Instructions for Paper Submissions to AISTATS 2023

Anonymous Author
Anonymous Institution

Abstract

The Abstract paragraph should be indented 0.25 inch (1.5 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The **Abstract** heading must be centered, bold, and in point size 12. Two line spaces precede the Abstract. The Abstract must be limited to one paragraph.

1 INTRODUCTION

Recently, ML has flourished in critical domains, such as healthcare and finance. In these areas, we need ML models that predict accurately but also with the ability to explain their predictions. Therefore, Explainable AI (XAI) is a rapidly growing field due to the interest in interpreting black box machine learning (ML) models. XAI literature distinguishes between local and global interpretation methods (Molnar et al., 2020). Local methods explain a specific prediction, whereas global methods explain the entire model behavior. Global methods provide a universal explanation, summarizing the numerous local explanations into a single interpretable outcome (number or plot). For example, if a user wants to know which features are significant (feature importance) or whether a particular feature has a positive or negative effect on the output (feature effect), they should opt for a global explainability technique. Aggregating the individual explanations for producing a global one comes at a cost. In cases where feature interactions are strong, the global explanation may obfuscate heterogeneous effects (Herbinger et al., 2022) that exist under the hood, a phenomenon called aggregation bias (Mehrabi et al., 2021).

Feature effect forms a fundamental category of global explainability methods, isolating a single feature's average impact on the output. Feature effect methods suffer from aggregation bias because the rationale behind the average effect might be unclear. For example, a feature with zero average effect may indicate that the feature has no effect on the output or, contrarily, it has a highly positive effect in some cases and a highly negative one in others.

There are two widely-used feature effect methods; Partial Dependence Plots (PDPlots)(Friedman, 2001) and Aggregated Local Effects (ALE)(Apley and Zhu, 2020). PDPlots have been criticized for producing erroneous feature effect plots when the input features are correlated due to marginalizing over out-of-distribution synthetic instances. Therefore, ALE has been established as the state-of-the-art feature effect method since it can isolate feature effects in situations where input features are highly correlated.

However, ALE faces two crucial drawbacks. First, it does not provide a way to inform the user about potential heterogeneous effects that are hidden behind the average effect. In contrast, in the case of PDPlots, the heterogeneous effects can be spotted by exploring the Individual Conditional Expectations (ICE)(Goldstein et al., 2015). Second, ALE requires an additional step, where the axis of the feature of interest is split in K fixed-size non-overlapping intervals, where K is a hyperparameter provided by the user. This splitting is done blindly, which can lead to inconsistent explanations.

In this paper, we extend ALE with a probabilistic component for measuring the uncertainty of the global explanation. The uncertainty of the global explanation expresses how certain we are that the global (expected) explanation is valid if applied to an instance drawn at random and informs the user about the level of heterogeneous effects hidden behind the expected explanation. Our method completes ALE, as ICE plots complement PDPlots, for revealing the heterogeneous effects.

Our method also automates the step of axis splitting into non-overlapping intervals. We, firstly, transform the bin splitting step into an unsupervised clustering problem and, second, find the optimal bin splitting for a robust estimation of (a) the global (expected) effect and (b) the uncertainty of the explanation from the limited samples of the training set. We formally prove that the objective of the clustering problem has as lower-bound the aggregated uncertainty of the global explanation. Our method works out of the box

Preliminary work. Under review by AISTATS 2023. Do not distribute.

without requiring any input from the user.

Contributions. The contributions of this paper are the following:

- We introduce Uncertainty DALE (UDALE), an extension of DALE that quantifies the uncertainty of the global explanation, i.e. the level of heterogeneous effects hidden behind the global explanation.
- We provide an algorithm that automatically computes the optimal bin splitting for robustly estimating the explanatory quantities, i.e., the global effect and the uncertainty.
- We formally prove that our method finds the optimal grouping of samples, minimizing the added uncertainty over the unavoidable heterogeneity that is the lower-bound of the objective.
- We provide empirical evaluation of the method in artificial and real datasets.

