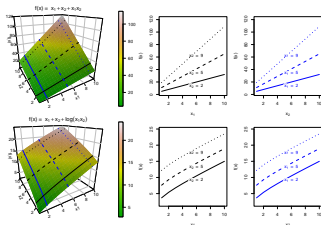


# Interpretable Machine Learning

## Feature Interactions

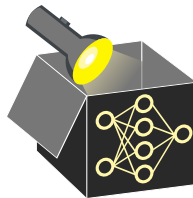


### Learning goals

- Feature interactions
- Difference to feature dependencies

# FEATURE INTERACTIONS

- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between  $X$  and  $\hat{f}(X)$  or  $X$  and  $Y = f(X)$ )
  - ↪ Feature dependencies may lead to feature interactions in a model



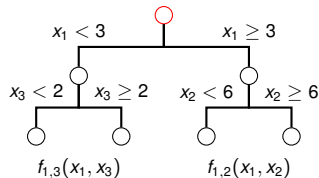
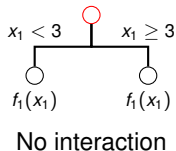
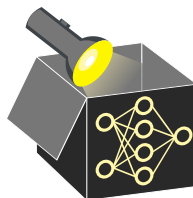
# FEATURE INTERACTIONS

- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between  $X$  and  $\hat{f}(X)$  or  $X$  and  $Y = f(X)$ )
  - ↪ Feature dependencies may lead to feature interactions in a model
- No. of potential interactions increases exponentially with no. of features
  - ↪ Difficult to identify interactions, especially when features are dependent



# FEATURE INTERACTIONS

- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between  $X$  and  $\hat{f}(X)$  or  $X$  and  $Y = f(X)$ )
  - ↪ Feature dependencies may lead to feature interactions in a model
- No. of potential interactions increases exponentially with no. of features
  - ↪ Difficult to identify interactions, especially when features are dependent
- Interactions: A feature's effect on the prediction depends on other features
  - ↪ Example:  $\hat{f}(\mathbf{x}) = x_1 x_2 \Rightarrow$  Effect of  $x_1$  on  $\hat{f}$  depends on  $x_2$  and vice versa



Interactions:  $x_1$  and  $x_3$ ,  
 $x_1$  and  $x_2$

No interactions:  $x_2$  and  $x_3$

# FEATURE INTERACTIONS

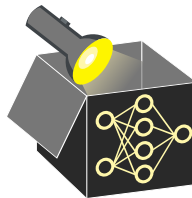
► Friedman and Popescu (2008)

**Definition:** A function  $f(\mathbf{x})$  contains an interaction between  $x_j$  and  $x_k$  if a difference in  $f(\mathbf{x})$ -values due to changes in  $x_j$  will also depend on  $x_k$ , i.e.:

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

⇒ If  $x_j$  and  $x_k$  do not interact,  $f(\mathbf{x})$  is sum of 2 functions, each independent of  $x_j$ ,  $x_k$ :

$$f(\mathbf{x}) = f_{-j}(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p) + f_{-k}(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_p)$$

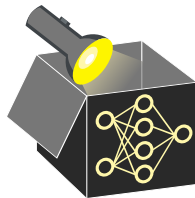


# FEATURE INTERACTIONS

Example:  $f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2$  (not separable)

$$\mathbb{E} \left[ \frac{\partial^2 (x_1 + x_2 + x_1 \cdot x_2)}{\partial x_1 \partial x_2} \right]^2 = \mathbb{E} \left[ \frac{\partial (1 + x_2)}{\partial x_2} \right]^2 = 1 > 0$$

