

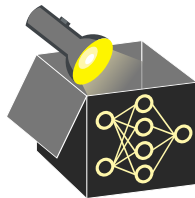
Interpretable Machine Learning

Additive Decomposition



Learning goals

- What are additive decomposition of prediction functions?
- Why are they useful?
- How do we obtain them?



FUNCTIONAL DECOMPOSITION

► Li and Rabitz (2011)

► Chastaing et al. (2012)

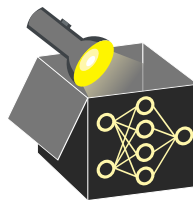
For interpretation purposes, one might be interested in decomposing a square-integrable function $\hat{f} : \mathbb{R}^p \mapsto \mathbb{R}$ into sum of components of different dimensions w.r.t. inputs x_1, \dots, x_p :

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

where

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

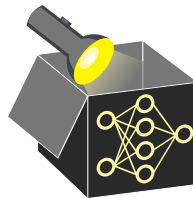
N.B.: A unique solution for the components only exists under certain assumptions



FUNCTIONAL DECOMPOSITION – ASSUMPTIONS

For independent inputs, the *vanishing condition* is required to obtain a unique solution:

$$\mathbb{E}_{\mathbf{x}_j}(g_S(\mathbf{x}_S)) = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0, \forall j \in S, \forall S \subseteq \{1, \dots, p\}$$



FUNCTIONAL DECOMPOSITION – ASSUMPTIONS

For independent inputs, the *vanishing condition* is required to obtain a unique solution:

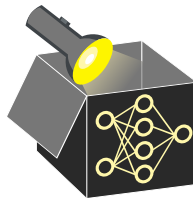
$$\mathbb{E}_{x_j}(g_S(\mathbf{x}_S)) = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0, \forall j \in S, \forall S \subseteq \{1, \dots, p\}$$

Vanishing condition has the following implications:

- Marginalizing out $x_j, \forall j \in S$ for component $g_S(\mathbf{x}_S)$ yields a constant 0
 \rightsquigarrow Makes sure that component $g_S(\mathbf{x}_S)$ does not contain effects of $x_j, \forall j \in S$
- Components are orthogonal (i.e., mutually independent and uncorrelated):

$$\mathbb{E}_X(g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)) = 0, \forall V \neq S$$

- Variance can be decomposed: $\text{Var}[\hat{f}(\mathbf{x})] = \sum_{S \subseteq \{1, \dots, p\}} \text{Var}[g_S(\mathbf{x}_S)]$



FUNCTIONAL DECOMPOSITION – ASSUMPTIONS

For independent inputs, the *vanishing condition* is required to obtain a unique solution:

$$\mathbb{E}_{x_j}(g_S(\mathbf{x}_S)) = \int g_S(\mathbf{x}_S) d\mathbb{P}(x_j) = 0, \forall j \in S, \forall S \subseteq \{1, \dots, p\}$$

Vanishing condition has the following implications:

- Marginalizing out $x_j, \forall j \in S$ for component $g_S(\mathbf{x}_S)$ yields a constant 0
 \rightsquigarrow Makes sure that component $g_S(\mathbf{x}_S)$ does not contain effects of $x_j, \forall j \in S$
- Components are orthogonal (i.e., mutually independent and uncorrelated):

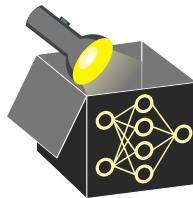
$$\mathbb{E}_X(g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)) = 0, \forall V \neq S$$

- Variance can be decomposed: $\text{Var}[\hat{f}(\mathbf{x})] = \sum_{S \subseteq \{1, \dots, p\}} \text{Var}[g_S(\mathbf{x}_S)]$

N.B.: For dependent inputs, [Hooker \(2007\)](#) showed the existence of a unique solution for the components under a “relaxed vanishing condition” which leads to a “hierarchical orthogonality”

$$\mathbb{E}_X(g_V(\mathbf{x}_V)g_S(\mathbf{x}_S)) = 0, \forall V \subset S$$

\rightsquigarrow Only components are orthogonal where features involved in $g_V(\mathbf{x}_V)$ also appear in $g_S(\mathbf{x}_S)$

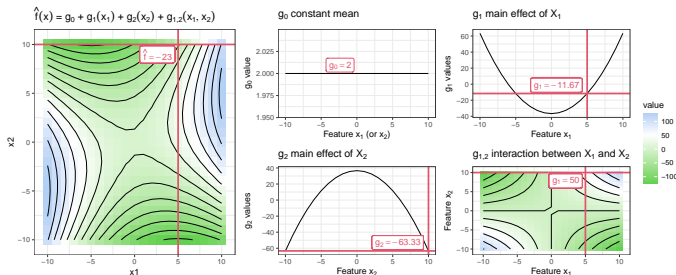


FUNCTIONAL DECOMPOSITION – EXAMPLE

Example: $\hat{f}(\mathbf{x}) = 2 + x_1^2 - x_2^2 + x_1 \cdot x_2$ (e.g., if $x_1 = 5$ and $x_2 = 10 \Rightarrow \hat{f}(\mathbf{x}) = -23$)

