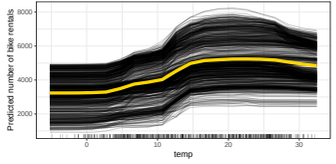
Interpretable Machine Learning

PDP - Comments and Extensions



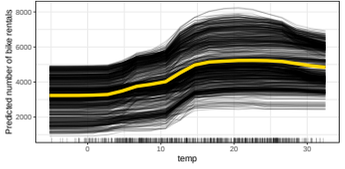
Learning goals

- Extrapolation and Interactions in PDPs
- Centered ICE and PDP

Interpretable Machine Learning

PDP - Comments and Extensions

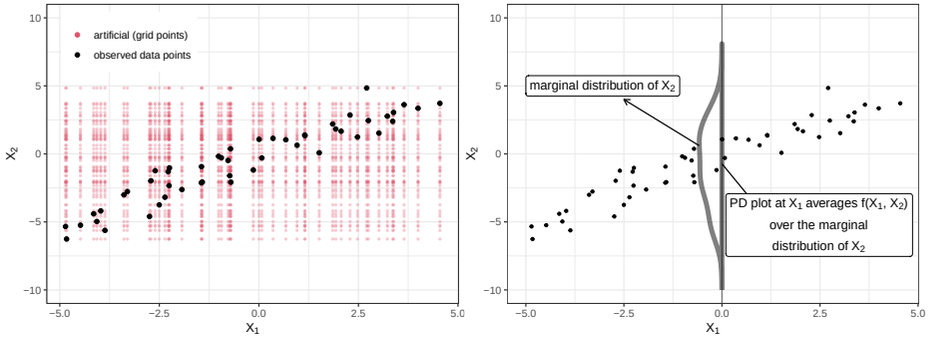
Feature Effects

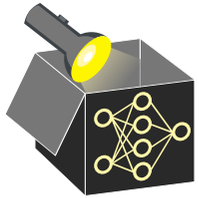


Learning goals

- Extrapolation and Interactions in PDPs
- Centered ICE and PDP

COMMENTS ON EXTRAPOLATION

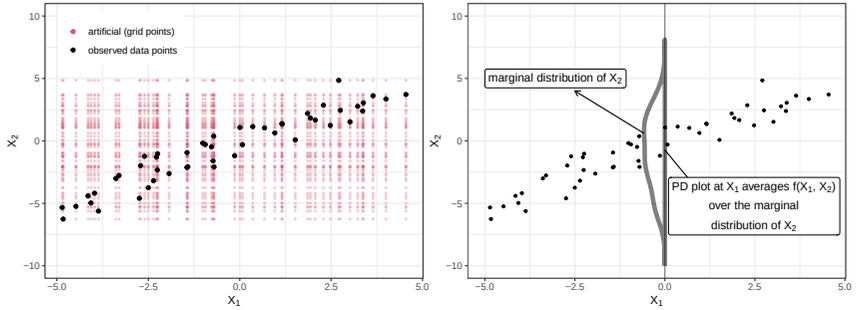




Extrapolation occurs in regions with few observations or if features are correlated

- **Example:** Features x_1 and x_2 are strongly correlated
- **Black points:** Observed points of the original data
- **Red:** Grid points to calculate ICE/PD (many unrealistic x_1, x_2 combinations)
  $\Rightarrow$ **PD at** $x_1 = 0$**:** Averages predictions over *full* marginal distribution of x_2
  $\Rightarrow$ **Issue:** Model may behave strangely outside training distribution
  $\Rightarrow$ Especially problematic for overfitted or interaction-heavy models

