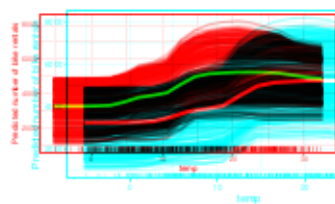


# Interpretable Machine Learning



## Partial Dependence (PD) plot



### Learning goals

#### Learning goals

- PD plots and relation to ICE plots
- Interpretation of PDP
- Extrapolation and Interactions in PDPs
- Centered ICE and PDP

**Definition:** PD function is expectation of  $\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})$  w.r.t. marginal distribution of features  $\mathbf{x}_{-S}$ :

$\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})$  w.r.t. marginal distribution of

features  $\mathbf{x}_{-S}$ :

$$f_{S,PD}(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_{-S}} \left( \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}) \right) = \int_{-\infty}^{\infty} \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}) dP(\mathbf{x}_{-S})$$

$$f_{S,PD}(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_{-S}} \left( \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}) \right)$$

**Estimation:** For a grid value  $\mathbf{x}_S^*$ , average ICE curves point-wise at  $\mathbf{x}_S^*$  over all observed  $\mathbf{x}_{-S}^{(i)}$ :

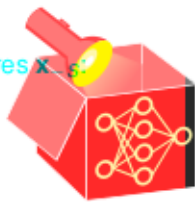
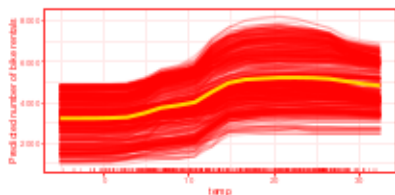
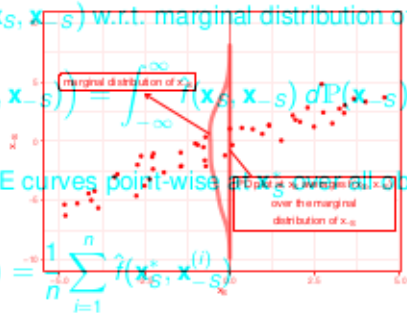
$$= \int_{-\infty}^{\infty} \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}) dP(\mathbf{x}_{-S})$$

$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \frac{1}{n} \sum_{i=1}^n \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$$

**Estimation:** For a grid value  $\mathbf{x}_S^*$ , average ICE curves point-wise at  $\mathbf{x}_S^*$  over all observed  $\mathbf{x}_{-S}^{(i)}$ :

$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \frac{1}{n} \sum_{i=1}^n \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$$

$$= \frac{1}{n} \sum_{i=1}^n \hat{f}_{S,ICE}^{(i)}(\mathbf{x}_S^*)$$



# PARTIAL DEPENDENCE PLOT FOR LINEAR MODEL



Assume a linear regression model with two features:

$x$	$x_1$	$x_2$	$x_3$	$f$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

$$\hat{f}(x) = \hat{f}(x_1, x_2) = \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \hat{\theta}_0$$

PD function for feature of interest  $S = \{1\}$  (with  $-S = \{2\}$ ) is:

$x$	$x_1$	$x_2$	$x_3$	$f$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

$$\frac{1}{3} \sum_{i=1}^3 f$$

$$\frac{1}{3} (0.4 + 0.6 + 0.1)$$

$$\frac{1}{3} (0.6 + 0.8 + 0.5)$$

$$\frac{1}{3} (0.7 + 0.9 + 0.6)$$

$x$	$x_1$	$x_2$	$x_3$	$f$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

$$f_{1,PD}(x_1) = \int_{-\infty}^{\infty} (\hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \hat{\theta}_0) dP(x_2)$$

$$= \hat{\theta}_1 x_1 + \hat{\theta}_2 \cdot \int_{-\infty}^{\infty} x_2 dP(x_2) + \hat{\theta}_0$$

$$= \hat{\theta}_1 x_1 + \hat{\theta}_2 \cdot E_{x_2}(x_2) + \hat{\theta}_0$$

$:= \text{const}$

Estimate PD function by **point-wise average of ICE curves at grid value**

$$x_S^* = x_1^* = 1$$

$\Rightarrow$  PD plot visualizes the function  $f_{1,PD}(x_1) \hat{=} \hat{\theta}_1 x_1 + \text{const}$  ( $\hat{=}$  feature effect of  $x_1$ ).

$$f_{1,PD}(x_1) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_1, x_{2,3})$$

