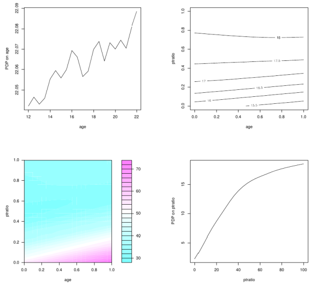
Interpretable Machine Learning

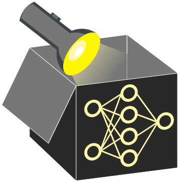
Functional Decompositions
Functional ANOVA



Learning goals

- One method for functional decomposition: Classical functional ANOVA (fANOVA)
- Algorithm for calculating the components in a fANOVA
- Variance decomposition in fANOVA

INTRODUCTION AND HISTORY OF FANOVA



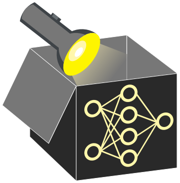
- One possible method to obtain functional decomposition
- Since 1940's: Developed under different names in mathematics and sensitivity analysis
- Since 1990's: Developed for probability distributions or statistical data
- Since 2000's: Applied to machine learning, subsequently alternatives developed extending applicability
- **Assumption**: Independent features

STANDARD FANOVA: IDEA

- Example:

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

STANDARD FANOVA: IDEA



- Example:

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

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.

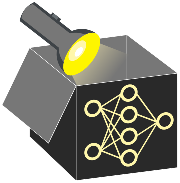
STANDARD FANOVA: IDEA

- Example:

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

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In 1st step, compute main effects using feat effect methods
  Here: PDP + more general PD-functions

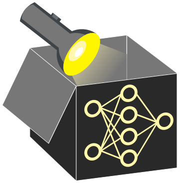
STANDARD FANOVA: IDEA



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

- **First idea:** Make sure higher-order terms don't contain lower-order terms
$\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In 1st step, compute main effects using feat effect methods
Here: PDP + more general PD-functions
- **Idea for fANOVA:** PD-function $\hat{f}_{S;PD}$ = sum of all components $g_{\tilde{S}}$ up to this order
$$\hat{f}_{S;PD}(\mathbf{x}_S) = \sum_{V \subseteq S} g_V(\mathbf{x}_V)$$

STANDARD FANOVA: IDEA



- Example:

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

- **First idea:** Make sure higher-order terms don't contain lower-order terms
  $\Rightarrow$ First compute lower-order terms, then higher-order terms.
- **Second idea:** In 1st step, compute main effects using feat effect methods
  Here: PDP + more general PD-functions
- **Idea for fANOVA:** PD-function $\hat{f}_{S;PD}$ = sum of all components $g_{\tilde{S}}$ up to this order

$$\hat{f}_{S;PD}(\mathbf{x}_S) = \sum_{V \subseteq S} g_V(\mathbf{x}_V)$$

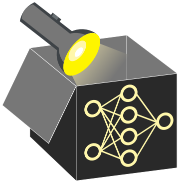
- **Remember:**
  Idea of PDPs or general PD-functions: Average out all other features

$\Rightarrow$ Total formula for calculating the components g_S in the fANOVA algorithm:

$$g_S(\mathbf{x}_S) = (\text{average out all features not contained in } S) \\ - (\text{All lower-order components})$$

FORMAL DEFINITION AND COMPUTATION

▸ HOOKER_2004



Definition

Recursive computation using PD-functions
(here $-S = \{1, \ldots, p\} \setminus S$ denotes all indices not contained in S):

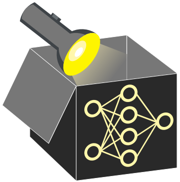
$$g_S(\mathbf{x}_S) = \hat{f}_{S;PD}(\mathbf{x}_S) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \mathbb{E}_{-S}\left[\hat{f}(\mathbf{x}_S, _{-S})\right] - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

$$= \int \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}) d\mathbb{P}(\mathbf{x}_{-S}) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

