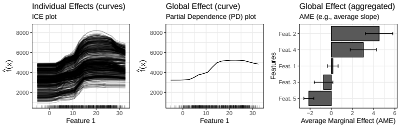
Interpretable Machine Learning

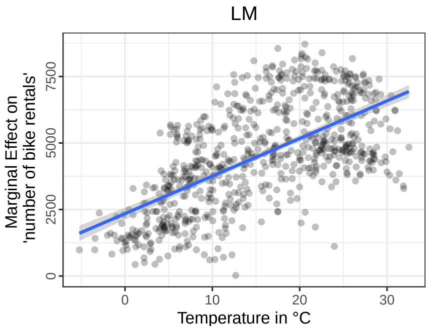
Introduction to Feature Effects



Learning goals

- Global Feature Effects
- Local Feature Effects

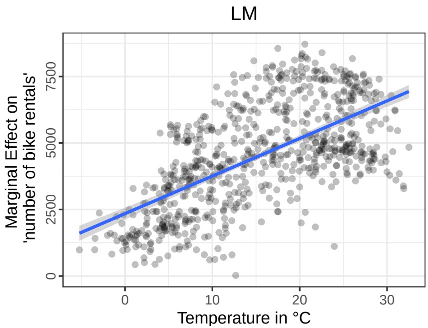
FEATURE EFFECTS - GLOBAL VIEW

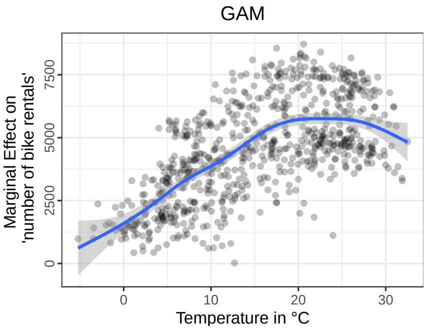


LM without interaction: $\hat{\theta}_j$ is linear effect of feature x_j (applies globally to all observations):

- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1\hat{\theta}_1$
- Single value $\hat{\theta}_1$ describes global effect

FEATURE EFFECTS - GLOBAL VIEW





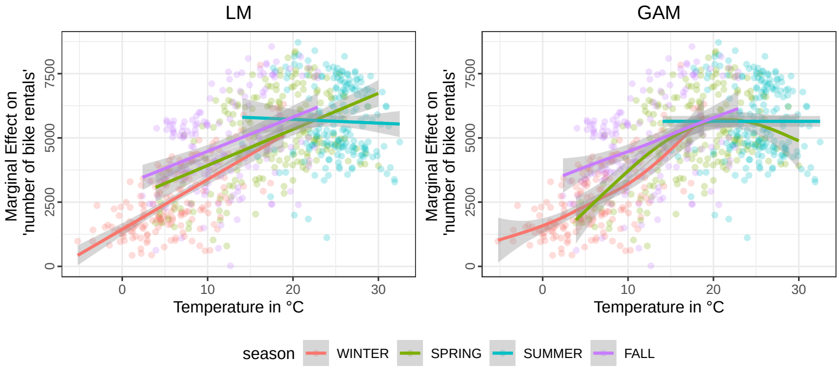
LM without interaction: $\hat{\theta}_j$ is linear effect of feature x_j (applies globally to all observations):

- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1\hat{\theta}_1$
- Single value $\hat{\theta}_1$ describes global effect

GAM without interaction: $\hat{f}_j(x_j)$ is non-linear effect of feature x_j (applies globally):

- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + \hat{f}_j(x_1)$
- Curve $\hat{f}_1$ describes global effect

FEATURE EFFECTS - LOCAL VIEW

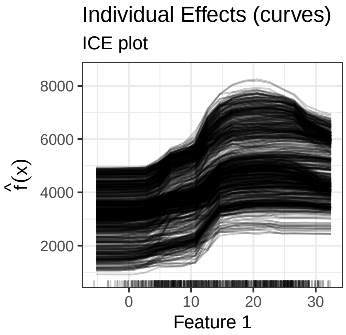


- Interactions: Feature effect is modified by other features and varies across observations
  $\Rightarrow$ Effect of temperature varies across seasons
  $\Rightarrow$ Multiple values / curves needed to describe effect
- ML models often model non-linear effects and complex interactions
  $\Rightarrow$ Need for local feature effect methods, e.g., analyze effect for individual observations
  $\Rightarrow$ Analyzing global effects by aggregating local effects

FEATURE EFFECTS

Feature effects visualize or quantify marginal contribution of a feature of interest w.r.t. predictions

- Similar to regression coefficients (LMs) or Splines (GAMs)
- Different aggregation levels for feature effects exist (simplification but information loss)
- Methods: ICE curves (local curves)

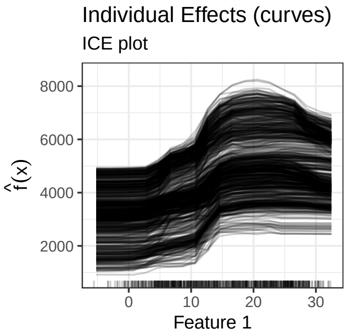


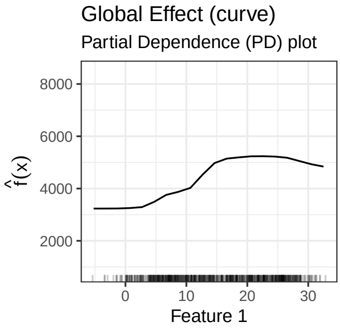
Individual (curves)

FEATURE EFFECTS

Feature effects visualize or quantify marginal contribution of a feature of interest w.r.t. predictions

- Similar to regression coefficients (LMs) or Splines (GAMs)
- Different aggregation levels for feature effects exist (simplification but information loss)
- Methods: ICE curves (local curves), PD and ALE plots (global curves)



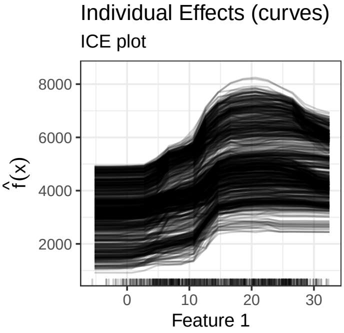


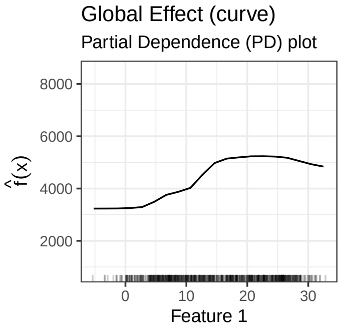
Individual (curves) $\xrightarrow[\text{curves}]{\text{aggregate}}$ Global (single curve)

