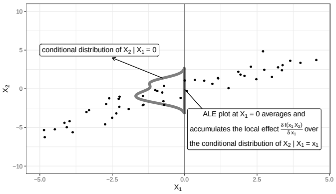
Interpretable Machine Learning

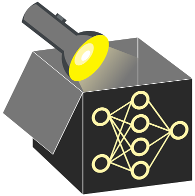
Regional Effects



Learning goals

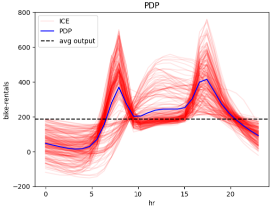
- Difference between feature effects and feature interactions
- REPID

WHY REGIONAL EXPLANATIONS?

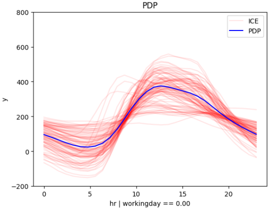


Problem: PD & ICE plots can be confounded by feature interactions.
Solution: Group homogeneous ICE curves in such a way that reduces the presence of individual interaction effects within a group $\rightsquigarrow$ Regional effect plots (REPs).

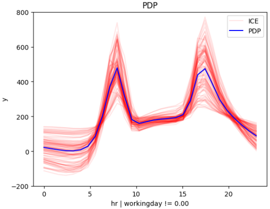
Global Effect



Regional Effect (1)

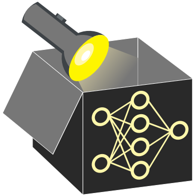


Regional Effect (2)



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

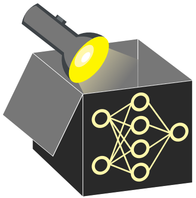
ICE CURVE: LOCAL FEATURE EFFECTS



Question: How do feature changes affect the prediction for **one observation**?
Idea: Split $\mathbf{x} = (x_j, \mathbf{x}_{-j})$ into x_j (feature of interest) and $\mathbf{x}_{-j}$ (remaining features)

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

ICE CURVE: LOCAL FEATURE EFFECTS

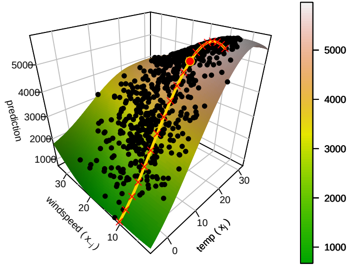


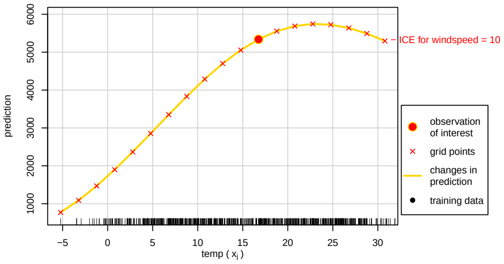
Question: How do feature changes affect the prediction for **one observation**?

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

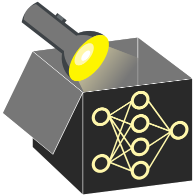
- Replace observed value x_j with grid values $\tilde{x}_j$ while keeping values $\mathbf{x}_{-j}$ fixed
- Visualize function $\hat{f}(\tilde{x}_j, \mathbf{x}_{-j})$ for varying $\tilde{x}_j$ (ICE)

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





PD PLOT - GLOBAL FEATURE EFFECTS



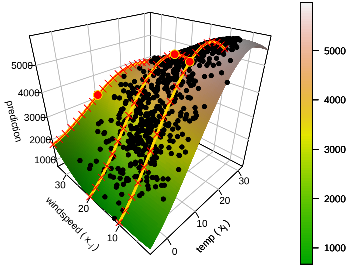
Question: How do changes of feature values affect model prediction **on average**?

