Regional effect plots for the interpretation of black box machine learning models

Χρήστος Δίου / Christos Diou

Department of Informatics and Telematics
Harokopio University of Athens

Statistics Seminars 2023-2024
Department of Statistics, Athens University of Economics and Business
19/04/2024
Contents

Introduction

Feature Effect

Differential Accumulated Local Effects (DALE)

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

Regional effect plots - effector
Hypothetical (?) scenarios

- The computer vision subsystem of an autonomous vehicle leads the vehicle to take a left turn, in front of a car moving in the opposite direction¹

---

¹https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco
Hypothetical (?) scenarios

• The computer vision subsystem of an autonomous vehicle leads the vehicle to take a left turn, in front of a car moving in the opposite direction¹

• The credit assessment system leads to the rejection of an application for a loan - the client suspects racial bias²

¹https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco
²https://www.technologyreview.com/2021/06/17/1026519/
Hypothetical (?) scenarios

- The computer vision subsystem of an autonomous vehicle leads the vehicle to take a left turn, in front of a car moving in the opposite direction¹
- The credit assessment system leads to the rejection of an application for a loan - the client suspects racial bias²
- A model that assesses the risk of future criminal offenses (and used for decisions on parole sentences) is biased against black prisoners³

---

¹ https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco
Questions

- Why did a model make a specific decision?
- What could we change so that the model will make a different decision?
- Can we summarize and predict the model’s behavior?

Today we focus on the last question
Figure: Timo Speith, “A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods”. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22), 2022 (Speith, 2022)
Contents

Introduction

Feature Effect

Differential Accumulated Local Effects (DALE)

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

Regional effect plots - effector
Interpretable models (ante-hoc)

- Some models afford explanations
  - interpretable-by-design
- Examples, (generalized) linear models, decision trees, $k$-NN
- Example: Linear regression

$$\hat{y} = w_1 x_1 + \ldots + w_p x_p + b$$
Interpretable models (ante-hoc)

• Result in the bike sharing dataset (model weights)

\[
\hat{y} = w_1 x_1 + \ldots + w_p x_p + b
\]

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2399.4</td>
<td>238.3</td>
<td>10.1</td>
</tr>
<tr>
<td>seasonSPRING</td>
<td>899.3</td>
<td>122.3</td>
<td>7.4</td>
</tr>
<tr>
<td>seasonSUMMER</td>
<td>138.2</td>
<td>161.7</td>
<td>0.9</td>
</tr>
<tr>
<td>seasonFALL</td>
<td>425.6</td>
<td>110.8</td>
<td>3.8</td>
</tr>
<tr>
<td>holidayHOLIDAY</td>
<td>-686.1</td>
<td>203.3</td>
<td>3.4</td>
</tr>
<tr>
<td>workingdayWORKING DAY</td>
<td>124.9</td>
<td>73.3</td>
<td>1.7</td>
</tr>
<tr>
<td>weathersitMISTY</td>
<td>-379.4</td>
<td>87.6</td>
<td>4.3</td>
</tr>
<tr>
<td>weathersitRAIN/SNOW/STORM</td>
<td>-1901.5</td>
<td>223.6</td>
<td>8.5</td>
</tr>
<tr>
<td>temp</td>
<td>110.7</td>
<td>7.0</td>
<td>15.7</td>
</tr>
<tr>
<td>hum</td>
<td>-17.4</td>
<td>3.2</td>
<td>5.5</td>
</tr>
<tr>
<td>windspeed</td>
<td>-42.5</td>
<td>6.9</td>
<td>6.2</td>
</tr>
<tr>
<td>days_since_2011</td>
<td>4.9</td>
<td>0.2</td>
<td>28.5</td>
</tr>
</tbody>
</table>

**Figure:** C. Molnar, IML book, 2022 (Molnar, 2022)
Interpretable models (ante-hoc)

- Feature effects (visualization)

\[ \text{effect}_j^{(i)} = w_j x_j^{(i)} \]

Figure: C. Molnar, IML book, 2022 Molnar, 2022
Feature effect methods (1)

- Black-box model \( f(\cdot) : \mathcal{X} \rightarrow \mathcal{Y} \), trained on \( \mathcal{D} \)
- Goal:
  - For single variable: Plot illustrating the effect of a feature \( x_s \) on \( f \) for all values of \( x_s \)
  - For pairs of variables: Plot illustrating the effect of pair \((x_s, x_l)\) on \( f \) for all values of \( x_s \) and \( x_l \)