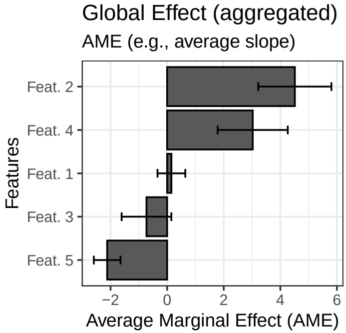
FEATURE EFFECTS

Feature effects visualize or quantify marginal contribution of a feature of interest w.r.t. predictions

- Similar to regression coefficients (LMs) or Splines (GAMs)
- Different aggregation levels for feature effects exist (simplification but information loss)
- Methods: ICE curves (local curves), PD and ALE plots (global curves), AME (global value)



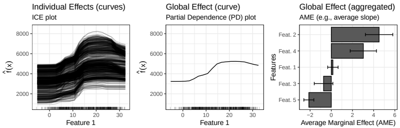




Individual (curves) $\xrightarrow[\text{curves}]{\text{aggregate}}$ Global (single curve) $\xrightarrow[\text{slopes}]{\text{aggregate}}$ Global (single value)

Interpretable Machine Learning

Individual Conditional Expectation (ICE) Plot



Learning goals

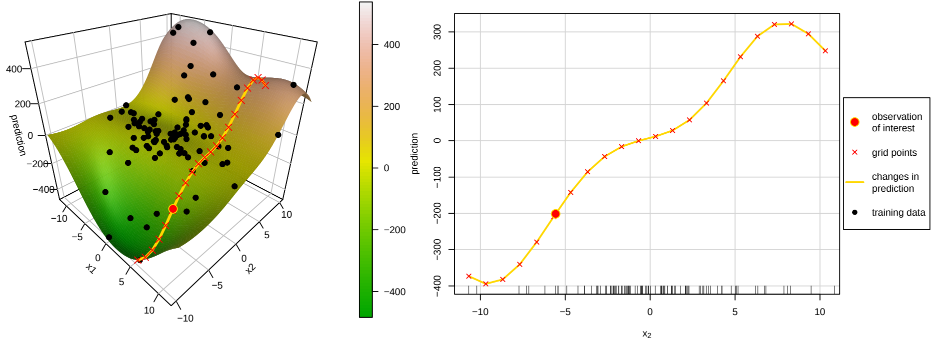
- ICE curves as local effect method
- How to sample grid points for ICE curves

MOTIVATION

Question: How does changing values of a single feature of an observation affect model prediction?

Idea: Change values of observation and feature of interest, and visualize how prediction changes

Example: Prediction surface of a model (left), select observation and visualize changes in prediction for different values of x_2 while keeping x_1 fixed $\Rightarrow$ **local interpretation**



INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

▸ Goldstein et. al (2013)

Partition each observation $\mathbf{x}$ into $\mathbf{x}_S$ (features of interest) and $\mathbf{x}_{-S}$ (remaining feat.)

$\rightsquigarrow$ In practice, $\mathbf{x}_S$ consists of one or two features (i.e., $|S| \leq 2$ and $-S = S^\complement$).

i	$\overbrace{\mathbf{x}_1}^{\mathbf{x}_S}$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	1	4	7
2	2	5	8
3	3	6	9

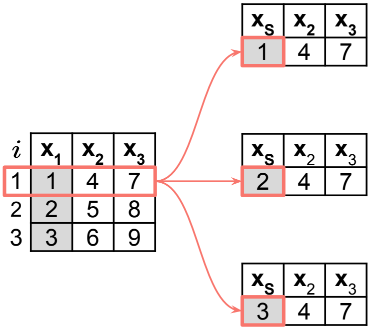
($\mathbf{x}_2, \mathbf{x}_3$: $\mathbf{x}_{-S}$)

Formal definition of ICE curves:

- Choose grid points $\mathbf{x}_S^* = \mathbf{x}_S^{*^{(1)}}, \dots, \mathbf{x}_S^{*^{(g)}}$ to vary $\mathbf{x}_S$
- Plot point pairs $\left\{ \left(\mathbf{x}_S^{*^{(k)}}, \hat{f}_S^{(i)}(\mathbf{x}_S^{*^{(k)}}) \right) \right\}_{k=1}^{g}$ where $\hat{f}_S^{(i)}(\mathbf{x}_S^*) = \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$
- For each k connect point pairs to obtain **ICE curve**

$\rightsquigarrow$ ICE curves visualize how prediction of i-th observation changes after varying its feature values indexed by S using grid points $\mathbf{x}_S^*$ while keeping all values in $-S$ fixed:

ICE CURVES - ILLUSTRATION



1. Step - Grid points:

Sample grid values $\mathbf{x}_S^{*(1)}, \ldots, \mathbf{x}_S^{*(g)}$ along feature of interest $\mathbf{x}_S$ and replace vector $\mathbf{x}^{(i)}$ in data with grid

$\Rightarrow$ Creates new artificial points for the i-th observation (here: $\mathbf{x}_S^* = x_1^* \in \{1, 2, 3\}$ is a scalar)

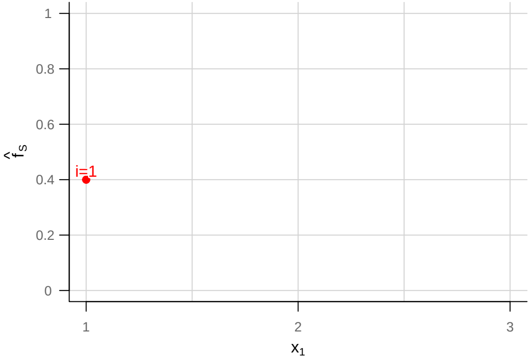
ICE CURVES - ILLUSTRATION

i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

x_S	x_2	x_3	$\hat{f}$
1	4	7	0.4

x_S	x_2	x_3	$\hat{f}$
2	4	7	0.6

x_S	x_2	x_3	$\hat{f}$
3	4	7	0.7



2. Step - Predict and visualize:

For each artificially created data point of i-th observation, plot prediction $\hat{f}_S^{(i)}(\mathbf{x}_S^*)$ vs. grid values $\mathbf{x}_S^*$:

$$\hat{f}_1^{(i)}(x_1^*) = \hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)}) \text{ vs. } x_1^* \in \{1, 2, 3\}$$