The implementation of our method and the code for reproducing all the experiments is provided in the submission and will become publicly available upon acceptance.

2 BACKGROUND AND RELATED WORK

Notation. We refer to random variables (rv) using uppercase X, whereas to simple variables with plain lowercase x. Bold denotes a vector; $\mathbf{x}$ for simple variables or $\mathbf{X}$ for rvs. Often, we partition the input vector $\mathbf{x} \in \mathbb{R}^D$ to the feature of interest $x_s \in \mathbb{R}$ and the rest of the features $\mathbf{x_c} \in \mathbb{R}^{D-1}$. For convenience we denote it as $(x_s, \curvearrowleft_c)$, but we clarify that it corresponds to the vector $(x_1, \cdots, x_s, \cdots, x_D)$. Equivalently, we denote the corresponding rv as $X = (X_s, \mathbf{X}_c)$. The black-box function is $f : \mathbb{R}^D \rightarrow \mathbb{R}$ and the feature effect of the s-th feature is $f_{<\texttt{method}>}(x_s)$, where $< \texttt{method} >$ is the name of the feature effect method. [1]

Feature Effect Methods. PDPlots formulate the feature effect of the s-th attribute as an expectation over the marginal distribution $\mathbf{X}_c$, i.e., $f_{\texttt{PDP}}(x_s) = \mathbb{E}_{\mathbf{X}_c}[f(x_s, \mathbf{X}_c)]$. MPlots formulate it as an expectation over the conditional $\mathcal{X}_c|\mathcal{X}_s$, i.e., $f_{\texttt{MP}}(x_s) = \mathbb{E}_{\mathcal{X}_c|\mathcal{X}_s=x_s}[f(x_s, \mathcal{X}_c)]$. ALE computes the global effect at x_s as an integration of local effects. The local effects are measured as the expected change on the output $\frac{\partial f(x_s, \mathcal{X}_c)}{\partial x_s}$ over the conditional distribution $\mathcal{X}_c|\mathcal{X}_s$. The formula that defines ALE is presented below:

$$f_{\texttt{ALE}}(x_s) = c + \int_{-\infty}^{x_s} \mathbb{E}_{\mathcal{X}_c|\mathcal{X}_s=z} \left[\frac{\partial f(z, \mathcal{X}_c)}{\partial z} \right] \partial z \qquad (1)$$

The constant c is used for centering the ALE plot. PDPlots integrate over unrealistic instances due to the use of the marginal distribution $p(\mathcal{X}_1)$. Therefore, they incorrectly result in a quadratic effect in the region $x_1 \in [0, 1]$. MPlots resolve this issue using the conditional distribution $\mathcal{X}_2|\mathcal{X}_1$ but suffer from computing combined effects. ALE plots resolve, both they have two drawbacks TODO add more info.

Qunatify the Heterogeneous Effects. It is crucial for feature effect methods to inform about the heterogeneous effects. Elaborate. Interpretation of the heterogeneous effects behind the global effect is available only for PDP, with three different approaches; (a) ICE and d-ICE plots provide a visual understanding of the heterogeneous effects. (b) grouping of ICE in homogeneous clusters, for spliting the input space into subspace(s) with homogeneous effects (c) Feature Interaction strength indexes, like H-statistic, provide a value indicating how much a feature interacts with the others (not the type of interaction). There is no method for quantifying the heterogeneous effects, based on ALE. Therefore, no method to exploit the advantages of ALE while, on the same time, informing about the heterogeneous effects.