Feature Effect: global, model-agnostic, outputs plot
Feature Effect methods (2)

\[ y = f(x_s) \rightarrow \text{plot showing the effect of } x_s \text{ on the output } y \]

Figure: C. Molnar, IML book, 2022 (Molnar, 2022)

Feature Effect is simple and intuitive.
Feature Effect Methods (3)

- $x_s \rightarrow$ feature of interest, $x_c \rightarrow$ other features
- How can we isolate $x_s$?
- Difficult task:
  - features are correlated
  - $f$ has learned complex interactions
PDP, MPlot and ALE

- PDP (Friedman, 2001)
  - \( f(x_s) = \mathbb{E}_{x_c}[f(x_s, x_c)] \)
  - **Unrealistic instances**
  - e.g. \( f(\text{age} = 20, \text{years contraceptives} = 20) = ?? \)

- MPlot (Apley and Zhu, 2020)
  - \( x_c | x_s: f(x_s) = \mathbb{E}_{x_c|z}[f(x_s, x_c)] \)
  - **Aggregated effects**
  - Real effect: \( \text{age} = 50 \rightarrow 10, \text{years contraceptives} = 20 \rightarrow 10 \)

- MPlot may assign 17 to both

- ALE (Apley and Zhu, 2020)
  - \( f(x_s) = R_{x_s x_{\text{min}}} \mathbb{E}_{x_c|z}[\partial f/\partial x_s(z, x_c)] \partial z \)
  - **Resolves both failure modes**
PDP, MPlot and ALE

- **PDP (Friedman, 2001)**
  - \( f(x_s) = \mathbb{E}_{x_c}[f(x_s, x_c)] \)
  - **Unrealistic instances**
  - e.g. \( f(x_{\text{age}} = 20, x_{\text{years contraceptives}} = 20) = ?? \)

- **MPlot (Apley and Zhu, 2020)**
  - \( x_c|x_s: f(x_s) = \mathbb{E}_{x_c|x_s}[f(x_s, x_c)] \)
  - **Aggregated effects**
  - Real effect: \( x_{\text{age}} = 50 \rightarrow 10, x_{\text{years contraceptives}} = 20 \rightarrow 10 \)
  - MPlot may assign 17 to both
PDP, MPlot and ALE

- **PDP** (Friedman, 2001)
  - $f(x_s) = \mathbb{E}_{x_c}[f(x_s, x_c)]$
  - **Unrealistic instances**
  - e.g. $f(x_{\text{age}} = 20, x_{\text{years contraceptives}} = 20) = ??$

- **MPlot** (Apley and Zhu, 2020)
  - $x_c|x_s: f(x_s) = \mathbb{E}_{x_c|x_s}[f(x_s, x_c)]$
  - **Aggregated effects**
  - Real effect: $x_{\text{age}} = 50 \rightarrow 10$, $x_{\text{years contraceptives}} = 20 \rightarrow 10$
  - MPlot may assign 17 to both

- **ALE** (Apley and Zhu, 2020)
  - $f(x_s) = \int_{x_{\text{min}}}^{x_s} \mathbb{E}_{x_c|z}[\frac{\partial f}{\partial x_s}(z, x_c)] \partial z$
  - **Resolves both failure modes**
ALE definition: $f(x_s) = \int_{x_{s,\text{min}}}^{x_s} \mathbb{E}_{x_c|z}[\frac{\partial f}{\partial x_s}(z, x_c)] \partial z$

ALE approximation: $f(x_s) = \sum_{k=1}^{k_x} \frac{1}{|S_k|} \sum_{i: x_i \in S_k} \left[ f(z_k, x_c^i) - f(z_{k-1}, x_c^i) \right]$

- point effect
- bin effect
ALE approximation:

\[ f(x_s) = \sum_{k=1}^{k_x} \frac{1}{|S_k|} \sum_{i : x^i \in S_k} [f(z_k, x^i_c) - f(z_{k-1}, x^i_c)] \]

