Regionally Additive Models: Explainable-by-design models minimizing feature interactions

Vasilis Gkolemis

June 7, 2023

Abstract

Generalized Additive Models (GAMs) are a popular class of explainable-by-design models that are widely used in practice. GAMs are based on the assumption that the effect of each feature on the target is independent of the values of the other features, however, in cases where this assumption is violated they may lead to poor performance. To address this limitation we propose Regionally Additive Models (RAMs), a novel class of explainable-by-design models, that fits multiple GAMs to subregions of the feature space where interactions are minimized. Our approach consists of two steps: first, we fit a black-box model and we identify the subregions where the black-box model is nearly locally additive, i.e., where the effect of each feature on the target is independent of the values of the other features. Secondly, we train a GAM specifically for each identified subregion.

We show that RAMs are more expressive than GAMs while they are still interpretable.

1 Introduction

To motivate about the use of RAMs, consider the black-box function $f(\mathbf{x}) = 8x_2 \mathbb{1}_{x_1>0} \mathbb{1}_{x_3=0}$ with $x_1, x_2 \sim \mathcal{U}(-1,1)$ and $x_3 \sim Bernoulli(0,1)$. The black-box function f involves an interaction term between the three features. In principle, such interactions cannot be captured by a GAM which is a linear combination of univariate functions. However, the black-box if we split the input space in two regions we observe that f is locally additive in each one, i.e.,

$$f(\mathbf{x}) = \begin{cases} 8x_2 & \text{if } x_1 > 0 \text{ and } x_3 = 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

which makes it trivial to fit a separate GAMs in each subregion.

Correctly identifying the regions where the black-box function is locally additive is the key idea behind RAMs. This can be handled

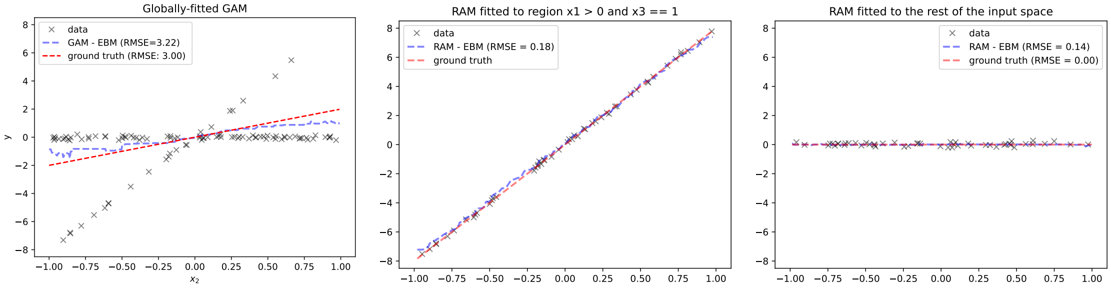

Figure 1: Caption

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, 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 $< \texttt{method} >$ defines the feature effect and $\hat{f}^{\texttt{<method>}}(x_s)$ how it estimates it from the training set.

3 RAM: Regionally Additive Models

4 Synthetic Examples

5 Real-World Datasets