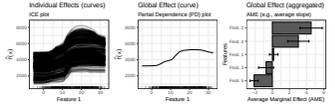
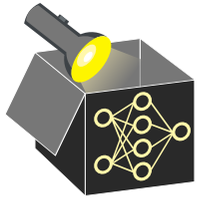
Interpretable Machine Learning

Introduction to Feature Effects



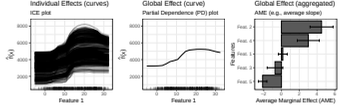
Learning goals

- Global Feature Effects
- Local Feature Effects



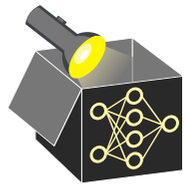
Interpretable Machine Learning

Feature Effects
Introduction to Feature Effects

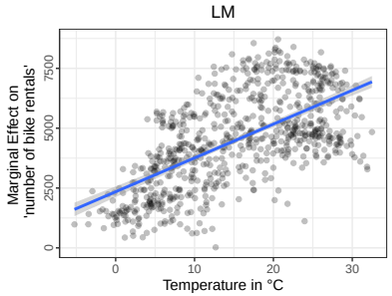


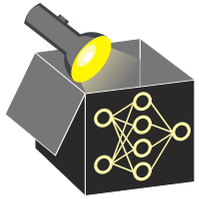
Learning goals

- Global Feature Effects
- Local Feature Effects



FEATURE EFFECTS - GLOBAL VIEW

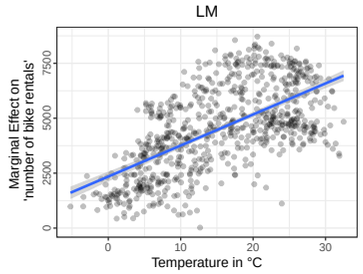


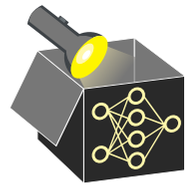


LM without interaction: $\hat{\theta}_j$ is linear effect of feature x_j (applies globally to all obs.):

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

FEATURE EFFECTS - GLOBAL VIEW

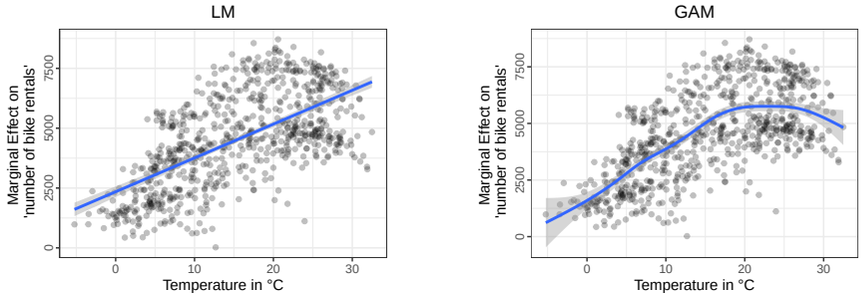


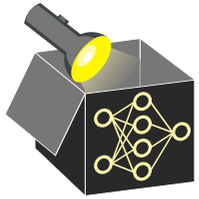


LM without interaction:
$\hat{\theta}_j$ is linear effect of feature x_j
(applies globally to all observations):

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

FEATURE EFFECTS - GLOBAL VIEW





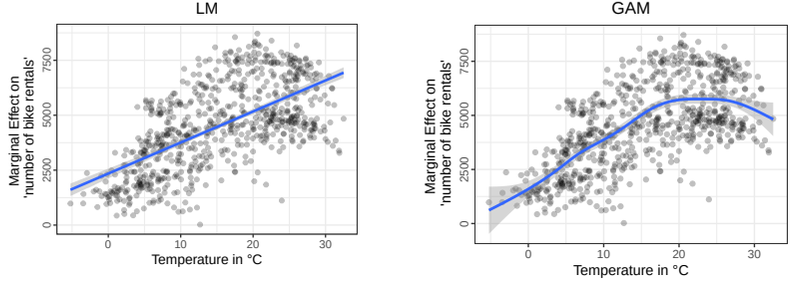
LM without interaction: $\hat{\theta}_j$ is linear effect of feature x_j (applies globally to all obs.):

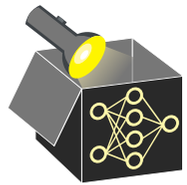
- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1\hat{\theta}_1$
- Scalar $\hat{\theta}_1$ describes global effect

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

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

FEATURE EFFECTS - GLOBAL VIEW





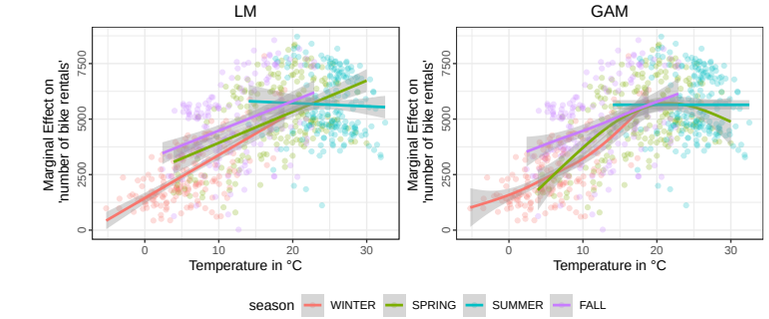
LM without interaction:
$\hat{\theta}_j$ is linear effect of feature x_j
(applies globally to all observations):

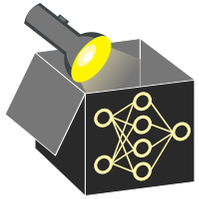
- Model equation: $\hat{f}(\mathbf{x}) = \hat{\theta}_0 + x_1\hat{\theta}_1$
- Scalar $\hat{\theta}_1$ describes global effect

GAM without interaction:
$\hat{f}_j(x_j)$ is non-lin. effect of feature x_j
(applies globally to all observations):

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

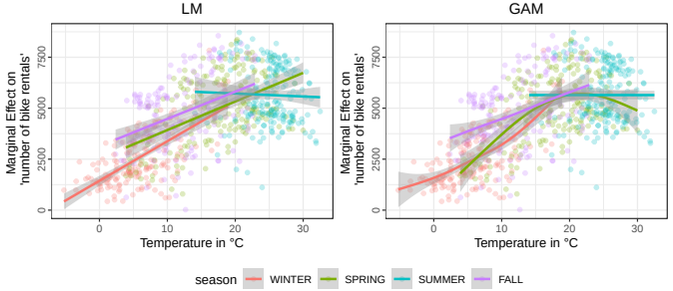
FEATURE EFFECTS - LOCALIZED VIEW

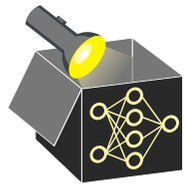




- **Interactions:** Feature effect depends on other features and varies across obs.
  ⇒ E.g., effect of **temperature** varies across **season**
  ⇒ Multiple values / curves needed to describe effect
- ML models capture non-linear effects and high-order interactions
  ⇒ Global view often misleading (single curve may fail to capture complexity)
  ⇒ Need for local feature effect methods to estimate effects for individual obs.
  ⇒ Global view can be reconstructed by aggregating local effects