Point effect

Bin effect

**Figure:** Image taken from Interpretable ML book (Molnar, 2022)

Bin splitting (parameter \( K \)) is crucial!
ALE approximation - weaknesses

\[ f(x_s) = \sum_{k}^{k_x} \frac{1}{|S_k|} \sum_{i:x^i \in S_k} \left[ f(z_k, \mathbf{x}_c^i) - f(z_{k-1}, \mathbf{x}_c^i) \right] \]

- **Point Effect** \(\Rightarrow\) evaluation at bin limits
  - 2 evaluations of \(f\) per point \(\rightarrow\) slow
  - change bin limits, pay again \(2 \times N\) evaluations of \(f\) \(\rightarrow\) restrictive
  - broad bins may create out of distribution (OOD) samples \(\rightarrow\) not-robust in wide bins
Contents

Introduction

Feature Effect

Differential Accumulated Local Effects (DALE)
  Dale is faster and more versatile
  DALE is more Accurate

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

Regional effect plots - effector
V. Gkolemis, T. Dalamagas and C. Diou, “DALE: Differential Accumulated Local Effects for efficient and accurate global explanations”, ACML 2022 (Gkolemis, Dalamagas, and Diou, 2023)

Work in collaboration with Vasilis Gkolemis (PhD student @ HUA) and Theodoros Dalamagas (Researcher, ATHENA RC)
Our proposal: Differential ALE

\[ f(x_s) = \Delta x \sum_{k}^{k_x} \frac{1}{|S_k|} \sum_{i:x^i \in S_k} \left[ \frac{\partial f}{\partial x_s}(x^i_s, x^i_c) \right] \]

- **Point Effect** ⇒ evaluation on instances
  - Fast → use of auto-differentiation, all derivatives in a single pass
  - Versatile → point effects computed once, change bins without cost
  - Secure → does not create artificial instances
  - Unbiased estimator of ALE (bias / variance proofs in the paper and supporting material)

For **differentiable** models, DALE resolves ALE weaknesses
DALE is faster and more versatile - theory

\[ f(x_s) = \Delta x \sum_k^{k_x} \frac{1}{|S_k|} \sum_{i : x_i \in S_k} \left[ \frac{\partial f}{\partial x_s}(x_s^i, x_c^i) \right] \]

- Faster
  - gradients wrt all features \( \nabla_x f(x^i) \) in a single pass (via the Jacobian)
  - auto-differentiation must be available (deep learning)

- Versatile
  - Change bin limits, with near zero computational cost

DALE is faster and allows redefintion of the bin limits
DALE is faster and versatile - Experiments

Figure: Light setup; small dataset ($N = 10^2$ instances), computationaely light $f$. Heavy setup; big dataset ($N = 10^5$ instances), computationally heavy $f$. $D$ is the number of dimensions.

DALE considerably accelerates the estimation
DALE uses on-distribution samples - Theory

\[
f(x_s) = \sum_k \frac{1}{|S_k|} \sum_{i : x^i \in S_k} \left[ \frac{\partial f}{\partial x_s}(x^i_s, x^i_c) \right]
\]

- point effect independent of bin limits
  - \( \frac{\partial f}{\partial x_s}(x^i_s, x^i_c) \) computed on real instances \( x^i = (x^i_s, x^i_c) \)

- bin limits affect only the resolution of the plot
  - wide bins \( \rightarrow \) low resolution plot, bin estimation from more points
  - narrow bins \( \rightarrow \) high resolution plot, bin estimation from less points

DALE enables wide bins without creating out of distribution instances
DALE uses on-distribution samples - Experiments

\[ f(x_1, x_2, x_3) = x_1x_2 + x_1x_3 \pm g(x) \]

\[ x_1 \in [0, 10], x_2 \sim x_1 + \epsilon, x_3 \sim \mathcal{N}(0, \sigma^2) \]

\[ f_{\text{ALE}}(x_1) = \frac{x_1^2}{2} \]

- point effects affected by \((x_1x_3)\) (\(\sigma\) is large)
- bin estimation is noisy (samples are few)

**Intuition**: we need wider bins (more samples per bin)
DALE vs ALE - 40 Bins

- **DALE**: on-distribution, noisy bin effect → poor estimation
- **ALE**: on-distribution, noisy bin effect → poor estimation
### DALE vs ALE - 40 Bins

- **DALE**: on-distribution, noisy bin effect → poor estimation
- **ALE**: on-distribution, noisy bin effect → poor estimation
DALE vs ALE - 20 Bins

