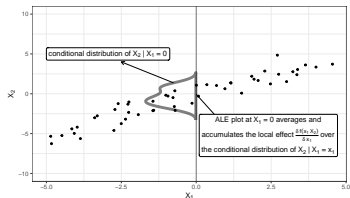
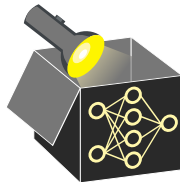


# Interpretable Machine Learning

## Regional Effects REPID



### Learning goals

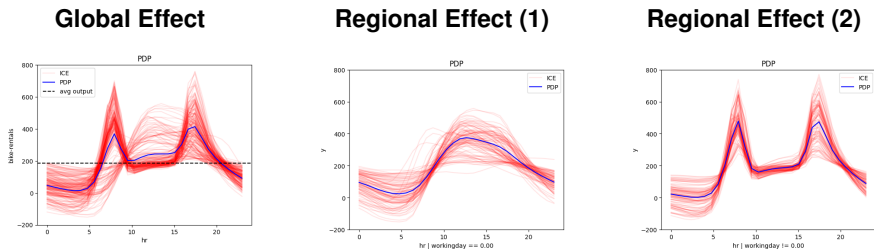
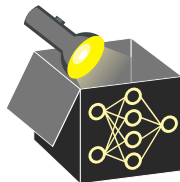
- Difference between feature effects and feature interactions
- REPID

# WHY REGIONAL EXPLANATIONS?

**Problem:** PD & ICE plots can be confounded by feature interactions.

**Solution:** Group homogeneous ICE curves in such a way that reduces the presence of individual interaction effects within a group

~> Regional effect plots (REPs).



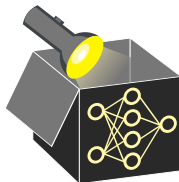
- Splitting by the "workingday" revealed 2 different patterns that we're clashed together in the initial plot
- $\cup_i(\text{regional\_explanation}_i) = \text{global\_explanation}$
- $\text{Fidelity}(\text{regional\_explanation}_i) > \text{Fidelity}(\text{global\_explanation})$

# ICE CURVE: LOCAL FEATURE EFFECTS

**Question:** How do feature changes affect the prediction for **one obs.**?

**Idea:** Split  $\mathbf{x} = (x_j, \mathbf{x}_{-j})$  into  $x_j$  (feat of interest) and  $\mathbf{x}_{-j}$  (remaining feats)

- Replace observed values  $x_j$  with **grid values**  $\tilde{x}_j$  while keeping  $\mathbf{x}_{-j}$  fixed
- Visualize function  $\hat{f}(\tilde{x}_j, \mathbf{x}_{-j})$  for varying  $\tilde{x}_j$  (ICE)



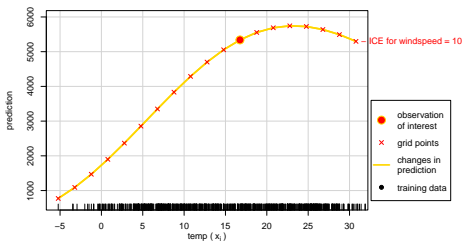
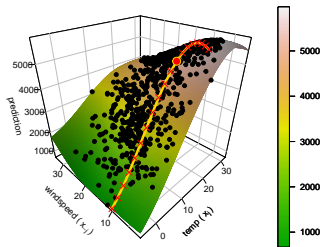
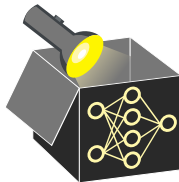
# ICE CURVE: LOCAL FEATURE EFFECTS

**Question:** How do feature changes affect the prediction for **one obs.**?

**Idea:** Split  $\mathbf{x} = (x_j, \mathbf{x}_{-j})$  into  $x_j$  (feat of interest) and  $\mathbf{x}_{-j}$  (remaining feats)

- Replace observed values  $x_j$  with **grid values**  $\tilde{x}_j$  while keeping  $\mathbf{x}_{-j}$  fixed
- Visualize function  $\hat{f}(\tilde{x}_j, \mathbf{x}_{-j})$  for varying  $\tilde{x}_j$  (ICE)

**Example:** SVM prediction surface (left), select obs. and visualize changes in prediction for varying  $x_2$  while keeping  $x_1$  fixed  $\Rightarrow$  **local interpretation**