FEATURE EFFECTS - LOCALIZED VIEW



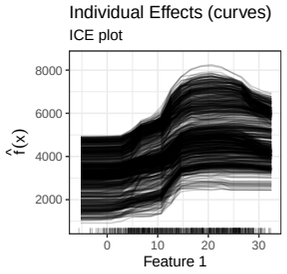


- **Interactions:** Feature effect depends on other features and varies across obs.
  ⇒ E.g., effect of **temperature** varies across **season**
  ⇒ Multiple values / curves needed to describe effect
- ML models capture non-linear effects and high-order interactions
  ⇒ Global view may mislead (single curve may fail to capture complexity)
  ⇒ Local feat. effect methods needed to estimate effects for individ. obs.
  ⇒ Global view can be reconstructed by aggregating local effects

FEATURE EFFECTS

Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

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

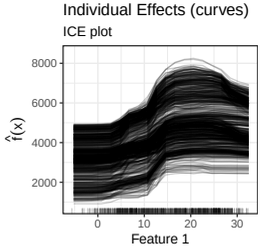


Individual (curves)

FEATURE EFFECTS

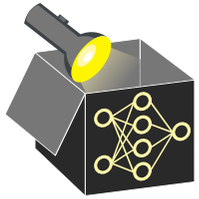
Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

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



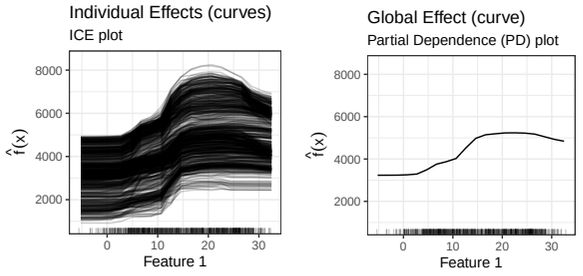
Individual (curves)

FEATURE EFFECTS



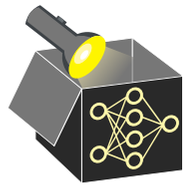
Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

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



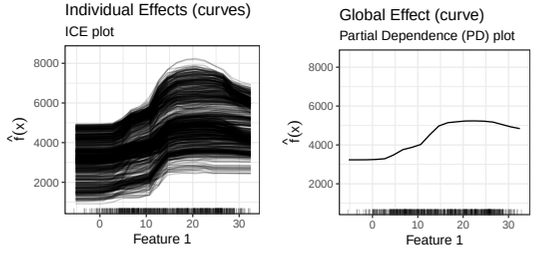
Individual (curves) $\xrightarrow[\text{curves}]{\text{aggregate}}$ Global (single curve)

FEATURE EFFECTS



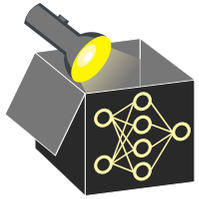
Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

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



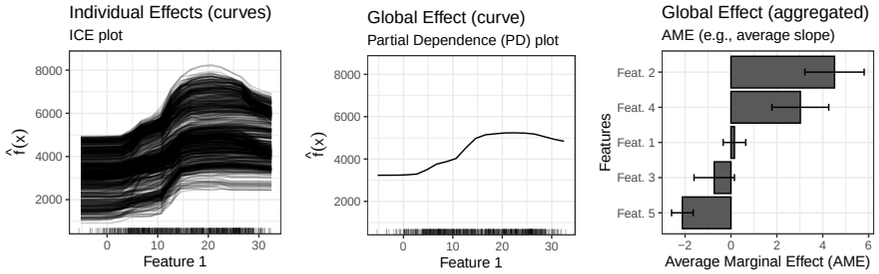
Individual (curves) $\xrightarrow[\text{curves}]{\text{aggregate}}$ Global (single curve)

FEATURE EFFECTS



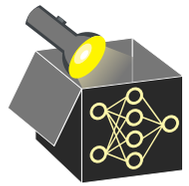
Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

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



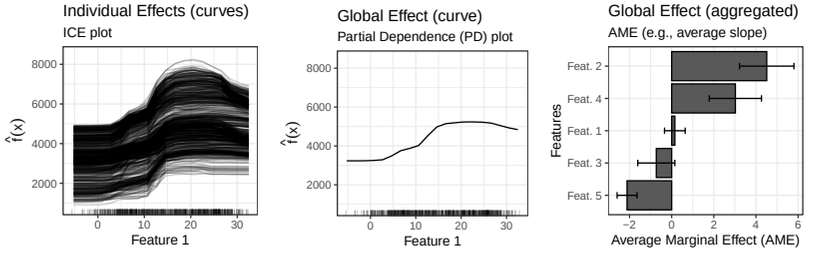
Individual (curves) $\xrightarrow[\text{curves}]{\text{aggregate}}$ Global (single curve) $\xrightarrow[\text{slopes}]{\text{aggregate}}$ Global (single value)

FEATURE EFFECTS



Feature effects visualize or quantify how model predictions change as a single feature varies, while all other features are held fixed.

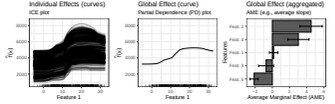
- Analogous to regression coefficients (LMs) or Splines (GAMs)
- Different aggregation levels exist (simplification but information loss)
- Methods: ICE curves (local curves), PD and ALE plots (global curves), AME (global value)



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

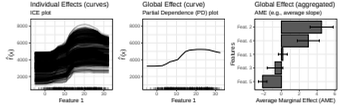


Learning goals

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

Interpretable Machine Learning

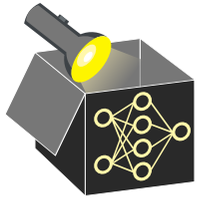
Feature Effects
Individual Conditional Expectation (ICE) Plot



Learning goals

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

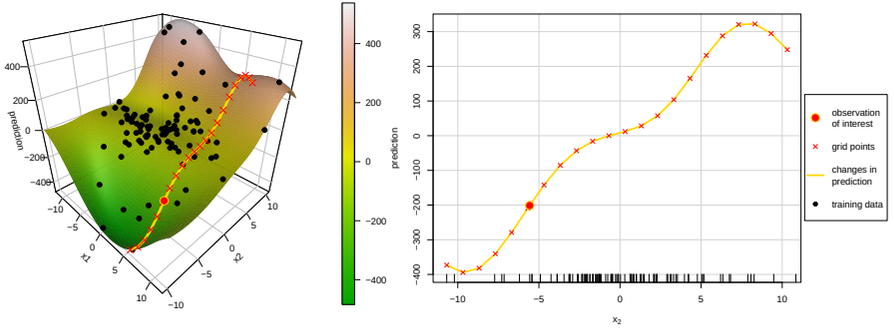
MOTIVATION



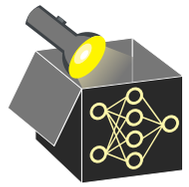
Question: How does varying a single feature of an obs. affect its predicted outcome?