- DALE: on-distribution, noisy bin effect → poor estimation
- ALE: on-distribution, noisy bin effect → poor estimation
• DALE: on-distribution, noisy bin effect → poor estimation
• ALE: on-distribution, noisy bin effect → poor estimation
• DALE: on-distribution, noisy bin effect → poor estimation
• ALE: starts being OOD, noisy bin effect → poor estimation
DALE vs ALE - 10 Bins

- DALE: on-distribution, noisy bin effect $\rightarrow$ poor estimation
- ALE: starts being OOD, noisy bin effect $\rightarrow$ poor estimation
DALE vs ALE - 5 Bins

- DALE: on-distribution, robust bin effect $\rightarrow$ good estimation
- ALE: completely OOD, robust bin effect $\rightarrow$ poor estimation
• **DALE**: on-distribution, robust bin effect $\rightarrow$ good estimation
• **ALE**: completely OOD, robust bin effect $\rightarrow$ poor estimation
DALE vs ALE - 3 Bins

- DALE: on-distribution, robust bin effect → good estimation
- ALE: completely OOD, robust bin effect → poor estimation
DALE vs ALE - 3 Bins

- **DALE**: on-distribution, robust bin effect → good estimation
- **ALE**: completely OOD, robust bin effect → poor estimation
Real Dataset Experiments - Efficiency

- Bike-sharing dataset (Fanaee-T and Gama, 2013)
- $y \rightarrow$ daily bike rentals
- $x$: 10 features, most of them characteristics of the weather

<table>
<thead>
<tr>
<th>Number of Features</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>DALE</td>
<td>1.17</td>
<td>1.19</td>
<td>1.22</td>
<td>1.24</td>
<td>1.27</td>
<td>1.30</td>
<td>1.36</td>
<td>1.32</td>
<td>1.33</td>
<td>1.37</td>
<td>1.39</td>
</tr>
<tr>
<td>ALE</td>
<td>0.85</td>
<td>1.78</td>
<td>2.69</td>
<td>3.66</td>
<td>4.64</td>
<td>5.54</td>
<td>6.85</td>
<td>7.73</td>
<td>8.86</td>
<td>9.9</td>
<td>10.9</td>
</tr>
</tbody>
</table>

DALE requires almost same time for all features
Real Dataset Experiments - Accuracy

- Difficult to compare in real world datasets
- We do not know the ground-truth effect
- In most features, DALE and ALE agree.
- Only $X_{\text{hour}}$ is an interesting feature

Figure: (Left) DALE (Left) and ALE (Right) plots for $K = \{25, 50, 100\}$
Introduction

Feature Effect

Differential Accumulated Local Effects (DALE)

**RHALE: Robust and Heterogeneity-aware Accumulated Local Effects**

Regional effect plots - effector
V. Gkolemis, T. Dalamagas, E. Ntoutsi and C. Diou, “RHALE: Robust and Heterogeneity-aware Accumulated Local Effects”, ECAI 2023 (Gkolemis, Dalamagas, Ntoutsi, et al., 2023)

Work in collaboration with Vasilis Gkolemis (PhD student @ HUA), Theodoros Dalamagas (Researcher, ATHENA RC) and Eirini Ntoutsi (Prof, Universität der Bundeswehr, München)
Next step: Heterogeneity and optimal bin selection

Using DALE, one has the computational margin to worry about additional issues:

- Computation of heterogeneity of local effects (i.e., standard error of the mean)
- Optimal selection of bins such that the effect does not have a high variation within the bin

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

- Robust: Automatic bin splitting (result does not depend on arbitrary bin selection)
- Heterogeneity aware: $\pm$ from the average
Example (based on Goldstein et al., 2015)

Aggregation bias

\[ Y = 0.2X_1 - 5X_2 + 10X_2 \mathbb{1}_{X_3 > 0} + \mathcal{E} \]

\[ \mathcal{E} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad X_1, X_2, X_3 \overset{i.i.d.}{\sim} \mathcal{U}(-1, 1) \]
Definitions and Approximations - Main effect

ALE main effect definition

\[ f_{\text{ALE}}(x_s) = \int_{x_s, \text{min}}^{x_s} \mathbb{E}_{X_s | X_c = z} [f^s(z, X_c)] \partial z \]