# PD PLOT - GLOBAL FEATURE EFFECTS

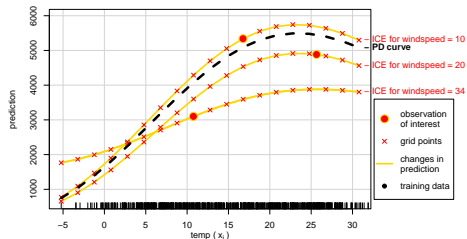
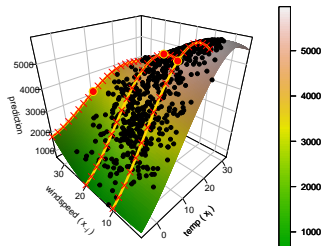
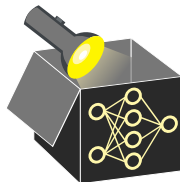
**Question:** How do changes of feat values affect model prediction **on avg.**?

- **PD function:** Integrate out effect of  $X_{-j}$  to obtain marginal effect of  $x_j$

$$f_j^{PD}(\tilde{x}_j) = \mathbb{E}_{X_{-j}}[\hat{f}(\tilde{x}_j, X_{-j})] = \int \hat{f}(\tilde{x}_j, X_{-j}) d\mathbb{P}(X_{-j})$$

- **Estimate (MC integration):** Average ICE curves at grid points  $\tilde{x}_j$

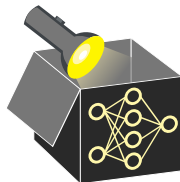
$$\hat{f}_j^{PD}(\tilde{x}_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}(\tilde{x}_j, \mathbf{x}_{-j}^{(i)})$$



# FEATURE INTERACTIONS

Hooker (2004, 2007): Functional ANOVA decomp. of a function

$$\hat{f}(\mathbf{x}) = g_0 + \underbrace{\sum_{j=1}^p g_j(x_j)}_{\text{main effect}} + \underbrace{\sum_{j \neq k} g_{j,k}(x_j, x_k)}_{\text{two-way interaction effect}} + \cdots + \underbrace{g_{1,2,\dots,p}(\mathbf{x})}_{\text{p-way interaction effect}}$$



Friedman and Popescu (2008):

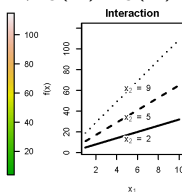
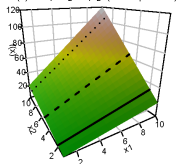
⇒ If  $x_j$  and  $\mathbf{x}_{-j}$  don't interact, we can decomp.  $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

⇒ If  $x_j$  and  $x_k$  don't interact, decomposition:  $f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

**Example:** Not additively separable:

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

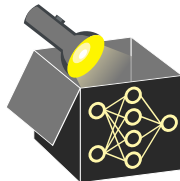
$f(\mathbf{x}) = x_1 + x_2 + x_1 x_2$  (not separable)



# FEATURE INTERACTIONS

Hooker (2004, 2007): Functional ANOVA decomp. of a function

$$\hat{f}(\mathbf{x}) = g_0 + \underbrace{\sum_{j=1}^p g_j(x_j)}_{\text{main effect}} + \underbrace{\sum_{j \neq k} g_{j,k}(x_j, x_k)}_{\text{two-way interaction effect}} + \cdots + \underbrace{g_{1,2,\dots,p}(\mathbf{x})}_{\text{p-way interaction effect}}$$



Friedman and Popescu (2008):

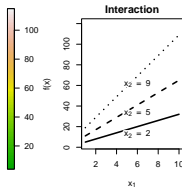
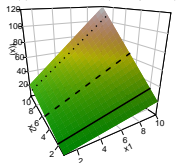
⇒ If  $x_j$  and  $\mathbf{x}_{-j}$  don't interact, we can decomp.  $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

⇒ If  $x_j$  and  $x_k$  don't interact, decomposition:  $f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

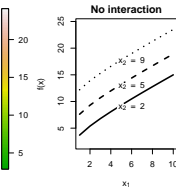
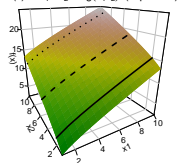
**Example: Separable:**

$$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = (x_1 + \log(x_1)) + (x_2 + \log(x_2)) = g_1(x_1) + g_2(x_2)$$

$f(x) = x_1 + x_2 + x_1 x_2$  (not separable)



$f(x) = x_1 + x_2 + \log(x_1 x_2)$  (separable)



# REPID: REGIONAL EFFECT PLOTS

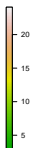
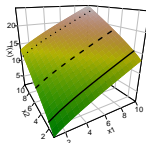
► “Herbinger et al.” 2022

**Recall:** Different shapes of ICE curves  $\Rightarrow$  interactions (ignore vertical shifts)  
 $\Rightarrow$  Focus on shape differences of **mean-centered ICE curves**.

