Effector: A Python package for regional explanations

Vasilis Gkolemis[1,2], Christos Diou[1], Eirini Ntoutsi[3], Theodore Dalamagas[2], Bernd Bischl[4], Julia Herbinger[4] and Giuseppe Casalicchio[4]

[1]Harokopio University of Athens
[2]ATHENA Research Center
[3]University of the Bundeswehr Munich
[4]Munich Center for Machine Learning (MCML), Department of Statistics, LMU Munich

Abstract

Global feature effect methods explain a model outputting one plot per feature. The plot shows the average effect of the feature on the output, like the effect of age on the annual income. However, average effects may be misleading when derived from local effects that are heterogeneous, i.e., they significantly deviate from the average. To decrease the heterogeneity, regional effects provide multiple plots per feature, each representing the average effect within a specific subspace. For interpretability, subspaces are defined as hyperrectangles defined by a chain of logical rules, like age's effect on annual income separately for males and females and different levels of professional experience. We introduce `Effector`, a Python library dedicated to regional feature effects. `Effector` implements well-established global effect methods, assesses the heterogeneity of each method and, based on that, provides regional effects. `Effector` automatically detects subspaces where regional effects have reduced heterogeneity. All global and regional effect methods share a common API, facilitating comparisons between them. Moreover, the library's interface is extensible so new methods can be easily added and benchmarked. The library has been thoroughly tested, ships with many tutorials (https://xai-effector.github.io/) and is available under an open-source license at PyPi https://pypi.org/project/effector/ and Github https://github.com/givasile/effector.

Keywords

Explainability, Interpretability, Feature Effect, Regional Effect, Global Explanations

1. Introduction

The increasing adoption of machine learning (ML) in high-stakes domains like healthcare and finance has raised the demand for explainable AI (XAI) techniques [**?** **?**]. Global feature effect methods explain a black-box model through a set of plots, where each plot is the effect of a feature on the output, as in Figure **??**.

Global effects may be misleading when the black-box model $f(\cdot)$ exhibits interactions between features. An interaction between two features, x_s and x_k, exists when the difference in the output $f(\mathbf{x})$ as a result of changing the value of x_s *depends* on the value of x_k [**?**]. Global effects are often computed as averages over local effects. When feature interactions are present, local effects become heterogeneous, i.e., they significantly deviate from the average (global) effect. In these cases, the global effect may be misleading, a phenomenon known as *aggregation bias* [**?** **?**].

Regional [**?** **?** **?** **?** **?** **?**] or cohort explanations [**?**], partition the input space into subspaces and compute a regional explanation within each. The partitioning aims at subspaces with homogeneous local effects, i.e., with reduced feature interactions, yielding regional effects with minimized aggregation bias [**?**]. By combining these regional explanations, users can interpret the model's behavior across the entire input space. Several libraries focus on XAI, but none targets on regional effect methods (Table **??**). Therefore, we present `Effector`, a Python library dedicated to regional explainability methods, which:

- implements well-established global and regional effect methods, accompanied by an intuitive way to visualize the heterogeneity of each plot.
- follows a consistent and modular software design. Existing methods share a common API and novel methods can be easily added and compared to existing ones.
- demonstrates through tutorials the use of regional effects in real and synthetic datasets.

2. Conclusion and Future Work

The 2nd World Conference on eXplainable Artificial Intelligence, July 17–19, 2024, Malta, Valetta