

## 1. Notation List

We set the following notation rules: We refer to random variables (r.v.) using uppercase and calligraphic font  $\mathcal{X}$ , whereas to simple variables with plain lowercase  $x$ . Bold  $\mathbf{x}$  denotes a vector variable,  $\mathcal{X}_s$  the r.v. of the feature of interest and  $\mathcal{X}_c$  the rest of the features so that  $\mathcal{X} = (\mathcal{X}_s, \mathcal{X}_c)$  represents the input space. The black-box function is notated as  $f$  and the feature effect of the  $s$ -th feature as  $f_{\langle \text{method} \rangle}(x_s)$ , where  $\langle \text{method} \rangle$  is the name of the feature effect method. The extensive list of symbols used in the paper is:

- $s$ , index of the feature of interest
- $\mathcal{X}_s$ , feature of interest as a r.v.
- $\mathcal{X}_c = (\mathcal{X}_{/s}, )$ , the rest of the features in as a r.v.
- $\mathcal{X} = (\mathcal{X}_s, \mathcal{X}_c)$ , all input features as r.v.
- $x_s$ , feature of interest
- $\mathbf{x}_c$ , the rest of the features
- $\mathbf{x} = (x_s, \mathbf{x}_c)$ , all the input features
- $\mathbf{X}$ , training set
- $f(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}$ , black box function
- $D$ , dimensionality of the input
- $N$ , number of training examples
- $\mathbf{x}^i$ ,  $i$ -th training example
- $x_s^i$ ,  $s$ -th feature of the  $i$ -th training example
- $\mathbf{x}_c^i$ , the rest of the features of the  $i$ -th training example
- $f_{\text{ALE}}^s(x) : \mathbb{R} \rightarrow \mathbb{R}$ , feature effect computed by ALE for the  $s$ -th feature  $s$
- $f_{\text{DALE}}^s(x) : \mathbb{R} \rightarrow \mathbb{R}$ , feature effect computed by DALE for the  $s$ -th feature  $s$
- $\hat{f}_{\text{ALE}}^s(x) : \mathbb{R} \rightarrow \mathbb{R}$ , unnormalized feature effect computed by ALE for the  $s$ -th feature  $s$
- $f_s(\mathbf{x}) = \frac{\partial f(x_s, \mathbf{x}_c)}{\partial x_s}$ , the partial derivative of the  $s$ -th feature
- $z_{k-1}$ , the left limit of the  $k$ -th bin
- $z_k$ , the right limit of the  $k$ -th bin
- $\mathcal{S}_k = \{\mathbf{x}^i : x_s^i \in [z_{k-1}, z_k)\}$ , the set of training points that belong to the  $k$ -th bin
- $k_x$  the index of the bin that  $x$  belongs to
- $\hat{\mu}_k^s$ , DALE approximation of the mean value inside a bin, equals  $\frac{1}{|\mathcal{S}_k|} \sum_{i: x^i \in \mathcal{S}_k} f_s(\mathbf{x}^i)$
- $(\hat{\sigma}_k^s)^2$ , DALE approximation of the variance inside a bin, equals  $\frac{1}{|\mathcal{S}_k|-1} \sum_{i: x^i \in \mathcal{S}_k} (f_s(\mathbf{x}^i) - \hat{\mu}_k^s)^2$

## 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:

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

$$p(\mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2) = \begin{cases} 1 & x_1 \in [0, 1], x_2 = x_1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$p(\mathcal{X}_1 = x_1) = \begin{cases} 1 & 0 \leq x_1 \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$p(\mathcal{X}_2 = x_2) = \begin{cases} 1 & 0 \leq x_2 \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$p(\mathcal{X}_2 = x_2 | \mathcal{X}_1 = x_1) = \delta(x_2 - x_1) \quad (5)$$

PDPLOTS.

The feature effect computed by PDP plots is:

$$\begin{aligned} f_{\text{PDP}}(x_1) &= \\ &= \mathbb{E}_{\mathcal{X}_2}[f(x_1, \mathcal{X}_2)] \\ &= \int_{x_2} f(x_1, x_2) p(x_2) \partial x_2 \\ &= \int_0^{1-x_1} (1 - x_1 - x_2) \partial x_2 + \int_{1-x_1}^1 0 \partial x_2 \\ &= \int_0^{1-x_1} 1 \partial x_2 + \int_0^{1-x_1} -x_1 \partial x_2 + \int_0^{1-x_1} -x_2 \partial x_2 \\ &= (1 - x_1) - x_1(1 - x_1) - \frac{(1 - x_1)^2}{2} \\ &= (1 - x_1)^2 - \frac{(1 - x_1)^2}{2} \\ &= \frac{(1 - x_1)^2}{2} \end{aligned} \quad (6)$$

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:

$$\begin{aligned}
 f_{\text{MP}}(x_1) &= \\
 &= \mathbb{E}_{\mathcal{X}_2|\mathcal{X}_1=x_1}[f(x_1, \mathcal{X}_2)] \\
 &= \int_{x_2} f(x_1, x_2)p(x_2|x_1)\partial x_2 \\
 &= f(x_1, x_1) = \\
 &= \begin{cases} 1 - 2x_1, & x_1 \leq 0.5 \\ 0, & \text{otherwise} \end{cases}
 \end{aligned} \tag{8}$$

Due to symmetry:

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

ALE

The feature effect computed by ALE is:

$$\begin{aligned}
 \hat{f}_{\text{ALE}}(x_1) &= \\
 &= \int_{z_0}^{x_1} \mathbb{E}_{\mathcal{X}_2|\mathcal{X}_1=z} \left[ \frac{\partial f}{\partial z}(z, \mathcal{X}_2) \right] \partial z \\
 &= \int_{z_0}^{x_1} \int_{x_2} \frac{\partial f}{\partial z}(z, x_2)p(x_2|z)\partial x_2 \partial z = \\
 &= \int_{z_0}^{x_1} \frac{\partial f}{\partial z}(z, z)\partial z = \\
 &= \begin{cases} \int_{z_0}^{x_1} -1 \partial z & x_1 \leq 0.5 \\ \int_{z_0}^{0.5} -1 \partial z + \int_{.5}^{x_1} 0 \partial z & x_1 > 0.5 \end{cases} \\
 &= \begin{cases} -x_1 & x_1 \leq 0.5 \\ -0.5 & x_1 > 0.5 \end{cases}
 \end{aligned} \tag{10}$$

The normalization constant is:

$$\begin{aligned}
 c &= -\mathbb{E}[\hat{f}_{\text{ALE}}(x_1)] \\
 &= -\int_{-\infty}^{\infty} \hat{f}_{\text{ALE}}(x_1) \\
 &= -\int_0^{0.5} -z \partial z - \int_{0.5}^1 -0.5 \partial z \\
 &= \frac{0.25}{2} + 0.25 = 0.375
 \end{aligned} \tag{11}$$

Therefore, the normalized feature effect is:

$$y = f_{ALE}(x_1) = \begin{cases} 0.375 - x_1 & 0 \leq x_1 \leq 0.5 \\ -0.125 & 0.5 < x_1 \leq 1 \end{cases} \quad (12)$$

Due to symmetry:

$$y = f_{ALE}(x_2) = \begin{cases} 0.375 - x_2 & 0 \leq x_2 \leq 0.5 \\ -0.125 & 0.5 < x_2 \leq 1 \end{cases} \quad (13)$$

### 3. First-order and Second-order DALE approximation

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

$$f_{DALE}^s(x) = \Delta x \sum_{k=1}^{k_x} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} [f_s(\mathbf{x}^i)] \quad (14)$$

For keeping the equation compact, we ommit 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

$$\begin{aligned} f_{DALE}^s(x) = \Delta x \sum_{k=1}^{k_x-1} \frac{1}{|\mathcal{S}_k|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_k} [f_s(\mathbf{x}^i)] \\ + (x - z_{(k_x-1)}) \frac{1}{|\mathcal{S}_{k_x}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_{k_x}} [f_s(\mathbf{x}^i)] \end{aligned} \quad (15)$$

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

$$\begin{aligned} f_{DALE}^{l,m}(x_l, x_m) = \Delta x_l \sum_{p=1}^{p_x-1} \Delta x_m \sum_{q=1}^{q_x-1} \frac{1}{|\mathcal{S}_{k,q}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_{k,q}} f_{l,m}(\mathbf{x}^i) \\ + (x_l - z_{(p_x-1)})(x_m - z_{(q_x-1)}) \frac{1}{|\mathcal{S}_{p_x,q_x}|} \sum_{i:\mathbf{x}^i \in \mathcal{S}_{p_x,q_x}} f_{l,m}(\mathbf{x}^i) \end{aligned} \quad (16)$$

### 4. Second-order ALE definition

The second-order ALE plot definition is

$$f_{\text{ALE}}^{l,m}(x_l, x_m) = c + \int_{x_{l,\min}}^{x_l} \int_{x_{m,\min}}^{x_m} \mathbb{E}_{\mathcal{X}_c | X_l=z_l, X_m=z_m} [f_{l,m}(\mathbf{x})] \partial z_l \partial z_m \quad (17)$$

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

## 5. DALE variance inside each bin

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

$$\begin{aligned} \text{Var}[\hat{\mu}_k^s] &= \text{Var}\left[\frac{1}{|\mathcal{S}_k|} \sum_{i: \mathbf{x}^i \in \mathcal{S}_k} f_s(\mathbf{x}^i)\right] \\ &= \frac{1}{|\mathcal{S}_k|^2} \sum_{i: \mathbf{x}^i \in \mathcal{S}_k} \text{Var}[f_s(\mathbf{x}^i)] \\ &= \frac{|\mathcal{S}_k|}{|\mathcal{S}_k|^2} \text{Var}[f_s(\mathbf{x})] \\ &= \frac{(\sigma_k^s)^2}{|\mathcal{S}_k|} \end{aligned} \quad (18)$$

## 6. Attributes description in the bike-sharing dataset

In the final experiment, we use 11 features from the bike-sharing dataset. In the following list we quickly explain each one;

- $X_{\text{year}}$ : (0 = 2011, 1 = 2012)
- $X_{\text{month}}$ : (1=January, ..., 12=December)
- $X_{\text{hour}}$ : (0, ..., 23)
- $X_{\text{holiday}}$ : (0 = non-holiday, 1 = holiday)
- $X_{\text{weekday}}$ : (0 = Sunday, ..., 6 = Saturday)
- $X_{\text{workingday}}$ : (0 = non-workingday, 1 = workingday)
- $X_{\text{weather-situation}}$ : (1 = best weather situation, ..., 4 = worst weather situation)
- $X_{\text{temp}}$ : temperature in Celsius
- $X_{\text{atemp}}$ : feeling temperature in Celsius
- $X_{\text{hum}}$ : humidity  $X_{\text{windspeed}}$ : windspeed

The target value we want to predict are the bike rentals counts  $Y_{\text{count}}$ .