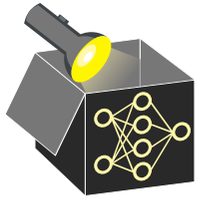
COMMENTS ON EXTRAPOLATION



Extrapolation occurs in regions with few obs. or if features are correlated

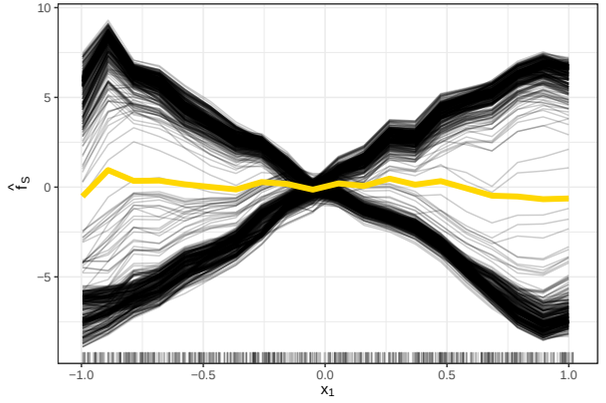
- **Example:** Features x_1 and x_2 are strongly correlated
- **Black points:** Observed points of the original data
- **Red:** Grid points to calculate ICE/PD (many unrealistic x_1, x_2 combinations)
  $\Rightarrow$ **PD at** $x_1 = 0$**:** Averages predictions over *full* marginal distribution of x_2
  $\Rightarrow$ **Issue:** Model may behave strangely outside training distribution
  $\Rightarrow$ Especially problematic for overfitted or interaction-heavy models

COMMENTS ON INTERACTIONS

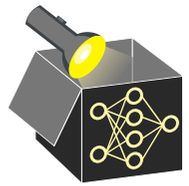


PD plots average ICE curves $\Rightarrow$ May **obscure heterogeneous effects** (interactions)

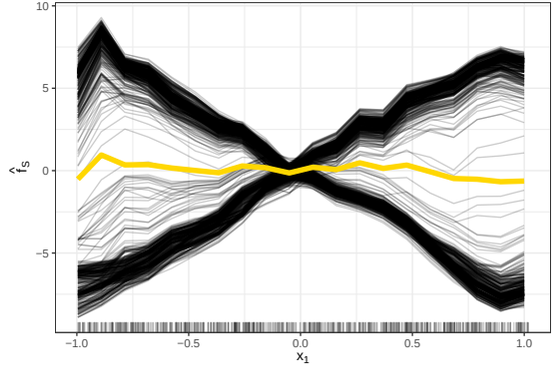
- **Example:** Feature x_1 = treatment dosage; x_2 = gender
  $\Rightarrow$ Males (increasing) and females (decreasing) respond differently to dosage
  $\Rightarrow$ PD curve (yellow) hides this divergence
- Plotting ICE and PD together helps detect interaction
- Diverse ICE shapes suggest interaction (but not with which feature)



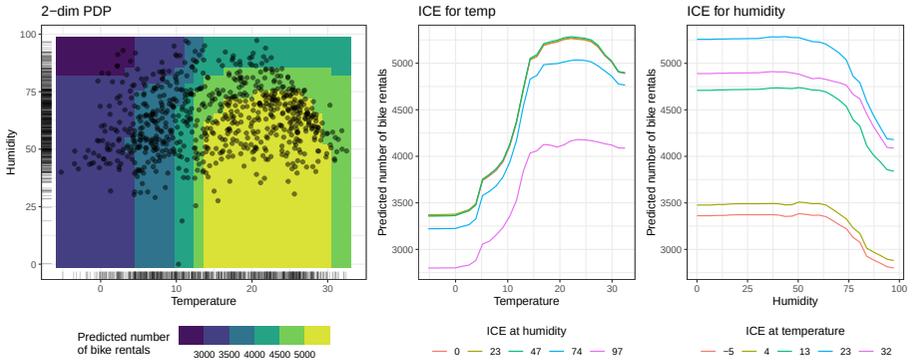
COMMENTS ON INTERACTIONS

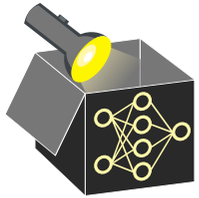


PD plots average ICE curves
$\rightsquigarrow$ May **obscure heterogeneous effects** (interactions)
- **Example:** Feature x_1 = treatment dosage; x_2 = gender
  $\Rightarrow$ Males ($\nearrow$) and females ($\searrow$) respond differently to dosage
  $\Rightarrow$ PD curve (yellow) hides this divergence
- Plotting ICE and PD together helps detect interaction
- Diverse ICE shapes suggest interaction (but not with which feature)



