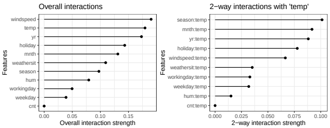
Interpretable Machine Learning

Friedman's H-Statistic



Learning goals

- Friedman's H-statistic with two purposes:
- Measure general *k*-way interactions between arbitrary features
- Measure a single feature's overall interaction strength

IDEA ▸ Friedman and Popescu (2008)

2-way interaction:

- Two features j and k do not interact, if their 2-way interaction component in functional decomposition $g_{\{j,k\}}$ is 0
- Idea from standard fANOVA: PD-function contains all components:

$$\hat{f}_{\{jk\},PD}(x_j, x_k) = g_{\emptyset} + g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

- Here: **Centered PD-functions** $\hat{f}^c_{S,PD}(\mathbf{x}_S) = \hat{f}_{S,PD}(\mathbf{x}_S) - g_{\emptyset}$

$$\Rightarrow \hat{f}^c_{\{jk\},PD}(x_j, x_k) = g_j(x_j) + g_k(x_k) + g_{\{j,k\}}(x_j, x_k)$$

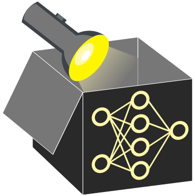
- **Definition:** A function $\hat{f}$ contains no 2-way interactions between j and k, if there exists a decomposition

$$\begin{aligned} & \hat{f}^c_{\{jk\},PD}(x_xj, x_k) = g_j(x_j) + g_k(x_k) \\ \Leftrightarrow \quad & \hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \end{aligned}$$

- This means: There are interactions
  $\Leftrightarrow$ Every possible decomposition must contain some non-zero term $g_{\{j,k\}}(x_j, x_k)$
- Again: remember GAMs

IDEA

▸ Friedman and Popescu (2008)



3-way interaction:

- **Definition:** $\hat{f}$ contains no 3-way interactions between features i, j, k, if corresponding 3-dimensional PD-function can be decomposed into lower-order terms:

$$\hat{f}_{\{ijk\},PD}(x_i, x_j, x_k) = g_\emptyset + g_i(x_i) + g_j(x_j) + g_k(x_k) \\ + g_{\{i,j\}}(x_i, x_j) + g_{\{i,k\}}(x_i, x_k) + g_{\{i,j\}}(x_j, x_k)$$

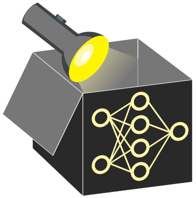
- **Example**:

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

- **Note:** Again using centered PD-functions $\hat{f}^c_{S,PD}$ instead of components g_S $\rightsquigarrow$ things get complicated, e.g. for 3 features, definition becomes:

$$\hat{f}^c_{\{ijk\},PD}(x_i, x_j, x_k) = \hat{f}^c_{\{ij\},PD}(x_i, x_j) + \hat{f}^c_{\{ik\},PD}(x_i, x_k) + \hat{f}^c_{\{jk\},PD}(x_j, x_k) \\ - \hat{f}^c_{i,PD}(x_i) - \hat{f}^c_{j,PD}(x_j) - \hat{f}^c_{k,PD}(x_k)$$

IDEA ▸ Friedman and Popescu (2008)



k-way interaction:

- **Analogous** for general k-way interactions between features $S = \{i_1, i_2, \ldots, i_k\}$: No interactions, if

$$\hat{f}_{S,PD}(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \sum_{V \subsetneq S} g_V(\mathbf{x}_V) = \sum_{\substack{V \subseteq S \\ |V| < k}} g_V(\mathbf{x}_V)$$

Overall interaction:

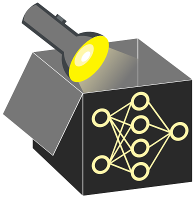
- Question: Does feature j interact with any other feature at all?

$\Rightarrow$ H-statistic analogous to 2-way interactions, but for feature sets $S = \{j\}$ and $-S = \{1, \ldots, p\} \setminus \{j\}$ instead of two single features:

$$\hat{f}(\mathbf{x}) - g_\emptyset = \hat{f}^c_{\{1,\ldots,p\},PD}(\mathbf{x}) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{-j,PD}(\mathbf{x}_{-j})$$

- $-j$ denotes $-S = \{1, \ldots, p\} \setminus \{j\}$, i.e. all other features
- $\hat{f}_{-j,PD}(\mathbf{x}_{-j})$: $(p-1)$-dim PD function of all p features except feature j

2-WAY INTERACTION STRENGTH



- **Question:** How to measure interaction strength without computing functional decomposition components g_S?
- **Idea:** Only use centered PD-functions

$$\hat{f}^c_{\{jk\},PD}(x_j, x_k) = \hat{f}^c_{j,PD}(x_j) + \hat{f}^c_{k,PD}(x_k) \text{ ?}$$

- **H-statistic** for 2-way interaction between feature j and k:

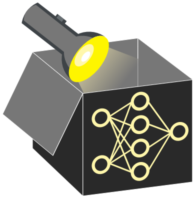
$$H^2_{jk} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{k,PD}(x_k^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c_{jk,PD}(x_j^{(i)}, x_k^{(i)})\right)^2}$$