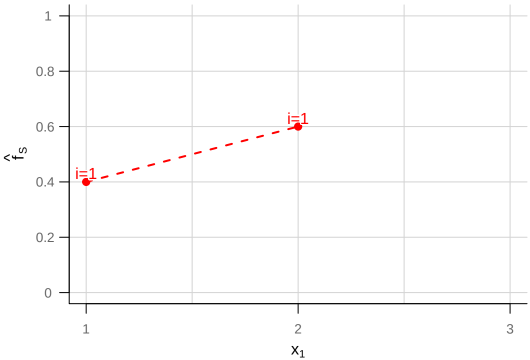
ICE CURVES - ILLUSTRATION

i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

x_S	x_2	x_3	$\hat{f}$
1	4	7	0.4

x_S	x_2	x_3	$\hat{f}$
2	4	7	0.6

x_S	x_2	x_3	$\hat{f}$
3	4	7	0.7



2. Step - Predict and visualize:

For each artificially created data point of i-th observation, plot prediction $\hat{f}_S^{(i)}(\mathbf{x}_S^*)$ vs. grid values $\mathbf{x}_S^*$:

$$\hat{f}_1^{(i)}(x_1^*) = \hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)}) \text{ vs. } x_1^* \in \{1, 2, 3\}$$

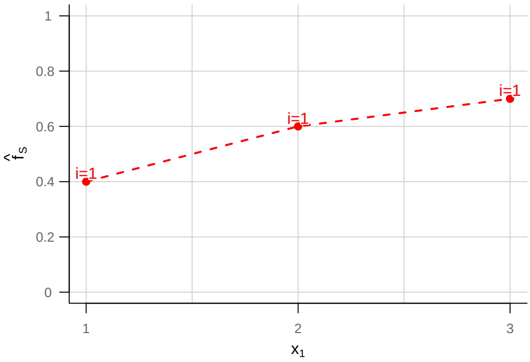
ICE CURVES - ILLUSTRATION

i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

x_S	x_2	x_3	$\hat{f}$
1	4	7	0.4

x_S	x_2	x_3	$\hat{f}$
2	4	7	0.6

x_S	x_2	x_3	$\hat{f}$
3	4	7	0.7



2. Step - Predict and visualize:

For each artificially created data point of i-th observation, plot prediction $\hat{f}_S^{(i)}(\mathbf{x}_S^*)$ vs. grid values $\mathbf{x}_S^*$:

$$\hat{f}_1^{(i)}(x_1^*) = \hat{f}(x_1^*, \mathbf{x}_{2,3}^{(i)}) \text{ vs. } x_1^* \in \{1, 2, 3\}$$

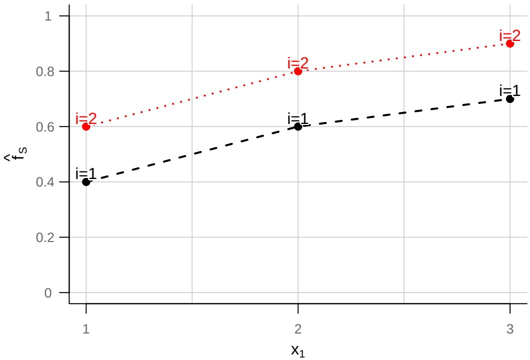
ICE CURVES - ILLUSTRATION

i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

x_S	x_2	x_3	$\hat{f}$
1	4	7	0.4
1	5	8	0.6

x_S	x_2	x_3	$\hat{f}$
2	4	7	0.6
2	5	8	0.8

x_S	x_2	x_3	$\hat{f}$
3	4	7	0.7
3	5	8	0.9



3. Step - Repeat for other observations:
ICE curve for $i = 2$ connects all predictions at grid values associated to i-th observation.

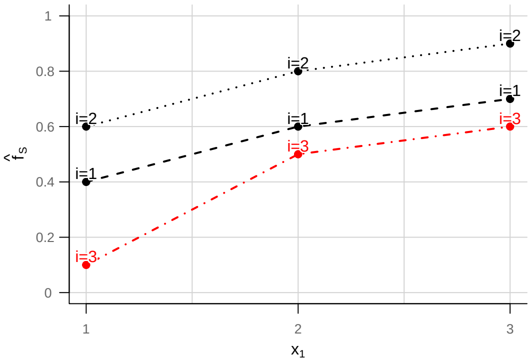
ICE CURVES - ILLUSTRATION

i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

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

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

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



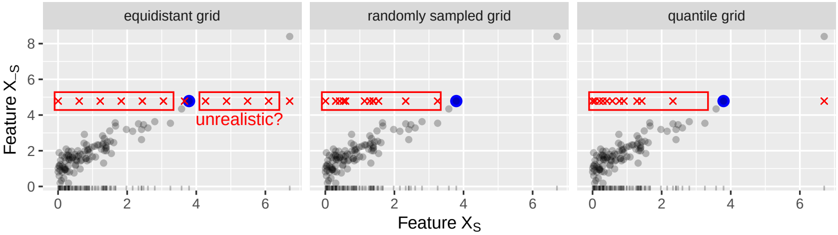
3. Step - Repeat for other observations:

ICE curve for $i = 3$ connects all predictions at grid values associated to i-th observation.

COMMENTS ON GRID VALUES

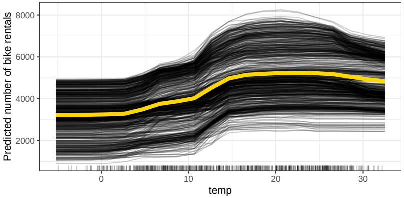
- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$ that are visualized on the x-axis
- Common choices for grid values are
  - equidistant grid values within feature range
  - randomly sampled values or quantile values of observed feature values
- Except equidistant grid, the other two options preserve (approximately) the marginal distribution of feature of interest $\Rightarrow$ Avoids unrealistic feature values for distributions with outliers

Grid points for X_S (red) for highlighted observation (blue)



Interpretable Machine Learning

Partial Dependence (PD) plot



Learning goals

- PD plots and relation to ICE plots
- Interpretation of PDP
- Extrapolation and Interactions in PDPs
- Centered ICE and 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 $\mathbf{x}_{-S}^{(i)}$:

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

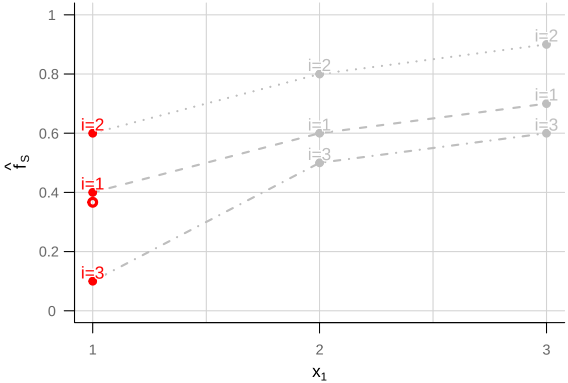
PARTIAL DEPENDENCE

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

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

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



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

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

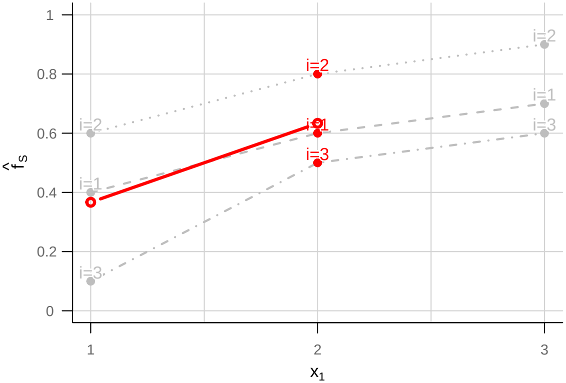
PARTIAL DEPENDENCE

i	$\mathbf{x_S}$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
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 $\boxed{\mathbf{x}_S^* = x_1^* = 2}$:

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

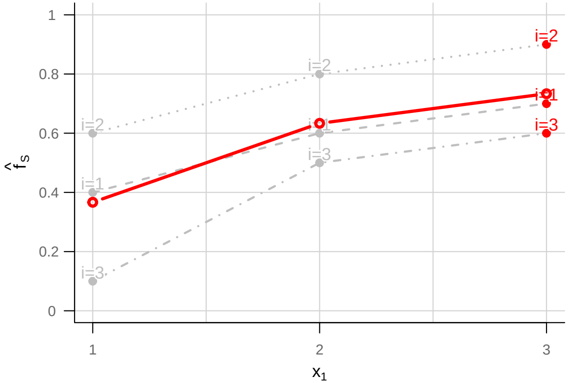
PARTIAL DEPENDENCE

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

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

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



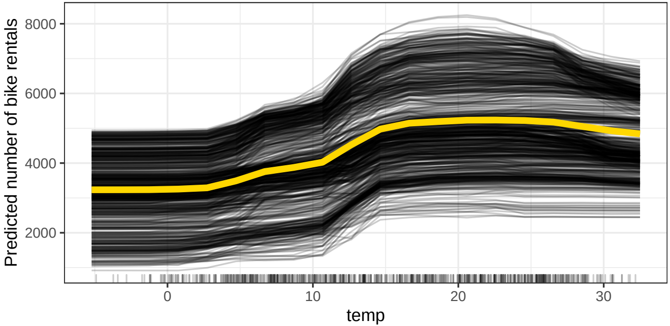
Estimate PD function by **point-wise** average of ICE curves at grid value $\boxed{\mathbf{x}_S^* = x_1^* = 3}$:

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

INTERPRETATION: PD AND ICE

If feature varies:

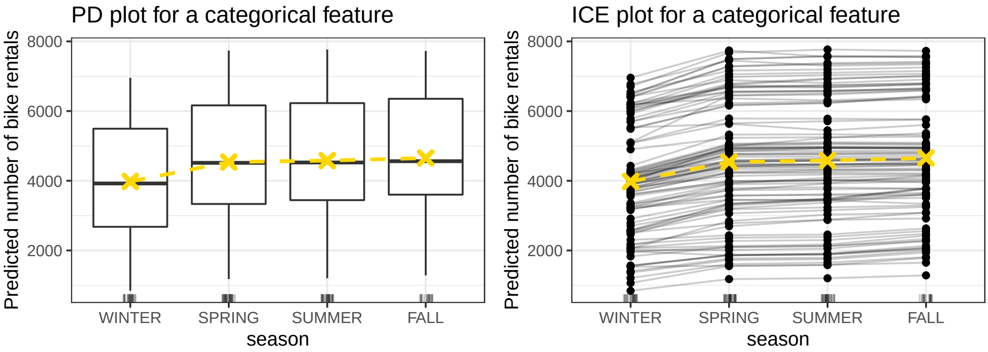
- **ICE:** How does **prediction of individual observation** change? $\Rightarrow$ **local** interpretation
- **PD:** How does **average effect / expected prediction** change? $\Rightarrow$ **global** interpretation



Insights from bike sharing data:

- Parallel ICE curves = homogeneous effect
- Warmer $\Rightarrow$ more rented bikes
- Too hot $\Rightarrow$ slightly less bikes

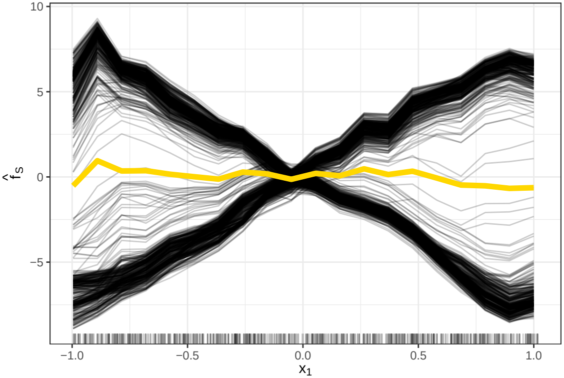
INTERPRETATION: CATEGORICAL FEATURES



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

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 does not tell with which feature)



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 $x' \sim \mathbb{P}(\mathbf{x}_S)$, 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. grid point $\mathbf{x}^*_S$

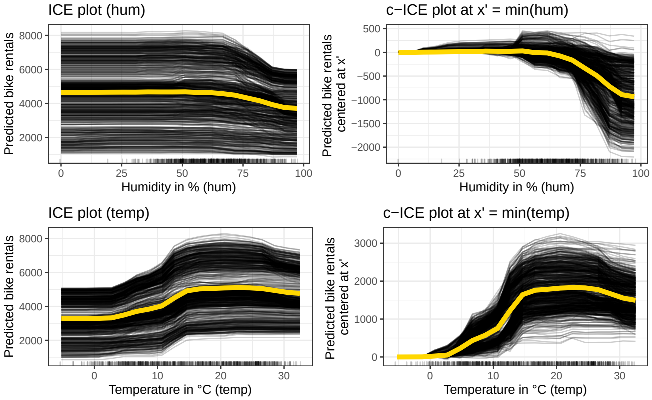
CENTERED ICE PLOT (C-ICE)