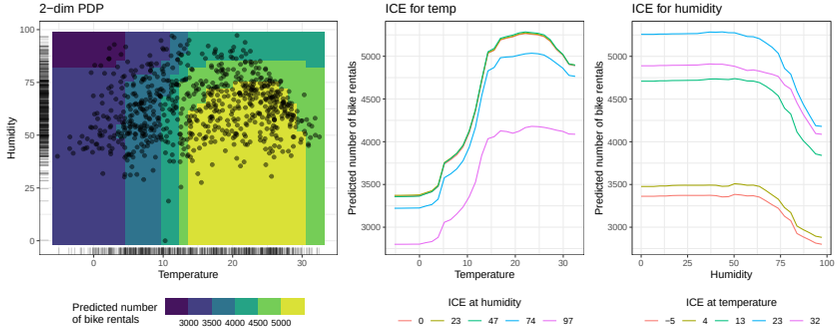
COMMENTS ON INTERACTIONS - 2D PD PLOT

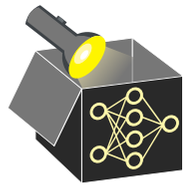




- Humidity and temperature interact at high values (see shape difference)
  ⇝ Shape of ICE curves changes across different (higher) values of other feature
  - ICE (temp): At high humidity, temp effect flattens (pink line)
  - ICE (hum): At high temperature, humidity effect falls steeper (blue/pink)
- Most rentals occur at *high temperature* and *low to medium humidity*

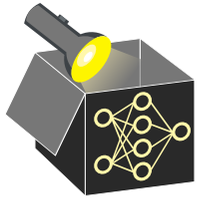
COMMENTS ON INTERACTIONS - 2D PD PLOT





- Humidity and temperature interact at high values (see shape difference)
  ⇝ ICE curve shape changes across different (higher) values of other feat.
  - ICE (temp): At high humidity, temp effect flattens (pink line)
  - ICE (hum): At high temp., humidity effect falls steeper (blue/pink)
- Most rentals occur at *high temperature* and *low to medium humidity*

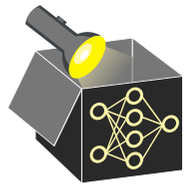
CENTERED ICE PLOT (C-ICE) ▸ Goldstein et al. (2015)



Issue: Varying-intercept (stacked) ICE curves obscure shape heterogeneity
Solution: Center ICE curves at fixed reference value, often $x' = \min(\mathbf{x}_S)$
$\Rightarrow$ Easier to identify heterogeneous shapes with c-ICE curves

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

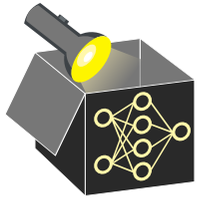
CENTERED ICE PLOT (C-ICE) ▸ GOLDSTEIN_2015



Issue: Varying-intercept (stacked) ICE curves obscure shape heterogeneity
Solution: Center ICE curves at fixed reference value, often $x' = \min(s)$
$\Rightarrow$ Easier to identify heterogeneous shapes with c-ICE curves

$$\overset{(i)}{s,cICE}(s) = (s, \xi_{-s}) - (x', \xi_{-s}) = \overset{(i)}{s}(s) - \overset{(i)}{s}(x')$$

