DALE: Differential Accumulated Local Effects for efficient and accurate global explanations

Vasilis Gkolemis [1, 2] Theodore Dalamagas [1] Christos Diou [2]

[1]ATHENA Research Center [2]Harokopio University of Athens

TL;DR

DALE is a better approximation to ALE, the SotA feature effect method. By better, we mean faster and more accurate.

keywords: eXplainable AI, global, model-agnostic, deep learning

Motivation

Feature effect (FE) plots are simple and intuitive; they isolate the impact of a single feature x_s on the output y. By inspecting a FE plot, a non-expert can quickly understand whether a feature has positive/negative impact (and to what extent) on the target variable.

This simplicity comes at a cost; isolating the effect of a single variable on the output is tricky because normally, features are correlated and the black-box function learns complex input-output mappings. ALE (Apley and Zhu 2020) is the SotA feature effect method because it handles well correlated features. However, ALE estimation, i.e., the approximation of ALE from the instances of the training set, has some drawbacks; it becomes inefficient in high-dimensional datasets and it is vulnerable to creating synthetic out-of-distribution instances.

In this work, we analyze these drawbacks and propose Differential ALE (DALE), a novel approximation, that we addresses them.

DALE vs ALE

ALE definition

$$f(x_s) = \int_{x_{s,min}}^{x_s} \mathbb{E}_{\boldsymbol{x_c}|z} \underbrace{[\frac{\partial f}{\partial x_s}(z, \boldsymbol{x_c})]}_{\text{point effect}} \partial z$$

ALE defines the effect at $x_s = z$ as the expected change (derivative) on the output over the conditional distribution $\boldsymbol{x_c}|z$. The feature effect plot is the accumulation of the expected changes.

ALE approximation, i.e., estimating ALE from the training set $\mathcal{D}$, requires partitioning the s-th axis in K equisized bins. The value of the parameter K has crucial consequences on the final curve. If K is high (narrow bins), we get a high-resolution plot but with limited samples per bin (noisy estimation). If K is low (wide bins), we get a low-resolution plot but with more samples per bin (robust estimation).

ALE approximation

$$f(x_s) = \sum_{k}^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \underbrace{[f(z_k, \boldsymbol{x_c^i}) - f(z_{k-1}, \boldsymbol{x_c^i})]}_{\text{point effect}}}_{\text{bin effect}}$$

ALE computes the point effects by evaluating f at the bin limits: $[f(z_k, \boldsymbol{x_c^i}) - f(z_{k-1}, \boldsymbol{x_c^i})]$. This approach has some drawbacks:

1. requires $\mathcal{O}(N * D)$ evaluations of f
2. recomputes all effects from scratch if changing K
3. creates artificial samples, that my become OOD when the bin size is large

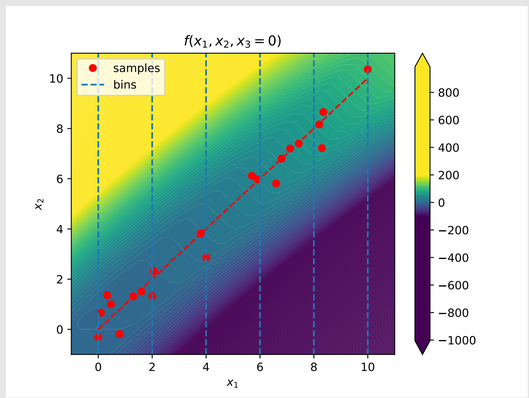
DALE approximation

$$f(x_s) = \Delta x \sum_{k}^{k_x} \underbrace{\frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} \underbrace{[\frac{\partial f}{\partial x_s}(x_s^i, \boldsymbol{x_c^i})]}_{\textbf{point effect}}}_{\text{bin effect}}$$

For adressing these issues we propose DALE, an alternative approximation that computes the point effects by evaluating the derivatives $\frac{\partial f}{\partial x_s}$ on the dataset instances.

1. requires $\mathcal{O}(N)$ evaluations of f
2. local effects are decoupled from bin size K
3. does not create artificial samples

DALE is accurate



Consider the following case; (a) we have limited samples (b) high variance in some features and (c) the black-box function changes abruptply outside of the data manifold. For example, $f(x_1, x_2, x_3) = x_1x_2 + x_1x_3 \pm g(x)$, with $x_1 \in [0, 10]$, $x_2 = x_1 + \epsilon$ and $x_3 \sim \mathcal{N}(0, \sigma^2)$. The term x_1x_3 makes estimations from limited samples (narrow bins) noisy, see Figure 1. If we use larger bins (more $\frac{\text{points}}{\text{bin}}$), DALE leads to a good estimation whereas ALE fails due to OOD samples.

