Interpretable Machine Learning

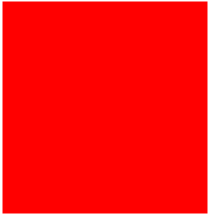
Interaction

Learning goals

- Feature interactions

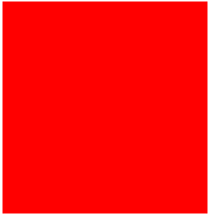
FEATURE INTERACTIONS

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



FEATURE INTERACTIONS

- While feature dependencies concern data distribution, feature interactions occur in structure of model or DGP (e.g., functional relationship between X and $\hat{f}(X)$ or X and $Y = f(X)$)
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model
- Number of potential interactions in a model increases exponentially with number of features
  $\rightsquigarrow$ Interactions are difficult to identify, especially if feature dependencies are also present



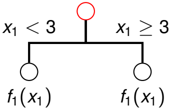
FEATURE INTERACTIONS

- While feature dependencies concern data distribution, feature interactions occur in structure of model or DGP (e.g., functional relationship between X and $\hat{f}(X)$ or X and $Y = f(X)$)

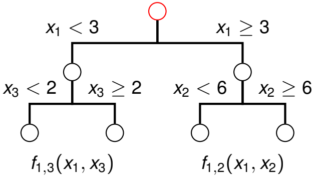
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model
- Number of potential interactions in a model increases exponentially with number of features

  $\rightsquigarrow$ Interactions are difficult to identify, especially if feature dependencies are also present
- With interactions present, a feature's effect on the prediction depends on other features

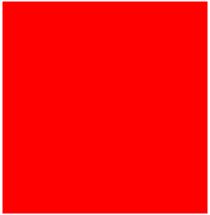
  $\rightsquigarrow$ $\hat{f}(\mathbf{x}) = x_1 x_2 \Rightarrow$ Effect of x_1 on $\hat{f}$ depends on x_2 and vice versa



No interaction



Interactions: x_1 and x_3,
x_1 and x_2



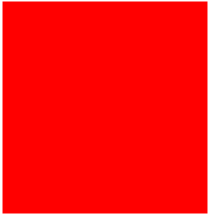
FEATURE INTERACTIONS ▸ Friedman and Popescu (2008)

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.:

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

$\Rightarrow$ If x_j and x_k do not interact, $f(\mathbf{x})$ is a sum of two functions, each independent from x_j and 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 ▸ 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.:

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

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

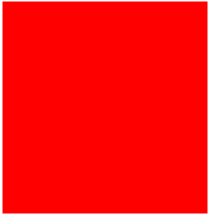
$$f(\mathbf{x}) = f_{-j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) + f_{-k}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)$$

Example ($f(\mathbf{x})$ not separable):

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

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

$\Rightarrow$ interaction between x_1 and x_2



FEATURE INTERACTIONS ▸ 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.:

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

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

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

Example ($f(\mathbf{x})$ not separable):

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

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

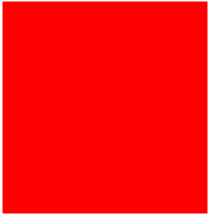
$\Rightarrow$ interaction between x_1 and x_2

Example ($f(\mathbf{x})$ separable):

$$\begin{aligned} f(\mathbf{x}) &= x_1 + x_2 + \log(x_1 \cdot x_2) \\ &= x_1 + x_2 + \log(x_1) + \log(x_2) \\ &= f_1(x_1) + f_2(x_2), \text{ with} \end{aligned}$$

$f_1(x_1) = x_1 + \log(x_1)$ and
$f_2(x_2) = x_2 + \log(x_2)$

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

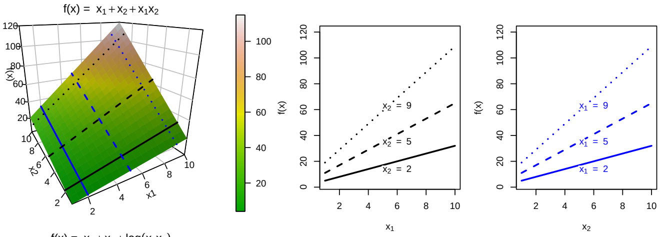


FEATURE INTERACTIONS

Interaction:

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

$\Rightarrow$ Different slopes



No interaction:

- 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