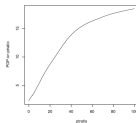
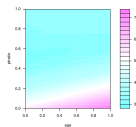
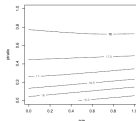
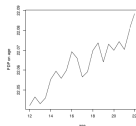


Interpretable Machine Learning

Introduction to Functional Decomposition



Learning goals

- Basic idea of additive functional decompositions
- Motivation and usefulness of functional decompositions
- Difficulty of obtaining or even defining a functional decomposition
- Several examples

PRELIMINARIES

Recap: Interactions

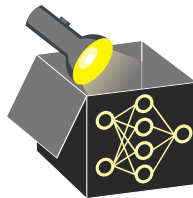
- Interactions between features: Effect of one feature on the prediction output depends on (one or more) other features
- Definition: f has interacting features x_j, x_k , if

$$\mathbb{E} \left[\left(\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k} \right)^2 \right] > 0$$

Recap: GAMs

- Decomposition into only main effects
- Do not contain any interactions

$$\hat{f}(\mathbf{x}) = \theta_0 + g_1(x_1) + g_2(x_2) + \dots + g_p(x_p)$$



FIRST EXAMPLE: ADDITIVE DECOMPOSITION

Example

Consider

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2$$

- Idea: Additive decomposition depending on which features used:

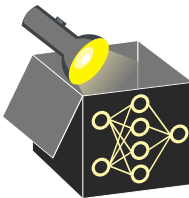
$g_0(x_1, x_2) = 4$ Part depending on no features at all (intercept)

$\left. \begin{aligned} g_1(x_1, x_2) &= 2x_1 \\ g_2(x_1, x_2) &= 0.3e^{x_2} \end{aligned} \right\}$ Parts depending on a single feature (main effects) (1)

$g_{1,2}(x_1, x_2) = |x_1|x_2$ Part depending on both features (interaction)

↪ single terms with immediate interpretation, full understanding of the model

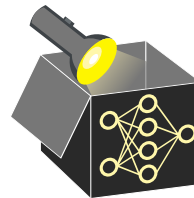
↔ Not possible with effects of single features (e.g. PDPs) or GAM surrogate model (miss interaction part)



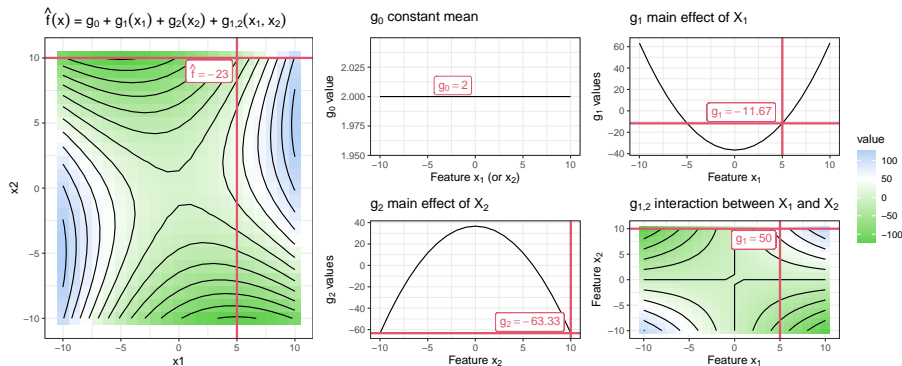
ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION

Goal in general: Given a black-box model $\hat{f} : \mathbb{R}^2 \rightarrow \mathbb{R}$, find a decomposition

$$\hat{f}(x_1, x_2) = g_0 + g_1(x_1) + g_2(x_2) + g_{1,2}(x_1, x_2) \quad (2)$$



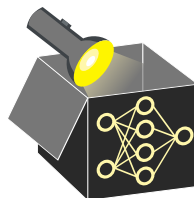
Example



→ More details on this example later

ANOTHER EXAMPLE: ADDITIVE DECOMPOSITION

Example



$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

Again, read additive decomposition from formula:

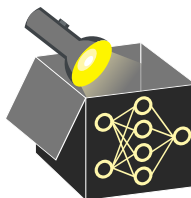
$$\begin{aligned} g_{\emptyset}(x_1, x_2, x_3) &= 1 && \text{constant part, no effects} \\ \left. \begin{aligned} g_1(x_1, x_2, x_3) &= -2x_1 \\ g_2(x_1, x_2, x_3) &= 0 \\ g_3(x_1, x_2, x_3) &= -2\sin(x_3) \end{aligned} \right\} && \text{main effects, no interactions} \\ \left. \begin{aligned} g_{1,2}(x_1, x_2, x_3) &= |x_1|x_2 \\ g_{1,3}(x_1, x_2, x_3) &= 0 \\ g_{2,3}(x_1, x_2, x_3) &= -\sin(x_2x_3) \end{aligned} \right\} && \text{2-way interactions (depending on 2 features)} \\ g_{1,2,3}(x_1, x_2, x_3) &= 0.5x_1x_2x_3 && \text{3-way interactions} \end{aligned} \tag{3}$$

\Rightarrow 8 components in total, but some empty \rightsquigarrow Certain interactions not present

GENERAL FORM OF FUNCTIONAL DECOMPOSITION

► Li and Rabitz (2011)

► Chastaing et al. (2012)



Definition

Functional decomposition: Additive decomposition of a function $\hat{f} : \mathbb{R}^p \mapsto \mathbb{R}$ into a sum of components of different dimensions w.r.t. inputs x_1, \dots, x_p :

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \sum_{S \subseteq \{1, \dots, p\}} g_S(\mathbf{x}_S) \\ &= g_{\emptyset} + g_1(x_1) + g_2(x_2) + \dots + g_p(x_p) + \\ &\quad g_{1,2}(x_1, x_2) + \dots + g_{p-1,p}(x_{p-1}, x_p) + \dots + \\ &\quad g_{1,2,3}(x_1, x_2, x_3) + \dots + g_{1,2,3,4}(x_1, x_2, x_3, x_4) + \dots + g_{1,\dots,p}(x_1, \dots, x_p)\end{aligned}$$

\rightsquigarrow one component for every possible combination S of indices

Sort terms according to degree of interaction:

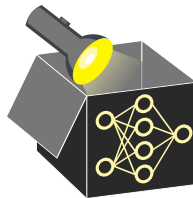
- $g_{\emptyset} \hat{=}$ Constant mean (intercept)
- $g_j \hat{=}$ first-order or main effect of j -th feature alone on $\hat{f}(\mathbf{x})$
- $g_{j,k} \hat{=}$ second-order interaction effect of features j and k w.r.t. $\hat{f}(\mathbf{x})$
- $g_S(\mathbf{x}_S) \hat{=}$ $|S|$ -order effect, depends **only** on features in S

PROPERTIES OF THE DEFINITION

- **Interpretability:** Extremely powerful decomposition, reveals complete interaction structure
- Compare to GAM: Same decomposition, but without interactions
⇒ Any GAM already comes with its decomposition

$$\hat{f}(\mathbf{x}) = g_0 + g_1(x_1) + g_2(x_2) + \dots + g_p(x_p)$$

- Same for LMs: Decomposition explicitly modeled
- Easy decomposition also for decision trees and tree ensembles (see below)
- **Problem 1:** Calculating decomposition extremely difficult, often infeasible
 - For p features: Decomposition with 2^p terms → too many different terms, difficult to interpret
- **Problem 2:** Definition not complete: Decomposition not unique, many trivial decompositions not useful
 - More requirements or constraints needed to ensure decomposition is meaningful
 - Even worse once features are dependent or correlated (see later)



PROBLEM 2: DEFINITION NOT ENOUGH

Example

Again consider

$$\hat{f}(x_1, x_2, x_3) = -2x_1 - 2\sin(x_3) + |x_1|x_2 + 0.5x_1x_2x_3 - \sin(x_2x_3) + 1$$