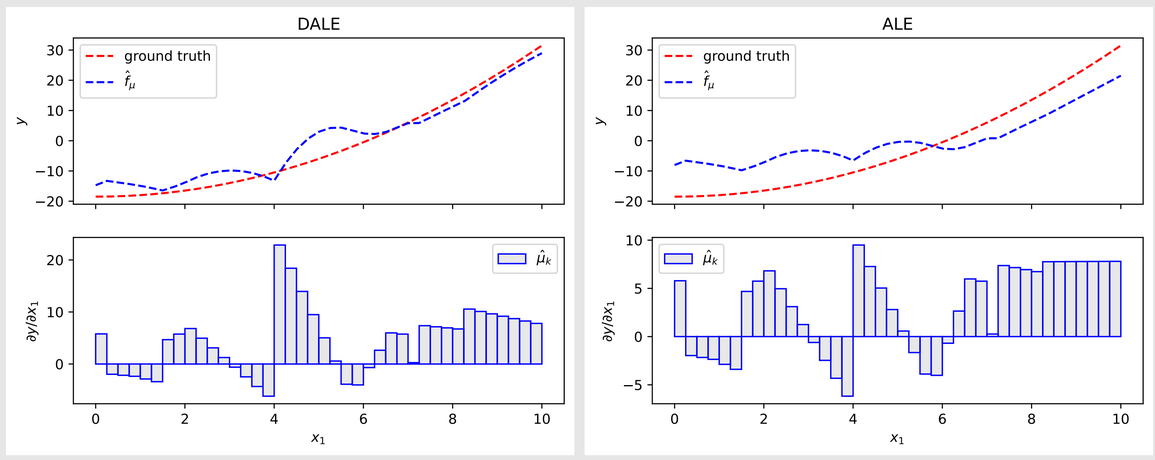


Figure 1. Narrow bins ($K = 40$) $\Rightarrow$ limited $\frac{\text{samples}}{\text{bin}}$ $\Rightarrow$ both plots are noisy

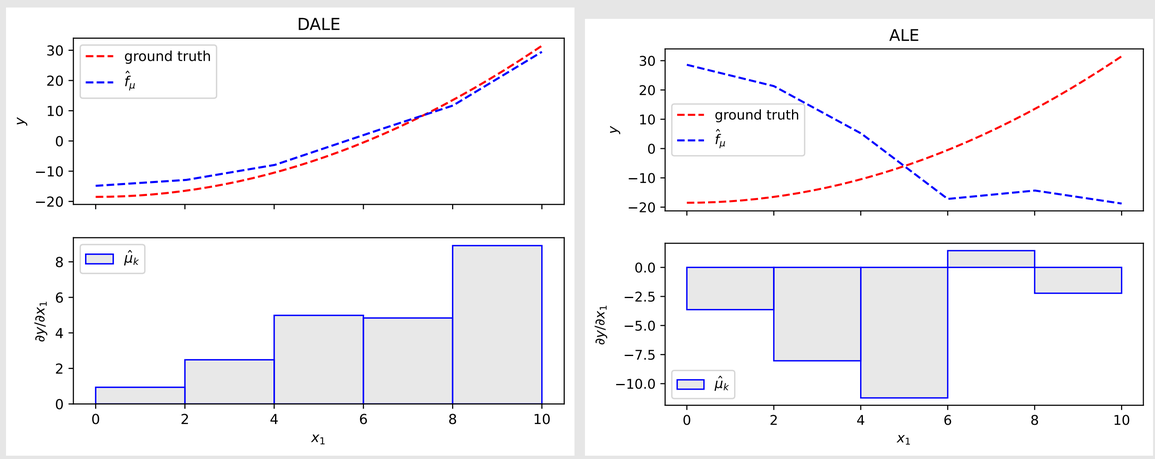


Figure 2. Wide bins ($K = 5$) $\Rightarrow$ many $\frac{\text{samples}}{\text{bin}}$ $\Rightarrow$ DALE is accurate, ALE is affected by OOD

DALE is fast

In a large and high dimensional dataset, ALE needs 10 mins, DALE some seconds! We test DALE vs ALE in two setups (Figure 3). The light and heavy setup differ in the size of the dataset ($N = 10^2$ vs $N = 10^5$ instances) and the cost of evaluating f (light vs heavy). In both cases, DALE scales much better wrt dimensionality D.

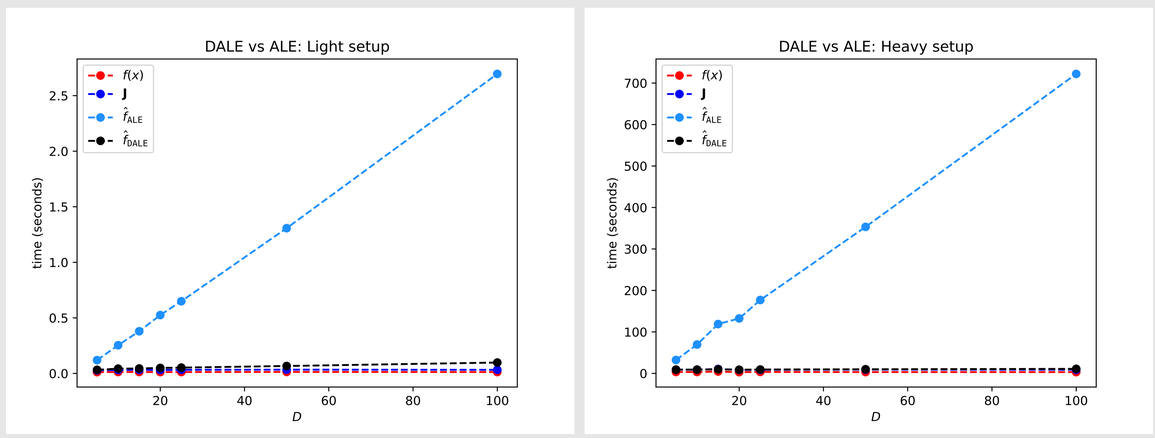


Figure 3. Light setup; small dataset ($N = 10^2$ instances), light f. Heavy setup; big dataset ($N = 10^5$ instances), heavy f

Conclusion

In case you work with a differentiable model, as in Deep Learning, use DALE to:

- compute the effect of all features efficiently
- test the FE plot for many different bin sizes K, without computational cost
- ensure on-distribution estimation, irrespectively of the bin size

References

Apley, D. W. and J. Zhu (2020). "Visualizing the effects of predictor variables in black box supervised learning models". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82.

- givasile.github.io, twitter.com/givasile1