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 not only predict accurately but also explain their predictions. Therefore, Explainable AI (XAI), the field that provides interpretations for the prediction mechanism of complex black-box models, has increased interest. 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, usually a number or a plot. 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 concise global one comes at a cost. Under strong feature interactions, 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(Grömping, 2020) 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 in others. There are three widely-used feature effect methods; Partial Dependence Plots (PDP)(Friedman, 2001), Marignal Plots (MP)(Apley and Zhu, 2020) and Aggregated Local Effects (ALE)(Apley and Zhu, 2020). PDP and MP have been criticized for computing erroneous effects when the input features are (highly) correlated, which is the case in most ML setups. Therefore, ALE has been established as the state-of-the-art feature effect method.

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 PDP, 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. So far, this splitting is done in a blind way, i.e., the user does not have an indication about the parameter K, which often leads 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 behind the expected explanation. The probabilistic extension completes ALE, as ICE plots complement PDP, for revealing the heterogeneous effects. Our method also transforms the axis-splitting step into an unsupervised clustering problem, i.e., we search for the optimal splitting given the instances of the training set. We, therefore, relieve the user from the obligation of defining a parameter K by providing an automated method for finding the optimal bin splitting to robustly estimate (a) the global (expected) effect and (b) the uncertainty of the explanation, from the limited samples of the training set.

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

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 axis-splitting for robustly estimating the 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 $(x_s, \mathbf{x}_c)$, 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 $f : \mathbb{R}^D \to \mathbb{R}$ 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. As stated at the Introduction, the three well-known feature effect methods are: PDPs, MPs and ALE. PDPs formulate the feature effect of the s-th attribute as an expectation over the marginal distribution $\mathbf{X}_c$, i.e., $f^{\texttt{PDP}}(x_s) = \mathbb{E}_{\mathbf{X}_c}[f(x_s, \mathbf{X}_c)]$. MPs formulate it as an expectation over the conditional $\mathbf{X}_c|X_s$, i.e., $f^{\texttt{MP}}(x_s) = \mathbb{E}_{\mathbf{X}_c|X_s=x_s}[f(x_s, \mathbf{X}_c)]$. ALE computes the global effect at x_s as the accumulation of the averaged local effects, which are defined as the change on the output, i.e. $\frac{\partial f(x_s, \mathbf{X}_c)}{\partial x_s}$:

$$f^{\texttt{ALE}}(x_s) = \int_{z_{s,min}}^{x_s} \mathbb{E}_{\mathbf{X}_c|X_s=z} \left[\frac{\partial f(z, \mathbf{X}_c)}{\partial z} \right] \partial z \qquad (1)$$

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

ALE has specific advantages which gain particular value in cases of corelated input features. In these cases, PDPs integrate over unrealistic instances, due to the use of the marginal distribution $p(\mathbf{X}_c)$, and MPlots compute aggregated effects, i.e., impute the combined effect of sets of features to a single feature. ALE manages to resolve both issues, and is therefore the only trustable method in cases of correlated features.

Quantify the Heterogeneous Effects. Feature effect methods answer the question *what happens (effect) to the output, if I increase/decrease the value of a specific feature*. Having answered the question above, it comes naturally to also wonder *how certain we are about the change on the output*. For this reason, a lot of interest is given lately for quantifying the level of uncertainty, along with the expected effect. The level of uncertainty is mostly quantified by measuring the existence of heterogeneous effects, i.e. whether there are local explanations that deviate from the expected global effect. ICE and d-ICE plots provide a visual understanding of the heterogeneous effects on top of PDPs. Another approach targets on grouping the heterogeneous effects, e.g., allocating ICE plots in homogeneous clusters, by dividing the input space. Some other approaches, like H-Statistic, Greenwel, move a step behind and try to quantify the level of interaction between the input features, a possible cause of heterogeneous effects. In this case, the interpretation is indirect, since a strong interaction index is only an indicator of heterogeneous effects. The aforementioned approaches face two pathogenies; They either do not quantify the uncertainty of the feature effect directly or they are based on PDPs, and, therefore, they are subject to the failure modes of PDPs in cases of correlated features. To the best of our knowledge, no method so far targets on quantify the uncertainty of the feature effect as it is modelled by ALE.

