Paper summary

Vasilis Gkolemis

July 2021

1 Introduction

Description. An introductory paragraph for interpretable ML

XAI literature distinguishes between local and global interpretation methods [4]. Local methods explain a specific prediction, whereas global methods explain the entire model behavior. Global methods can 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 [2] that exist under the hood, a phenomenon called aggregation bias [3].

Feature effect forms a fundamental category of global explainability methods. The goal of the feature effect is to isolate the average impact of a single feature 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 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 due to isolating feature effects in situations where input features are highly correlated.

However, ALE faces two crucial drawbacks. First, it lacks a way to inform the user about potential heterogeneous effects that hide behind the average effect. In contrast, in Partial Dependence Plots (PDP), the heterogeneous effects can be spotted by exploring the Individual Conditional Expectations (ICE). Second, ALE require an additional step, where the input space 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 results.

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 a local individual drawn at random, or, in other words, to what extent heterogeneous effects are hidden behind the expected explanation. The uncertainty of the global explanation emerges from the natural characteristics of the experiment, i.e., the data generating distribution and the black-box function.

Furthermore, we automate the step of splitting the axis into non-overlapping intervals, transforming it into an unsupervised clustering problem, which we solve automatically using dynamic programming. The objective of the clustering problem has as lower-bound the (unavoidable) heterogeneity, i.e., the aggregated uncertainty of the global explanation. The restriction of the clustering is to create groups of samples for a robust estimation estimating the feature effect (expected value) and the uncertainty of the explanation (variance), we reside on the limited instances of the training set. Therefore, we aim to find the optimal grouping of samples that adds the slightest uncertainty over the unavoidable heterogeneity, but with enough samples. We finally solve the minimization problem by finding the global optimum using dynamic programming. Our method works out of the box without requiring any input by the user. We provide a theoretical and empirical evaluation of our method.

Contribution. The contribution of our paper in bullets:

- Reformulation of ALE method to quantify the uncertainty of the global explanation
- Formal definition of the variable-size interval splitting as an unsupervised clustering problem
- Method for finding a global optimum in the clustering problem
- Theoretical evaluation of our method (e.g. show that the objective's lower bound is the unavoidable heterogeneity due to the characteristics of the problem, highlight specific cases, e.g. specific generative distribution and specific black-box function)
- Empirical evaluation of the method in artificial and real datasets

2 Mathematical formulation

Effect at point x_s. In the intro, we described the *uncertainty of the global explanation* as a metric of how certain we are that the global explanation is valid if applied to a randomly-drawn individual. ALE plots measure the s-th feature effect at point x_s as the **expected** change in the output y, if we slightly change the value of the feature of interest x_s:

$$\mu(x_s) = \mathbb{E}_{\mathbf{x_c}|x_s}\left[\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c})\right] \tag{1}$$

We model the s-th feature effect at point x_s (change in the output y wrt a slight change in the feature of interest x_s) as the random variable $\mathbf{\Delta Y}; x_s$. For notation convenience, we will refer to the s-th feature effect as $\mathbf{\Delta Y}$, ommiting the $; xs$ part. The randomness has its origins in the ignorance of the values of the rest of the features, denoted with $\mathbf{X}_c$. Therefore, the s-th feature effect is defined as:

$$\mathbf{\Delta Y} = g(x_s) = \frac{\partial f}{\partial x_s}(x_s, \mathbf{X}_c) \tag{2}$$

As shown in Eq. (1), ALE is only interested in the expected value of $\mathbf{\Delta Y}$. Instead, we are also interested in the variance of $\mathbf{\Delta Y}$ for measuring the uncertainty of the local change:

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

The variance in Eq. (3) informs us about the heterogeneous effects hiding behind the explanation.