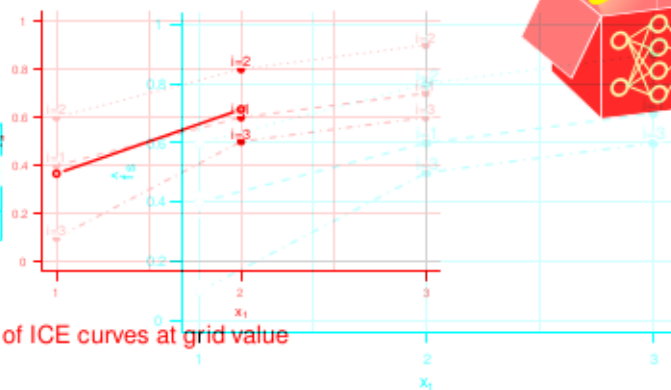
# PARTIAL DEPENDENCE

$i$	$x_1$	$x_2$	$x_3$	$f$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

$i$	$x_1$	$x_2$	$x_3$	$f$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

$i$	$x_1$	$x_2$	$x_3$	$f$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

$f$	$\frac{1}{3} \sum_{i=1}^3 f$
0.4	$\frac{1}{3}(0.4+0.6+0.1)$
0.6	$\frac{1}{3}(0.6+0.8+0.5)$
0.1	$\frac{1}{3}(0.7+0.9+0.6)$



Estimate PD function by **point-wise average of ICE curves at grid value**

$x_1^*$	$x_2$	$x_3$	$f$
1	3	4	0.7
2	5	8	0.6
3	6	9	0.6

Estimate PD function by **point-wise average of ICE curves at grid value**  $x_1^* = x_1^* = 1$ :

$$\hat{f}_{1,PD}(x_1^*) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

# PARTIAL DEPENDENCE

$i$	$x_1$	$x_2$	$x_3$	$f$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

$i$	$x_1$	$x_2$	$x_3$	$f$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

$i$	$x_1$	$x_2$	$x_3$	$f$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

$f$
0.4
0.6
0.1

$f$
$\frac{1}{3} \sum_{i=1}^3 f$
$\frac{1}{3}(0.4+0.6+0.1)$
$\frac{1}{3}(0.6+0.8+0.5)$
$\frac{1}{3}(0.7+0.9+0.6)$
0.6
0.8
0.5

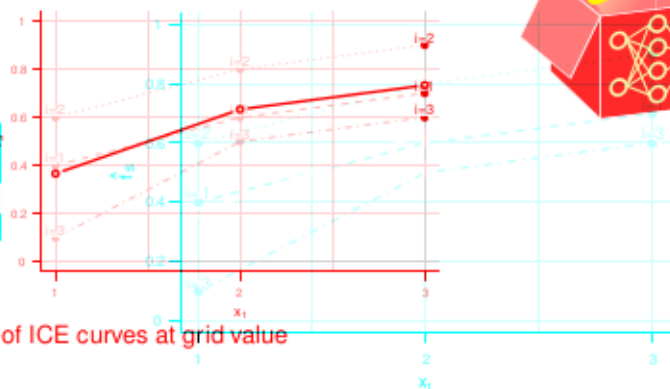
$f$
0.7
0.9
0.6

$$\frac{1}{3} \sum_{i=1}^3 f$$

$$\frac{1}{3} \sum_{i=1}^3 f$$

$$\frac{1}{3} \sum_{i=1}^3 f$$

$$\frac{1}{3} \sum_{i=1}^3 f$$



Estimate PD function by **point-wise average of ICE curves at grid value**

$x_1^*$	$x_2$	$x_3$	$f$
2	4	7	0.6
2	5	8	0.8
2	6	9	0.5

Estimate PD function by **point-wise average of ICE curves at grid value**  $x_1^* = x_1^* = 2$ :

$$\hat{f}_{1,PD}(x_1^*) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

# EXAMPLE: PD FOR LINEAR MODEL

Assume a linear regression model with two features:

$i$	$x_1$	$x_2$	$x_3$	$f$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

$$\hat{f}(\mathbf{x}) = \hat{f}(x_1, x_2) = \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \hat{\theta}_0$$

PD function for feature of interest  $S = \{1\}$  (with  $-S = \{2\}$ ) is:

$i$	$x_1$	$x_2$	$x_3$	$f$
1	2	4	7	0.5
2	2	5	8	0.8
3	2	6	9	0.5

$$\begin{aligned}
 f_{1,PD}(x_1) &= \mathbb{E}_{x_2} \left( \hat{f}(x_1, x_2) \right) = \int_{-\infty}^{\infty} (\hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \hat{\theta}_0) dP(x_2) \\
 &= \hat{\theta}_1 x_1 + \hat{\theta}_2 \cdot \int_{-\infty}^{\infty} x_2 dP(x_2) + \hat{\theta}_0 \\
 &= \hat{\theta}_1 x_1 + \underbrace{\hat{\theta}_2 \cdot \mathbb{E}_{x_2}(x_2)}_{:=const} + \hat{\theta}_0
 \end{aligned}$$

$i$	$x_1$	$x_2$	$x_3$	$f$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6

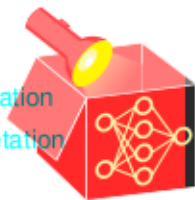
⇒ PD plot visualizes the function  $f_{1,PD}(x_1) = \hat{\theta}_1 x_1 + const$  ( $\hat{=}$  feature effect of  $x_1$ ).

Estimate PD function by **point-wise** average of ICE curves at grid value  $x_1^* = x_1 = 3$ :

$$\hat{f}_{1,PD}(x_1^*) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)})$$

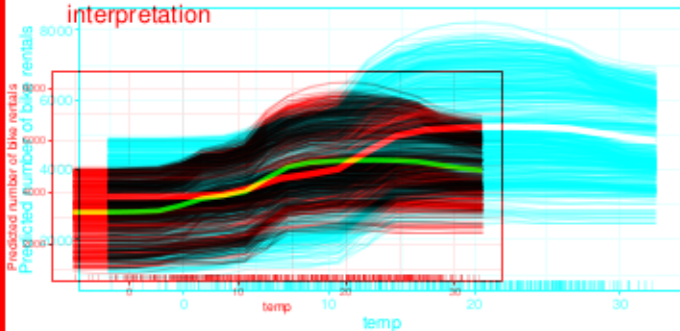


# INTERPRETATION: PD AND ICE

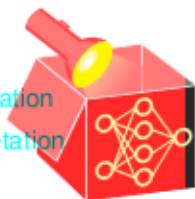


If feature varies:

- ICE: How does prediction of individual observation change? ⇒ **local** interpretation
- PD: How does average effect / expected prediction change? ⇒ **global** interpretation
- PD: How does average effect / expected prediction change? ⇒ **global** interpretation

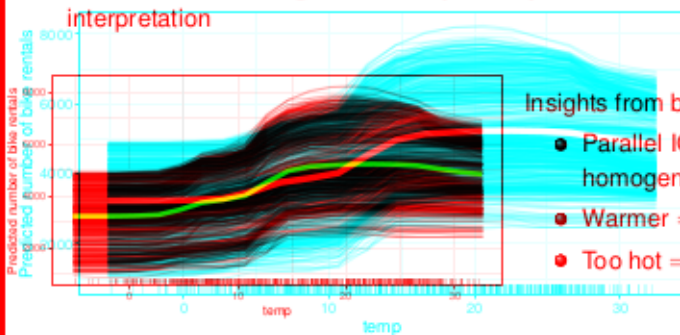


# INTERPRETATION: PD AND ICE



If feature varies:

- **ICE**: How does prediction of individual observation change? ⇒ **local** interpretation
- **PD**: How does average effect / expected prediction change? ⇒ **global** interpretation
- **PD**: How does average effect / expected prediction change? ⇒ **global** interpretation



insights from bike sharing data:

insights from bike sharing data:

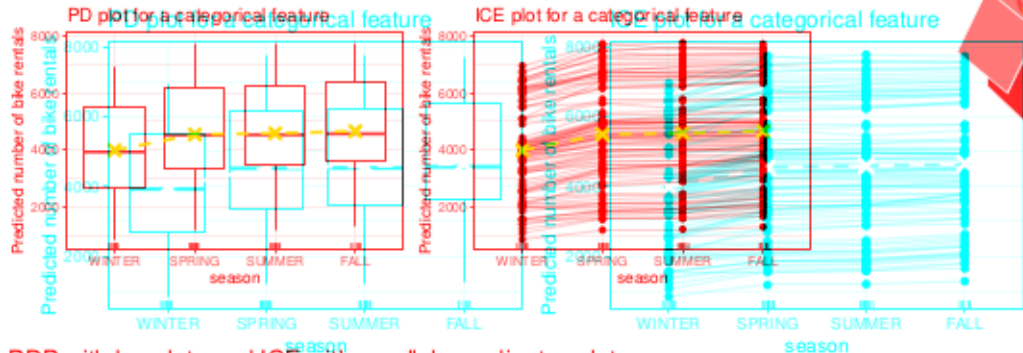
● Parallel ICE curves = homogeneous effect