$\Rightarrow$  interaction between  $x_1$  and  $x_2$

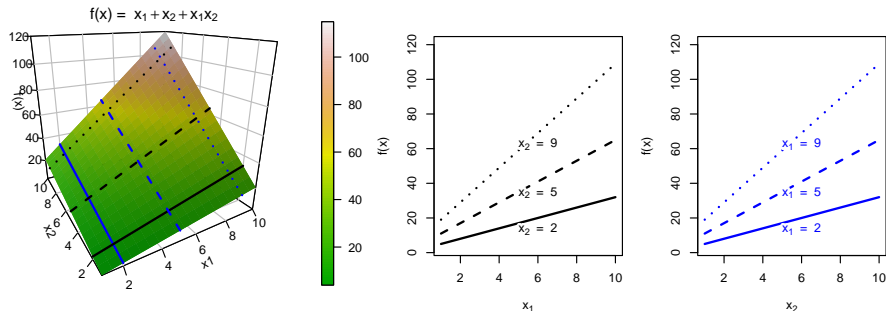


# FEATURE INTERACTIONS

Example:  $f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2$  (not separable)

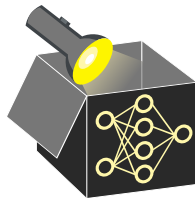
$$\mathbb{E} \left[ \frac{\partial^2 (x_1 + x_2 + x_1 \cdot x_2)}{\partial x_1 \partial x_2} \right]^2 = \mathbb{E} \left[ \frac{\partial (1 + x_2)}{\partial x_2} \right]^2 = 1 > 0$$

⇒ interaction between  $x_1$  and  $x_2$



- Effect of  $x_1$  on  $f(\mathbf{x})$  varies with  $x_2$  (and vice versa)

⇒ Different slopes



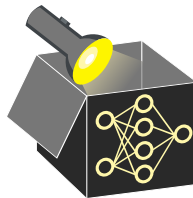
# FEATURE INTERACTIONS

Example of separable function:

$$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = x_1 + x_2 + \log(x_1) + \log(x_2)$$

$$\Rightarrow f(\mathbf{x}) = f_1(x_1) + f_2(x_2) \text{ with } f_1(x_1) = x_1 + \log(x_1) \text{ and } f_2(x_2) = x_2 + \log(x_2)$$

$$\Rightarrow \text{no interactions due to separability, also } \mathbb{E} \left[ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} \right]^2 = 0$$





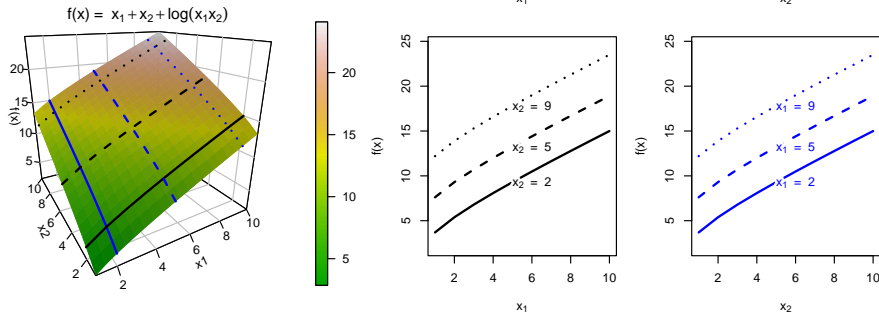
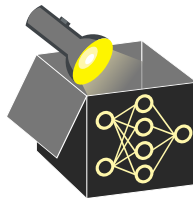
# FEATURE INTERACTIONS

Example of separable function:

$$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = x_1 + x_2 + \log(x_1) + \log(x_2)$$

$$\Rightarrow f(\mathbf{x}) = f_1(x_1) + f_2(x_2) \text{ with } f_1(x_1) = x_1 + \log(x_1) \text{ and } f_2(x_2) = x_2 + \log(x_2)$$

$$\Rightarrow \text{no interactions due to separability, also } \mathbb{E} \left[ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} \right]^2 = 0$$



- Effect of  $x_1$  on  $f(\mathbf{x})$  stays the same for different  $x_2$  values (and vice versa)

$\Rightarrow$  Parallel lines at different horizontal (blue) or vertical (black) slices