Idea: For a given observation, change the value of the feature of interest, and visualize how prediction changes

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



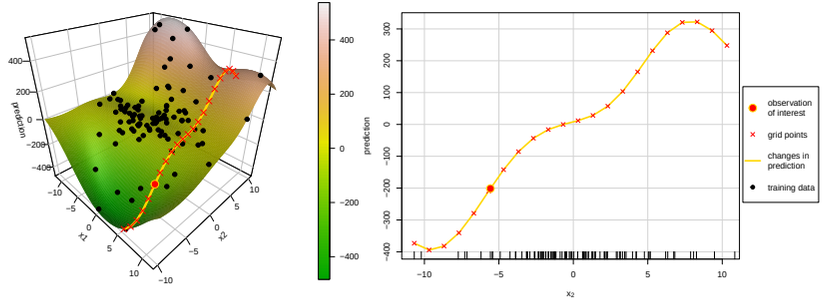
MOTIVATION



Question: How does varying a single feature of an observation affect its predicted outcome?

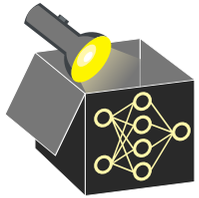
Idea: For a given observation, change the value of the feature of interest, and visualize how prediction changes

Example: On model prediction surface (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 $\mathbf{x}_S$ (feature(s) of interest) and $\mathbf{x}_{-S}$ (remaining features)

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

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

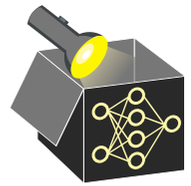
Formal definition of ICE curves:

- Define grid points $\mathbf{x}_S^* = \mathbf{x}_S^{*(1)}, \ldots, \mathbf{x}_S^{*(g)}$ to vary $\mathbf{x}_S$
- Plot point pairs $\left\{ \left(\mathbf{x}_S^{*(k)}, \hat{f}_{S,ICE}^{(i)}(\mathbf{x}_S^{*(k)}) \right) \right\}_{k=1}^{g}$
  where $\hat{f}_{S,ICE}^{(i)}(\mathbf{x}_S^*) = \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$
- 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

INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

▸ GOLDSTEIN_2013



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

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

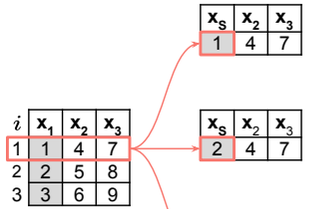
$\rightsquigarrow$ In practice, $\mathbf{x}_S$ consists of one or two features (i.e., $|S| \leq 2$ and $-S = S^{\complement}$).

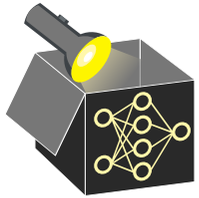
Formal definition of ICE curves:

- Define grid points $\mathbf{x}_S^* = \mathbf{x}_S^{*(1)}, \ldots, \mathbf{x}_S^{*(g)}$ to vary $\mathbf{x}_S$
- Plot point pairs $\left\{ \left(\mathbf{x}_S^{*(k)}, S^{(i)}(\mathbf{x}_S^{*(k)}) \right) \right\}_{k=1}^{g}$
  where $S^{(i)}(\mathbf{x}_S^*) = \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$
- 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



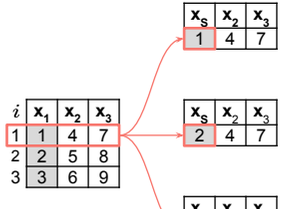


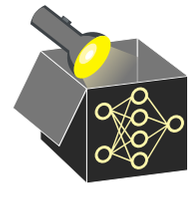
1. Step - Grid points:

- Sample grid values $\mathbf{x}_S^{*(1)}, \ldots, \mathbf{x}_S^{*(g)}$ along possible values of feature S ($|S| = 1$)
- For $\mathbf{x}^{(i)} = (\mathbf{x}_S, \mathbf{x}_{-S})$, replace $\mathbf{x}_S$ with those grid values

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

ICE CURVES - ILLUSTRATION



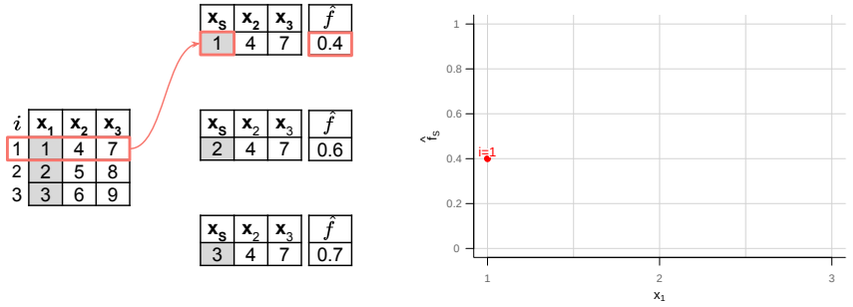


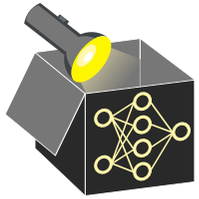
1. Step - Grid points:

- Sample grid values $\mathbf{x}_S^{*(1)}, \ldots, \mathbf{x}_S^{*(g)}$ along possible values of feature S ($|S| = 1$)
- For $\mathbf{x}^{(i)} = (\mathbf{x}_S, \mathbf{x}_{-S})$, replace $\mathbf{x}_S$ with those grid values

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

ICE CURVES - ILLUSTRATION



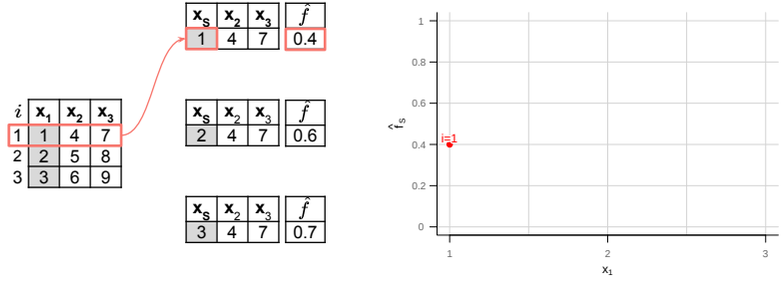


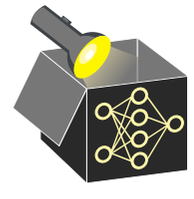
2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION



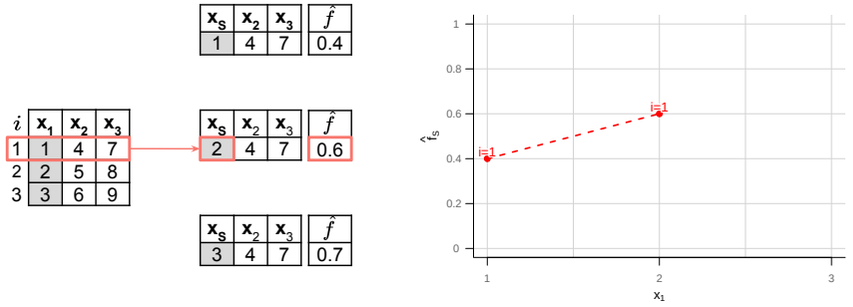


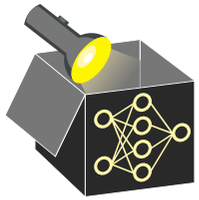
2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION



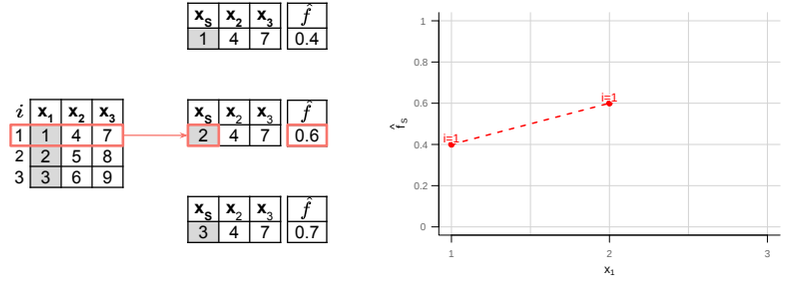


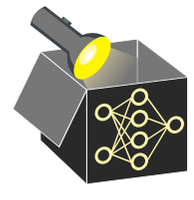
2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION



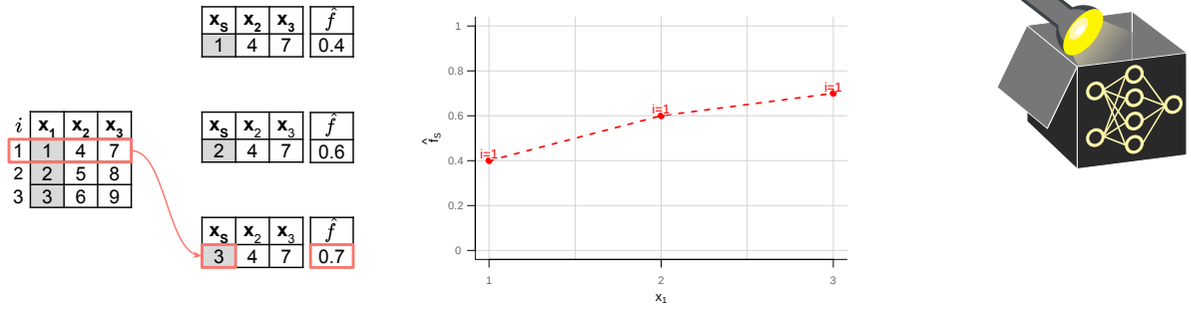


2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION

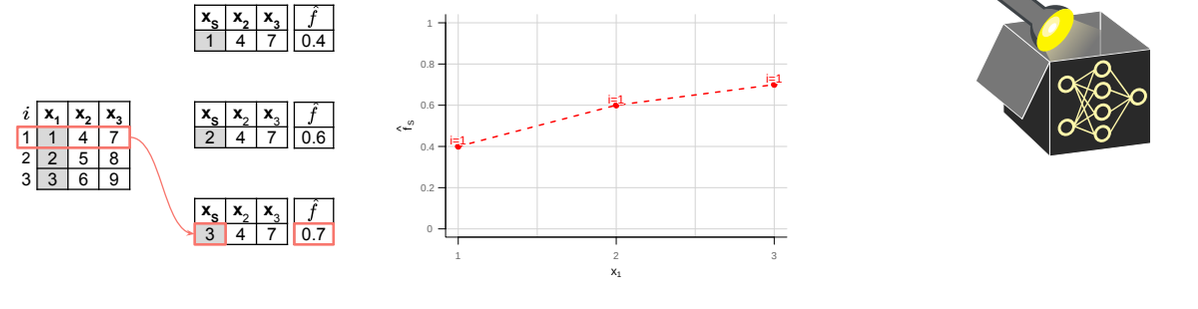


2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION

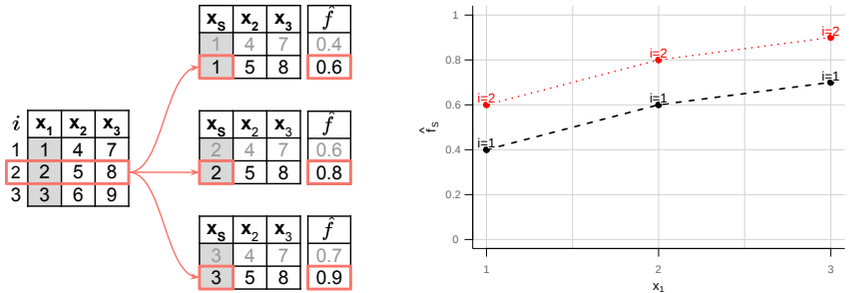


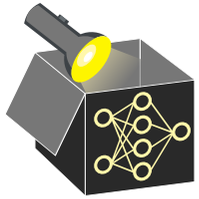
2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION

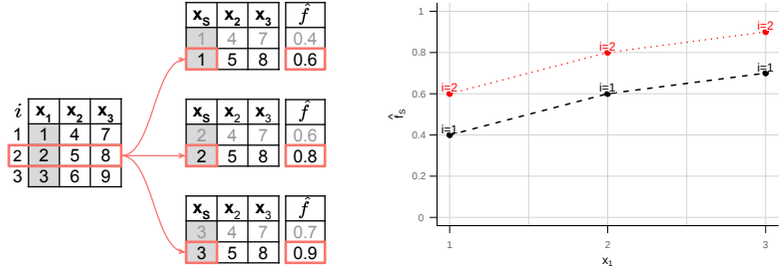


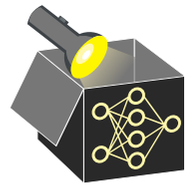


3. Step - Repeat for other observations:

ICE curve for $i = 2$ connects all predictions at grid values associated to i-th obs.

ICE CURVES - ILLUSTRATION

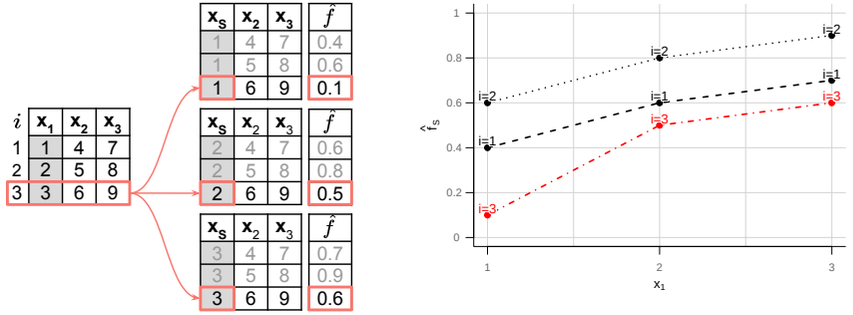


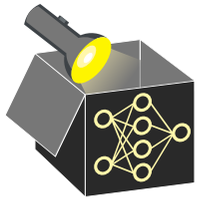


3. Step - Repeat for other observations:

ICE curve for $i = 2$ connects all predictions at grid values associated to the i-th observation.