ALE main effect approximation

\[ \hat{f}_{\text{ALE}}(x_s) = \Delta x \sum_k^{k_s} \frac{1}{|S_k|} \sum_{i : x^i \in S_k} \left[ \frac{\partial f}{\partial x_s}(x^i_s, x^i_c) \right] \]

bin effect: \( \hat{\mu}(z) \)
Simple but wrong: ALE + Heterogeneity

**Figure:** Left: approximation with narrow bin-splitting (5 bins) and (Right) with dense-bin splitting

- Fixed-size bin splitting can ruin the estimation of the heterogeneity
Definitions and Approximations - Heterogeneity

ALE heterogeneity definition

\[
\sigma(x_s) = \sqrt{\int_{x_{s,\text{min}}}^{x_s} \left( \mathbb{E}_{X_c|X_s=z} \left[ (f^s(z, X_c) - \mu(z))^2 \right] \right) \partial z}
\]

ALE heterogeneity approximation

\[
\text{STD}(x_s) = \sqrt{\sum_{k=1}^{k_x} (z_k - z_{k-1})^2 \left( \frac{1}{|S_k| - 1} \sum_{i: x^i \in S_k} \left( f^s(x^i) - \hat{\mu}(z_{k-1}, z_k) \right)^2 \right) \sigma^2(z)}
\]
In the paper we formally prove

1. the conditions under which the above definition is an unbiased estimator of the heterogeneity
2. the conditions under which a bin splitting minimizes the estimator variance

Based on the above, we formulate bin-splitting as an optimization problem and propose an efficient solution using dynamic programming.
RHALE: Robust and Heterogeneity-aware ALE

Figure: Variable bin size leads to improved estimation

Simple but correct:
- Automatically finds the optimal bin-splitting
- Optimal $\Rightarrow$ best approximation of the average (ALE) effect
- Optimal $\Rightarrow$ best approximation of the heterogeneity
Impact

In case you work with a differentiable model, as in Deep Learning, use the combination of DALE and RHALE to:

- compute ALE fast, for multiple bin sizes in one pass
- quantify the heterogeneity of the ALE plot, i.e., the deviation of the instance-level effects from the average effect
- get a robust approximation of (a) the main ALE effect and (b) the heterogeneity, using automatic bin-splitting
Contents

Introduction

Feature Effect

Differential Accumulated Local Effects (DALE)

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

Regional effect plots - effector
Effector - A python package for global and regional feature effects


https://xai-effector.github.io

Installation, for python version 3.7+:

pip install effector
Regional effects

Similar to the way one can select optimal bin splits to minimize heterogeneity, one can also identify optimal subregions of the features $x_c$ where the effect is homogeneous.
Regional effect plots - Process

• Combines two methods:
  • RHALE
  • Regional effects (Herbinger, Bischl, and Casalicchio, 2023)

• Idea:
  • Feature effect is the average effect of each feature $x_s$ on the output $y$
  • It is computed by averaging the instance-level effects
  • Heterogeneity $\mathcal{H}$ measures the deviation of the instance-level effects from the average effect due to feature interactions
  • Split the dataset in subgroups in order to minimize the heterogeneity

• Concretely:

$$\mathcal{H}(f_i(x_i)) \gg \mathcal{H}(f_i(x_i|x_j > \tau)) + \mathcal{H}(f_i(x_i|x_j \leq \tau))$$

$\mathcal{H}$ before split

sum of $\mathcal{H}$ after split
Regional effect plots - Objective

\[
\begin{align*}
\text{minimize} & \quad \mathcal{L}_s = \sum_{t=1}^{T_s} \frac{|\mathcal{D}_{st}|}{|\mathcal{D}|} H^m_{st} \\
\text{subject to} & \quad \bigcup_{t=1}^{T} \mathcal{R}_{st} = \mathcal{X}_c \\
& \quad \mathcal{R}_{st} \cap \mathcal{R}_{s\tau} = \emptyset, \quad \forall t \neq \tau
\end{align*}
\]
Algorithm 1: Detect subspaces

Input: Heterogeneity function $H_s$, Maximum depth $L$

Output: subspaces $\{R_{st}\}_{t=1}^{T_s}$, where $T_s \in \{0, 2, \ldots, 2^L\}$

1. $H_{s0}$; \hspace{1cm} // Compute the level of interactions before any split
2. $D = \{(x^i, y^i)\}_{i=1}^N$; \hspace{1cm} // Initial dataset
3. $T_s = 0$; \hspace{1cm} // Initialize the number of splits for feature $s$