Effect at interval $[z_{k-1}, z_k]$. In real scenarios, we have ignorance about the generative distribution $p(x_s, \mathbf{x}_c)$, residing to Monte-Carlo approximations of $\mu(x_s)$, $\sigma^2(x_s)$ using the samples of the training set. Therefore, it is impossible to estimate Eqs. (1), (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, in the limit where $h \to 0$. Therefore, we are obliged to estimate the local effect in larger intervals ($h > 0$). We refer to the expected value and the variance of the feature effect at the interval $[z_{k-1}, z_k)$, as:

$$\mu_k = \mu(z_{k-1}, z_k) = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \mathbb{E}_{\mathbf{x_c}|x_s=z}\left[\frac{\partial f}{\partial x_s}\right] \partial z \tag{4}$$

$$\sigma_k^2 = \sigma^2(z_{k-1}, z_k) = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \mathbb{E}_{\mathbf{x_c}|x_s=z}\left[(\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c}) - \mu_k)^2\right] \partial z \tag{5}$$

We prove that the interval-variance σ_k^2 is:

$$\sigma_k^2 = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \sigma^2(z) + \rho^2(z) \partial z \tag{6}$$

where $\rho(x_s) = \mu(x_s) - \mu_k$. The proof is at Section 5. We observe that the interval-variance is the mean point-variance of the points inside the interval $[z_{k-1}, z_k]$, plus the mean of the residual term ρ. The mean point-variance is the unavoidable variance, due to the uncertainty of the global explanation. The mean of the residual term is an extra variance (uncertainty) due to limiting the granularity of the effect at the bin level.

Approximation of effect at interval $[z_{k-1}, z_k]$. Eqs. (4), (5) are approximated by the instances of the training set that lie inside the k-th interval, i.e. $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^i < z_k\}$:

$$\hat{\mu}(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[\frac{\partial f}{\partial x_s}(\mathbf{x}^i) \right] \tag{7}$$

$$\hat{\sigma}_k(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[\frac{\partial f}{\partial x_s}(\mathbf{x}^i) - \hat{\mu}_k(z_{k-1}, z_k) \right]^2 \tag{8}$$

Uncertainty of the global effect. Eq. (8) 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. The approximation is unbiased only if the points are uniformly distributed in $[z_{k-1}, z_k]$. (TODOs: Check what happens otherwise).

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:

$$\mathcal{L} = \sum_{k=0}^{K} \hat{\sigma}_k(z_{k-1}, z_k) \tag{9}$$

The constraints are that all bins must include more than τ points, i.e., $|\mathcal{S}_k| \geq \tau$.
TODOS. Show theoretically that $\mathcal{L} \geq \int_{x_{s,\min}}^{x_{s,\max}} \sigma^2(x_s) \partial x_s$

Uncertainty of the approximation. In all experiments, it also important to measure the uncertainty of the approximation. The uncertainty of the approximation can be quantified with two approaches:

- Splitting the dataset in many folds and (re)estimating $\hat{\mu}(z_{k-1}, z_k), \hat{\sigma}_k(z_{k-1}, z_k)$
- Using the central limit theorem, we can (under assumptions) say that the standard error of the approximation in eq. (7) is std_error $= \frac{\hat{\sigma}_k}{\sqrt{|\mathcal{S}_k|}}$.

3 Toy Example

Example 1. We use the following example:

$$\begin{gathered} f(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 \\ x_1 \sim \mathcal{U}(0, 1) \\ x_2 = x_1 + \epsilon, \epsilon \sim \mathcal{N}(\mu = 0, \sigma_2^2) \end{gathered} \tag{10}$$

Therefore, according to Eq. (1) $\mu_(x_1) = b_1 + b_3 x_1$ and according to Eq. (3) $\sigma^2(x_1) = b_3^2 \sigma_2^2 \Rightarrow \sigma(x_1) = b_3 \sigma_2$.

ALE with uncertainty. In figure 1 we observe the ALE effect with uncertainty for 3 different values of $\sigma_2 = \{0.01, 0.1, 1\}$. In all case the mean effect (ALE) is the same, but the uncertainty of the global explanation quantifies the impact of the heterogeneous effects hiding behind the global explanation.

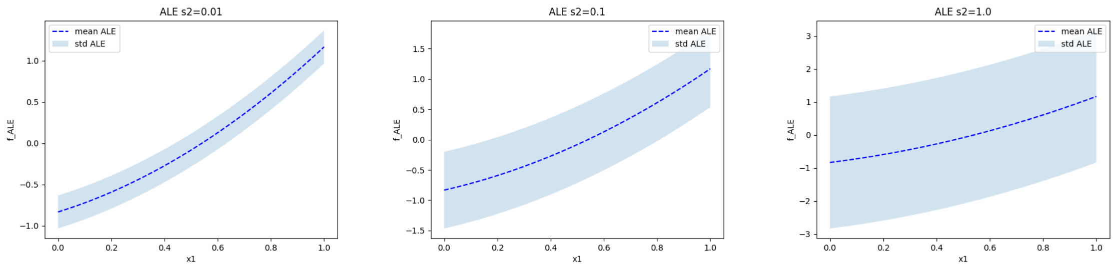

Figure 1: ALE (a) $\sigma_2^2 = 0.01$, (b) $\sigma_2^2 = 0.1$, (c) $\sigma_2^2 = 1$.

