Universität Konstanz

Interpretable Machine Learning: Regional Approaches, Counterfactuals and Efficient Model Selection

Prof. Dr. Bernd Bischl

Tuesday, 16 April 2024, 08:15, ZT 702

Supervised machine learning models are widely used due to their good predictive performance, but they are often difficult to interpret, especially if models are non-parametric and highly non-linear.

Post-hoc interpretation methods from the field of interpretable machine learning can help to draw conclusions about the inner processes of these models. For tabular data, these approaches are usually structured into global, regional and local methods. My presentation will focus on the latter two, by outlining several new algorithms for a) regional effect estimation (based on the well-known partial dependence, ALE plot and SHAP) to handle and also detect feature interactions; b) multi-objective counterfactual generation and hyperbox-based semi-factuals for local explanations.

I will conclude by at least providing a glimpse into how methods from IML can be combined with (usually multi-objective) Bayesian optimization, to restrain complexity in a model agnostic fashion.

Faculty of Sciences
Department of Computer and Information Science