Séminaire de Probabilités et Statistique

Jeudi 11 Octobre à 14h00

Laboratoire Dieudonné
Salle de conférence - Bâtiment Fizeau, 5-ième étage

Audrey Poterie

IRMAR-INSA Rennes

Decision trees and random forests for grouped variables

Supervised learning consists in explaining and/or predicting an output y by using some inputs x. In many problems, inputs have a known and/or obvious group structure. These groups can naturally exist or they can be defined to capture the underlying input associations. For instance, in biology, when we want to study the chemical composition of a serum based on spectrometry data, the inputs, which are functional, can be clustered into groups representing the different parts of the curve. In this context, elaborating a prediction rule that takes into account the group structure can be more relevant than using an approach based only on the individual variables for both prediction accuracy and interpretation. Some supervised algorithms which build prediction rules based on groups of inputs have been already proposed. One of the best-known methods is certainly the Group Lasso. The goal of this work is to develop some tree-based methods adapted to grouped variables. Indeed, tree-based approaches are commonly used in statistics. These methods allow to readily construct prediction rules easily understandable. Moreover, many aggregation algorithms such that the booting methods and the random forests are based on decision trees. These algorithms are often part of the list of the most successful methods currently use to handle prediction problems. Here, we propose two new tree-based approaches which use the group structure to build decision trees. These two methods begin with constructing a maximal tree by means of recursive partitioning of the data space. The first approach allows to build binary decision trees for classification problems. A split of a node is defined according to the choice of both a splitting group and a linear combination of the inputs belonging to the splitting group. The second method, which can be used for prediction problems in both regression and classification, builds a non-binary tree in which each split is a binary tree. In these two methods, the maximal tree is next pruned. To this end, we propose two pruning strategies, one of which is a generalization of the minimal cost-complexity pruning algorithm to non-binary trees. Since decisions trees are known to be unstable, we also introduce a method of random forests that deals with groups of inputs. In addition to the prediction purpose, these three new methods can be also use to perform group variable selection thanks to the introduction for each of them of some measures of group importance. This thesis work is supplemented by an independent part in which we consider the unsupervised framework. We introduce a new clustering algorithm which is based on the hierarchical clustering algorithm named single linkage. Under some classical regularity and sparsity assumptions, we obtain the rate of convergence of the clustering risk for the proposed algorithm.