Bin Splitting. In figure 2, set up for $b_0 = 0, b_1 = 1, b_2 = 1, b_3 = 2, \sigma_2^2 = 0.1$. Therefore, the unavoidable uncertainty is $\sigma^2(x_1) = b_3^2 \sigma_2^2 = 0.4$. We observe that as the bins become denser the added uncertainty due to binning becomes less. But as the bins become denser, fewer points lie inside them and the estimation is poor. The vertical lines show the maximum number of bins for dataset sizes, if I want at least 25 points per bin.

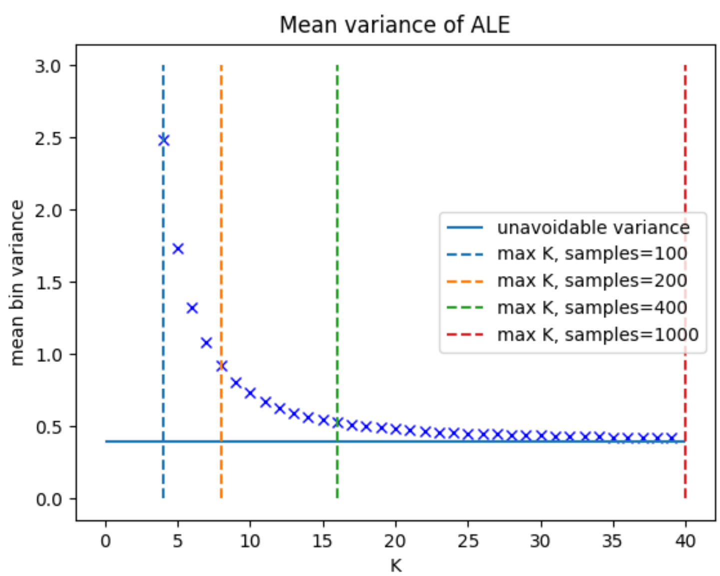

Figure 2:

4 Evaluation

Experiments.

Metrics.

5 Proofs

Given that:

$$\sigma_k^2 = \sigma^2(z_{k-1}, z_k) = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \mathbb{E}_{\mathbf{x_c}|x_s=z} \left[(\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c}) - \mu_k)^2 \right] \partial z$$

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

$$\rho(x_s) = \mu(x_s) - \mu_k$$

We want to prove that:

$$\sigma_k^2 = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \sigma^2(z) + \rho^2(z) \partial z$$

Proof:

$$\sigma_k^2 = \sigma^2(z_{k-1}, z_k) = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \mathbb{E}_{\mathbf{x_c}|x_s=z} \left[(\frac{\partial f}{\partial x_s}(z, \mathbf{x_c}) - \mu_k)^2 \right] \partial z \tag{11}$$

$$= \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \mathbb{E}_{\mathbf{x_c}|x_s=z} \left[(\frac{\partial f}{\partial x_s} - \mu(z) + \rho(z))^2 \right] \partial z \tag{12}$$

$$= \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \left(\mathbb{E}_{\mathbf{x_c}|x_s=z} \left[(\frac{\partial f}{\partial x_s} - \mu(z))^2 \right] + \mathbb{E}_{\mathbf{x_c}|x_s=z} \left[\rho(z)^2 \right] + \mathbb{E}_{\mathbf{x_c}|x_s=z} \left[2(\frac{\partial f}{\partial x_s} - \mu(z))\rho(z) \right] \right) \partial z \tag{13}$$

$$= \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \left(\sigma^2(x_s) + \rho^2(z) + 2(\mu(z) - \mu(z))\rho(z) \right) \partial z \tag{14}$$

$$= \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \sigma^2(x_s) + \rho^2(z) \partial z \tag{15}$$

$$\tag{16}$$

References

[1] 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*, 82:1059–1086, 2020. The paper that proposed ALE plots.

[2] Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. Repid: Regional effect plots with implicit interaction detection. 2 2022.

[3] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 8 2019.

[4] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges, 10 2020.

[5] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models, 2022.