- Expectation integrates $\hat{f}(\mathbf{x})$ over all input features except $\mathbf{x}_S$
- Subtract sum of g_V to remove all lower-order effects and center the effect
- **Note:** If no distribution given: Uniform distribution or plain integral

FORMAL DEFINITION AND COMPUTATION

▸ HOOKER_2004



Definition

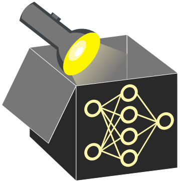
Recursive computation using PD-functions
(here $-S = \{1, \dots, p\} \setminus S$ denotes all indices not contained in S):

$$g_S(\mathbf{x}_S) = \hat{f}_{S;PD}(\mathbf{x}_S) - \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \mathbb{E}_{-S}\left[\hat{f}(\mathbf{x}_S, {}_{-S})\right] - \sum_{V \subsetneq S} g_V(\mathbf{x}_V)$$

- Recursive computation:

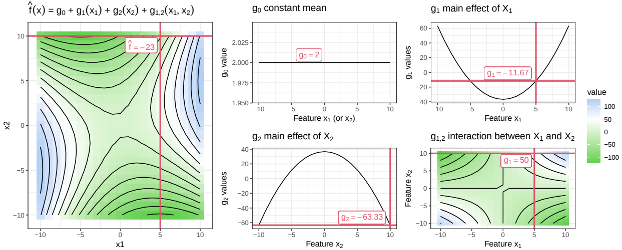
$$\begin{aligned} g_\emptyset &= \mathbb{E}_{[\hat{f}()]} \\ g_j(x_j) &= \mathbb{E}_{-j}\left[\hat{f}() \mid X_j = x_j\right] - g_\emptyset, \ \forall j \in \{1, \dots, p\} \\ &\vdots \\ g_{1,\dots,p}(\mathbf{x}) &= \hat{f}(\mathbf{x}) - \textstyle\sum_{S \subsetneq \{1,\dots,p\}} g_S(\mathbf{x}_S) \\ &= \hat{f}(\mathbf{x}) - g_{1,\dots,p-1}(x_1, \dots x_{p-1}) - \dots - g_{1,2}(x_1, x_2) \\ &\quad - g_p(x_p) - \dots - g_2(x_2) - g_1(x_1) - g_\emptyset \end{aligned}$$

STANDARD FANOVA – EXAMPLE



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

- Computation of components using feature values $x_1 = x_2 = (-10, -9, \dots, 10)^\top$ gives:



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

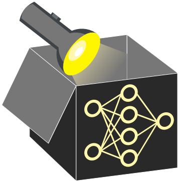
- $g_\emptyset = 2$
- $g_1(x_1) = -9.67$
- $g_2(x_2) = -65.33$
- $g_{1,2}(x_1, x_2) = 50$

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

STANDARD FANOVA - EXAMPLE

In-class task

STANDARD FANOVA - EXAMPLE REVISITED



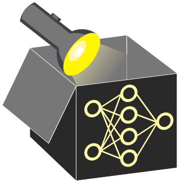
Example

$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0, 1]^2$ uniformly distributed

- **Intercept:**

$$g_\emptyset = \mathbb{E}\left[\hat{f}(x_1, x_2)\right] = \int_0^1 \int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_1 \, dx_2$$
$$= 4 - \left(\int_0^1 2x_1 \, dx_1\right) + \left(\int_0^1 0.3e^{x_2} \, dx_2\right) + \left(\int_0^1 |x_1| \, dx_1\right)\left(\int_0^1 x_2 \, dx_2\right)$$
$$= 4 - 1 + 0.3(e - 1) + 0.5^2 = 2.95 + 0.3e.$$

STANDARD FANOVA - EXAMPLE REVISITED



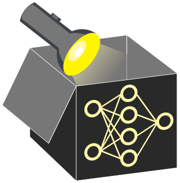
Example

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad \text{uniformly distributed}$$