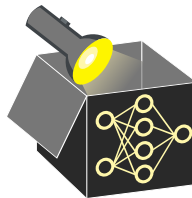
- Computation of components using feature values

$x_1 = x_2 = (-10, -9, \dots, 10)^T$ gives:



For $x_1 = 5$ and $x_2 = 10$:

- $g_0 = 2$
 - $g_1(x_1) = -9.67$
 - $g_2(x_2) = -65.33$
 - $g_{1,2}(x_1, x_2) = 50$
- $\Rightarrow \hat{f}(\mathbf{x}) = -23$

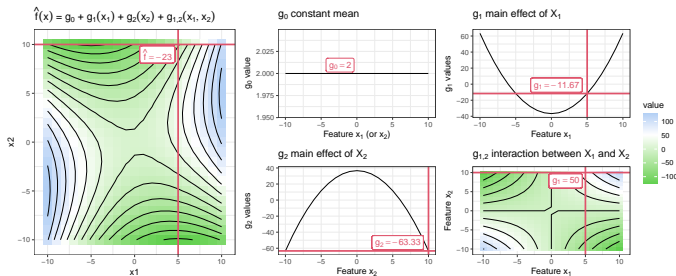


FUNCTIONAL DECOMPOSITION – EXAMPLE

Example: $\hat{f}(\mathbf{x}) = 2 + x_1^2 - x_2^2 + x_1 \cdot x_2$ (e.g., if $x_1 = 5$ and $x_2 = 10 \Rightarrow \hat{f}(\mathbf{x}) = -23$)

- Computation of components using feature values

$x_1 = x_2 = (-10, -9, \dots, 10)^T$ gives:



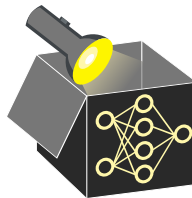
For $x_1 = 5$ and $x_2 = 10$:

- $g_0 = 2$
- $g_1(x_1) = -9.67$
- $g_2(x_2) = -65.33$
- $g_{1,2}(x_1, x_2) = 50$

$$\Rightarrow \hat{f}(\mathbf{x}) = -23$$

- Vanishing condition means:

- g_1 and g_2 are mean-centered w.r.t. marginal distribution of x_1 and x_2
- Integral of $g_{1,2}$ over marginal distribution x_1 (or x_2) is 0

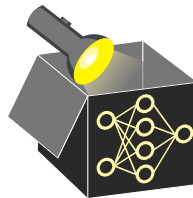


FUNCTIONAL DECOMPOSITION – COMPUTATION

Computation of components via recursive expectations (where $-S = \{1, \dots, p\} \setminus S$):

$$g_S(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_{-S}} \left[\hat{f}(\mathbf{x}) \mid \mathbf{x}_S \right] - \sum_{V \subset S} g_V(x_V)$$

- Expectation integrates $\hat{f}(\mathbf{x})$ over all input features except \mathbf{x}_S
- Subtract all components g_V with $V \subset S$ to remove all lower-order effects and center the effect



FUNCTIONAL DECOMPOSITION – COMPUTATION

Computation of components via recursive expectations (where $-S = \{1, \dots, p\} \setminus S$):

$$g_S(\mathbf{x}_S) = \mathbb{E}_{X_{-S}} \left[\hat{f}(\mathbf{x}) \mid x_S \right] - \sum_{V \subset S} g_V(x_V)$$

- Expectation integrates $\hat{f}(\mathbf{x})$ over all input features except \mathbf{x}_S
- Subtract all components g_V with $V \subset S$ to remove all lower-order effects and center the effect
- Recursive computation:

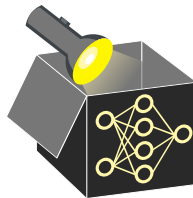
$$g_{\emptyset} = \mathbb{E}_X \left[\hat{f}(\mathbf{x}) \right]$$

$$g_j(x_j) = \mathbb{E}_{X_{-j}} \left[\hat{f}(\mathbf{x}) \mid x_j \right] - g_{\emptyset}, \forall j \in \{1, \dots, p\}$$

$$g_{j,k}(x_j, x_k) = \mathbb{E}_{X_{-\{j,k\}}} \left[\hat{f}(\mathbf{x}) \mid x_j, x_k \right] - g_k(x_k) - g_j(x_j) - g_{\emptyset}, \forall j < k$$