Bin Spliting. ALE also has the peculiarity of spliting the axis into intervals, allocating the instances of the training set in the intervals and compute a single (constant) effect in each interval. With DALE extension, bin spliting is decoupled from instant effect estimation. With our extension for measuring the heterogeneous effects, we transfrom interval spliting from a step to a clustering problem with a meaningful objective to minimise. We provide a thorough analysis, where we show that our objective has a consistent meaning. It can be split in two parts; the first part is the unavoidable uncertainty due to the natural characteristics of the experiment, i.e., the data generating distribution and the black-box function. The second part is an added uncertainty due to the limited-samples estimation, that enforces to create groups with constant main effect. We opt for minimizing the objective, i.e. sum of the two uncertainties, that given that the first uncertainty is independent of the bin spliting, therefore we want to minimize the added uncertainty. To conclude, we transform the axis-spliting into an unsupervised clustering problem with a principled objective. We a computationally-grounded solution that works out-of-the-box, relaxing the user from providing a hyperparameter without any indication which one is the correct. This step can be used independently of whether the user wants to explore the heterogeneous effects or not.

[1] An extensive list of all symbols used in the paper is provided in the helping material.

3 THE ... METHOD

3.1 ALE with Uncertainty Quantification

ALE defines the local effect of x_s on $f(\cdot)$ at $(x_s, \mathbf{x_c})$ as $\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c})$. Given that the black-box function $f(\cdot)$ is a deterministic predictor, the local effect is also a deterministic quantity. Global methods summarize the local explanations into a single quantity. Therefore, ALE summarizes all the local explanations at x_s by averaging the local effects accross all values of $\mathbf{x_c}$ weighting them by $p(\mathbf{x_c}|x_s)$. In other words, ALE globalizes the local explanations at x_s through the expected change in the output y wrt x_s:

$$\mu(x_s) = \mathbb{E}_{\mathbf{x_c}|x_s}\left[\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c})\right] \tag{2}$$

As described at the Introduction, limiting the explanation at this level does not shed light to possible heterogeneous effects behind the averaged explanation. Therefore, we model the model the uncertainty of the effect at x_s as the variance of the local explanations:

$$\sigma^2(x_s) = \mathrm{Var}_{\mathbf{x_c}|x_s}\left[\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c})\right] \tag{3}$$

The uncertainty of the explanation emerges from the natural characteristics of the experiment, i.e., the data generating distribution and the black-box function. It is important to clarify that the variance is only a way to model the uncertainty. Other statistical properties can also be used. Elaborate.

ALE computes the final effect at x_s by accumulating/integrating the averaged local effects $\mu(x_s = z)$ over all values of z from x_{min} until x_s, as show in Eq. (1). The choice for x_{min} is not important, as it only affects the centering of ALE plot along the vertical axis. Equivalently, we define the accumulated uncertainty until the point x_s, as the integral of the variance of local effects:

$$f^{\sigma^2}_{\mathtt{ALE}}(x_s) = \int_{z_1,min}^{x_s} \sigma^2(z)\partial z \tag{4}$$

In ALE plots the absolute value at a specific point is not an interpretable quantity (add ref to paper). The meaningful interpretation is the effect at a specific point, i.e. what happens to the output given a small change in the feature of interest. The integration takes place only for making the visual interpretation smooth, for example, to spot larger intervals where the effect of a **local** change is continuously negative. The same stands for the accumulated uncertainty. It is meaningful to measure, only, the uncertainty at at a specific point to understand the heterogeneous effects behind the averaged effect. In the experimental Section, we propose appropriate visualizations to concentrate the user to this effect. However, the accumulated uncertainty, i.e. the aggregate uncertainty allong all the points of the plot, will help us define a meaningfull objective for the interval spliting step. (TODO: check here if the idea about whether the accumulated uncertainty is a good metric for modeling the interaction strength)

3.2 Uncertainty Quantification at an interval

In real scenarios, we have ignorance about the data-generating distribution $p(x_s, \mathbf{x_c})$ and on estimations based on the limited instances of the training set. Therefore, it is impossible to estimate Eqs. (2), (3) at the granularity of a point x_s since the possibility to observe a sample in the interval $[x_s - h, x_s + h]$ is zero, when $h \to 0$. Therefore, for estimation purposes, we are obliged to create larger intervals ($h > 0$). The mean effect at the interval $[z_1, z_2)$ is the mean of the expected effects:

$$\mu(z_1, z_2) = \frac{1}{z_2 - z_1}\int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x_c}|x_s=z}\left[\frac{\partial f}{\partial x_s}\right]\partial z \tag{5}$$