Issue: Difficult to identify heterogenous ICE curves if curves have different intercepts (are stacked)
Solution: Center ICE curves at fixed reference value $x' \sim \mathbb{P}(\mathbf{x}_S)$, often $x' = \min(\mathbf{x}_S)$
$\Rightarrow$ Easier to identify heterogenous shapes with c-ICE curves

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

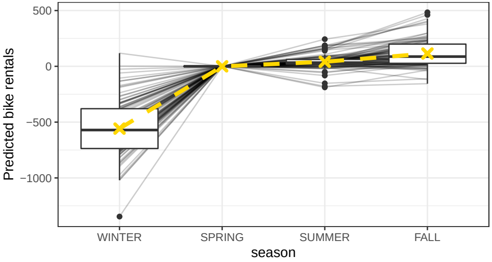
$\Rightarrow$ Visualize $\hat{f}^{(i)}_{S,cICE}(\mathbf{x}^*_S)$ vs. grid point $\mathbf{x}^*_S$

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

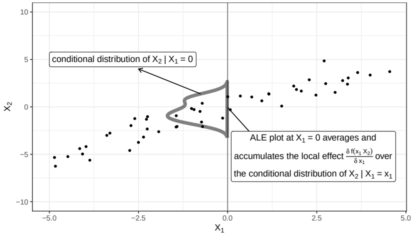


Interpretation:

- The reference category is $x' =$ 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 `WINTER` and is slightly higher in `SUMMER` and `FALL` compared to `SPRING`

Interpretable Machine Learning

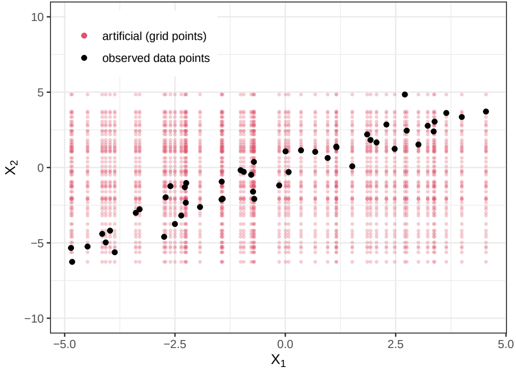
Accumulated Local Effect (ALE) plot

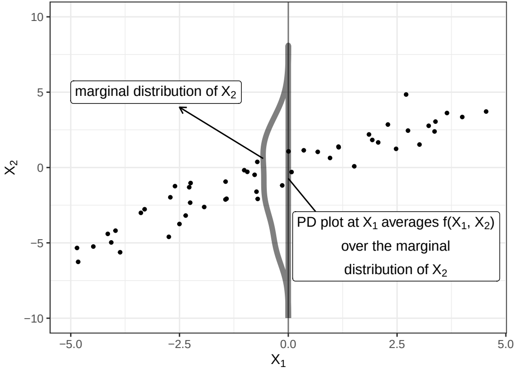


Learning goals

- PD plots and its extrapolation issue
- M plots and its omitted-variable bias
- Understand ALE plots

MOTIVATION - CORRELATED FEATURES

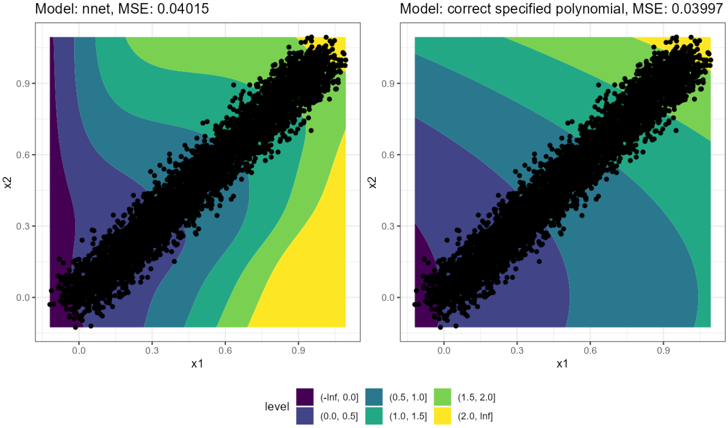




- PD plots **average over predictions** of artificial points that are out of distribution / unlikely (red)
  $\Rightarrow$ Can lead to misleading / biased interpretations, especially if model also contains interactions
- Not wanted if interest is to interpret effects within data distribution

MOTIVATION - CORRELATED FEATURES

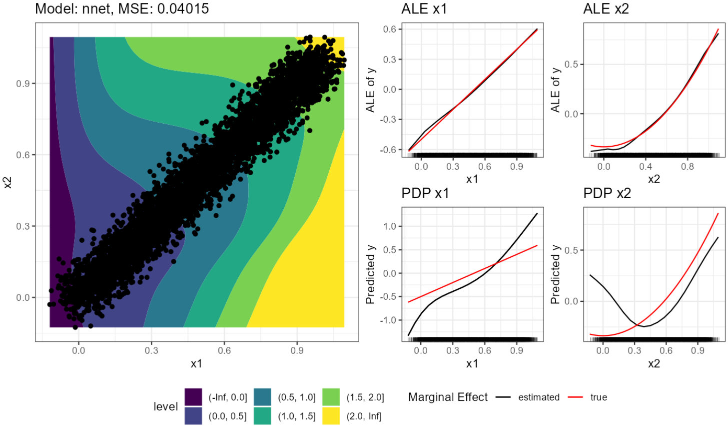
Example: Fit a NN to 5000 simulated data points with $x \sim Unif(0,1)$, $\epsilon \sim N(0, 0.2)$ and $y = x_1 + x_2^2 + \epsilon$, where $x_1 = x + \epsilon_1$, $x_2 = x + \epsilon_2$ and $\epsilon_1, \epsilon_2 \sim N(0, 0.05)$.



- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)

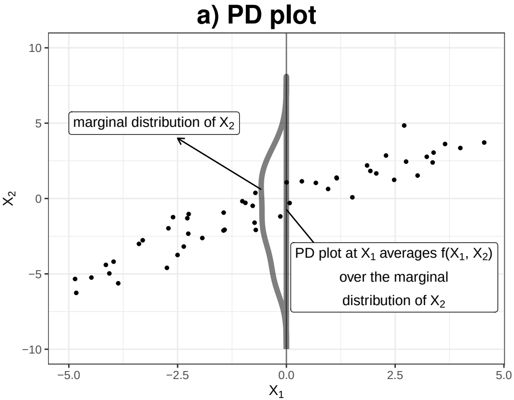
MOTIVATION - CORRELATED FEATURES

Example: Fit a NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and
$y = x_1 + x_2^2 + \epsilon$, where $x_1 = x + \epsilon_1$, $x_2 = x + \epsilon_2$ and $\epsilon_1, \epsilon_2 \sim N(0, 0.05)$.



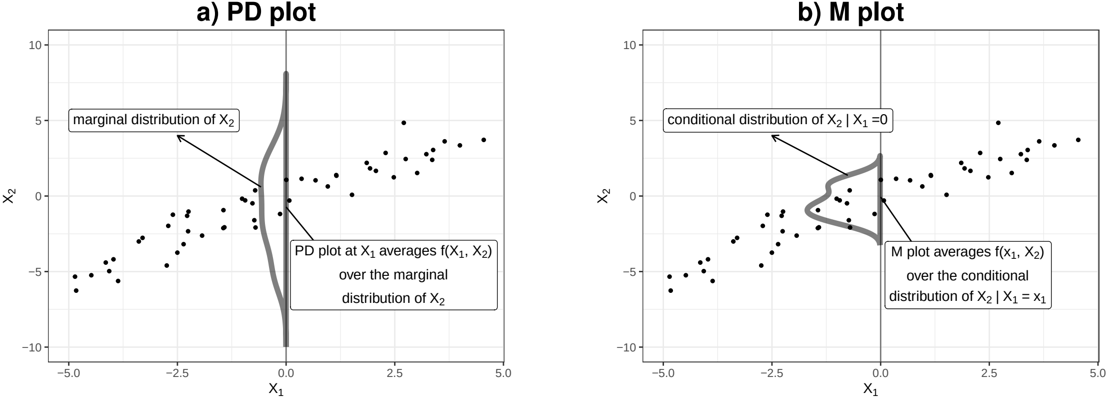
- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)
- ALE in line with ground truth
- PDP does not reflect ground truth effects of DGP well
  $\Rightarrow$ Due to interactions and averaging of points outside data distribution

M PLOT VS. PD PLOT



a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1, \mathbf{x}_2^{(i)})$

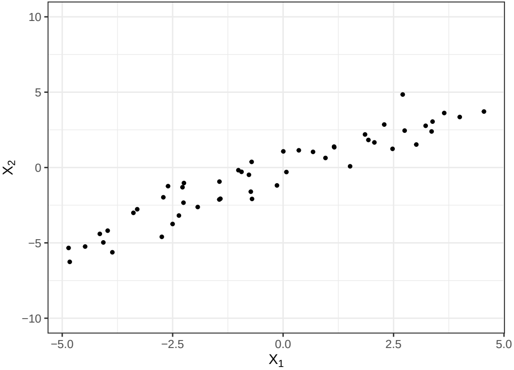
M PLOT VS. PD PLOT

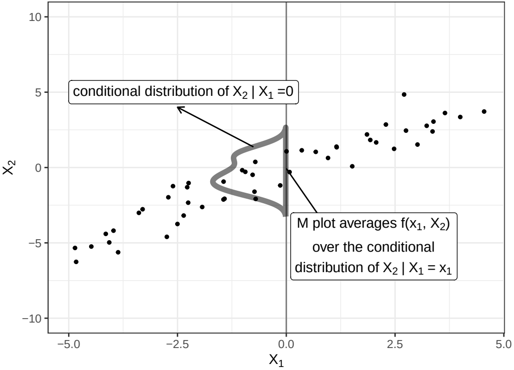


a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_1, \mathbf{x}_2^{(i)})$

b) M plot $\mathbb{E}_{\mathbf{x}_2|\mathbf{x}_1}\left(\hat{f}(x_1, \mathbf{x}_2)\middle|\mathbf{x}_1\right)$ is estimated by $\hat{f}_{1,M}(x_1) = \frac{1}{|N(x_1)|}\sum_{i \in N(x_1)} \hat{f}(x_1, \mathbf{x}_2^{(i)})$, where index set $N(x_1) = \{i : x_1^{(i)} \in [x_1 - \epsilon, x_1 + \epsilon]\}$ refers to observations with feature value close to x_1.

M PLOT VS. PD PLOT





- M plots average predictions over conditional distribution (e.g., $\mathbb{P}(\mathbf{x}_2|x_1)$)
  $\Rightarrow$ Averaging predictions close to data distribution avoid extrapolation issues
- **But:** M plots suffer from omitted-variable bias (OVB)
  - They contain effects of other dependent features
  - Useless in assessing a feature's marginal effect if feature dependencies are present

M PLOT VS. PD PLOT - OVB EXAMPLE

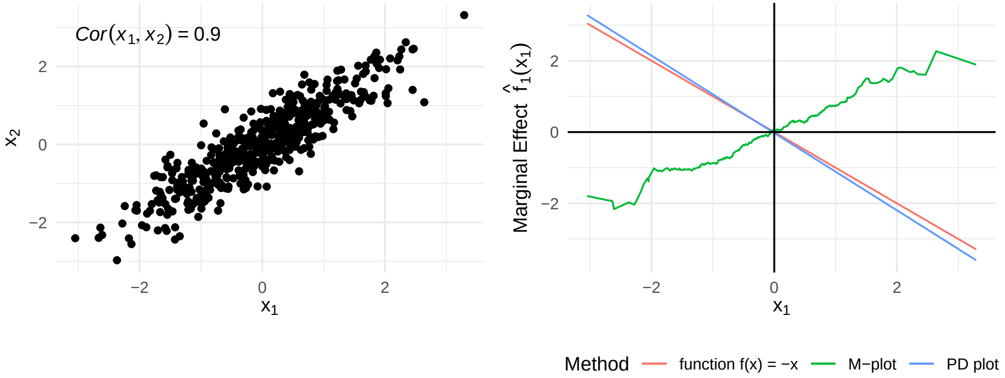


Illustration: Fit LM on 500 i.i.d. observations with features $x_1, x_2 \sim N(0, 1)$, $Cor(x_1, x_2) = 0.9$ and

$$y = -x_1 + 2 \cdot x_2 + \epsilon, \ \epsilon \sim N(0, 1).$$

Results: M plot of x_1 also includes marginal effect of all other dependent features (here: x_2)

