hidden correlations between different features. In a conventional machine learning approach, the process of model training starts after collecting a complete dataset at one place.

However, due to the privacy concerns and very large datasets, it is impractical to collect all the data in one place. These limitations hinder the required training process to develop effective and highly accurate models in production. To address these challenges, we have designed an offline decentralized approach for machine learning model training. The aim is to provide an alternative approach to the centralized model training process which requires all data to be available at one site. The proposed approach allows machine learning models to travel towards the available data, capture the insights (update weights according to the dataset), move to the next site that holds another set of data and continue the process till the last available dataset. The offline approach does not require all the participating sites to be simultaneously available during the training phase. The proposed approach offers privacy-preserved settings where data owners are not required to pool their datasets in one place. The proposed approach is promising, but there are numbers of open questions related to this approach and in this project, the task is to find answers to the fundamental questions that can affect the training process and result as a biased model. Following are the three selected questions:

1 - How do unbalanced datasets affect overall model training in a distributed environment?
2 - How IID and non-IID datasets affect model training in a distributed environment?
3 - How can we retain a balance between model training based on old datasets and new datasets?

The presented approach is similar to the concept of federated machine learning but we would like to investigate a loosely coupled model where all participating sites are not required to be online simultaneously. The project has similarities with the study presented in [1]. For this project, we will use the datasets based on data centre resource utilization. We have two real datasets, first from the UPPMAX data centre (Swedish academic service provider) and second from the CSC data centre (Finnish commercial service provider). Your task will be to simulate a federated environment, trained neural network models based on distributed settings and find the answers to the above mentioned questions.

Supervisors:

Salman Toor salman.toor@it.uu.se
Prashant Singh prashant.singh@it.uu.se

References:

1 - H. Brendan McMahan and Eider Moore and Daniel Ramage and Blaise Aguera y Arcas. Federated Learning of Deep Networks using Model Averaging. https://arxiv.org/abs/1602.05629