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

Vasilis Gkolemis

May 24, 2023

Abstract

Partial dependence (PD) plots are an established tool for visualizing main effects from general black box models. In recent years they have been supplemented with individual conditional expectation plots (ICE plots). Furthermore, they have been fundamentally criticized as being invalid for correlated features, and average local effects plots (ALE plots) have been proposed as a remedy. This paper discusses the properties of PD plots, ICE plots and ALE plots both in terms of their estimands for linear models with interactions and in terms of their performance for nonparametric models that do not extrapolate well. A stratified PD plot is introduced, which is particularly useful for the visualization of interactions between correlated features. Recommendations for the use of model-agnostic effects plots are given, with special emphasis to nonparametric machine learning models.

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 $s \in \{1, \dots, d\}$ for the feature of interest and $c = \{1, \dots, d\} - s$ for the rest. For convenience, we use $(x_s, \mathbf{x_c})$ to refer to $(x_1, \cdots, x_s, \cdots, x_D)$ and, equivalently, (X_s, X_c) instead of $(X_1, \cdots, X_s, \cdots, X_D)$ when we refer to random variables. 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 $<$ `method` $>$ defines the feature effect and $\hat{f}^{<\texttt{method}>}(x_s)$ how it estimates it from the training set.

3 Feature Effect

Feature Effect (FE) methods explain the ‘black-box‘ function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ with a set of $1D$ mappings $f_s(x_s) : \mathbb{R} \rightarrow \mathbb{R}$, each one being the effect of the s-th feature on the output. For example, [running example]..

In this study, we propose interpreting FE as global surrogate models, i.e., as algorithms that take as input (a) the black-box function f and (b) the training set $\mathcal{D}$ and they output a generalized additive model $f_{\texttt{FE}}(\mathbf{x}) = c + \sum_{s=1}^{D} f_s(x_s)$. Using this perspective, we will discuss the different approaches for defining the feature effect and we will propose a quantitative framework for evaluating them.

3.1 Approaches

At Table 1, we present the most used feature effect methods, namely, (a) the Partial Dependence Plot (PDP), (b) the derivative of the PDP (d-PDP) (c) the M-Plots and (d) the Accumulated Local Effects (ALE) plots.

Table 1: Table Caption

Name	**Definition** $f(x_s)$	**Approximation** $\hat{f}(x_s)$
PDP	$\mathbb{E}_{X_c}[f(x_s, X_c)]$	$\frac{1}{N} \sum f(x_s, x_c^{(i)})$
d-PDP	$\mathbb{E}_{X_c}[\frac{\partial f(x_s, X_c)}{\partial x_s}]$	$\frac{1}{N} \sum \frac{\partial f(x_s, x_c^{(i)})}{\partial x_s}$
M-Plot	$\mathbb{E}_{X_c \mid x_s}[f(x_s, X_c)]$	$\frac{1}{N} \sum_{i: x_s^i \in B(x_s)} f(x_s, x_c^{(i)})$
ALE	$\int_{x_{s,\min}}^{x_s} \mathbb{E}_{X_c \mid X_s = z} \left[f^s(z, X_c) \right] \partial z$	$\Delta x \sum_{k=1}^{k_x} \frac{1}{\lvert \mathcal{S}_k \rvert} \sum_{i:\mathbf{x}^{(i)} \in \mathcal{S}_k} f^s(\mathbf{x}^i)$

Global surrogate methods are evaluated based on their fidelity, i.e., how well they can replicate the underlying black-box function.

4 Interaction Index

4.1 Approaches

5 Regional Effects

5.1 Approaches

6 A unified evaluation framework

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