Local to Global Explainable Features

Introduction

Deep learning-based systems have exploded in popularity since the onset of the decade. Tools like ChatGPT, DALL-E, and Stable Diffusion have assisted professionals, artists, and enthusiasts with their day-to-day activities. The large-scale adoption of deep learning in areas like healthcare and law has resulted in significant automation and load reduction.

However, with the use of such models in these critical areas comes additional challenges concerning accountability and explainability. Research has shown that machine learning models suffer from bias [1] and can substantially damage people and their livelihoods. Wielding AI tools hence comes as a double-edged sword wherein we need to maintain explainability while not compromising on model predictions.

As a result, research in explainable AI tools and models has exploded in hopes of explaining the reasons behind predictions of black-box models. Key categorizations in this include inherently explainable models, like decision trees, whose very structure captures the differences between classes and serves for explainable interpretations, and post hoc explainability methods that aim to explain a black-box model after it has been trained.

Many post hoc explainability tools seek to emulate the local area behind the point-in-question as an approximation of explainable models. LIME and SHAP [2, 3] seek to model the local area as an interpretable approximation. It has been seen, however, that these models suffer from large variations owing to the nature of point estimates. Bayesian interpretations like BayesLIME and BayesSHAP [4] also model the expected error behind local predictions, offering a view of the expected variance of the estimate.

Though work has been done to model the local interpretations, there has not been much progress in converting an aggregation of local estimates into one unified, global picture. [5] hierarchically merges local interpretations into a global perspective by merging similar logical theories using Jaccard similarities of coverage as a similarity measure. [6] aggregates local interpretations using absolute values but does not consider whether the scaling is negative or positive.

In this work, I propose a Bayesian framework based on Markov chain Monte Carlo and importance sampling to aggregate local interpretations. The contributions of this work are as follows:

1. Proposed a sampling-based approach for aggregating local interpretations to global perspectives.
2. Developed a benchmark inspired by Slack et al. (2021) for determining the effectiveness of the interpretations.

Methodology

The working of the algorithm is remarkably simple. Using the training data, a Gaussian mixture model is fit because of its ability to capture multiple modes in the data, which makes it suitable for classification tasks, as different classes might have different distributions. Points are sampled from the fit Gaussian mixture, and local explanations are constructed.

For the Monte Carlo variant, the mean of absolute values of explanations is taken. For the importance sampling variant, *n* points x_i are sampled from a standard normal distribution and weighted according to the equation:

$$t_i = \frac{q(x_i)}{p(x_i)} f(x_i)$$

where $q(x_i)$ is the Gaussian PDF, and $p(x_i)$ is the likelihood of the point from the mixture distribution. Then, absolute values of t_i from $i = 1$ *to* n are averaged to give a global perspective, proceeding as per [6]

Experimental Setup

No standard benchmark is available to evaluate the performance of aggregating local explainability to global context explainability of models using such frameworks. However, using the premise that removing the most important features should result in the maximal performance drop on the testing set after retraining the model on the limited set of features, an incremental dropping of features was performed, moving from 1 dropped feature to 6. The method that gave the maximum drop in performance was capturing the most important features regarding explainability.

The model trained was a decision tree classifier because of its fast fitting and proneness to overfitting, which should make it capture most patterns in the data. The training splits and testing split ratios were 0.7 and 0.3, respectively. The global aggregation functions were fit on the training data. Standard scaling practices were followed.

The model fitting was carried out for 30 iterations, and the mean and variance on the test set were taken as the criterion. Randomization was permitted in this case.

For the fitting, 1500 samples were used for the global interpretation estimation.

Testing was carried out on the German Credit risk and the Wine category dataset by UCI.

Results

Features Removed	**Dataset**	**(SHAP) Test acc**	**(SHAP) Test var**	**(MCMC) Test acc**	**(MCMC) Test var**	**(Imp. Samp.) Test acc**	**(Imp. Samp.) Test var**
1	German Credit	0.944	0.00E+00	0.944	0.00E+00	0.944	0.00E+00
2	German Credit	0.926	0.00E+00	**0.87**	0.00E+00	0.981	0.00E+00
3	German	0.907	0.00E+00	**0.852**	0.00E+00	**0.852**	0.00E+00

	Credit						
4	German Credit	0.852	0.00E+00	0.852	0.00E+00	0.852	0.00E+00
5	German Credit	0.889	0.00E+00	0.889	0.00E+00	0.889	0.00E+00
6	German Credit	0.833	0.00E+00	0.833	0.00E+00	0.833	0.00E+00

Performance on the German Credit Dataset

Features Removed	**Dataset**	**(SHAP) Test acc**	**(SHAP) Test var**	**(MCMC) Test acc**	**(MCMC) Test var**	**(Imp. Samp.) Test acc**	**(Imp. Samp.) Test var**
1	Wine	0.944	0.00E+00	0.944	0.00E+00	**0.944**	0.00E+00
2	Wine	0.926	0.00E+00	**0.87**	0.00E+00	**0.87**	0.00E+00
3	Wine	**0.796**	0.00E+00	0.852	0.00E+00	0.852	0.00E+00
4	Wine	**0.833**	0.00E+00	0.852	0.00E+00	0.852	0.00E+00
5	Wine	0.889	0.00E+00	0.889	0.00E+00	0.889	0.00E+00
6	Wine	0.833	0.00E+00	0.833	0.00E+00	0.833	0.00E+00

Performance on the Wine Quality Dataset

We observe good performance by the MCMC and importance sampling method when fewer features are removed, often beating the SHAP aggregations. We also note that the performance drop upon feature removal starts to converge with more number of removed features, which might refer to similar features being removed during those instances. More testing needs to be done with larger datasets and better testing methods to gauge the performance of the proposed method.

Conclusion and Future Work

We see good performance by Bayesian methods for aggregating local interpretations. We also note the need for a unified benchmark for converting local feature importances to global.

Possible extensions of this work include proposing a different importance for each classification. Additionally, the current model requires the main distribution of the data, which might inhibit getting 'post-hoc black box' interpretations.

An alternate route to achieve local to global would be to use a hierarchical encoder that models the distribution of $f(x|xi, w)$ and $f(xi|W)$, where the first depicts the local interpretation given a point xi and the second depicts the sampling of the point xi given the main distribution W.

References

1. Mikołajczyk-Bareła, Agnieszka, and Michał Grochowski. "A survey on bias in machine learning research." *arXiv preprint arXiv:2308.11254* (2023).
2. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" 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*. 2016.
3. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
4. Slack, Dylan, et al. "Reliable post hoc explanations: Modeling uncertainty in explainability." *Advances in neural information processing systems* 34 (2021): 9391-9404.
5. Setzu, Mattia, et al. "Glocalx-from local to global explanations of black box ai models." *Artificial Intelligence* 294 (2021): 103457.
6. van der Linden, Ilse, Hinda Haned, and Evangelos Kanoulas. "Global aggregations of local explanations for black box models." *arXiv preprint arXiv:1907.03039* (2019).