ICE CURVES - ILLUSTRATION

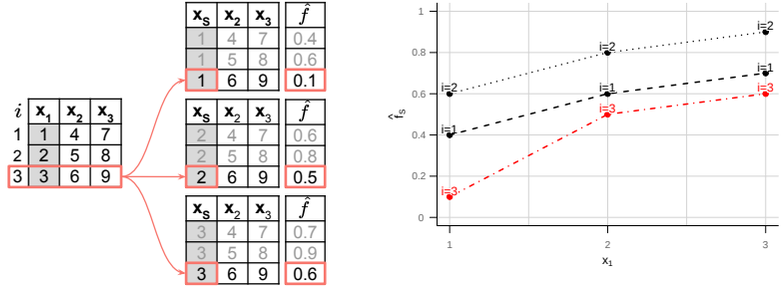


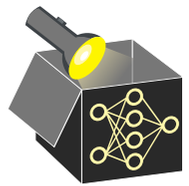


3. Step - Repeat for other observations:

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

ICE CURVES - ILLUSTRATION

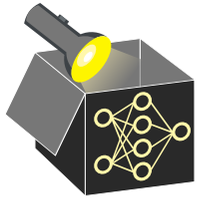




3. Step - Repeat for other observations:

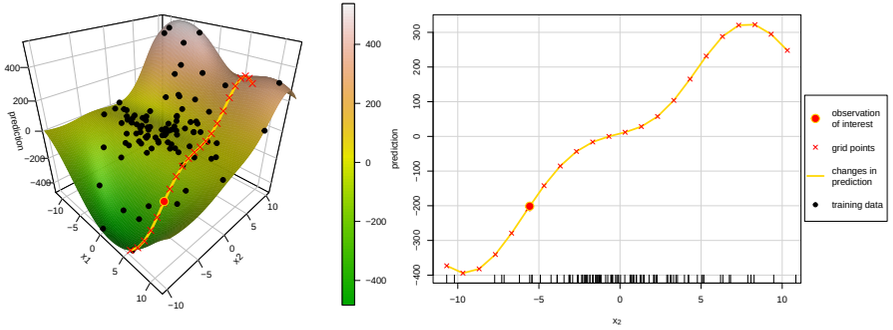
ICE curve for $i = 3$ connects all predictions at grid values associated to the i-th observation.

ICE CURVES - INTERPRETATION

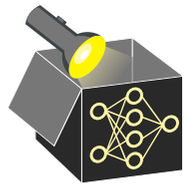


Example: Prediction surface of a model (left), select observation and visualize changes in prediction for different values of x_2 while keeping x_1 fixed

$\Rightarrow$ **local interpretation**

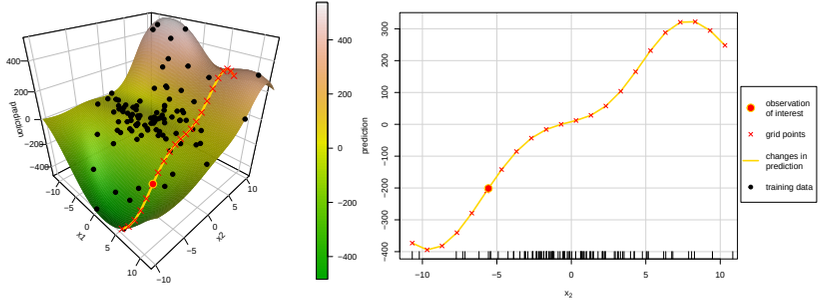


ICE CURVES - INTERPRETATION

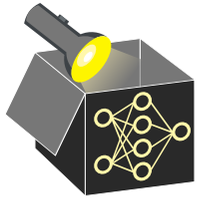


Example: Prediction surface of a model (left), select observation and visualize changes in prediction for different values of x_2 while keeping x_1 fixed

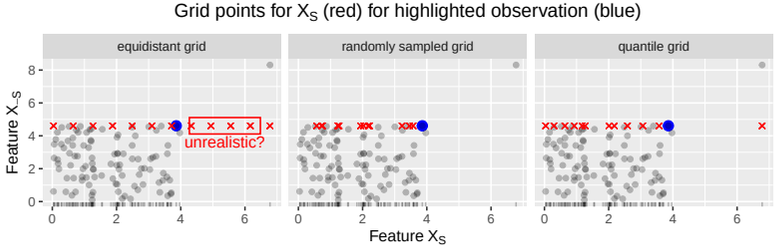
$\Rightarrow$ **local interpretation**



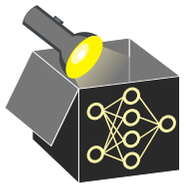
COMMENTS ON GRID VALUES



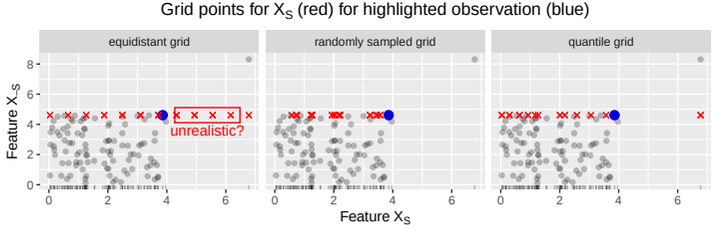
- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$; visualized on x-axis
- **Three common strategies** for grid definition:
  - Equidistant grid values within feature range
  - Random samples from observed feature values
  - Quantiles of observed feature values
- **Marginal realism:** Random and quantile grids better reflect the marginal distribution of $x_S \Rightarrow$ reduce unrealistic values along x_S



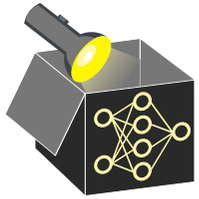
COMMENTS ON GRID VALUES



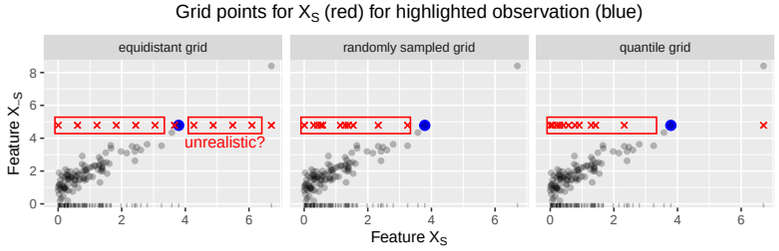
- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$; shown on x-axis
- **Three common strategies** for grid definition:
  - Equidistant grid values within feature range
  - Random samples from observed feature values
  - Quantiles of observed feature values
- **Marginal realism:** Random and quantile grids better reflect the marginal distribution of $x_S \Rightarrow$ reduce unrealistic values along x_S



