Kolloquium „Statistische Methoden in der empirischen Forschung“

Wann: 11. Februar 2020, 17:00 – 18:30 Uhr

Wo: Robert Koch-Institut | Nordufer 20 | 13353 Berlin (Wedding)

Giuseppe Casalicchio (LMU München)

Interpretable Machine Learning Methods

Machine learning algorithms are often considered to produce black-box models because they do not provide any direct explanation for their predictions. However, they often outperform simple and more interpretable models such as linear models or decision trees in terms of predictive performance as they can model more complex relationships in the data. Therefore, it is not surprising that the use of machine learning techniques is gaining more attraction in many different disciplines such as medicine (e.g., to support decisions on health care plans for patients with serious diseases), politics, criminology, ecology, and astrophysics.

However, in such critical areas, it is essential to use models that are interpretable and provide explanations for their decisions, especially if this can affect human life. Fortunately, a huge amount of model-agnostic methods to improve the transparency, trustability, and interpretability of machine learning models have been developed.

However, different notations and terminology have complicated the understanding, discussion and research of model-agnostic interpretation methods in machine learning. A unified view of these techniques has been missing. This talk gives an overview of several state-of-the-art interpretability methods and how they relate to each other.

Specifically, we propose a generalized framework of work stages for model-agnostic interpretability methods and demonstrate how several prominent methods can be embedded in our proposed SIPA (Sampling, Intervention, Prediction, Aggregation) framework.