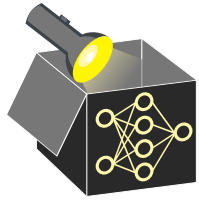
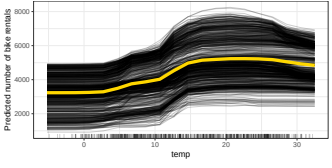
# Interpretable Machine Learning



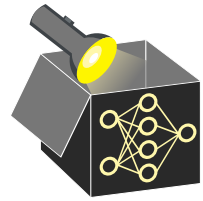
## PDP - Comments and Extensions



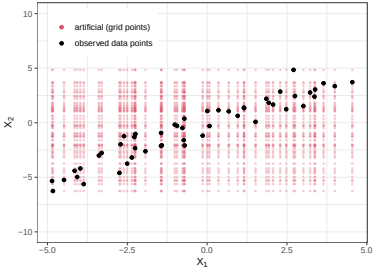

**Learning goals**

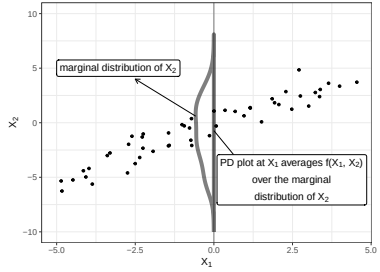
- PD plots and relation to ICE plots
- Interpretation of PDP
- Extrapolation and Interactions in PDPs
- Centered ICE and PDP

# COMMENTS ON EXTRAPOLATION



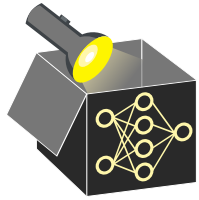
Extrapolation can cause issues in regions with few observations or if features are correlated



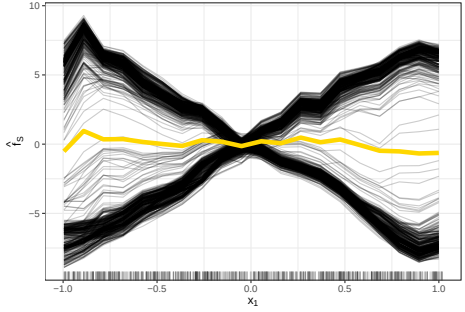




- **Example:** Features $x_1$ and $x_2$ are strongly correlated
- **Black points:** Observed points of the original data
- **Red:** Grid points used to calculate the ICE and PD curves (several unrealistic values)
  $\Rightarrow$ PD plot at $x_1 = 0$ averages predictions over the whole marginal distribution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

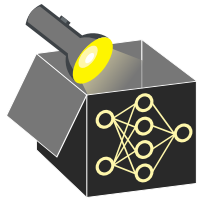
# COMMENTS ON INTERACTIONS

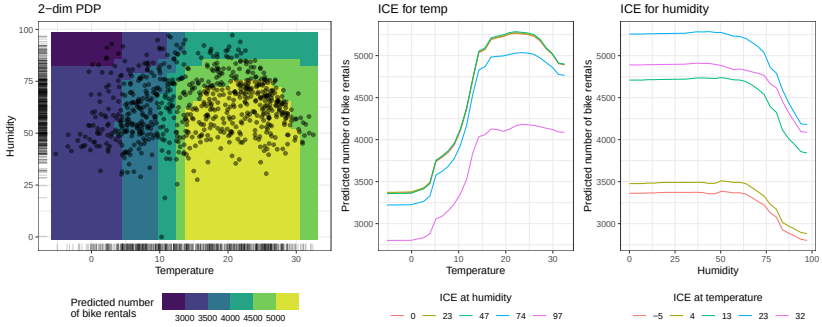


- PD plots: averaging of ICE curves might **obfuscate** heterogeneous effects and interactions
  $\Rightarrow$ Ideally plot ICE curves and PD plots together to uncover this fact
  $\Rightarrow$ Different shapes of ICE curves suggest interaction (but do not tell with which feature)



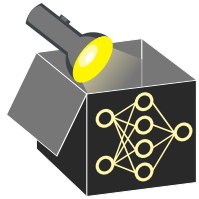
# COMMENTS ON INTERACTIONS - 2D PARTIAL DEPENDENCE






- Humidity and temperature interact with each other at high values (see shape difference)
  $\rightsquigarrow$ 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) ▸ Goldstein et al. (2015)



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

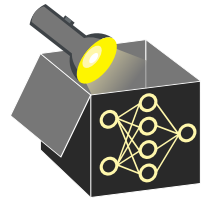
**Solution:** Center ICE curves at fixed reference value $x' \sim \mathbb{P}(\mathbf{x}_S)$, often $x' = \min(\mathbf{x}_S)$

$\Rightarrow$ Easier to identify heterogenous shapes with c-ICE curves

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

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

# CENTERED ICE PLOT (C-ICE) ▸ Goldstein et al. (2015)



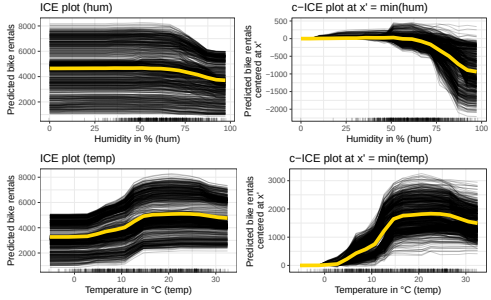
**Issue:** Difficult to identify heterogenous ICE curves if curves have different intercepts (are stacked)

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

$\Rightarrow$ Easier to identify heterogenous shapes with c-ICE curves

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

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

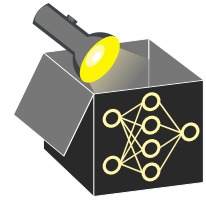



**Interpretation**

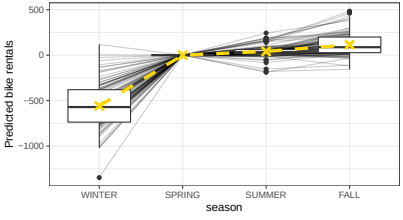
(yellow curve: analog to PDP the average of c-ICE curves):

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
  $\Rightarrow$ Number of bike rentals drops by $\sim 560$ in $\mathrm{WINTER}$ and is slightly higher in $\mathrm{SUMMER}$ and $\mathrm{FALL}$ compared to $\mathrm{SPRING}$