● Warmer ⇒ more rented bikes

● Too hot ⇒ slightly less bikes



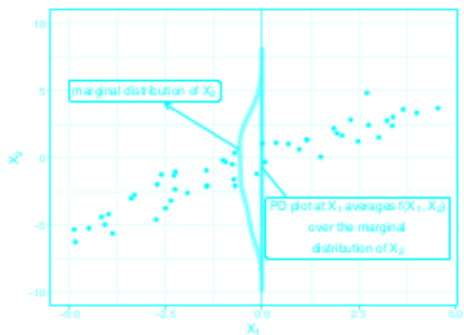
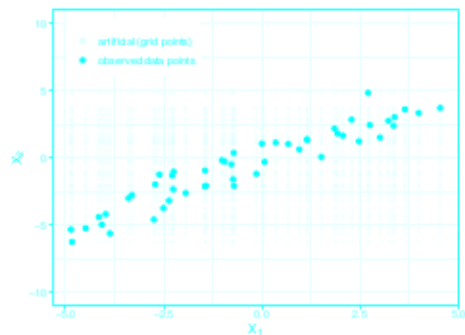
# INTERPRETATION: CATEGORICAL FEATURES



- PDP with boxplots and ICE with parallel coordinates plots
- NB: Categories can be unordered, if so, rather compare pairwise
  - PDP with boxplots and ICE with parallel coordinates plots
  - NB: Categories can be unordered, if so, rather compare pairwise

# COMMENTS ON EXTRAPOLATION

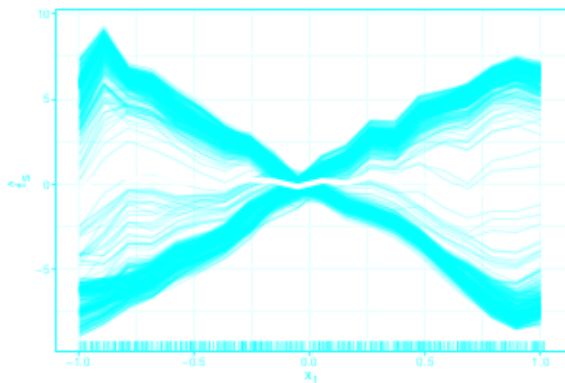
Extrapolation can cause issues in regions with few observations or if features are correlated



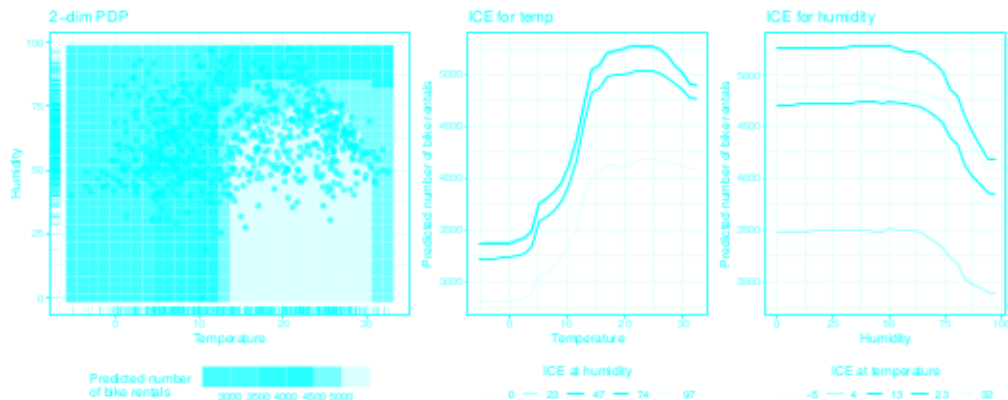
- **Example:** Features  $x_1$  and  $x_2$  are strongly correlated
- **Black points:** Observed points of the original data
- Grid points used to calculate the ICE and PD curves (several unrealistic values)
  - ⇒ PD plot at  $x_1 = 0$  averages predictions over the whole marginal distribution of feature  $x_2$
  - ⇒ May be problematic if model behaves strange outside training distribution

