Séminaire de Probabilités et Statistique

Mardi 11 janvier à 14h00

Salle Fizeau (5ème étage)

Clément Benard

Université Pierre et Marie Curie / Safran

MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA

Variable importance measures are the main tools to analyze the black-box mechanisms of random forests. Although the mean decrease accuracy (MDA) is widely accepted as the most efficient variable importance measure for random forests, little is known about its statistical properties. In fact, the exact MDA definition varies across the main random forest software. In this article, our objective is to rigorously analyze the behavior of the main MDA implementations. Consequently, we mathematically formalize the various implemented MDA algorithms, and then establish their limits when the sample size increases. In particular, we break down these limits in three components: the first one is related to Sobol indices, which are well-defined measures of a covariate contribution to the response variance, widely used in the sensitivity analysis field, as opposed to the third term, whose value increases with dependence within covariates. Thus, we theoretically demonstrate that the MDA does not target the right quantity when covariates are dependent, a fact that has already been noticed experimentally. To address this issue, we define a new importance measure for random forests, the Sobol-MDA, which fixes the flaws of the original MDA. We prove the consistency of the Sobol-MDA and show that the Sobol-MDA empirically outperforms its competitors on both simulated and real data. An open source implementation in R and C++ is available online.