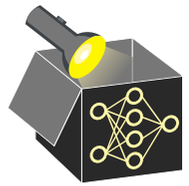
COMMENTS ON GRID VALUES



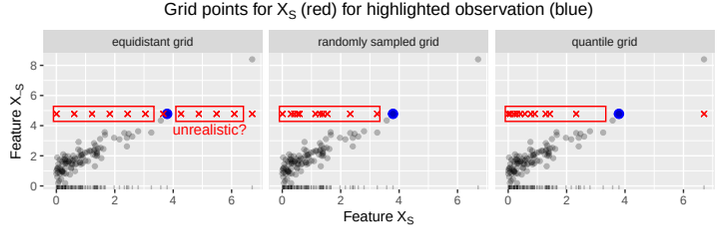
- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$; visualized on x-axis
- **Three common strategies** for grid definition:
  - Equidistant grid values within feature range
  - Random samples from observed feature values
  - Quantiles of observed feature values
- **Marginal realism:** Random and quantile grids better reflect the marginal distribution of $x_S \Rightarrow$ reduce unrealistic values along x_S
- **However:** For **correlated features**, extrapolation remains:



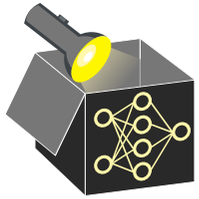
COMMENTS ON GRID VALUES



- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$; shown on x-axis
- **Three common strategies** for grid definition:
  - Equidistant grid values within feature range
  - Random samples from observed feature values
  - Quantiles of observed feature values
- **Marginal realism:** Random and quantile grids better reflect the marginal distribution of $x_S \Rightarrow$ reduce unrealistic values along x_S
- **However:** For **correlated features**, extrapolation remains:



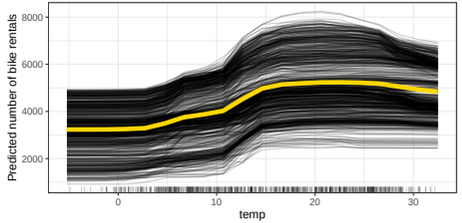
PRACTICAL CONSIDERATIONS



- **Grid resolution** (instances $\times$ grid over feature of interest)
  - Too coarse $\Rightarrow$ may miss sharp nonlinearities or discontinuities
  - Too fine $\Rightarrow$ high runtime (without gaining much)
  - Fix: cap at $\approx 50 - 100$ grid points; vectorize predictions by feeding the model a single data frame containing all grid-modified instances
- **ICE curves** (number of instances/curves visualized)
  - Too few $\Rightarrow$ hides variability across instances, misses subgroup differences
  - Too many $\Rightarrow$ visual overload (many overlapping curves), time intensive
  - Fix: Stratified or cluster-based subsample (e.g., 100); facet by subgroup

Default values for popular libraries:

Library	Grid	ICE curves
sklearn (Py)	100	1 000 (random)
PDPbox (Py)	10	num. rows
iml (R)	20	num. rows
pdp (R)	51	num. rows



ICE curves (**black lines**) and their point-wise average across the grid (**yellow line**)

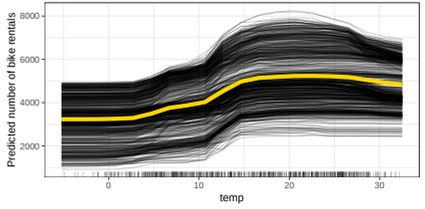
PRACTICAL CONSIDERATIONS

Grid resolution (instances $\times$ grid over feature of interest)
- Too coarse $\Rightarrow$ may miss sharp nonlinearities or discontinuities
- Too fine $\Rightarrow$ high runtime (without gaining much)
- Fix: cap at $\approx 50 - 100$ grid points; vectorize predictions by feeding the model a single data frame containing all grid-modified instances

ICE curves (number of instances/curves visualized)
- Too few $\Rightarrow$ hides instance variability, misses subgroup differences
- Too many $\Rightarrow$ visual overload (many overlapping curves), time intensive
- Fix: Stratified or cluster-based subsample (e.g., 100); facet by subgroup

Default values for popular libraries:

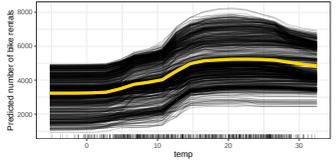
Library	Grid	ICE curves
sklearn (Py)	100	1 000 (random)
PDPbox (Py)	10	num. rows
iml (R)	20	num. rows
pdp (R)	51	num. rows



ICE curves (**black lines**) and their point-wise average across the grid (**yellow line**)

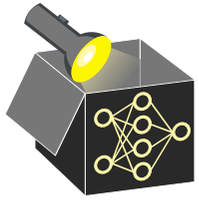
Interpretable Machine Learning

Partial Dependence (PD) plot



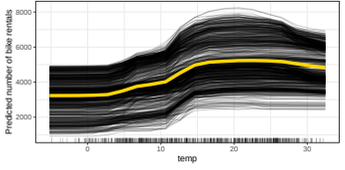
Learning goals

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



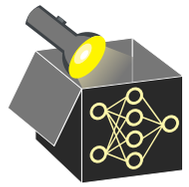
Interpretable Machine Learning

Feature Effects
Partial Dependence (PD) plot

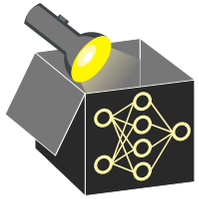


Learning goals

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

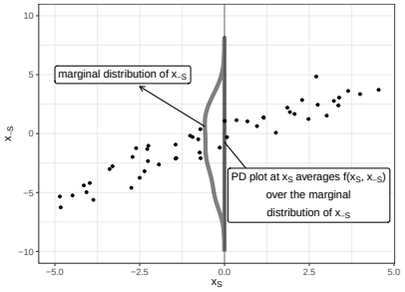


PARTIAL DEPENDENCE (PD) ▸ Friedman (2001)



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

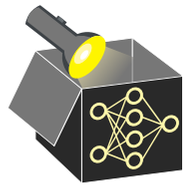
$$f_{S,PD}(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\right)$$
$$= \int_{-\infty}^{\infty} \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\, d\mathbb{P}(\mathbf{x}_{-S})$$



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

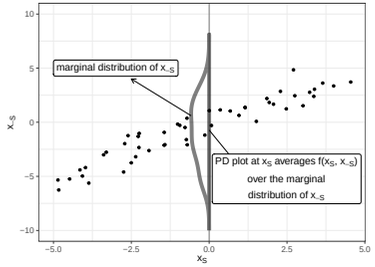
$$\hat{f}_{S,PD}(\mathbf{x}_S^*) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$$
$$= \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{S,ICE}^{(i)}(\mathbf{x}_S^*)$$

PARTIAL DEPENDENCE (PD) ▸ FRIEDMAN_2001



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

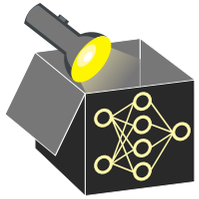
$$f_{S,PD}(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\right)$$
$$= \int_{-\infty}^{\infty} \hat{f}(\mathbf{x}_S, \mathbf{x}_{-S})\, d\mathbb{P}(\mathbf{x}_{-S})$$