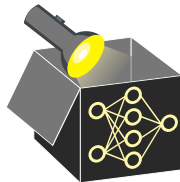
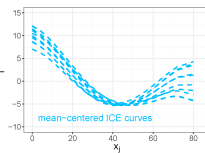
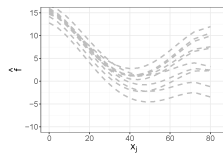
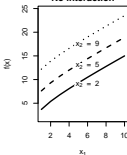
Mean-centered ICE curve for obs.  $\mathbf{x}$  evaluated at  $m$  grid points  $\tilde{x}_j^{(1)}, \dots, \tilde{x}_j^{(m)}$  is:

$$\hat{f}^c(\tilde{x}_j, \mathbf{x}_{-j}) = \hat{f}(\tilde{x}_j, \mathbf{x}_{-j}) - \frac{1}{m} \sum_{k=1}^m \hat{f}(\tilde{x}_j^{(k)}, \mathbf{x}_{-j})$$

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



No interaction





# REGIONAL EFFECTS - SYNTHETIC EXAMPLE

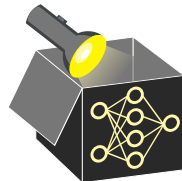
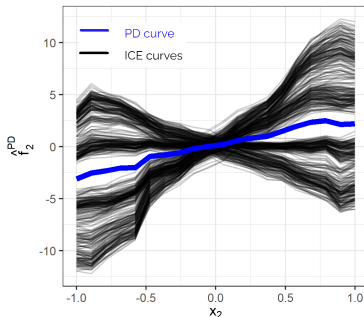
**Example:**  $X_1, X_2, X_6 \sim \mathcal{U}(-1, 1)$ ,  $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$  (all iid)

$\rightsquigarrow$  Ground truth:  $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{I}_{(X_1>0)} + 16X_2\mathbb{I}_{(X_3=0)} + \epsilon$

$\rightsquigarrow$  Model: Random forest

**Problem:**

- PD curve of  $X_2$  is misleading due to interactions  $\rightsquigarrow$  ICE
- ICE curves do not identify the interacting features



# REGIONAL EFFECTS - SYNTHETIC EXAMPLE

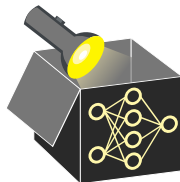
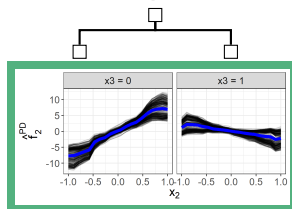
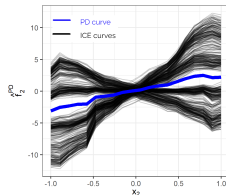
**Example:**  $X_1, X_2, X_6 \sim \mathcal{U}(-1, 1)$ ,  $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$  (all iid)

$\rightsquigarrow$  Ground truth:  $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{I}_{(X_1>0)} + 16X_2\mathbb{I}_{(X_3=0)} + \epsilon$

$\rightsquigarrow$  Model: Random forest

**Problem:**

- PD curve of  $X_2$  is misleading due to interactions  $\rightsquigarrow$  ICE
- ICE curves do not identify the interacting features



**Idea:** Find regions with similar ICE curves and aggregate them to regional effects

# REGIONAL EFFECTS - SYNTHETIC EXAMPLE

**Example:**  $X_1, X_2, X_6 \sim \mathcal{U}(-1, 1)$ ,  $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$  (all iid)

$\rightsquigarrow$  Ground truth:  $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{I}_{(X_1>0)} + 16X_2\mathbb{I}_{(X_3=0)} + \epsilon$

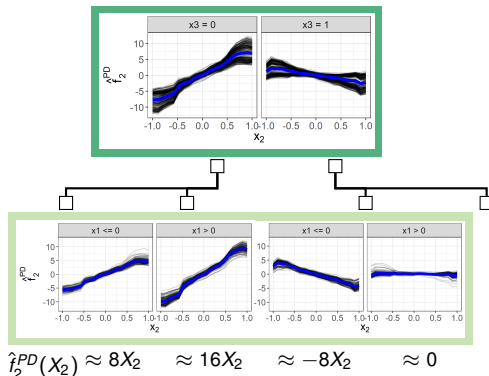
$\rightsquigarrow$  Model: Random forest

