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

Vasilis Gkolemis  
GKOLEMIS@HUA.GR, VGKOLEMIS@ATHENARC.GR  
Harokopio University of Athens, IMIS ATHENA Research Center

Theodore Dalamagas  
DALAMAG@ATHENARC.GR  
IMIS ATHENA Research Center

Christos Diou  
DIOU@HUA.GR  
Harokopio University of Athens

Editors: Emtiyaz Khan and Mehmet Gönen

1. Notation List

- $s$, index of the feature of interest
- $X_s$, feature of interest as a r.v.
- $X_c = (X/s)$, the rest of the features in as a r.v.
- $X = (X_s, X_c) = (X_1, \ldots, X_s, \ldots, X_D)$, all input features as r.v.
- $x_s$, feature of interest
- $x_c$, the rest of the features
- $x = (x_s, x_c) = (x_1, \ldots, x_s, \ldots, x_D)$, all the input features
- $X$, design matrix/training set
- $f(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}$, black box function
- $f_s(x) = \frac{\partial f(x_s, x_c)}{\partial x_s}$, the partial derivative of the $s$-th feature
- $D$, dimensionality of the input
- $N$, number of training examples
- $x^i$, $i$-th training example
- $x_s^i$, $s$-th feature of the $i$-th training example
- $x_c^i$, the rest of the features of the $i$-th training example
- $f_{ALE}(x_s) : \mathbb{R} \rightarrow \mathbb{R}$, ALE definition for the $s$-th feature
- $\hat{f}_{DALE}(x_s) : \mathbb{R} \rightarrow \mathbb{R}$, DALE approximation for the $s$-th feature

© 2022 V. Gkolemis, T. Dalamagas & C. Diou.
- $\hat{f}_{\text{ALE}}(x_s) : \mathbb{R} \rightarrow \mathbb{R}$, ALE approximation for the $s$-th feature
- $z_{k-1}, z_k$, the left and right limit of the $k$-th bin
- $\mathcal{S}_k = \{x^i : x^i_s \in [z_{k-1}, z_k]\}$, the set of training points that belong to the $k$-th bin
- $k_x$ the index of the bin that $x$ belongs to
- $\hat{\mu}_s^k$, DALE approximation of the mean value inside a bin, equals $\frac{1}{|\mathcal{S}_k|} \sum_{i : x^i \in \mathcal{S}_k} f_s(x^i)$
- $(\hat{\sigma}_s^k)^2$, DALE approximation of the variance inside a bin, equals $\frac{1}{|\mathcal{S}_k|-1} \sum_{i : x^i \in \mathcal{S}_k} (f_s(x^i) - \hat{\mu}_s^k)^2$

2. Derivation of equations in the Background section

In this section, we present the derivations for obtaining the feature effect at the Background.

**Example Definition.**

The black-box function and the generating distribution are:

$$f(x_1, x_2) = \begin{cases} 1 - x_1 - x_2, & \text{if } x_1 + x_2 \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$p(X_1 = x_1, X_2 = x_2) = \begin{cases} 1 & x_1 \in [0, 1], x_2 = x_1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$p(X_1 = x_1) = \begin{cases} 1 & 0 \leq x_1 \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$p(X_2 = x_2) = \begin{cases} 1 & 0 \leq x_2 \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$p(X_2 = x_2 | X_1 = x_1) = \delta(x_2 - x_1) \quad (5)$$

**PDPlots.**

The feature effect computed by PDP plots is:
$$f_{\text{PDP}}(x_1) =$$
$$= \mathbb{E}_{X_2}[f(x_1, X_2)]$$
$$= \int_{X_2} f(x_1, x_2) p(x_2) \, dx_2$$
$$= \int_0^{1-x_1} (1 - x_1 - x_2) \, dx_2 + \int_{1-x_1}^1 0 \, dx_2$$
$$= \int_0^{1-x_1} 1 \, dx_2 + \int_0^{1-x_1} -x_1 \, dx_2 + \int_0^{1-x_1} -x_2 \, dx_2$$
$$= (1 - x_1) - x_1(1 - x_1) - \frac{(1 - x_1)^2}{2}$$
$$= (1 - x_1)^2 - \frac{(1 - x_1)^2}{2}$$
$$= \frac{(1 - x_1)^2}{2}$$

Due to symmetry:
$$y = f_{\text{PDP}}(x_2) = \frac{(1 - x_2)^2}{2} \quad (7)$$

MPlots.