Cluster Instances with homogeneous effects. In real ML scenarios, the expected feature effect and the uncertainty are estimated from the limited instances of the training set. ALE approximation requires an additional step, where the axis of the s-th feature is split into a sequence of non-overlaping bins and a single effect (expectation and uncertainty) is computed from the population of instances that lie inside each bin. (Apley and Zhu, 2020) proposed estimating the local effects in each bin by evaluating the black box-funtion at the bin limits:

$$\hat{f}^{\texttt{ALE}}(x_s) = \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[f(z_k, \mathbf{x_c}^i) - f(z_{k-1}, \mathbf{x_c}^i) \right] \qquad (2)$$

We denote as k_x the index of the bin that x_s belongs to, i.e. $k_x : z_{k_x-1} \leq x_s < z_{k_x}$ and $\mathcal{S}_k$ is the set of training

instance that lie in the k-th bin, i.e. $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^i < z_k\}$. In contrast, (cite) proposed the Differential ALE (DALE) estimation for quantifying the local effects on the training-set instances, instead of the bin limits:

$$\hat{f}^{\texttt{DALE}}(x_s) = \Delta x \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \frac{\partial f}{\partial x_s}(\mathbf{x}^i) \tag{3}$$

Their method has the advantages of remaining on-distribution even when bins become wider and, most importantly, it allows the recomputation of the accumulated effect with different bin-splitting in near-zero computational overhead. However, none of the approximations above deals with the crucial problem of the optimal bin-spliting. They partition the axis in K equally-sized bins without considering the properties of the underlying data, which can lead to erroneous approximations.

Instead, we propose treating the axis-spliting step as an unsupervised clustering problem. The objective of the clustering problem should fulfill in the best way to contradictory ojectives. First, secure robust estimations of the expected effect and the uncertainty inside each bin given the limited instances of the training set and, second, create bins with as homogeneous local effects as possible, for not losing fine-grain resolution feature effects due to wide bins.

3 THE NAME METHOD

The NAME method extends the traditional ALE method with a component for uncertainty quantification and improves the ALE estimation by automatically discovering the optimal bin-splitting. In Section 3.1 we define the component for the uncertainty quantification. In Section 3.2, we show how to estimate NAME from the limited samples of the training set and we make an important proof about the aggregated variance defined over a bin. In Section 3.3, we define and solve the problem of optimal bin-splitting. Finally, in Section 3.4, we illustrate the appropriate visualization of NAME for facilitating its interpretation by a non-expert and we discuss important aspects of the method

3.1 Uncertainty Quantification

ALE defines the local effect of the s-th feature on $f(\cdot)$ at point $(x_s, \mathbf{x_c})$ as $\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c})$. All the local explanations at x_s are, then, weigthed by the conditional distribution $p(\mathbf{x_c}|x_s)$ and are averaged, to produce the summarized effect at x_s:

$$\mu(x_s) = \mathbb{E}_{\mathbf{X}_c|x_s}\left[\frac{\partial f}{\partial x_s}(x_s, \mathbf{X}_c)\right] \tag{4}$$

The feature effect at x_s is the accumulation of the averaged local effects from $x_{s,min}$ until x_s, i.e. $f^{\texttt{ALE}}(x_s) = \int_{x_{s,min}}^{x_s} \mu(z)\partial z$. As described at the Introduction, limiting the explanation to the exepected value level does not shed light to possible heterogeneous effects behind the averaged explanation. Therefore, we model the uncertainty of the local effects at $\mathcal{H}(x_s)$ as the variance of the local explanations:

$$\mathcal{H}(x_s) := \sigma^2(x_s) = \mathrm{Var}_{\mathbf{X}_c|x_s}\left[\frac{\partial f}{\partial x_s}(x_s, \mathbf{X}_c)\right] \tag{5}$$

The uncertainty of the explanation emerges from the natural characteristics of the experiment, i.e., the data generating distribution and the properties of the black-box function. In Section 3.4, we propose appropriate visualizations for easier interpretation of Eq. (5). In ALE, the feature effect at x_s is the accumulation of the averaged local effects from x_{min} until x_s, as shown in Eq. (1). Equivalently, we define the accumulated uncertainty (variance) until x_s, as the integral of the variances of the local effects:

$$f_{\sigma^2}^{\texttt{ALE}}(x_s) = \int_{z_{s,min}}^{x_s} \sigma^2(z)\partial z \tag{6}$$

The accumulated uncertainty is not a directly interpretable quantity. It only helps us define a sensible objective for the interval spliting step, as we discuss in Section 3.3.