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1 x_2$$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- Partial derivative of $\hat{f}$ w.r.t. x_1: $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- Partial derivative of $\hat{f}$ w.r.t. x_1: $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of partial derivative ($z_0 = \min(x_1)$):

$$\int_{z_0}^{x} \frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} dx_1 = [2x_1 - 4x_1x_2]_{z_0}^{x}$$

IDEA: INTEGRATING PARTIAL DERIVATIVES

Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ Computing the partial derivative of $\hat{f}$ w.r.t. $\mathbf{x}_j$ removes other main effects

$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1 x_2$$

- Partial derivative of $\hat{f}$ w.r.t. x_1: $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of partial derivative ($z_0 = \min(x_1)$):

$$\int_{z_0}^{x} \frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} dx_1 = [2x_1 - 4x_1 x_2]_{z_0}^{x}$$

- We removed the main effect of x_2, which was our goal

ACCUMULATED LOCAL EFFECTS (ALE) ▸ Apley, Zhu (2020)

ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$

ACCUMULATED LOCAL EFFECTS (ALE) ▸ Apley, Zhu (2020)

ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$
2. averaging local effects over conditional distribution $\mathbb{P}(\mathbf{x}_{-S} | x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation issue

ACCUMULATED LOCAL EFFECTS (ALE) ▸ Apley, Zhu (2020)

ALE plots use the idea of integrating partial derivatives. They do not suffer from the extrapolation issue of PD plots and the OVB issue of M plots when features are dependent.

Concept of ALE plots is based on

1. estimating local effects $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences) evaluated at certain points $(x_S = z_S, \mathbf{x}_{-S})$
2. averaging local effects over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation issue
3. integrating averaged local effects up to a specific value $x \sim \mathbb{P}(x_S)$
   $\Rightarrow$ Accumulates local effects to estimate global main effect of x_S
   $\Rightarrow$ Avoids OVB issue as other unwanted main effects were removed in (1)

FIRST ORDER ALE

- Let x_S be feature of interest with $z_0 = \min(x_S)$ and $\mathbf{x}_{-S}$ all other features (complement of S)
- Uncentered first order ALE $\tilde{f}_{S,ALE}(x)$ at feature value $x \sim \mathbb{P}(x_S)$ is defined as:

$$\tilde{f}_{S,ALE}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S}}_{(2)} \left(\underbrace{\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}}_{(1)} \middle| x_S = z_S \right) dz_S$$

FIRST ORDER ALE

- Let x_S be feature of interest with $z_0 = \min(x_S)$ and $\mathbf{x}_{-S}$ all other features (complement of S)
- Uncentered first order ALE $\tilde{f}_{S,ALE}(x)$ at feature value $x \sim \mathbb{P}(x_S)$ is defined as:

$$\tilde{f}_{S,ALE}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S}}_{(2)} \left(\underbrace{\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}}_{(1)} \middle| x_S = z_S \right) dz_S$$

- Substract average of uncentered ALE curve (constant) to obtain centered ALE curve $f_{S,ALE}(x)$ with zero mean regarding marginal distribution of feature of interest x_S:

$$f_{S,ALE}(x) = \tilde{f}_{S,ALE}(x) - \underbrace{\int_{-\infty}^{\infty} \tilde{f}_{S,ALE}(x_S)\, d\mathbb{P}(x_S)}_{:= constant}$$

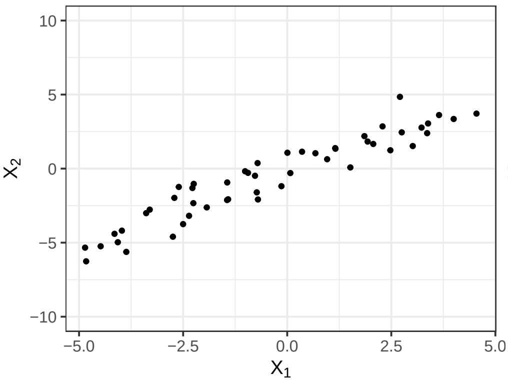
ALE ESTIMATION

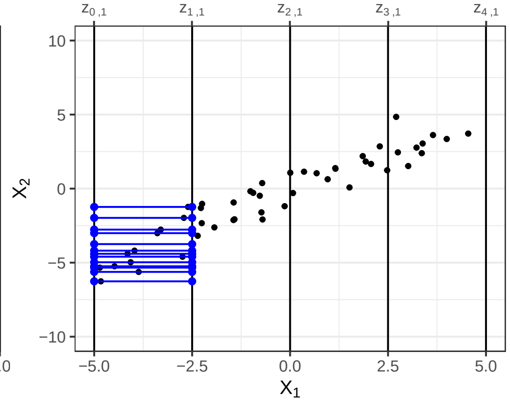
- Partial derivatives not useful for all models (e.g., tree-based methods as random forests)
- Approximate partial derivatives by finite differences of predictions within K intervals for $\mathbf{x}_S$:

$$
\begin{aligned}
x \in [\min(\mathbf{x}_S), \max(\mathbf{x}_S)] \iff & x \in [z_{0,S}, z_{1,S}] \\
& \vee x \in]z_{1,S}, z_{2,S}] \\
& \quad \dots \\
& \vee x \in]z_{K-1,S}, z_{K,S}]
\end{aligned}
$$

- A simple way to create K intervals for feature $\mathbf{x}_S$ is to use its quantile distribution with $K - 1$ quantiles as interval bounds $z_{1,S}, \dots, z_{K-1,S}$ (not counting the 0% and 100% quantiles)

