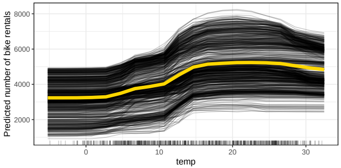
Interpretable Machine Learning

Feature Effects

Partial Dependence (PD) plot

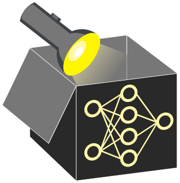


Learning goals

- PD plots and relation to ICE plots
- Interpretation of PDP

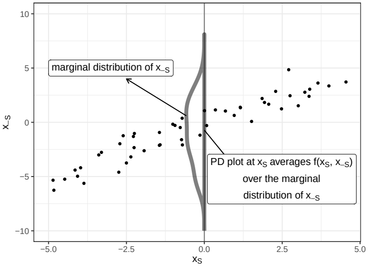
PARTIAL DEPENDENCE (PD)

▸ "Friedman" 2001



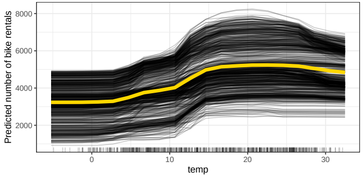
Definition: PD function is expectation of $\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})$ w.r.t. marginal distribution of features $\mathbf{x}_{-S}$:

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



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

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



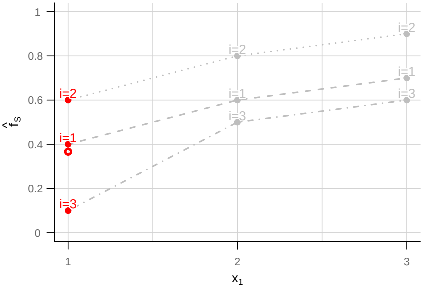
PARTIAL DEPENDENCE

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6



Estimate PD function by **point-wise** average of ICE curves at grid value $\boxed{\mathbf{x}_S^* = x_1^* = 1}$:

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

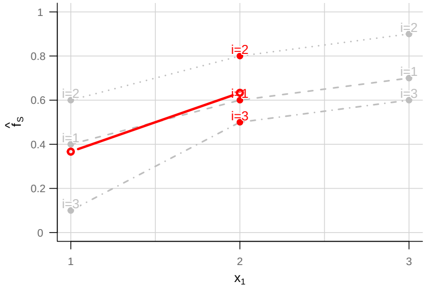
PARTIAL DEPENDENCE

i	x_S	x_2	x_3	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

i	x_S	x_2	x_3	$\hat{f}$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)

i	x_S	x_2	x_3	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6



Estimate PD function by **point-wise** average of ICE curves at grid value $\boxed{\mathbf{x}_S^* = x_1^* = 2}$:

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

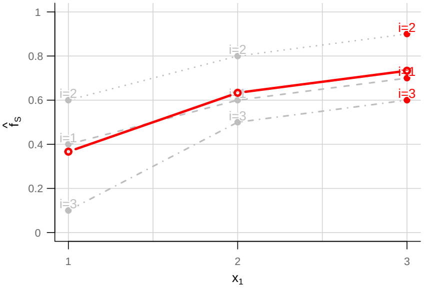
PARTIAL DEPENDENCE

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	2	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)

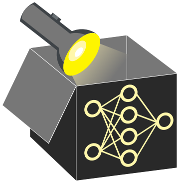
i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6



Estimate PD function by **point-wise** average of ICE curves at grid value $\boxed{\mathbf{x}_S^* = x_1^* = 3}$:

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

EXAMPLE: PD FOR LINEAR MODEL



Assume a linear regression model with two features:

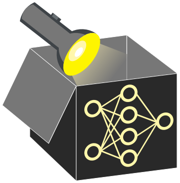
$$\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}_1, \mathbf{x}_2) = \hat{\theta}_1 \mathbf{x}_1 + \hat{\theta}_2 \mathbf{x}_2 + \hat{\theta}_0$$

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

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

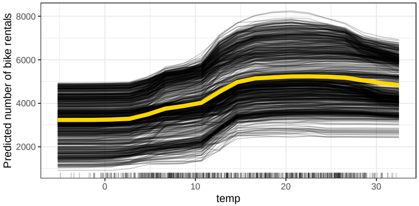
$\Rightarrow$ PD plot visualizes function $f_{1,PD}(\mathbf{x}_1) = \hat{\theta}_1 \mathbf{x}_1 + const$ ($\hat{=}$ feature effect of $\mathbf{x}_1$).

INTERPRETATION: PD AND ICE



If feature varies:

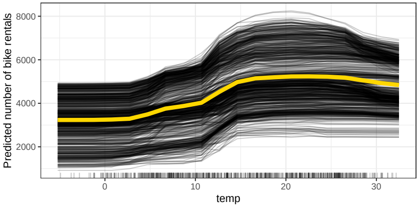
- **ICE:** How does **prediction of individual observation** change?
  $\Rightarrow$ **local** interpretation
- **PD:** How does **average effect / expected prediction** change?
  $\Rightarrow$ **global** interpretation



INTERPRETATION: PD AND ICE

If feature varies:

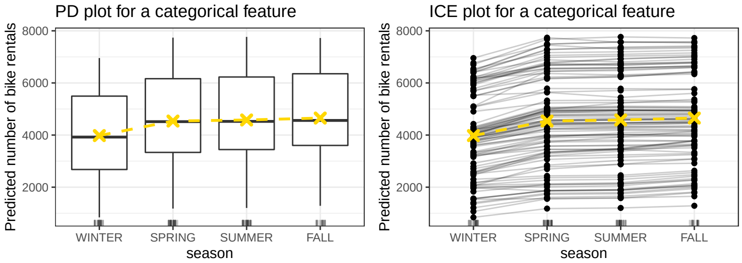
- **ICE:** How does **prediction of individual observation** change?
  $\Rightarrow$ **local** interpretation
- **PD:** How does **average effect / expected prediction** change?
  $\Rightarrow$ **global** interpretation

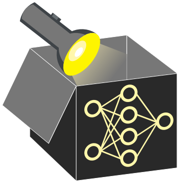


Insights from bike sharing data:

- Parallel ICE curves = homogeneous effect across obs.
- Warmer $\Rightarrow$ more rented bikes
- Too hot $\Rightarrow$ slightly less bikes
- Steepest increase in rentals occurs as temperature rises from 10℃ to 15℃.

INTERPRETATION: CATEGORICAL FEATURES





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