Machine Learning sur Big Data

ECTS : 3

Volume horaire : 24

Description du contenu de l'enseignement :

This course focuses on the typical, fundamental aspects that need to be dealt with in the design of machine learning algorithms that can be executed in a distributed fashion, on Hadoop clusters, in order to deal with big data sets, by taking into account scalability and robustness.

Machine learning algorithms are more and more used today, and there is an ever increasing demand of machine learning algorithms that scales over massives data sets.
This course focuses on the typical, fundamental aspects that need to be dealt with in the design of machine learning algorithms that can be executed in a distributed fashion, on Hadoop clusters, in order to deal with big data sets, by taking into account scalability and robustness. So the course will focus on a bunch of main-stream, sequential machine learning algorithms, by taking into account the following crucial and complex aspects. The first one is the re-design of algorithms by relying on programming paradigms for distribution and parallelism based on map-reduce, to this end Spark will be used. The second aspect is experimental analysis of the Spark implementation of designed algorithms in order to test their scalability and precision. The third aspect concerns the study and application of optimisation techniques in order to overcome lack of scalability and to improve execution time of designed algorithm.

The attention will be on machine learning technique for dimension reduction, clustering and classification, whose underlying implementation techniques are transversal and find application in a wide range of machine learning algorithms. For some of the studied algorithms, the course will present techniques for a from-scratch implementation in Spark core, while for other algorithms Spark ML will be used and end-to-end pipelines will be designed. In both cases algorithms will be analysed and optimised on real life data sets, by relaying on a local Hadoop cluster, as well as on a cluster on the Amazon WS cloud.

Bibliographie, lectures recommandées :

References:
- Mining of Massive Datasets
http://www.mmds.org
- High Performance Spark - Best Practices for Scaling and Optimizing Apache Spark
Holden Karau, Rachel Warren
O'Reilly