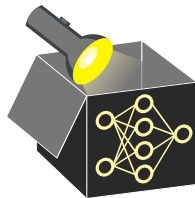
\vdots

$$g_{1,\dots,p}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \sum_{S \subseteq \{1,\dots,p-1\}} g_S(\mathbf{x}_S)$$



VARIANCE DECOMPOSITION

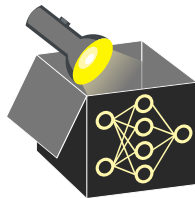
- Decomposition of $\hat{f}(\mathbf{x})$ allows to conduct functional analysis of variance (fANOVA)



VARIANCE DECOMPOSITION

- Decomposition of $\hat{f}(\mathbf{x})$ allows to conduct functional analysis of variance (fANOVA)
- If features are independent, variance can be additively decomposed without covariances:

$$\begin{aligned} \text{Var} [\hat{f}(\mathbf{x})] &= \text{Var} [g_{\emptyset} + g_1(x_1) + \dots + g_{1,2}(x_1, x_2) + \dots + g_{1,\dots,p}(\mathbf{x})] \\ &= \text{Var} [g_{\emptyset}] + \text{Var} [g_1(x_1)] + \dots + \text{Var} [g_{1,2}(x_1, x_2)] + \dots + \text{Var} [g_{1,\dots,p}(\mathbf{x})] \end{aligned}$$



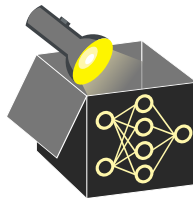
VARIANCE DECOMPOSITION

- Decomposition of $\hat{f}(\mathbf{x})$ allows to conduct functional analysis of variance (fANOVA)
- If features are independent, variance can be additively decomposed without covariances:

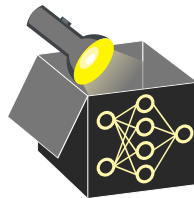
$$\begin{aligned} \text{Var} [\hat{f}(\mathbf{x})] &= \text{Var} [g_{\emptyset} + g_1(x_1) + \dots + g_{1,2}(x_1, x_2) + \dots + g_{1,\dots,p}(\mathbf{x})] \\ &= \text{Var} [g_{\emptyset}] + \text{Var} [g_1(x_1)] + \dots + \text{Var} [g_{1,2}(x_1, x_2)] + \dots + \text{Var} [g_{1,\dots,p}(\mathbf{x})] \end{aligned}$$

- Dividing by the prediction variance, yields fraction of variance explained by each term:

$$1 = \frac{\text{Var} [g_{\emptyset}]}{\text{Var} [\hat{f}(\mathbf{x})]} + \frac{\text{Var} [g_1(x_1)]}{\text{Var} [\hat{f}(\mathbf{x})]} + \dots + \frac{\text{Var} [g_{1,2}(x_1, x_2)]}{\text{Var} [\hat{f}(\mathbf{x})]} + \dots + \frac{\text{Var} [g_{1,\dots,p}(\mathbf{x})]}{\text{Var} [\hat{f}(\mathbf{x})]}$$



VARIANCE DECOMPOSITION



- Decomposition of $\hat{f}(\mathbf{x})$ allows to conduct functional analysis of variance (fANOVA)
- If features are independent, variance can be additively decomposed without covariances:

$$\begin{aligned}\text{Var} [\hat{f}(\mathbf{x})] &= \text{Var} [g_{\emptyset} + g_1(x_1) + \dots + g_{1,2}(x_1, x_2) + \dots + g_{1,\dots,p}(\mathbf{x})] \\ &= \text{Var} [g_{\emptyset}] + \text{Var} [g_1(x_1)] + \dots + \text{Var} [g_{1,2}(x_1, x_2)] + \dots + \text{Var} [g_{1,\dots,p}(\mathbf{x})]\end{aligned}$$

- Dividing by the prediction variance, yields fraction of variance explained by each term:

$$1 = \frac{\text{Var} [g_{\emptyset}]}{\text{Var} [\hat{f}(\mathbf{x})]} + \frac{\text{Var} [g_1(x_1)]}{\text{Var} [\hat{f}(\mathbf{x})]} + \dots + \frac{\text{Var} [g_{1,2}(x_1, x_2)]}{\text{Var} [\hat{f}(\mathbf{x})]} + \dots + \frac{\text{Var} [g_{1,\dots,p}(\mathbf{x})]}{\text{Var} [\hat{f}(\mathbf{x})]}$$

- Fraction of variance explained by a component $g_V(\mathbf{x}_V)$ is the Sobol index:

$$S_V = \frac{\text{Var}[g_V(\mathbf{x}_V)]}{\text{Var}[\hat{f}(\mathbf{x})]}$$

\rightsquigarrow Importance measure of component $g_V(\mathbf{x}_V)$

\rightsquigarrow Small S_V values \Rightarrow Component g_V does not explain much of total variance of \hat{f}