3.2 Interval-Based Estimation

In real scenarios, we have ignorance about the data-generating distribution $p(x_s, \mathbf{x}_c)$, so, the estimations are based on the limited instances of the training set. Estimating Eqs. (4), (5) at the granularity of a point is impossible, because the probability of observing a sample inside the interval $[x_s - h, x_s + h]$ tends to zero, when $h \to 0$. We are, therefore, obliged to split the axis of x_s into a sequence of non-overlaping intervals (bins) and estimate the mean and the variance from the samples that lie inside each bin. The mean effect at an interval $[z_1, z_2)$ is defined as the mean of the expected effects:

$$\mu(z_1, z_2) = \frac{1}{z_2 - z_1}\int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x_c}|x_s=z}\left[\frac{\partial f}{\partial x_s}\right]\partial z \tag{7}$$

Similarly, the accumulated variance at an interval $[z_1, z_2)$ is defined as:

$$\sigma^2(z_1, z_2) = \int_{z_1}^{z_2} \mathbb{E}_{\mathbf{x_c}|x_s=z}\left[(\frac{\partial f}{\partial x_s} - \mu(z_1, z_2))^2\right]\partial z \tag{8}$$

Theorem 1. If we define the residual $\rho(z)$ as the difference between the expected effect at x_s and the mean expected effect at the interval, i.e $\rho(z) = \mu(z) - \mu(z_1, z_2)$,

then, the accumulated variance at an interval $[z_1, z_2)$ is the accumulation of the all variances plus the accumulation of squared residuals inside the interval:

$$\sigma^2(z_1, z_2) = \int_{z_1}^{z_2} \sigma^2(z) + \rho^2(z)\partial z \tag{9}$$

The proof is in the Appendix. Theorem 1 decouples the accumulated variance at an interval into two terms. The first term $\int_{z_1}^{z_2} \sigma^2(z)\partial z$, quantifies the aggregated uncertainty due to the natural characteristics of the experiment $\mathcal{H}$ and the second term adds extra nuisance uncertainty due to the limited resolution $\mathcal{H}_n$. In other words, enforcing the computation of a single effect for all points in $[z_1, z_2)]$, burdens the estimation inside the interval with a nuisance uncertainty of:

$$\mathcal{H}_{bin}(z_1, z_2) = \int_{z_1}^{z_2} \mathcal{H}(z) + \mathcal{H}_n(z)\partial z \tag{10}$$

Eqs. (7), (8) can be directly estimated from the set $\mathcal{S}$ of the dataset instances with the s-th feature lying inside the interval, i.e., $\mathcal{S} = \{\mathbf{x}^i : z_1 \leq x_s^i < z_2\}$. The mean effect at the interval, Eq. (7) is approximated by:

$$\hat{\mu}(z_1, z_2) = \frac{1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} \left[\frac{\partial f}{\partial x_s}(\mathbf{x}^i) \right] \tag{11}$$

and the accumulated variance, Eq. (8) can be approximated by

$$\hat{\sigma}^2(z_1, z_2) = \frac{z_2 - z_1}{|\mathcal{S}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}} \left(\frac{\partial f}{\partial x_s}(\mathbf{x}^i) - \hat{\mu}(z_1, z_2) \right)^2 \tag{12}$$

The approximation is unbiased only if the points are uniformly distributed in $[z_1, z_2]$. Elaborate.

3.3 Interval Spliting As A Clustering Problem

We formulate the axis-splitting as an unsupervised clustering problem. We search for the optimal bin splitting, i.e., the number and size of consecutive non-overlapping intervals that minimizes the accumulated variance. The optimization problem is defined as follows:

$$\begin{aligned} \min_{\{z_0,\ldots,z_K\}} \quad & \mathcal{L} = \sum_{k=1}^{K} \hat{\sigma}^2(z_{k-1}, z_k) \\ \text{s.t.} \quad & |\mathcal{S}_k| \geq N \end{aligned} \tag{13}$$

The objective of the optimization problem is the sum of the accumulated variances in each bin, estimated by the instances of the training set. As we prove in Theorem 1, $\mathcal{L} = \sum_{k=1}^{K} \mathcal{H}_{bin}(z_{k-1}, z_k)$ is the sum of the uncertainty of the local explanations $\mathcal{H}$ due to heterogeneous effects and the added nuisance uncertainty due to bin-splitting $\mathcal{H}_n$. Given that $\mathcal{H}$ is not affected by interval splitting, optimizing $\mathcal{L}$ equals to finding the sequence of non-overlaping bins that add the minimum nuisance uncertainty. The restriction of (13) secures that each bin is populated with enough samples for a robust estimation, which is set by the parameter N. In other words, the user makes the following proposal; *find the bin sequence that adds the minimum nuisance uncertainty, given that I want at least N points per bin*.

3.3.1 Algorithm For Solving The Clustering Problem

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:

$$\mathcal{L} = \sum_{k=0}^{K} \hat{\sigma}_k(z_{k-1}, z_k) \tag{14}$$

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 Visualization of NAME Method and Discussion

- Discuss about the meaning of ALE, to find intervals with some effect

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 statisti-

cal learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

Ulrike Grömping. Model-agnostic effects plots for interpreting machine learning models, 03 2020.

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.