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

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

PARTIAL DEPENDENCE

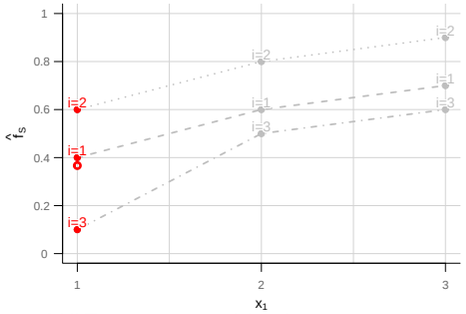


i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

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

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

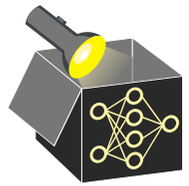
$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



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

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

PARTIAL DEPENDENCE

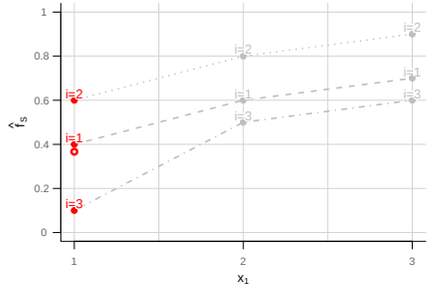


i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

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

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

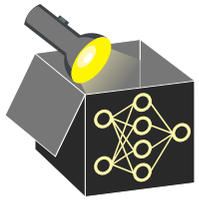
$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



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

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

PARTIAL DEPENDENCE

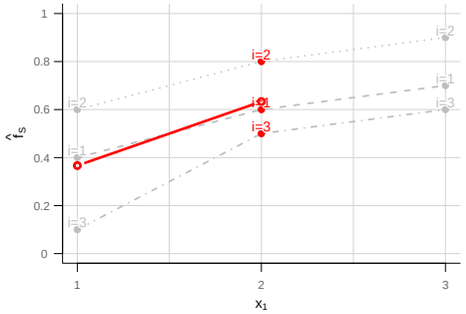


i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

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)

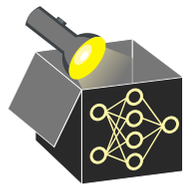
i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6



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

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

PARTIAL DEPENDENCE

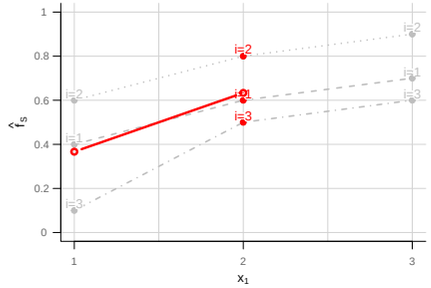


i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

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)

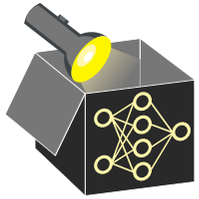
i	$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6



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

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

PARTIAL DEPENDENCE

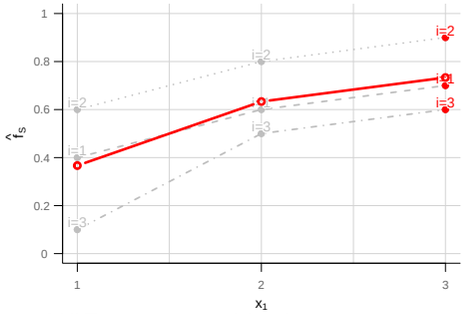


i	x_S	x_2	x_3	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

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

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

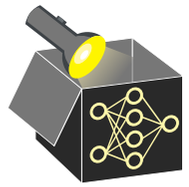
$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



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

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

PARTIAL DEPENDENCE

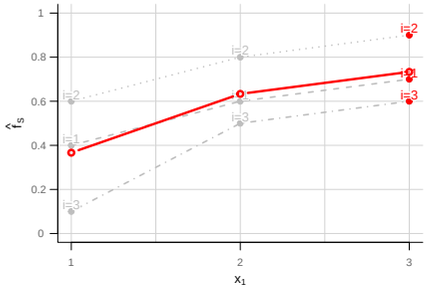


i	x_S	x_2	x_3	$\hat{f}$
1	1	4	7	0.4
2	1	5	8	0.6
3	1	6	9	0.1

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

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

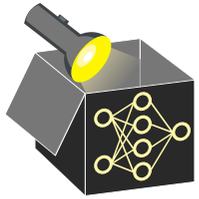
$\frac{1}{3}\sum_{i=1}^{3}\hat{f}$
⅓ (0.4 + 0.6 + 0.1)
⅓ (0.6 + 0.8 + 0.5)
⅓ (0.7 + 0.9 + 0.6)



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

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

EXAMPLE: PD FOR LINEAR MODEL



Assume a linear regression model with two features:

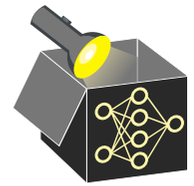
$$\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}_1, \mathbf{x}_2) = \hat{\theta}_1 \mathbf{x}_1 + \hat{\theta}_2 \mathbf{x}_2 + \hat{\theta}_0$$

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

EXAMPLE: PD FOR LINEAR MODEL



Assume a linear regression model with two features:

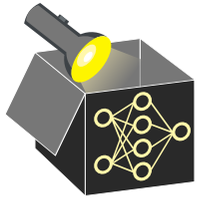
$$\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}_1, \mathbf{x}_2) = \hat{\theta}_1 \mathbf{x}_1 + \hat{\theta}_2 \mathbf{x}_2 + \hat{\theta}_0$$

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

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

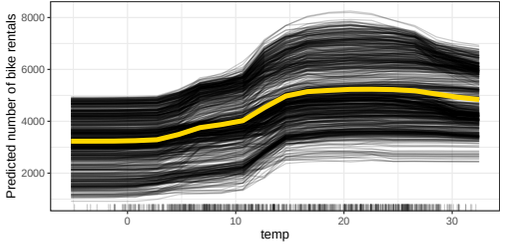
$\Rightarrow$ PD plot visualizes function $f_{1,PD}(\mathbf{x}_1) = \hat{\theta}_1 \mathbf{x}_1 + const$ ($\hat{=}$ feature effect of $\mathbf{x}_1$).

