Provable recovery of feature interactions from Random Forest with applications to heterogeneous treatment effect estimation

Prof. Dr. Merle Behr
(University Regensburg)

July 03, 2024

The rapid generation of data in the healthcare sector, exemplified by electronic patient records and genetic data, has revolutionized our approach to information extraction in this field. Initially, many problems in this domain resemble classical prediction questions, such as predicting whether a patient with a specific genetic profile will develop a particular disease. However, beyond prediction, the interpretation and explanation of these models are often of greater interest. Random forests, known for their robust predictive capabilities, can manage large-scale, high-dimensional data effectively. This talk explores how the tree structures within random forest ensembles can be systematically analyzed to identify key features and their interactions that drive the prediction models. This methodology provides interpretable explanations for complex models. We will demonstrate the application of this approach to uncover genetic interactions relevant to genotype-phenotype relationships. Additionally, we will discuss its application in estimating heterogeneous treatment effects in causal inference settings using observational data, such as electronic health records.