UNIVERSITY OF WARSAW
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
Munich Center for Machine Learning
Warsaw University of Technology

ECML PKDD 2024

On the Robustness of Global Feature Effect Explanations

Hubert Baniecki, Giuseppe Casalicchio, Bernd Bischl, Przemyslaw Biecek

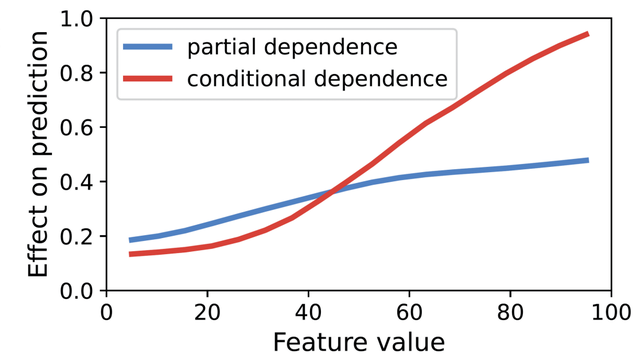
TL;DR: Theoretical bounds for the robustness of feature effect explanations to data and model perturbations.

0. Feature effect explanations (partial dependence, ALE)

$$\text{PD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) \coloneqq \mathbb{E}_{\mathbf{X}_{\bar{s}} \sim p_{\mathbf{X}_{\bar{s}}}} \left[f(\mathbf{x}_s, \mathbf{X}_{\bar{s}}) \right] \coloneqq \int f(\mathbf{x}_s, \mathbf{x}_{\bar{s}}) p_{\mathbf{X}_{\bar{s}}}(\mathbf{x}_{\bar{s}}) d\mathbf{x}_{\bar{s}}$$

$$\text{CD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) \coloneqq \mathbb{E}_{\mathbf{X}_{\bar{s}} \sim p_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}} \left[f(\mathbf{x}_s, \mathbf{X}_{\bar{s}}) \right] \coloneqq \int f(\mathbf{x}_s, \mathbf{x}_{\bar{s}}) p_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}(\mathbf{x}_{\bar{s}} | \mathbf{x}_s) d\mathbf{x}_s$$

Example:



1. Robustness to data perturbations

Assumption 1. *We assume that the model f has bounded predictions, i.e., there exists a constant B such that $|f(\mathbf{x})| \leq B$ for all $\mathbf{x} \in \mathbb{R}^p$.*

Theorem 2. *The robustness of partial dependence and conditional dependence to **data perturbations** is given by the following formulas*

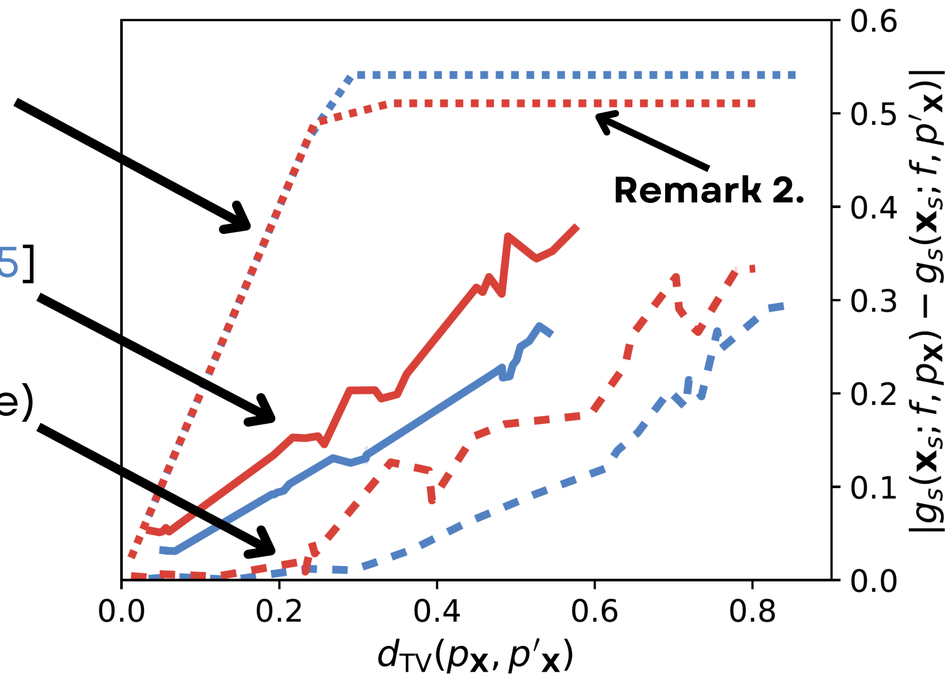
$$\left| \text{PD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \text{PD}_s(\mathbf{x}_s; f, p'_{\mathbf{X}}) \right| \leq 2B \cdot d_{\text{TV}}\left(p_{\mathbf{X}_{\bar{s}}}, p'_{\mathbf{X}_{\bar{s}}}\right),$$

$$\left| \text{CD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \text{CD}_s(\mathbf{x}_s; f, p'_{\mathbf{X}}) \right| \leq 2B \cdot d_{\text{TV}}\left(p_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}, p'_{\mathbf{X}_{\bar{s}} | \mathbf{X}_s = \mathbf{x}_s}\right),$$

where the total variation distance d_{TV} is defined via the l_1 functional distance.

