Large scale machine learning in Spark

ECTS : 3

Volume horaire : 24

Description du contenu de l'enseignement :

Les algorithmes d'apprentissage automatique sont de plus en plus utilisés de nos jours, et il existe une demande croissante d'algorithmes d'apprentissage qui sont capable de passer à l'échelle et de traiter des données massives.
Plutôt qu'offrir une introduction exhaustive à l'apprentissage automatique, ce cours se concentre sur les aspects typiques qui doivent être traités dans la conception d'algorithmes distribués pour l'apprentissage, et qui peuvent être exécutés sur les clusters Hadoop, afin d'analyser des grands jeux de données, en tenant compte l'adaptabilité à la croissance du volume des données ainsi que la robustesse en cas de pannes.

Le focus sera sur des algorithmes de réduction de dimension, de clustering et de classification, en tenant compte les aspects suivants. Le premier est la conception d'algorithmes en s'appuyant sur des paradigmes basés sur map-reduce, à cette fin Spark sera utilisé. Le second aspect est l'analyse expérimentale des algorithmes implémentés en Spark, afin de tester leur capacité de passer à l'échelle (scalabilité). Le troisième aspect concerne l'étude et l'application de techniques d'optimisation afin de pallier le manque éventuel de scalabilité.
Bien que le cours se focalise sur certains algorithmes d'apprentissage, les techniques étudiées sont transversales et trouvent application dans un large éventail d'algorithmes d'apprentissage automatique. Pour certains des algorithmes étudiés le cours présentera des techniques pour une implémentation à partir de zéro en Spark-core, tandis que pour d'autres algorithmes Spark ML sera utilisé, et des pipelines de bout en bout seront conçus. Dans les deux cas, les algorithmes seront analysés et optimisés sur des jeux de données réels, sur un cluster Hadoop local, ainsi que sur un cluster sur le cloud Amazon WS.

Compétence à acquérir :

Savoir concevoir des applications efficaces pour l'apprentissage machine sur les données massives.

Mode de contrôle des connaissances :

CC et Projet