Accordingly, the accumulated uncertainty (variance) at the interval $[z_1, z_2)$ is:

$$\sigma^2(z_1, z_2) = \int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x_c}|x_s=z}\left[(\frac{\partial f}{\partial x_s} - \mu(z_1, z_2))^2\right]\partial z \tag{6}$$

Theorem 1. The accumulated variance at the interval $[z_1, z_2)$ is the accumulation of the all variances inside the interval plus the accumulated squared residuals:

$$\sigma^2(z_1, z_2) = \int_{z_1}^{z_2} \sigma^2(z) + \rho^2(z)\partial z \tag{7}$$

where $\rho(z) = \mu(z) - \mu(z_1, z_2)$. The proof is at the Appendix. Theorem 1 decouples the (observable) variance in the interval $[z_1, z_2)$ into two terms. The first term $\int_{z_1}^{z_2}\sigma^2(z)\partial z$ accumulates the variance of the local explanations (Eq. (3)) for all points inside the interval. The second term accumulates the squared differences between the mean effect at the interval and mean of the local explanations for all points inside the interval. The first term accumulates the the uncertainty due to the natural characteristics of the experiment, whereas, the second term adds extra uncertainty due to limiting the granularity of the effect at the bin level.

3.3 Estimation of Effect at an Interval

As stated before, Eqs. (5), (6) can be estimated from the set $\mathcal{S}$ of the instances of the training set with the s-th feature lying inside the interval, i.e., $\mathcal{S} = \{\mathbf{x}^i : z_1 \leq x^i_s < z_2\}$. The mean effect at the interval, Eq. (5), can be approximated by:

$$\hat{\mu}(z_1, z_2) = \frac{1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} \left[\frac{\partial f}{\partial x_s}(\mathbf{x}^i) \right] \quad (8)$$

and the accumulated variance at the interval, Eq. (6) can be approximated by

$$\hat{\sigma}_k(z_1, z_2) = \frac{1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} \left(\frac{\partial f}{\partial x_s}(\mathbf{x}^i) - \hat{\mu}_k(z_1, z_2) \right)^2 \quad (9)$$

The approximation is unbiased only if the points are uniformly distributed in $[z_{k-1}, z_k]$. (TODOs: Check what happens otherwise).

3.4 Bin Spliting as a Clustering Problem

Uncertainty of the global effect. Eq. (9) gives an approximation of the uncertainty of the bin effect. The uncertainty of the global effect is simply the sum of the uncertainties in the bin effects.

Minimizing the uncertainty Solving the problem of finding (a) the optimal number of bins K and (b) the optimal bin limits for each bin $[z_{k-1}, z_k] \forall k$ to minimize:

$$\mathcal{L} = \sum_{k=0}^{K} \hat{\sigma}_k(z_{k-1}, z_k) \quad (10)$$

The constraints are that all bins must include more than τ points, i.e., $|\mathcal{S}_k| \geq \tau$.

TODOS. Show theoretically that $\mathcal{L} \geq \int_{x_{s,\min}}^{x_{s,\max}} \sigma^2(x_s) \partial x_s$

3.4.1 Methodology

3.4.2 Algorithms

4 SYNTHETIC EXAMPLES

5 REAL-WORLD EXAMPLES

Acknowledgements

All acknowledgments go at the end of the paper, including thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support. To preserve the anonymity, please include acknowledgments *only* in the camera-ready papers.

References

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Repid: Regional effect plots with implicit interaction detection. In *International Conference on Artificial Intelligence and Statistics*, pages 10209–10233. PMLR, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer, 2020.

Appendix

5.1 Proof for variance of the bin