Contribution ID: **22** Type: **not specified**

Green LIME: Improving AI Explainability through Design of Experiments

Friday, 30 May 2025 10:05 (20 minutes)

In artificial intelligence (AI), the complexity of many models and processes often surpasses human interpretability, making it challenging to understand why a specific prediction is made. This lack of transparency is particularly problematic in critical fields like healthcare, where trust in a model's predictions is paramount. As a result, the explainability of machine learning (ML) and other complex models has become a key area of focus.
Efforts to improve model interpretability often involve experimenting with AI systems and approximating their behavior through simpler mechanisms. However, these procedures can be resource-intensive. Optimal design of experiments, which seeks to maximize the information obtained from a limited number of observations, offers promising methods for improving the efficiency of these interpretability techniques.

To demonstrate this potential, we explore Local Interpretable Model-agnostic Explanations (LIME), a widely used method introduced by Ribeiro et al. 2016. LIME provides explanations by generating new data points near the instance of interest and passing them through the model. While effective, this process can be computationally expensive, especially when predictions are costly or require many samples.
LIME is highly versatile and can be applied to a wide range of models and datasets. In this work, we focus on models involving tabular data, regression tasks, and linear models as interpretable local approximations.

By utilizing optimal design of experiments' techniques, we reduce the number of function evaluations of the complex model, thereby reducing the computational effort of LIME by a significant amount. We consider this modified version of LIME to be energy-efficient or "green".

Type of presentation

Contributed Talk

Primary author: Ms STADLER, Alexandra (JKU Linz)

Co-authors: MUELLER, Werner G. (Johannes Kepler University); Prof. HARMAN, Radoslav (Comenius University Bratislava)

Presenter: MUELLER, Werner G. (Johannes Kepler University)

Session Classification: Session

Track Classification: Spring Meeting