**Problem:**

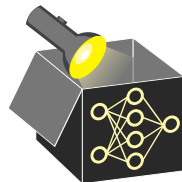
- PD curve of  $X_2$  is misleading due to interactions  $\rightsquigarrow$  ICE
- ICE curves do not identify the interacting features

**Idea:** Find regions with similar ICE curves and aggregate them to regional effects

**Regional effect** (blue curves)  $\hat{=}$  Estimate PD curve in each region



$\Rightarrow$  Additive decomposition of global feat effect

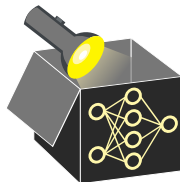
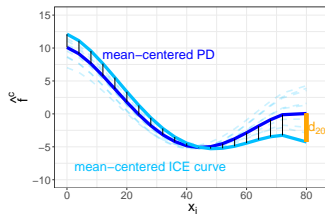


# REGIONAL EFFECTS - DETAILS

**Question:** How to split curves into regions?

Define risk as L2 loss of mean-centered ICE curves:

$$\mathcal{R}_j(\mathcal{N}) = \sum_{\mathbf{x} \in \mathcal{N}} \sum_{k=1}^m \underbrace{(\hat{f}^c(\tilde{x}_j^{(k)}, \mathbf{x}_{-j}) - \hat{f}_{j|\mathcal{N}}^{PD,c}(\tilde{x}_j^{(k)}))^2}_{d_k}$$



with the avg. feature effect in region  $\mathcal{N} \subseteq \mathcal{X}$ :

$$\hat{f}_{j|\mathcal{N}}^{PD,c}(\tilde{x}_j) = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{x} \in \mathcal{N}} \hat{f}^c(\tilde{x}_j, \mathbf{x}_{-j})$$

↪ Measures interaction-related heterogeneity (variance) of ICE curves in  $\mathcal{N}$

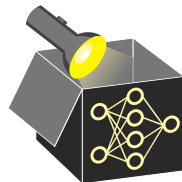
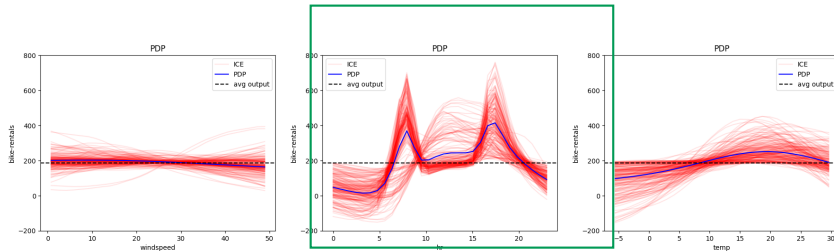
↪ Recursive partitioning (CART): Find best feat-split combo that solves

$$\arg \min_{z,t} \mathcal{R}_j(\mathcal{N}_{left}) + \mathcal{R}_j(\mathcal{N}_{right})$$

- $\mathcal{N}_{left} = \{\mathbf{x} \in \mathcal{N} | x_z \leq t\}$
- $\mathcal{N}_{right} = \{\mathbf{x} \in \mathcal{N} | x_z > t\}$
- Split point  $t$  for feature  $x_z, z \in -j$

**Intuition:** Is another feature  $x_z$  responsible for the heterogeneity (measured by  $\mathcal{R}_j$ )?

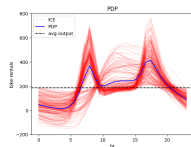
# REGIONAL EFFECT PLOTS - REAL EXAMPLE



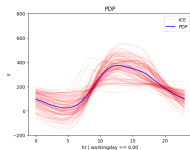
- Identify feature with highly heterogeneous local effects  
    ~> **hour: Most important; highly heterogeneous feat (highest variance)**
- Find regions in feature space where this heterogeneity is minimal  
    ~> Partition feature space using CART to minimize variance of mean-centered ICE curves within each region

# REGIONAL EFFECT PLOTS - REAL EXAMPLE

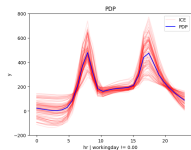
Regional effects of hour



workingday=False

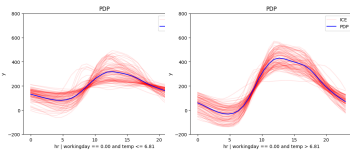


workingday=True



temp=cold /

temp=hot



temp=cold /

temp=hot