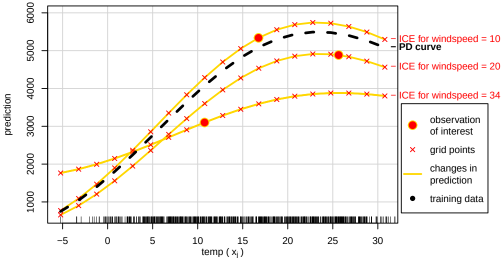
- **PD function**: Integrate out effect of X_{-j} to obtain marginal effect of x_j

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

- **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 decomposition of a function

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

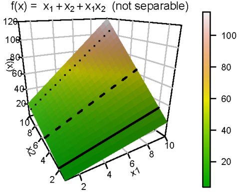
Friedman and Popescu (2008):

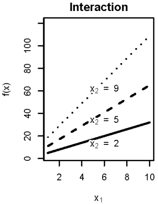
$\Rightarrow$ If x_j and $\mathbf{x}_{-j}$ do not interact, we can decompose $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

$\Rightarrow$ If x_j and x_k do not interact, we can decompose $f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

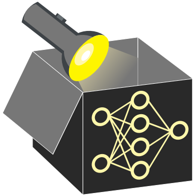
Example: Not additively separable:

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





FEATURE INTERACTIONS



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

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

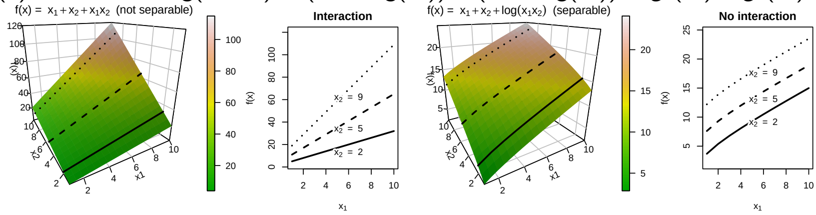
Friedman and Popescu (2008):

$\Rightarrow$ If x_j and $\mathbf{x}_{-j}$ do not interact, we can decompose $f(\mathbf{x}) = g_j(x_j) + g_{-j}(\mathbf{x}_{-j})$

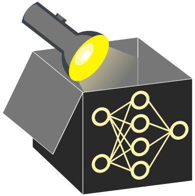
$\Rightarrow$ If x_j and x_k do not interact, we can decompose $f(\mathbf{x}) = g_{-j}(\mathbf{x}_{-j}) + g_{-k}(\mathbf{x}_{-k})$

Example: Separable:

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



REPID: REGIONAL EFFECT PLOTS ▸ Herbinger et al. (2022)

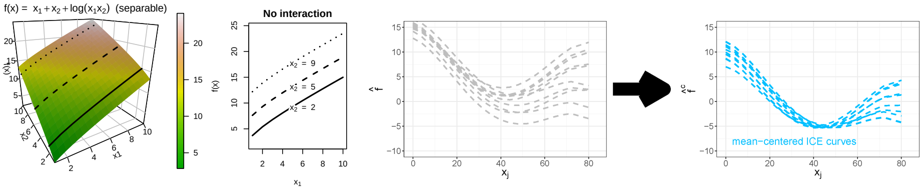


Recall: Different shapes of ICE curves indicate interactions (ignore vertical shifts)

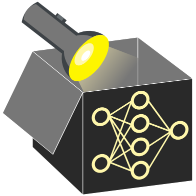
$\Rightarrow$ Focus on shape differences of **mean-centered ICE curves**.

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

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



REGIONAL EFFECTS - SYNTHETIC EXAMPLE



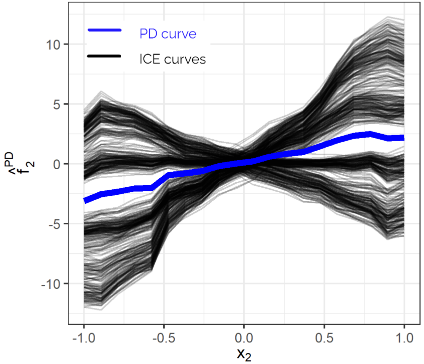
Example: $X_1, X_2, X_6 \sim \mathcal{U}(-1,1)$, $X_3, X_4, X_5 \sim \mathcal{B}(n, 0.5)$ (all iid)

$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 \boxed{-8X_2 + 8X_2 \mathbb{1}_{(X_1>0)} + 16X_2 \mathbb{1}_{(X_3=0)}} + \epsilon$