$\Rightarrow$ H^2_{jk} measures strength of this interaction quantitatively
H^2_{jk} small (close to 0) for weak interaction, close to 1 for strong interaction

- **Note:** Again, definition also usable without any probabilities or data distribution

H-STATISTIC: EXAMPLES



Note: Again, definition also usable without any probability or data distribution

Example

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

$$\hat{f}^c_{1,PD}(x_1) = -2x_1 + 0.5|x_1| + 0.75$$

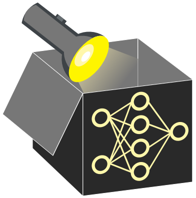
$$\hat{f}^c_{2,PD}(x_2) = 0.3e^{x_2} + 0.5x_2 - 0.3e + 0.05$$

$$\hat{f}^c_{1,2;PD}(x_1, x_2) = 1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e$$

$$\implies H^2_{12} = \frac{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{k,PD}(X_k)\right]}{\text{Var}\left[\hat{f}^c_{jk,PD}(X_j, X_k)\right]}$$

$$= \frac{\mathbb{E}\left[\left(|x_1|x_2 - 0.5|x_1| - 0.5x_2 + 0.25\right)^2\right]}{\mathbb{E}\left[\left(1.05 - 2x_1 + 0.3e^{x_2} + |x_1|x_2 - 0.3e\right)^2\right]} > 0$$

3-WAY INTERACTION STRENGTH



- Same idea as for 2-way, but different formula (see before):

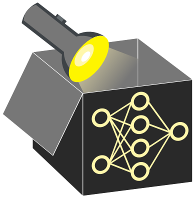
$$\hat{f}^c_{\{ijk\},PD}(x_i, x_j, x_k) = \hat{f}^c_{\{ij\},PD}(x_i, x_j) + \hat{f}^c_{\{ik\},PD}(x_i, x_k) + \hat{f}^c_{\{jk\},PD}(x_j, x_k) - \hat{f}^c_{i,PD}(x_i) - \hat{f}^c_{j,PD}(x_j) - \hat{f}^c_{k,PD}(x_k)$$

$\Rightarrow$ H-statistic for a 3-way interaction between features i, j and k:

$$H^2_{ijk} = \frac{\mathrm{Var}\left[\begin{array}{c}\hat{f}^c_{ijk,PD}(X_i, X_j, X_k) - \hat{f}^c_{ij,PD}(X_i, X_j) - \hat{f}^c_{ik,PD}(X_i, X_k) - \hat{f}^c_{jk,PD}(X_j, X_k) \\ +\hat{f}^c_{i,PD}(X_i) + \hat{f}^c_{j,PD}(X_j) + \hat{f}^c_{k,PD}(X_k)\end{array}\right]}{\mathrm{Var}\left[\hat{f}^c_{ijk,PD}(X_i, X_j, X_k)\right]}$$

- Analogous for higher order interactions, but more complicated

OVERALL INTERACTION STRENGTH



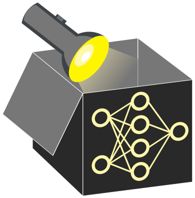
- Measure overall strength of interactions between feature j and all other features

$\Rightarrow$ **H-statistic** analogous to 2-way interaction:

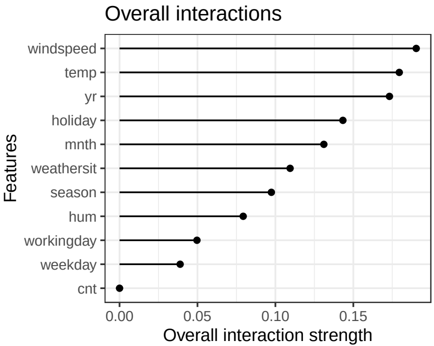
$$H_j^2 = \frac{\text{Var}\left[\hat{f}^c(\mathbf{X}) - \hat{f}^c_{j,PD}(X_j) - \hat{f}^c_{-j,PD}(\mathbf{X}_{-j})\right]}{\text{Var}\left[\hat{f}^c(\mathbf{X})\right]}$$

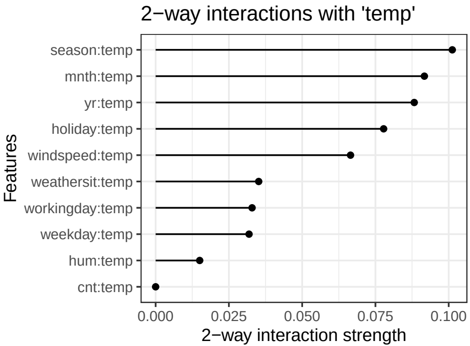
$$= \frac{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)}) - \hat{f}^c_{j,PD}(x_j^{(i)}) - \hat{f}^c_{-j,PD}(\mathbf{x}_{-j}^{(i)})\right)^2}{\sum_{i=1}^n \left(\hat{f}^c(\mathbf{x}^{(i)})\right)^2}$$

H-STATISTIC: EXAMPLE



Measure interactions of a random forest for the bike data set





Remarks and Conclusion:

- H-statistic provides **general definition of interactions** + an **algorithm for computation**
  Also adjustable to categorical / discrete features and / or function values
- For interaction order k still needs ¸ $\approx 2^k$ PD-functions
- Statistical test for whether interactions are present using this statistic