The feature effect computed by PDP plots is:

$$f_{\text{MP}}(x_1) =$$
$$= \mathbb{E}_{X_2|x_1=x_1}[f(x_1, X_2)]$$
$$= \int_{X_2} f(x_1, x_2) p(x_2|x_1) \, dx_2$$
$$= f(x_1, x_1) =$$
$$= \begin{cases} 
1 - 2x_1, & x_1 \leq 0.5 \\
0, & \text{otherwise} 
\end{cases} \quad (8)$$

Due to symmetry:
$$y = f_{\text{MP}}(x_2) = \begin{cases} 
1 - 2x_2, & x_2 \leq 0.5 \\
0, & \text{otherwise} 
\end{cases} \quad (9)$$

ALE

The feature effect computed by ALE is:
The normalization constant is:

\[
c = -E[\hat{f}_{ALE}(x_1)]
= -\int_{-\infty}^{\infty} \hat{f}_{ALE}(x_1)\,dz
\]

\[
= -\int_{0}^{0.5} -z\,dz - \int_{0.5}^{1} -0.5\,dz
= 0.25 + 0.25 = 0.375
\]

Therefore, the normalized feature effect is:

\[
y = f_{ALE}(x_1) = \begin{cases} 
0.375 - x_1 & 0 \leq x_1 \leq 0.5 \\
-0.125 & 0.5 < x_1 \leq 1 
\end{cases}
\]

Due to symmetry:

\[
y = f_{ALE}(x_2) = \begin{cases} 
0.375 - x_2 & 0 \leq x_2 \leq 0.5 \\
-0.125 & 0.5 < x_2 \leq 1 
\end{cases}
\]

3. First-order and Second-order DALE approximation

In the main part of the paper, we presented the first order ALE approximation as

\[
f_{DALE}(x_s) = \Delta x \sum_{k=1}^{k_s} \frac{1}{|S_k|} \sum_{i: x^i \in S_k} [f_s(x^i)]
\]
For keeping the equation compact, we omit a small detail about the manipulation of the last bin. In reality, we take complete $\Delta x$ steps until the $k_x - 1$ bin, i.e. the one that prepends the bin where $x$ lies in. In the last bin, instead of a complete $\Delta x$ step, we move only until the position $x$. Therefore, the exact first-order DALE approximation is

$$f_{DALE}(x_s) = \Delta x \sum_{k=1}^{k_x-1} \frac{1}{|S_k|} \sum_{i:x^i \in S_k} [f_s(x^i)] + (x - z(k_x-1)) \frac{1}{|S_{k_x}|} \sum_{i:x^i \in S_{k_x}} [f_s(x^i)]$$

Following a similar line of thought we define the complete second-order DALE approximation as

$$f_{DALE}(x_l, x_m) = \Delta x_l \sum_{p=1}^{p_x-1} \Delta x_m \sum_{q=1}^{q_x-1} \frac{1}{|S_{k,q}|} \sum_{i:x^i \in S_{k,q}} f_{l,m}(x^i) + (x_l - z(p_x-1))(x_m - z(q_x-1)) \frac{1}{|S_{p_x,q_x}|} \sum_{i:x^i \in S_{p_x,q_x}} f_{l,m}(x^i)$$

4. Second-order ALE definition

The second-order ALE plot definition is

$$f_{ALE}(x_l, x_m) = c + \int_{x_{l,min}}^{x_l} \int_{x_{m,min}}^{x_m} \mathbb{E}_{X_i|X_l=x_i,X_m=z_m} \frac{\partial^2 f(x)}{\partial x_i \partial x_m}$$

where $f_{l,m}(x) = \frac{\partial^2 f(x)}{\partial x_l \partial x_m}$.

5. DALE variance inside each bin

In this section, we show that the variance of the local effect estimation inside a bin, i.e. $\text{Var}[\hat{\mu}_k]$ equals with $\frac{(\sigma_k^*)^2}{|S_k|}$, where $\frac{(\sigma_k^*)^2}{|S_k|} = \text{Var}[f_s(x)]$.

$$\text{Var}[\hat{\mu}_k] = \text{Var} \left[ \frac{1}{|S_k|} \sum_{i:x^i \in S_k} f_s(x^i) \right]$$

$$= \frac{1}{|S_k|^2} \sum_{i:x^i \in S_k} \text{Var}[f_s(x^i)]$$

$$= \frac{|S_k|}{|S_k|^2} \text{Var}[f_s(x)]$$

$$= \frac{(\sigma_k^*)^2}{|S_k|}$$