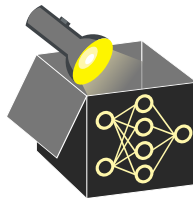
Two possible decompositions (valid according to definition):

$$g_{1,\dots,p}(x_1, \dots, x_p) := \hat{f}(\mathbf{x}) \text{ and for all other terms } g_S(\mathbf{x}_S) := 0,$$

or:

$$\begin{aligned} g_{\emptyset} &= 1; \quad g_1(x_1) = x_1; \quad g_2(x_2) = 2x_2; \quad g_3(x_3) = 3x_3; \\ g_{1,2}(x_1, x_2) &= \frac{1}{2}x_1x_2; \quad g_{1,3}(x_1, x_3) = \frac{1}{3}x_1x_3; \quad g_{2,3}(x_2, x_3) = \frac{2}{3}x_2x_3; \\ \text{and } g_{1,2,3}(x_1, x_2, x_3) &= \hat{f}(x_1, x_2, x_3) - \sum_{S \subsetneq \{1,2,3\}} g_S(\mathbf{x}_S) \end{aligned}$$

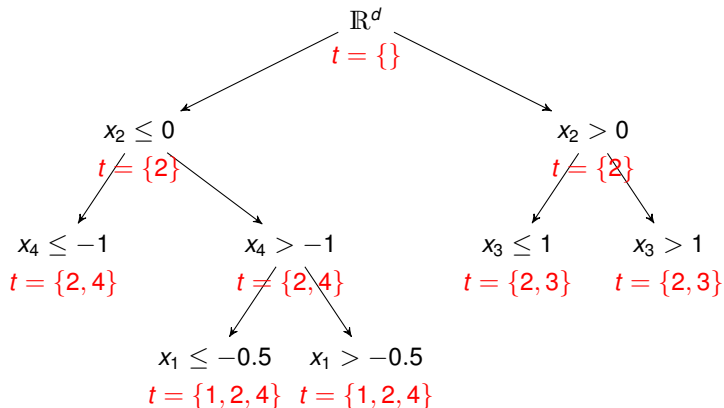
\Rightarrow Definition of decomposition not unique



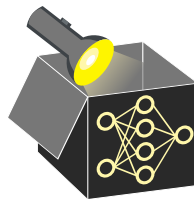
EXAMPLE: DECISION TREES

Define *interaction type* t of a node: subset of features involved in constructing this node.

Example:



⇒ Degree of interaction in each node is $|t|$.



DECOMPOSITION FOR DECISION TREES

Here: Decomposition via indicator functions

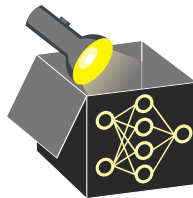
$$\hat{f}(\mathbf{x}) = g_{\emptyset} + g_{2,4}(x_2, x_4) + g_{2,3}(x_2, x_3) + g_{1,2,4}(x_1, x_2, x_4)$$

⇒ Decomposition no main effect, but model certainly contains an effect of e.g. x_2

⇒ Lower-order effects “hidden” inside higher-order terms

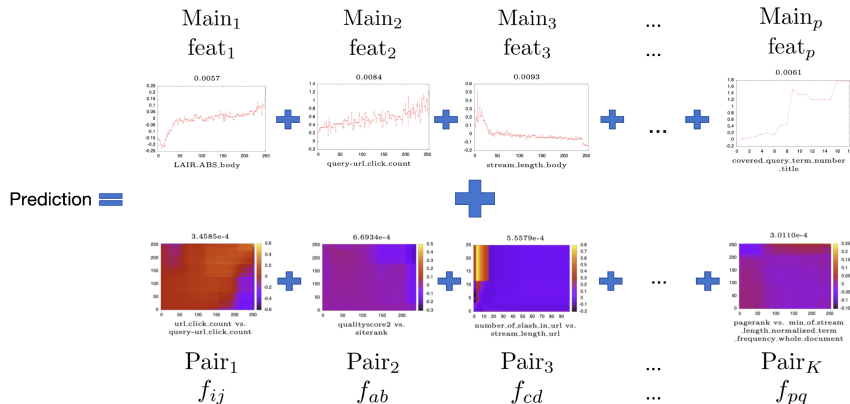
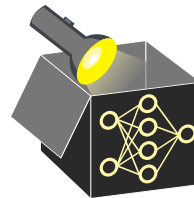
↪ reading from decision tree not enough, “bad decomposition”

Note: ▶ Yang (2024) propose a (quite complicated) solution for this case



EXAMPLE: EBM

- See before: **GAMs** have functional decomposition by definition
- **EBMs**: Sum of the final one- and two-dimensional components



- In general: Model with functional decomposition up to max. order 2 is always “inherently interpretable”
- **Reason:** Visualization of all components