- **First-order components:**

$$\begin{aligned}
g_1(x_1) &= \hat{f}_{1;PD}(x_1) - g_\emptyset = \left(\int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_2 \right) - g_\emptyset \\
&= 4 - 2x_1 + 0.3(e-1) + |x_1| \cdot \frac{1}{2} - (2.95 + 0.3e) \\
&= -2x_1 + 0.5|x_1| + 0.75 \\
g_2(x_2) &= \hat{f}_{2;PD}(x_2) - g_\emptyset = \left(\int_0^1 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \, dx_1 \right) - g_\emptyset \\
&= 4 - 1 + 0.3e^{x_2} + \frac{1}{2} \cdot x_2 - (2.95 + 0.3e) \\
&= 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05
\end{aligned}$$

STANDARD FANOVA - EXAMPLE REVISITED



Example

$$\hat{f}(x_1, x_2) = 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 \qquad (x_1, x_2) \in [0,1]^2 \qquad \text{uniformly distributed}$$

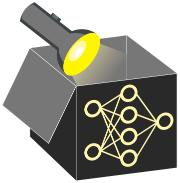
- **Second-order component:**

$$\begin{aligned} g_{12}(x_1, x_2) &= \hat{f}_{\{1,2\};PD}(x_1, x_2) - g_\emptyset - g_1(x_1) - g_2(x_2) \\ &= 4 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - (2.95 + 0.3e) \\ &\quad - (-2x_1 + 0.5|x_1| + 0.75) - (0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05) \\ &= |x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25 \end{aligned}$$

$\Rightarrow$ All components shifted to have mean 0

$\Rightarrow$ Parts of $|x_1|x_2$, which intuitively seems to be the “interaction term”, is attributed to the main effects (correctly, depends on distribution!)

ESTIMATE FANOVA IN PRACTICE



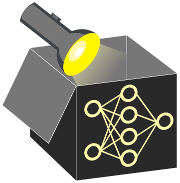
Main part: Calculate all PD-functions $\rightarrow 2^p$ **many PD-functions**
Estimation of a single PD-function: **Sampling**
(so-called **Monte-Carlo integration**)

- Same idea as for PDPs: Fix **grid values** for features x_S
  Here: Same grid for all features over the whole algorithm
- Estimate integral by sampling: for grid value $\mathbf{x}_S^*$:

$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \mathbb{E}_{-S}\left[\,\hat{f}(\mathbf{x}_S^*, {-S}\,)\,\right] \approx \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\mathbf{x}_S^*, \mathbf{x}^{(i)}_{-S})$$

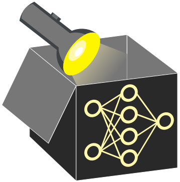
- Or: for each grid value $\mathbf{x}_S^*$, sample only $n_s < n$ many random samples (e.g. sampling uniformly)

VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)

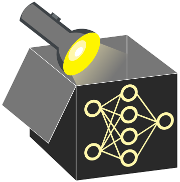
VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

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

VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



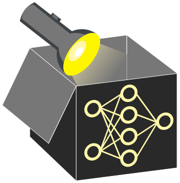
- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

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

- In other words: Single components uncorrelated (see later)

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

VARIANCE DECOMPOSITION - WHY "FUNCTIONAL ANOVA"?



- Decomp. of $\hat{f}(\mathbf{x})$ allows for "functional analysis of variance" (fANOVA)
- One can prove: If features independent $\Rightarrow$ additive decomposition of variance of $\hat{f}$ possible without covariances:

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

- In other words: Single components uncorrelated (see later)

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

$\rightarrow$ **Sobol index**: Fraction of variance explained by some component $g_V(\mathbf{x}_V)$:

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

$\rightsquigarrow$ Usable as importance measure of component $g_V(\mathbf{x}_V)$