IE 2022 Tutorial

From Centralized to Distributed Machine Learning

18th International Conference on Intelligent Environments (IE 2022)

1 Title

From Centralized to Distributed Machine Learning

2 Names and Affiliations of Speakers

- **Mikel Larrea**. University of the Basque Country UPV/EHU, Spain.
- **Iñigo Perona**. University of the Basque Country UPV/EHU, Spain.

3 Abstract, objectives and motivation

Machine Learning (ML) is a branch of Artificial Intelligence that aims to develop algorithms that "learn" behaviors by using data or through experience. In many applications, data is collected massively, and in order to run ML algorithms in a reasonable time it is necessary to run them in a distributed computational environment. Moreover, in many current application scenarios (Internet of Things, Smart City, social networks, blockchain...) the information to be treated is inherently distributed. All this makes that many centralized ML solutions (based on a single computer) are being adapted to distributed environments. In turn, within distributed solutions, some solutions are designed to be executed in a cluster of computers, i.e., in a High Performance Computation environment, where one node is selected as a master and the others are workers. Other solutions are designed to be executed on top of distributed computers over the Internet. In the latter scenario, the most developed topology schema, according to scientific literature, is the one composed of one central aggregator node and multiple participant nodes. Nevertheless, there are solutions from a centralized schema, all the way through a schema where all the nodes have equivalent capabilities.

In this tutorial a survey of this decentralizing path of ML algorithms will be presented, pointing out the algorithms that have been milestone and their most maintained implementations, including some demonstrations of these solutions. These demonstrations will show most promising frameworks to use in a distributed and heterogeneous system, which is the scenario where the Intelligent Environments are created. In this sense, some cases of use will be presented using a Federated Learning framework and also using a peer-to-peer solution.

4 Keywords

Machine Learning, Internet of Things, Distributed Machine Learning, Scalable Machine Learning, Federated Learning

5 Intended Audience

The tutorial is intended for researchers who are interested in modelling the behaviours presented in the data for both purposes, for describing and for predicting them. The level of the tutorial will be introductory, it will go over the basis. This tutorial specially will focus on Federated Learning, in which the final model is generated among different participants in such a way that they do not have to share data among them. This learning process opens new opportunities to generate superior models by means of the collaboration among different entities that they have strict privacy policies over the data they own, such as, different smart city halls or different medical institutions. In this tutorial a survey of different machine learning algorithms and topologies of the distributed systems will be carried out, paying attention to communication pattern, performance, scalability, failure resiliency and security. Moreover, some simple source codes will be presented and some demonstrations will be made.

6 Content outline

1. Introduction to Machine Learning
2. Introduction to distributed systems
3. Distributed ML on supercomputers
4. Federated Learning
5. Conclusions

7 Description

In the first part of the tutorial, Introduction to Machine Learning (ML), the basic methods for ML will be presented: simple probabilistic modeling (Naive

Bayes), construction of decision trees (ID3), linear models (logistic regression), instance-based learning (K-NN), clustering (k-means) and artificial neuronal networks.

Many applications collect data massively, and moreover, the information to be treated is inherently distributed (Internet of Things, Smart City, social networks, blockchain...). All this makes centralized solutions based on a single computer neither scalable nor robust. In order to parallelize the computing, to be able to work with distributed data, and to create failure resilient systems, the machine learning community has embraced ideas from the field of Distributed Systems and High Performance Computing (HPC). Therefore, in the second part of this tutorial, Introduction to distributed systems, the new problems and challenges that the distributed ML field needs to take into account will be presented: performance, scalability, resilience to failures and security. Moreover, when designing a framework for distributed machine learning, several aspects must be taken into account, such as the communication pattern of the algorithm to be distributed, how to parallelize the data, and the model and topology of the distributed system.

In the third part, Distributed ML on supercomputers, the concept of MapReduce will be introduced because it is the technique that firstly made popular the field of distributed ML.

In the forth part, Federated Learning (FL) will be introduced, which lays the foundation of implementing ML on a distributed landscape, where heterogeneous machines can participate in a collaborative manner. This method achieves that all the participants will benefit from common learning, without knowing about all the underlying data. Currently, the available implementations have centralized design where there are one aggregator node and various participants. Therefore, meanwhile the aggregator node is operating the system will keep going, however, when it fails all the system will be down, so it becomes a critical point. In response to this problem, decentralized FL solutions (e.g., peer-to-peer FL) are being proposed. Nevertheless, FL takes care of privacy, security and anonymity. Besides, the concept of Federated Computing is arising by the implementation of Map and Reduce operations in a federated ecosystem.

8 Teaching mode

This tutorial is planned to be face-to-face. However, if needed, it will be possible to set up a video-conference session in order to be possible to follow this tutorial remotely.

9 Materials

The slides, bibliographic references and code snippets that will be used in the tutorial will be put available on a conference website and on a Github repository.

10 Bio-sketches

- **Mikel Larrea** graduated in computer engineering from the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland, in 1995. He received the PhD degree in computer science from the University of the Basque Country UPV/EHU, Spain, in 2000. He is currently an associate professor at the Department of Computer Architecture and Technology, University of the Basque Country UPV/EHU. Previously, he was on the faculty of the Public University of Navarre. He is the co-founder and director of the Distributed Systems Group (`http://dsg.eus/`). His research interests include distributed algorithms and systems, dependability and ubiquitous computing.

- **Iñigo Perona** graduated in both technical and superior computer engineering from the University of the Basque Country UPV/EHU, Spain, in 2006 and 2008 respectively. He received the PhD degree in computer science from the same university in 2016. He is an assistant professor at the Department of Computer Architecture and Technology, University of the Basque Country UPV/EHU. He is a member of the Algorithms, Data Mining and Parallelism research group (`http://www.aldapa.eus/`). His research interests include data/web mining and machine learning in contexts of intrusion detection systems, tourism, accesibility, eGovernability, eHealth and smart cities.