Green LIME: Improving AI Explainability through Design of Experiments

Alexandra Stadler[1], Werner G. Müller[1], and Radoslav Harman[2]

[1]*Johannes Kepler University Linz, Austria*
[2]*Comenius University Bratislava, Slovakia*

Abstract

In artificial intelligence (AI), the complexity of many models and processes surpasses human understanding, making it challenging to determine 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 explainability often involve experimenting with AI systems and approximating their behavior through interpretable surrogate 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 explainability techniques.

To demonstrate this potential, we explore Local Interpretable Model-agnostic Explanations (LIME), a widely used method introduced by [1]. 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".

Keywords

Model Interpretability, Optimal Design of Experiments, Explainable Artificial Intelligence, Post-hoc Explanation, Local Regression.

Acknowledgements

We acknowledge the use of ChatGPT (OpenAI) to assist in editing the English grammar and improving language clarity in the abstract. The content and ideas are solely the authors' own, unless cited or stated otherwise.

References

[1] Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos, (2016).“Why Should I Trust You?”: Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144).