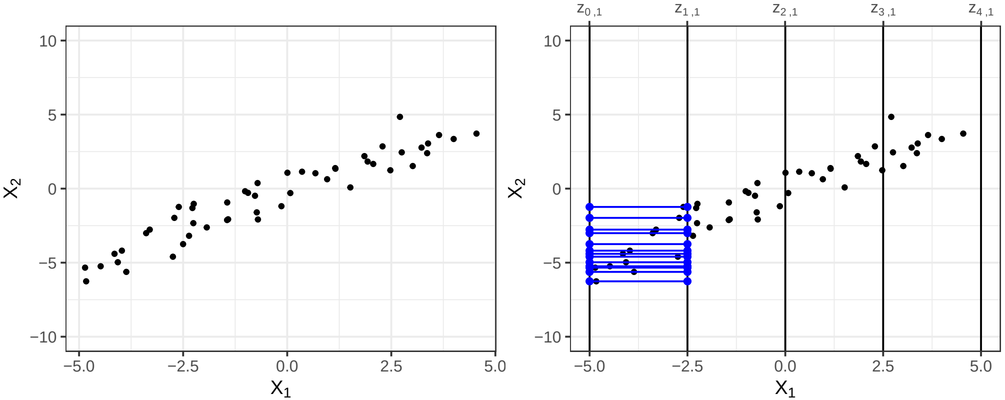
2-D ILLUSTRATION





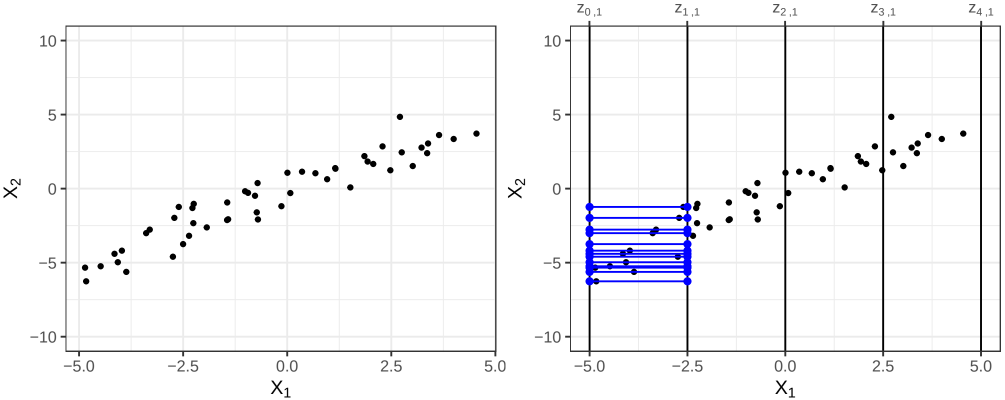
- Divide feature of interest into intervals (vertical lines)
- For all points within an interval, compute **prediction difference** when we replace feature value with upper/lower interval bound (blue points) while keeping other feature values unchanged
- These **finite differences** (approximate local effect) are accumulated & centered $\Rightarrow$ ALE plot

2-D ILLUSTRATION



- For $\mathbf{x}^{(i)} = (x_S^{(i)}, \mathbf{x}_{-S}^{(i)})$, value $x_S^{(i)}$ is located within k-th interval of $\mathbf{x}_S$ ($x_S^{(i)} \in]z_{k-1,S}, z_{k,S}]$)
- Replace $x_S^{(i)}$ by upper/lower interval bound while all other feature values $\mathbf{x}_{-S}^{(i)}$ are kept constant
- Finite differences correspond to $\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$

2-D ILLUSTRATION



- Estimate local effect of $\mathbf{x}_S$ within each interval by averaging all observation-wise finite differences $\hat{=}$ Approximation of inner integral that integrates over local effects w.r.t. $\mathbb{P}(\mathbf{x}_{-S}|z_S)$.
- Sum up local effects of all intervals up to point of interest $\hat{=}$ Estimates outer integral

ALE ESTIMATION: FORMULA

- Estimated uncentered first order ALE $\hat{\tilde{f}}_{S,ALE}(x)$ at point x:

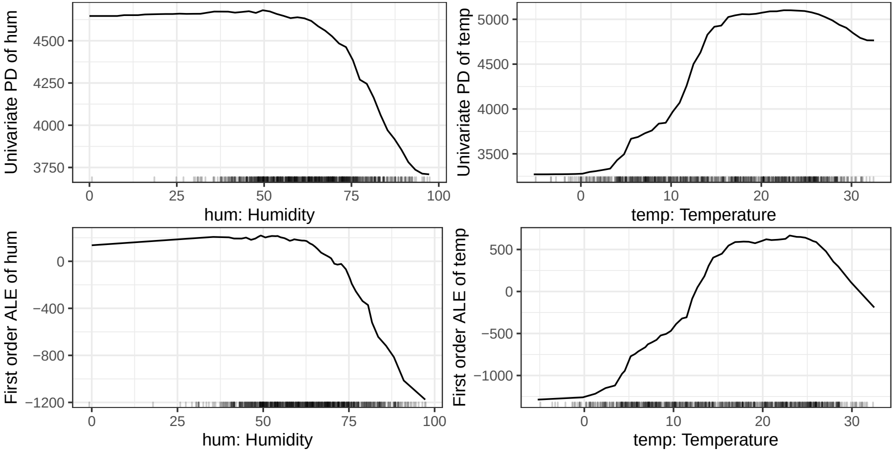
$$\hat{\tilde{f}}_{S,ALE}(x) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i:\ x_S^{(i)} \in\ ]z_{k-1,S}, z_{k,S}]} \left[\hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})\right]$$

- $k_S(x)$ denotes the interval index a feature value $x \in \mathbf{x}_S$ falls in
- $n_S(k)$ denotes the number of observations inside the k-th interval of $\mathbf{x}_S$
- Substract average of estimated uncentered ALE to obtain centered ALE estimate:

$$\hat{f}_{S,ALE}(x) = \hat{\tilde{f}}_{S,ALE}(x) - \frac{1}{n} \sum_{i=1}^{n} \hat{\tilde{f}}_{S,ALE}(x_S^{(i)})$$

BIKE SHARING DATASET

Shape of PD plot (left) often looks similar to (centered) first order ALE plot (right) but on different y-axis scale. In case of correlated features, ALE might be better due to PD's extrapolation issue.



PD VS. ALE

PD:

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

ALE:

$$f_{S,ALE}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S}\left(\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}\Bigg| x_S = z_S\right) dz_S - const$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$

PD VS. ALE

PD:

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

ALE:

$$f_{S,ALE}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S}\left(\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}\Bigg| x_S = z_S\right) dz_S - const$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates partial derivatives of feature S over z_S
  $\rightsquigarrow$ isolates effect of feature S and removes main effect of other dependent features

PD VS. ALE

PD:

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

ALE:

$$f_{S,ALE}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S}\left(\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}\Bigg| x_S = z_S\right) dz_S - \textit{const}$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- Difference 1: ALE averages the
  - change of predictions (via partial derivatives approximated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates partial derivatives of feature S over z_S
  $\rightsquigarrow$ isolates effect of feature S and removes main effect of other dependent features
- Difference 3: ALE is centered so that $\mathbb{E}_{x_S}\left(f_{S,ALE}(x)\right) = 0$