$\rightsquigarrow$ Model: Random forest

Problem:

- PD curve of $\boxed{X_2}$ 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)

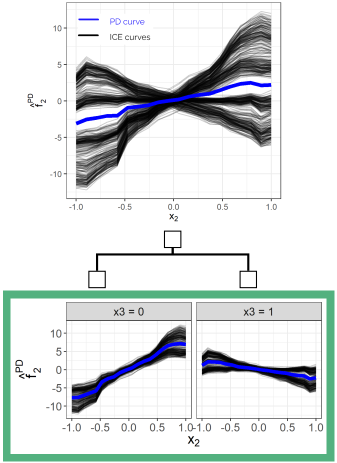
$\rightsquigarrow$ Ground truth: $f(X) = 0.2X_1 - 8X_2 + 8X_2 \mathbb{1}_{(X_1>0)} + \boxed{16X_2 \mathbb{1}_{(X_3=0)}} + \epsilon$

$\rightsquigarrow$ Model: Random forest

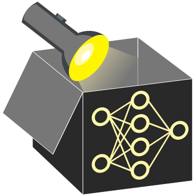
Problem:

- 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: $f(X) = 0.2X_1 - 8X_2 + 8X_2\mathbb{1}_{(X_1>0)} + 16X_2\mathbb{1}_{(X_3=0)} + \epsilon$

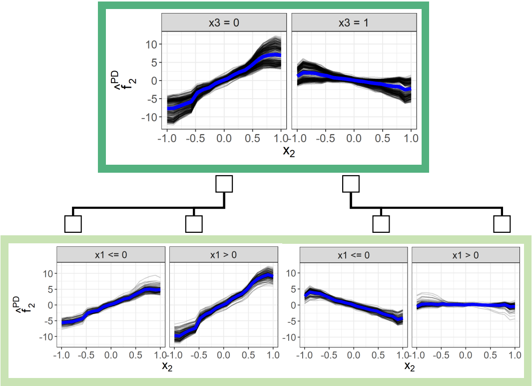
$\rightsquigarrow$ Model: Random forest

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



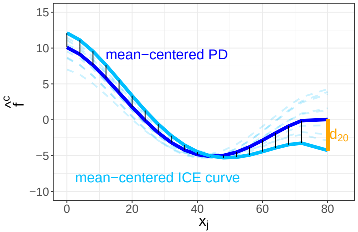
$\hat{f}_2^{PD}(X_2) \approx 8X_2 \quad \approx 16X_2 \quad \approx -8X_2 \quad \approx 0$

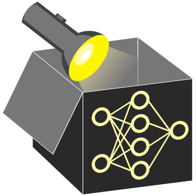
$\Rightarrow$ Additive decomposition of global feature effect

REGIONAL EFFECTS - DETAILS

Question: How to split the curves into regions?
Define risk as L2 loss of mean-centered ICE curves:

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





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

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

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

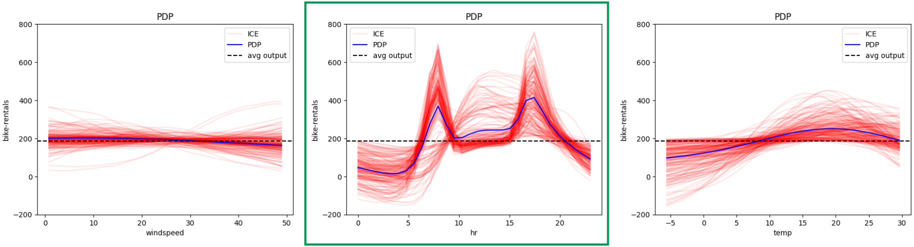
$\Rightarrow$ Recursive partitioning (CART): Find best feature-split combination that solves

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

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

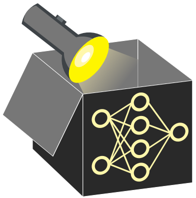
Intuition: Is another feature x_z responsible for the heterogeneity (measured by $\mathcal{R}_j$)?

REGIONAL EFFECT PLOTS - REAL EXAMPLE



- Identify feature with highly heterogeneous local effects
  ⇝ `hour`: Most important and highly heterogeneous feature (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`