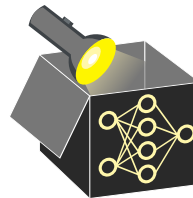


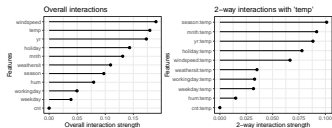
# Interpretable Machine Learning

## Friedman's H-Statistic



### Learning goals

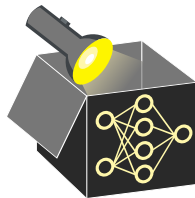
- Understand Friedman's H-statistic
- Measure 2-way interactions between pairs of features
- Measure a feature's overall interaction strength



**2-way interaction:** If two features  $j$  and  $k$  do not interact, their mean-centered PD function is

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

- $\hat{f}_{jk,PD}(x_j, x_k)$ : joint 2-dim PD function of feature  $j$  and  $k$
- $\hat{f}_{j,PD}(x_j)$  and  $\hat{f}_{k,PD}(x_k)$ : 1-dim PD functions of single features  $j$  and  $k$

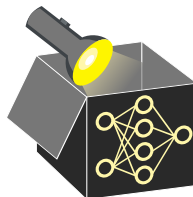


# IDEA

**Overall interaction:** If feature  $j$  does not interact with any other feature (denoted by index  $-j$ ), the mean-centered prediction function can be decomposed by

$$\hat{f}(\mathbf{x}) = \hat{f}_{j,PD}(x_j) + \hat{f}_{-j,PD}(\mathbf{x}_{-j})$$

- $\hat{f}(\mathbf{x})$ : mean-centered prediction function
- $\hat{f}_{j,PD}(x_j)$ : 1-dim PD function of feature  $j$
- $\hat{f}_{-j,PD}(\mathbf{x}_{-j})$ :  $(p - 1)$ -dim PD function of all  $p$  features except feature  $j$



## 2-WAY INTERACTION STRENGTH

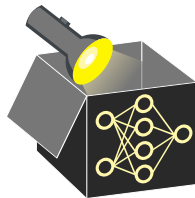
H-statistic measures interaction strength between feature  $j$  and  $k$  by

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

**Note:** The numerator is 0 if the two features  $x_j$  and  $x_k$  do not interact, i.e.,

$$\hat{f}_{jk,PD}(x_j, x_k) - \hat{f}_{j,PD}(x_j) - \hat{f}_{k,PD}(x_k) = 0.$$

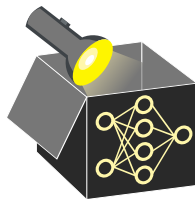
⇒ The smaller the values of  $H_{jk}^2$ , the weaker the interaction between  $x_j$  and  $x_k$ .



# OVERALL INTERACTION STRENGTH

Similarly, it is possible to measure whether a feature  $j$  interacts with any other feature (Overall interaction strength):

$$H_j^2 = \frac{\sum_{i=1}^n \left[ \hat{f}(x^{(i)}) - \hat{f}_{j,PD}(x_j^{(i)}) - \hat{f}_{-j,PD}(x_{-j}^{(i)}) \right]^2}{\sum_{i=1}^n \left[ \hat{f}(x^{(i)}) \right]^2}$$



**Example:** Inspect interactions of a random forest for the bike data