## COMMENTS ON INTERACTIONS

- PD plots: averaging of ICE curves might **obfuscate** heterogeneous effects and interactions
  - ⇒ Ideally plot ICE curves and PD plots together to uncover this fact
  - ⇒ Different shapes of ICE curves suggest interaction (but does not tell with which feature)



# COMMENTS ON INTERACTIONS - 2D PARTIAL DEPENDENCE



- Humidity and temperature interact with each other at high values (see shape difference)  
~> Shape of ICE curves at different horizontal and vertical slices varies (for high values)
- Low to medium humidity and high temperature  $\Rightarrow$  many rented bikes

## CENTERED ICE PLOT (C-ICE)

**Issue:** Difficult to identify heterogenous ICE curves if curves have different intercepts (are stacked)

**Solution:** Center ICE curves at fixed reference value  $x' \sim P(\mathbf{x}_S)$ , often  $x' = \min(\mathbf{x}_S)$

⇒ Easier to identify heterogenous shapes with c-ICE curves

$$\begin{aligned}\hat{f}_{S,cICE}^{(i)}(\mathbf{x}_S) &= \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}^{(i)}) - \hat{f}(x', \mathbf{x}_{-S}^{(i)}) \\ &= \hat{f}_S^{(i)}(\mathbf{x}_S) - \hat{f}_S^{(i)}(x')\end{aligned}$$

⇒ Visualize  $\hat{f}_{S,cICE}^{(i)}(\mathbf{x}_S^*)$  vs. grid point  $\mathbf{x}_S^*$

# CENTERED ICE PLOT (C-ICE)

**Issue:** Difficult to identify heterogenous ICE curves if curves have different intercepts (are stacked)

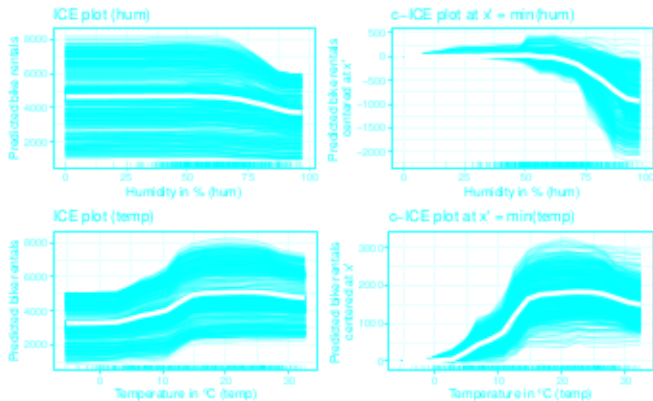
**Solution:** Center ICE curves at fixed reference value  $x' \sim P(\mathbf{x}_S)$ , often  $x' = \min(\mathbf{x}_S)$

⇒ Easier to identify heterogenous shapes with c-ICE curves

$$\begin{aligned}\hat{f}_{S, cICE}^{(i)}(\mathbf{x}_S) &= \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S}^{(i)}) - \hat{f}(x', \mathbf{x}_{-S}^{(i)}) \\ &= \hat{f}_S^{(i)}(\mathbf{x}_S) - \hat{f}_S^{(i)}(x')\end{aligned}$$

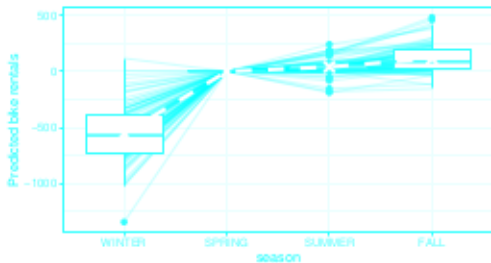
⇒ Visualize  $\hat{f}_{S, cICE}^{(i)}(\mathbf{x}_S^*)$  vs. grid point  $\mathbf{x}_S^*$

**Interpretation** (yellow curve in c-ICE):  
On average, the number of bike rentals  
at  $\sim 97\%$  humidity decreased by 1000  
bikes compared to a humidity of 0%



# CENTERED ICE PLOT (C-ICE)

For categorical features, c-ICE plots can be interpreted as in LMs due to reference value



## Interpretation:

- The reference category is  $x' = \text{SPRING}$
- Golden crosses: Average number of bike rentals if we jump from **SPRING** to any other season  
⇒ Number of bike rentals drops by  $\sim 560$  in **WINTER** and is slightly higher in **SUMMER** and **FALL** compared to **SPRING**