From global to regional effects: a comparison of the different approaches

Vasilis Gkolemis

May 24, 2023

1 Introduction

2 Background

Let $\mathcal{X} \in \mathbb{R}^d$ be the d-dimensional feature space, $\mathcal{Y}$ the target space and $f(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$ the black-box function. We use index $s \in \{1, \ldots, d\}$ for the feature of interest and $c = \{1, \ldots, d\} - s$ for the rest. For convenience, to denote the input vector, we use $(x_s, \mathbf{x_c})$ instead of $(x_1, \cdots, x_s, \cdots, x_D)$ and, for random variables, (X_s, X_c) instead of $(X_1, \cdots, X_s, \cdots, X_D)$. The training set $\mathcal{D} = \{(\mathbf{x}^i, y^i)\}_{i=1}^N$ is sampled i.i.d. from the distribution $\mathbb{P}_{X,Y}$. Finally, $f^{\texttt{<method>}}(x_s)$ denotes how $< \texttt{method} >$ defines the feature effect and $\hat{f}^{\texttt{<method>}}(x_s)$ how it estimates it from the training set.

3 Feature Effect

The purpose of any feature effect (FE) method is to explain the 'black-box' function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ using a Generalized Additive Model $f_{\texttt{<method>}}(x) = c + f_1(x_1) + \cdots + f_D(x_D)$, as a global surrogate.

Table 1: Table Caption

Name	**Definition** (f)	**Approximation** ($\hat{f}$)
PDP	$\mathbb{E}_{X_c}[f(x_s, X_c)]$	$\frac{1}{N}\sum f(x_s, x_c^{(i)})$
dPDP	$\mathbb{E}_{X_c}[\frac{\partial f(x_s, X_c)}{\partial x_s}]$	$\frac{1}{N}\sum \frac{\partial f(x_s, x_c^{(i)})}{\partial x_s}$

3.1 Approaches

4 Interaction Index

4.1 Approaches

5 Regional Effects

5.1 Approaches

6 Can we evaluate the approaches?

6.1 Idea 1

We may split every $f : \mathbb{R}^D \rightarrow \mathbb{R}$ into a model without interaction between $\mathbf{x_c}$ and x_s, i.e., $f_{ni}(\mathbf{x}) = f^{(x_s)}(x_s) + f^{(\mathbf{x_c})}(\mathbf{x_c})$, and the interaction term $\kappa(\mathbf{x_c}, x_s)$:

$$f(\mathbf{x}) = \underbrace{f^{(x_s)}(x_s) + f^{(\mathbf{x_c})}(\mathbf{x_c})}_{f_{ni}(\mathbf{x})} + \kappa(\mathbf{x_c}, x_s)$$

A simple approach is defining f to be a Neural Network and f_{ni} a Neural Additive Model without interaction between x_s and $\mathbf{x_c}$. Then $\kappa(\mathbf{x_c}, x_s) = f(\mathbf{x}) - f_{ni}(\mathbf{x})$ and we quantify the importance of κ as $\mathbb{E}_{X_c, X_s}\left[|\kappa(X_c, X_s)|\right] \approx \sqrt{\frac{1}{N}\sum_i \kappa^2(\mathbf{x_c}, x_s)}$.

6.2 Idea 2