LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN INSTITUT FÜR STATISTIK

Provable Boolean interaction recovery from tree ensemble obtained via random forests

Merle Behr

(University of Regensburg)

29/05/2024, 4.15pm

Department of Statistics, Ludwigstr. 33, Room 144
and online via Zoom (Link)
(Meeting-ID: 683 0699 4223; Password: StatsCol23)

Random Forests (RFs) are at the cutting edge of supervised machine learning in terms of prediction performance, especially in genomics. Iterative RFs (iRFs) use a tree ensemble from iteratively modified RFs to obtain predictive and stable nonlinear or Boolean interactions of features. They have shown great promise for Boolean biological interaction discovery that is central to advancing functional genomics and precision medicine. However, theoretical studies into how tree-based methods discover Boolean feature interactions are missing. Inspired by the thresholding behavior in many biological processes, we first introduce a discontinuous nonlinear regression model, called the "Locally Spiky Sparse" (LSS) model. Specifically, the LSS model assumes that the regression function is a linear combination of piecewise constant Boolean interaction terms. Given an RF tree ensemble, we define a quantity called "Depth-Weighted Prevalence" (DWP) for a set of signed features S. Intuitively speaking, DWP(S) measures how frequently features in S appear together in an RF tree ensemble. We prove that, with high probability, DWP(S) attains a universal upper bound that does not involve any model coefficients, if and only if S corresponds to a union of Boolean interactions under the LSS model. Consequentially, we show that a theoretically tractable version of the iRF procedure, called LSSFind, yields consistent interaction discovery under the LSS model as the sample size goes to infinity. Finally, simulation results show that LSSFind recovers the interactions under the LSS model, even when some assumptions are violated.

Biography:

M. Behr is a Professor of Machine Learning at the Faculty of Informatics and Data Science, University of Regensburg, since October 2022. Previously, she worked in industry for almost two years at Bayer AG in Pharmaceuticals R&D. From 2018 to 2020, M. Behr was a Postdoctoral Researcher and Neyman Visiting Assistant Professor at the University of California, Berkeley. She completed a PhD in Mathematics at the University of Göttingen in 2018 under the supervision of Prof. A. Munk, following studies in mathematics at the University of Göttingen and the University of Edinburgh from 2009 to 2014. Her research focuses on statistical machine learning, particularly in bio-medicine, genetics, and the natural sciences, with interests in tree ensemble methods, blind source separation, change-point detection, and statistical methodologies related to tree structures.