Distributed Machine Learning

Mots clés :
- Directeur de thèse : STEPHAN CLEMENCON
- Co-encadrant(s) :
- Unité de recherche : Laboratoire Traitement et Communication de l'Information
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

In most analyses carried out in the field of statistical learning theory, the practical constraints related to the data acquisition/storage/access system and inherent to processing speed, memory and computing capacity are generally ignored or incorporated into the mathematical framework in a very stylized manner so far. With the advent of highly complex digital network infrastructures and the pressing necessity of sharing resources and distributing computing power [14, 1], this facet of the machine-learning environment is however becoming more and more essential from a technological perspective and is receiving now increasing attention, see [8, 9, 10, 13, 7, 12, 11, 4, 6, 5] for instance. Motivated by the recent developments in the architecture of data repositories and computer systems, it is the main purpose of this research project to investigate machine-learning problems in a distributed, and possibly on-line, context, accounting for certain real-life situations, more and more currently encountered in the near future. More precisely, we will consider statistical frameworks, where the training data are not stored in some central memory but split into distinct clusters, individually processed by independent agents (e.g. processors). To process Big Data, one generally distribute data subsamples over a network of processors communicating with each other. Precisely, it is assumed that the agents can exchange a limited amount of information per unit of time only, through a communication structure modeled by a graph of which they form the nodes. Hence, due to these capacity constraints, merging all training sets at any node is unfeasible and a distributed approach, limiting the network overhead, is required. Here, by “distributed”, it is meant that both the learning and prediction stages are performed by the means of local computations of the agents and sparse communications between them: each agent simultaneously processes the data set it has been assigned to and shares some information with its neighbors in order to build a local decision rule. At the end of the learning procedure, all agents use the consensus classifier to predict labels assigned to test data, with no need for further communications. The nature of the problem we investigate through this project is very different, it is not of the type “distributed consensus”. It should be noticed that, unlike most works on “distributed learning”, agents are not assumed exchangeable in the framework we consider. First, we assume that the collection of local decision rules may vary from an agent to another: the issue at stake is not to seek to achieve a consensus between the agents but to learn how to aggregate efficiently the local decisions, typically through a majority vote or a well-chosen weighted average. Hence, both learning and test phases shall require distributed computations, relying on the whole network of agents. Expected results. Whereas most works documented in the literature consist in proposing ad-hoc methods so as to scale up empirical risk minimization techniques (generally based on the stochastic gradient descent paradigm) by means of a distributed strategy, relying on a preliminary data-splitting stage and proving asymptotic results or rate bounds, our goal is here to investigate the issue of optimizing under appropriate constraints (governing information transmission, the connectivity of the underlying network, the amount of information stored in each node, etc.). We will formulate issues such as: is there an optimal way of splitting the data, allocating them to the various nodes and transmitting the local updates in terms of convergence rate (respectively, rate bound, asymptotic covariance)? Such problems could be formulated in terms of model selection and solved by oracle inequalities. Adaptive procedures could also be considered. For instance, based on survey theory results, the problem of sampling data in a nearly optimal fashion has been recently investigated in [3, 2] in a centralized framework from an asymptotic perspective. This paves the way for establishing more ambitious results, extending the latter to a distributed setting and taking also into account the computational cost of sampling algorithms.