4. for $l = 1$ to $L$ do 
5. \hspace{1cm} if $H_{s}^{l-1} = 0$ then 
6. \hspace{2cm} break; \hspace{1cm} // Stop if the heterogeneity is zero
7. \hspace{1cm} end 

8. /* Iterate over all features $x_c$ and candidate split positions $p$ */
9. /* Find the optimal split with heterogeneity $H_s^l = \sum_{t=1}^{2^l} \frac{|D_{st}|}{|D|} H_{st}$ */
10. /* Define the subspaces $\{R_{st}\}_{t=1}^{2^l}$ and the datasets $\{D_{st}\}_{t=1}^{2^l}$ */

11. if $\frac{H_s^l}{H_s^{l-1}} < \epsilon$ then 
12. \hspace{1cm} break; \hspace{1cm} // Stop, if heterogeneity drop is small ($< \epsilon$)
13. \hspace{1cm} end 

14. $T_s = 2^l$; \hspace{1cm} // Update the number of splits for feature $s$

15. return $\{R_{st} | s \in \{1, \ldots, D\}, t \in \{1, \ldots, T_s\}\}$
### Effector - Implemented methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Equation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{f}_{\text{RH}}(x_s)$</td>
<td>Eq. (4)</td>
<td>$\sum_{k=1}^{K} \frac{z_k - z_{k-1}}{</td>
</tr>
<tr>
<td>$\hat{H}_{\text{RH}}$</td>
<td>Eq. (6)</td>
<td>$\sum_{k=1}^{K} \frac{z_k - z_{k-1}}{</td>
</tr>
<tr>
<td>$\hat{f}_{\text{ALE}}(x_s)$</td>
<td>Eq. (3)</td>
<td>$\sum_{k=1}^{K} \frac{1}{</td>
</tr>
<tr>
<td>$\hat{H}_{\text{ALE}}$</td>
<td>Eq. (5)</td>
<td>$\sum_{k=1}^{K} \frac{1}{</td>
</tr>
<tr>
<td>$\hat{f}_{\text{PDP}}(x_s)$</td>
<td>Eq. (7)</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} f(x_s, x_c^i)$</td>
</tr>
<tr>
<td>$\hat{H}_{\text{PDP}}$</td>
<td>Eq. (8)</td>
<td>$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{f}<em>{\text{ICE, centered}}(x_s, t) - \hat{f}</em>{\text{centered}}(x_s, t) \right]^2$</td>
</tr>
<tr>
<td>$\hat{f}_{\text{d-PDP}}(x_s)$</td>
<td>Eq. (9)</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} f(x_s, x_c^i)$</td>
</tr>
<tr>
<td>$\hat{H}_{\text{d-PDP}}$</td>
<td>Eq. (10)</td>
<td>$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{f}<em>{\text{ICE, centered}}(x_s, t) - \hat{f}</em>{\text{centered}}(x_s, t) \right]^2$</td>
</tr>
<tr>
<td>$\hat{f}_{\text{SHAP-DP}}(x_s)$</td>
<td>Eq. (13)</td>
<td>$\kappa(x_s), \ k(x_s)$ is a univariate spline fit to ${(x_s^i, \hat{f}<em>s^i)}</em>{i=1}^{N}$</td>
</tr>
<tr>
<td>$\hat{H}_{\text{SHAP-DP}}$</td>
<td>Eq. (14)</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \left[ \hat{f}<em>s^i - f</em>{\text{SHAP-DP}}(x_s^i) \right]^2$</td>
</tr>
</tbody>
</table>
Tutorial (Bike Sharing Dataset)
Colab notebook
Recap

- **DALE** can help with the computation of fast and accurate feature effect explanations for differentiable models
  - One can change the resolution of the explanation (i.e., number of bins $K$) for free
- **RHALE** can improve explanations by selecting variable bin splits, in an optimal way
  - Unbiased estimation of heterogeneity
  - Select optimal bin splits to minimize heterogeneity and improve the robustness of the explanation
- **Effector**
  - Implements all popular global effect plot methods
  - Extends these methods to regional effect plots
  - Has very fast implementation, especially for differentiable models (takes advantage of auto-differentiation)
Thank you!