Check out the paper:

2. Perturb the dataset



3. Robustness to model perturbations

Lemma 2. *The robustness of partial dependence and conditional dependence to **model perturbations** is given by the following formulas*

$$\left| \text{PD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \text{PD}_s(\mathbf{x}_s; f', p_{\mathbf{X}}) \right| \leq \|f - f'\|_{\infty},$$

$$\left| \text{CD}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \text{CD}_s(\mathbf{x}_s; f', p_{\mathbf{X}}) \right| \leq \|f - f'\|_{\infty, \mathcal{X}},$$

where $\|f\|_{\infty} \coloneqq \sup_{\mathbf{x} \in \mathbb{R}^p} |f(\mathbf{x})|$ denotes an infinity norm for a function and $\|f\|_{\infty, \mathcal{X}} \coloneqq \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})|$ is the same norm taken over the domain $\mathcal{X} \subseteq \mathbb{R}^p$.

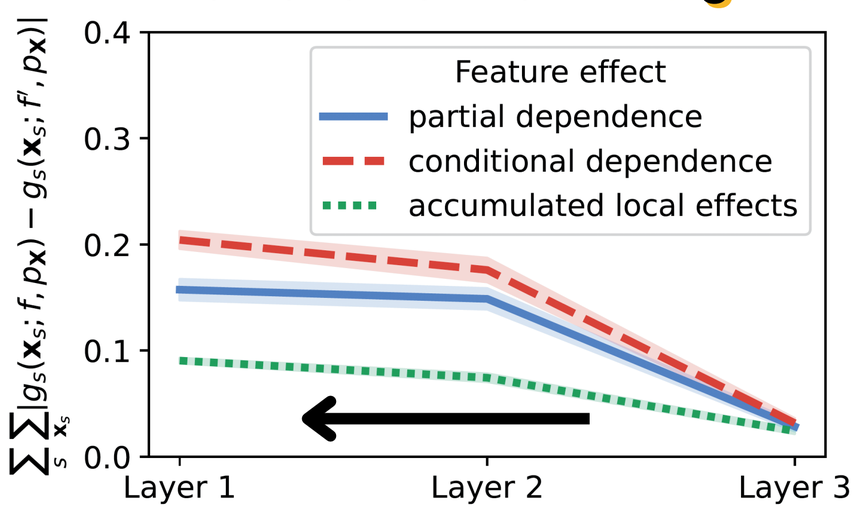
Proof. Follows directly from [26, lemmas 5 & 6].

Theorem 5. *The robustness of accumulated local effects to **model perturbations** is given by the following formula*

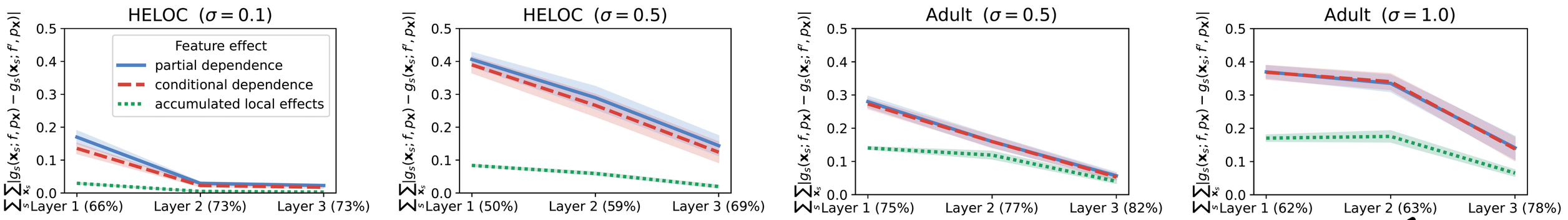
$$\left| \text{ALE}_s(\mathbf{x}_s; f, p_{\mathbf{X}}) - \text{ALE}_s(\mathbf{x}_s; f', p_{\mathbf{X}}) \right| \leq (\mathbf{x}_s - \mathbf{x}_{\min,s}) \cdot \|h - h'\|_{\infty, \mathcal{X}},$$

where $h \coloneqq \frac{\partial f}{\partial \mathbf{x}_s}$ and $h' \coloneqq \frac{\partial f'}{\partial \mathbf{x}_s}$ denote partial derivatives of f and f' respectively.

4. Perturb network weights



(Differential) **ALE does not pass** the model randomization test.



Accuracy drops, but the explanation remains similar.

References

[5] Baniecki et al. *Fooling Partial Dependence via Data Poisoning*. ECML PKDD 2022
[26] Lin et al. *On the Robustness of Removal-Based Feature Attributions*. NeurIPS 2023