Selective confidence intervals for tree-based models via parametric bootstrap

Tree-based regression models recursively partition the covariate space using binary splits to identify regions that differ most strongly with regard to the values of an outcome variable of interest. Classical approaches for inference on the model parameters are invalid for tree-based models as they negelect the uncertainty induced by the data-driven tree building. In particular, it is necessary to account for the fact that the interest in the specific parameters associated with a tree arises out of the fitted structure of that tree. This constitutes a so-called *selective inference* problem, where inference after data-driven model selection is of interest (Berk et al., 2013). Neufeld et al. (2022) proposed a selective inference framework tailored to regression trees. Their approach focuses on trees fitted with the classification and regression trees (CART) algorithm and is limited to metrically scaled outcome variables. More recently, Spuck et al. (2024) developed a more general parametric bootstrap-based approach to construct selective confidence intervals for tree-structured varying coefficients.

This project aims to further explore the approach by Spuck et al. (2024), which can be easily extended to other tree-based models. More specifically, the objective is to investigate whether the approach allows to construct confidence intervals with valid coverage proportions for parameters of tree-based models with non-Gaussian outcome variables (for example classification trees with categorical outcome variables). For more information, please contact Nikolai Spuck (spuck@imbie.uni-bonn.de) or Moritz Berger (moritz.berger@imbie.uni-bonn.de).

References

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41:802–837, 2013.

A. C. Neufeld, L. L. Gao, and D. M. Witten. Tree-values: Selective inference for regression trees. *Journal of Machine Learning and Research*, 23:1–43, 2022.

N. Spuck, M. Schmid, M. Monin, and M. Berger. Confidence interrvals for tree-structured varying coefficients. *arXiv:2406.19887*, 2024.