Distributed Algorithms for Scalable Nonparametric Learning

This is a joint academic-industrial research and development position between CIM-Uppsala (Centre for Interdisciplinary Mathematics, Uppsala University) and ACE-Combient (Analytics Centre of Excellence, Combient AB[1]) and will involve working in Uppsala and Stockholm each week.

Apache Spark has become a standard in the data industry for production-ready machine learning algorithms that scale over a cluster of commodity computers. The project will build on Spark's resilient distributed datasets that provide fault-tolerant distributed random-access-memory models, and take advantage of Spark's core libraries (MLlib, SQL, Streaming and GraphX) where possible. Thus, knowledge of the distributed computing framework of Apache Spark, scala and integration with C++ would be an advantage.

A major research goal of this position is to extend nonparametric data-driven density estimators involving space-partitioning trees [2] into the distributed framework of Apache Spark that is capable of supporting continuous streams of high-dimensional data in current industrial applications. Specifically, such estimators will be developed with a view towards straightforward adaptations in production-ready applications to solve industrial problems faced at ACE-Combient. Some industrial applications of density estimators that are of interest to ACE-Combient include, unsupervised clustering, anomaly detection and predictive maintenance. Our academic-industrial research and development goals will be accomplished by developing our distributed algorithms as a Spark dataset processing module that can be seamlessly plugged into various stages of MLlib's existing machine learning pipelines and thereby taking advantage of Spark's latest advances in dimensionality reduction, graph-frames and discrete streams. This is expected to nearly eliminate the usual gap between the deployment and production phase and the research and development phase.

More generally, the project will involve the development of distributed algorithms within the Hadoop ecosystem for novel nonparametric methods with desirable mathematical properties including, invariance to linear transformations, universal performance guarantees and being conducive to tree arithmetic, while remaining applicable to massive data streams. Some such extensions will involve collaborations between hyper-plane split trees and nearest neighbour methods. The project will result in Apache 2.0 licensed Spark packages that

[1] Nettovägen 6, 175 41 Järfälla, Stockholm

[2] Data-adaptive histograms through statistical regular pavings, Raazesh Sainudiin, Gloria Teng, Jennifer Harlow and Warwick Tucker, 2016 (under review) `http://lamastex.org/preprints/20161121optMAPMDE.pdf` and Posterior expectation of regularly paved random histograms, Raazesh Sainudiin, Gloria Teng, Jennifer Harlow and Dominic Lee, ACM Trans. Model. Comput. Simul. 23, 1, Article 6, 20 pages (Special Issue on Monte Carlo Methods in Statistics) `http://lamastex.org/preprints/SubPavingMCMC.pdf`

will deliver generic, novel and scalable nonparametric statistical methods that are readily applicable to a range of industrial problems.

Research aspects of the project will be conducted at CIM-Uppsala under the direction of Raazesh Sainudiin (Uppsala University) with mathematical statistical guidance from Luc Devroye (McGill University) and computer arithmetic guidance from Warwick Tucker (Uppsala University). The industrial component will include working regularly with data scientists and engineers at ACE-Combient in Stockholm under the direction of Annica Grimberg Lignell, VP and Head of ACE-Combient.