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

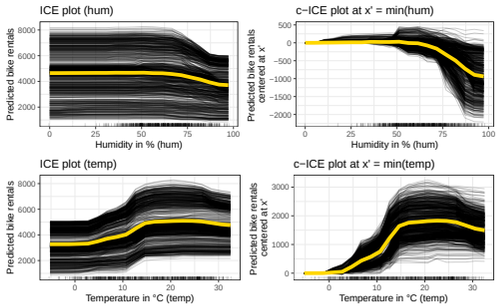


Issue: Varying-intercept (stacked) ICE curves obscure shape heterogeneity
Solution: Center ICE curves at fixed reference value, often $x' = \min(\mathbf{x}_S)$
$\Rightarrow$ Easier to identify heterogeneous shapes with c-ICE curves

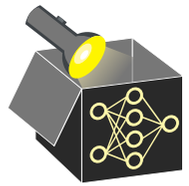
$$\hat{f}^{(i)}_{S,cICE}(\mathbf{x}_S) = \hat{f}(\mathbf{x}_S, \mathbf{x}^{(i)}_{-S}) - \hat{f}(x', \mathbf{x}^{(i)}_{-S}) = \hat{f}^{(i)}_S(\mathbf{x}_S) - \hat{f}^{(i)}_S(x')$$

Interpretation

- Yellow: c-PDP (mean of c-ICE)
- **c-PDP:** At 97% humidity, predicted rentals are 1000 fewer than at 0% humidity (on average)
- **Opening of c-ICE curves:** suggests interaction or varying effect across instances



CENTERED ICE PLOT (C-ICE) ▸ GOLDSTEIN_2015

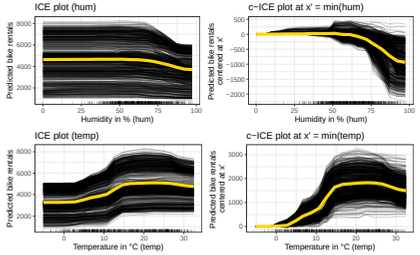


Issue: Varying-intercept (stacked) ICE curves obscure shape heterogeneity
Solution: Center ICE curves at fixed reference value, often $x' = \min(_S)$
$\Rightarrow$ Easier to identify heterogeneous shapes with c-ICE curves

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

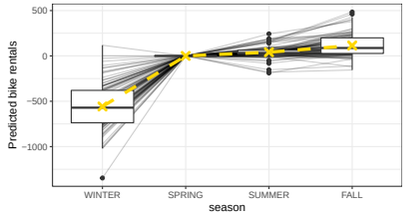
Interpretation

- Yellow: c-PDP (mean of c-ICE)
- **c-PDP:** At 97% humidity, predicted rentals are 1000 fewer than at 0% humidity (on average)
- **Opening of c-ICE curves:** suggests interaction or varying effect across instances



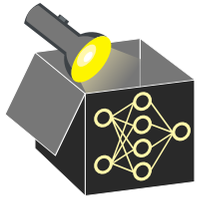
CENTERED ICE PLOT (C-ICE)

Categorical features: c-ICE plots can be interpreted as in LMs due to reference value



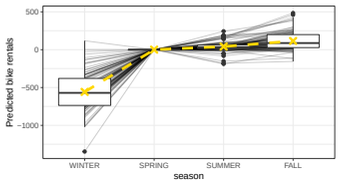
Interpretation:

- The reference category is $x' =$ SPRING
- Yellow crosses: Average rentals if we jump from SPRING to any other season $\Rightarrow$ Number of bike rentals drops by $\sim$ 560 in `WINTER` and is slightly higher in `SUMMER` and `FALL` compared to `SPRING`



CENTERED ICE PLOT (C-ICE)

Categorical features: c-ICE plots can be interpreted as in LMs due to reference value



Interpretation:

- The reference category is $x' =$ SPRING
- Yellow crosses: Average rentals if we jump from SPRING to any other season $\Rightarrow$ Number of bike rentals drops by $\sim$ 560 in `WINTER` and is slightly higher in `SUMMER` and `FALL` compared to `SPRING`