Paper summary

Vasilis Gkolemis

July 2021

1 Introduction

Description. XAI literature distinguishes between local and global interpretation methods ([4]). Global interpretation methods aim at explaining the overall behavior of an ML model. Many of these global interpretation methods are confounded by feature interactions, meaning that they can produce misleading explanations when feature interactions are present. In these cases, they often average individual effects of local interpretation methods and thereby obfuscate heterogeneous effects induced by feature interactions ([2]). This so-called aggregation bias ([3]) is responsible for producing global explanations that often conceal that individual instances may deviate from the global explanation. Therefore, global methods must quantify and inform about *the uncertainty of the global explanation*, i.e., how certain we are that a global explanation is valid if applied to a local individual drawn at random. 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.

In real ML scenarios, we do not know the data generating distribution for computing the expectations and the uncertainty. The local effects are estimated from the training set's limited instances, and the global effect is computed by aggregating them. Many methods, such as ALE, require an appropriate grouping of samples (partitioning of the feature space) for aggregating local effects that are as homogeneous as possible. If this grouping is not done correctly, it may erroneously mix heterogeneous local effects, not due to the unavoidable heterogeneity of the data generating mechanism but due to improper grouping.

ALE is a SotA method for measuring the global feature effect, but so far, it has two limitations. First, it does not quantify the uncertainty of the global explanation. Second, like most global explanation methods, it requires grouping together samples before computing the global effect. So far, this grouping is done by blindly splitting the feature space in K fixed-size non-overlapping intervals, where K is a hyperparameter provided by the user.

In this paper, first, we extend the ALE definition for quantifying the uncertainty of the global explanation. Second, we readjust ALE to work for variable-size bins and formulate the partitioning in non-overlapping intervals as an unsupervised clustering problem. In the unsupervised clustering problem, we minimize an objective which has as lower-bound the (unavoidable) heterogeneity, i.e. the aggregated uncertainty of the global explanation. Therefore, we aim to minimize the added uncertainty induced by the wrong grouping of samples. In other words, we aim to find the optimal grouping that adds no uncertainty over the unavoidable heterogeneity. 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.

Problem Statement - 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:

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

We argue that it is also crucial to compute the variance of the change as a measure of the uncertainty of the local change:

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

The variance in eq. (2) informs us about the heterogeneous effects hiding behind the explanation. The heterogeneous effects is a consequence of the uncertainty about the values of $\mathbf{x_c}$, i.e. the uncertainty introduced by the conditional distribution $p(\mathbf{x_c}|x_s)$, and the formula of the black-box function f.

Effect at interval $[z_{k-1}, z_k]$. In real scenarios, we approximate $\mu(x_s)$, $\sigma^2(x_s)$ using the finite number of samples the are available in the training set. Since the possibility to have a training instance in the interval $[x_s - h, x_s + h]$ is zero in the limit where $h \to 0$, we are obliged to estimate the local effect in larger neighboors. The mean and variance of the interval-effect defined in $z_{K-1} \leq x_s < z_k$ are:

$$\mu_k(z_{k-1}, z_k) = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} \mathbb{E}_{\mathbf{x_c}|x_s}\left[\frac{\partial f}{\partial x_s}\right] \partial x_s \tag{3}$$

$$\sigma^2(z_{k-1}, z_k) = \frac{1}{z_k - z_{k-1}} \int_{z_{k-1}}^{z_k} p(\mathbf{x_c}|x_s)\left(\frac{\partial f}{\partial x_s}(x_s, \mathbf{x_c}) - \mu_k(z_{k-1}, z_k)\right)^2 \partial x_s \tag{4}$$

Eq. (4) a second term to the uncertainty of the global explanation defined in Eq.(2); the uncertainty about the exact position of x_S. Since we quantify the local effect up to the resolution of the interval, we mix together all the effects in the interval $z_{K-1} \leq x_s < z_k$.

Approximation of effect at interval $[z_{k-1}, z_k]$. Eqs. (3), (4) are approximated by the instances of the training set that lie inside the k-th interval, i.e. $\mathcal{S}_k = \{\mathbf{x}^i : z_{k-1} \leq x_s^i < z_k\}$:

$$\hat{\mu}(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[\frac{\partial f}{\partial x_s}(\mathbf{x}^i)\right] \tag{5}$$

$$\hat{\sigma}_k(z_{k-1}, z_k) = \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \left[\frac{\partial f}{\partial x_s}(\mathbf{x}^i) - \hat{\mu}_k(z_{k-1}, z_k)\right]^2 \tag{6}$$

Uncertainty of the global effect. Eq. (6) 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:

$$\mathcal{L} = \sum_{k=0}^{K} \hat{\sigma}_k(z_{k-1}, z_k) \tag{7}$$

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. (5) is std_error $= \frac{\hat{\sigma}_k}{\sqrt{|\mathcal{S}_k|}}$.

3 Evaluation

Experiments.

Metrics.

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.