INTERPRETATION: PD AND ICE

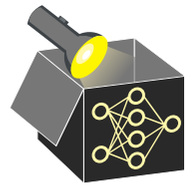


If feature varies:

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

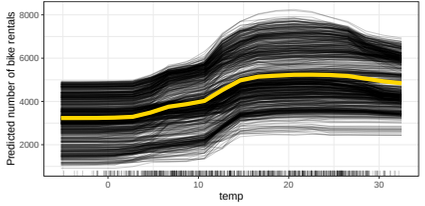


INTERPRETATION: PD AND ICE

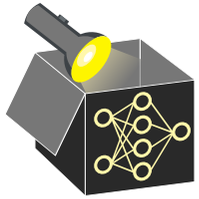


If feature varies:

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

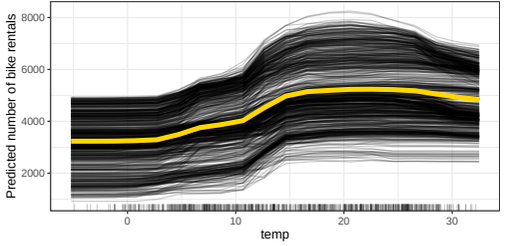


INTERPRETATION: PD AND ICE



If feature varies:

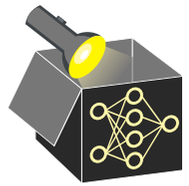
- **ICE:** How does **prediction of individual observation** change?
  $\Rightarrow$ **local** interpretation
- **PD:** How does **average effect** / **expected prediction** change?
  $\Rightarrow$ **global** interpretation



Insights from bike sharing data:

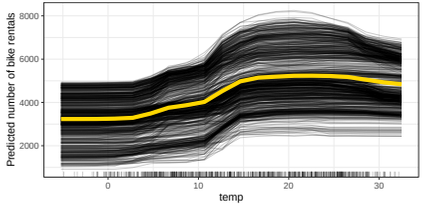
- Parallel ICE curves = homogeneous effect across obs.
- Warmer $\Rightarrow$ more rented bikes
- Too hot $\Rightarrow$ slightly less bikes
- Steepest increase in rentals occurs as temperature rises from 10℃ to 15℃.

INTERPRETATION: PD AND ICE



If feature varies:

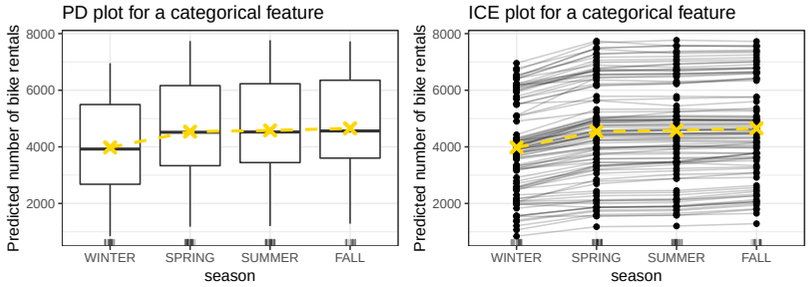
- **ICE:** How does **prediction of individual observation** change?
  $\Rightarrow$ **local** interpretation
- **PD:** How does **average effect** / **expected prediction** change?
  $\Rightarrow$ **global** interpretation

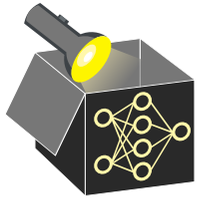


Insights from bike sharing data:

- 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℃.

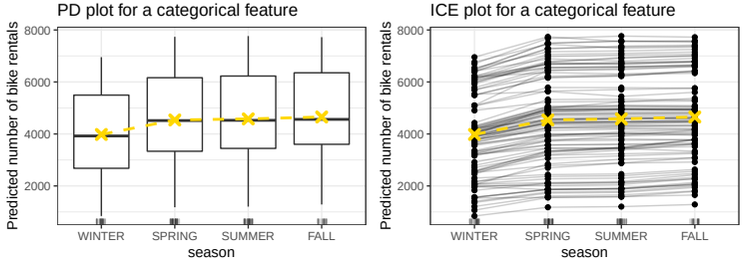
INTERPRETATION: CATEGORICAL FEATURES

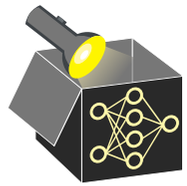




- PDP with boxplots and ICE with parallel coordinates plots
- NB: Categories can be unordered, if so, rather compare pairwise

INTERPRETATION: CATEGORICAL FEATURES

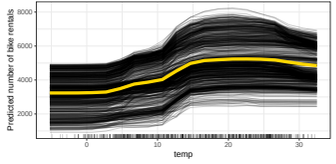




- PDP with boxplots and ICE with parallel coordinates plots
- NB: Categories can be unordered, if so, rather compare pairwise

Interpretable Machine Learning

PDP - Comments and Extensions



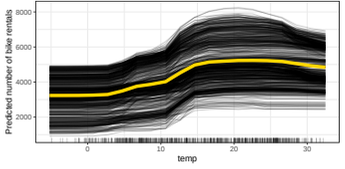
Learning goals

- Extrapolation and Interactions in PDPs
- Centered ICE and PDP

Interpretable Machine Learning

PDP - Comments and Extensions

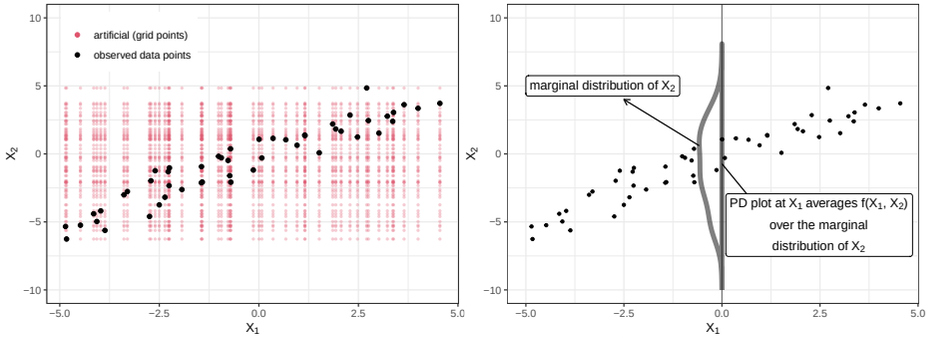
Feature Effects

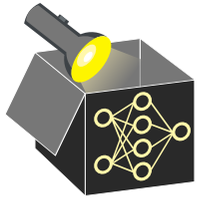


Learning goals

- Extrapolation and Interactions in PDPs
- Centered ICE and PDP

COMMENTS ON EXTRAPOLATION

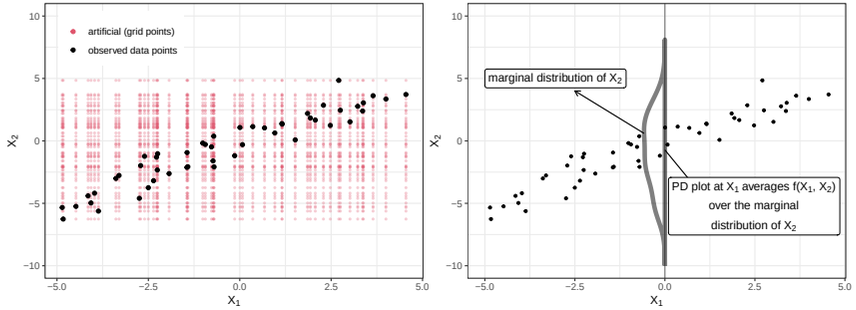




Extrapolation occurs in regions with few observations or if features are correlated

- **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

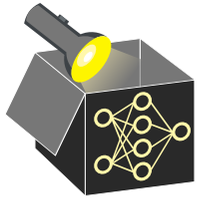
COMMENTS ON EXTRAPOLATION



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

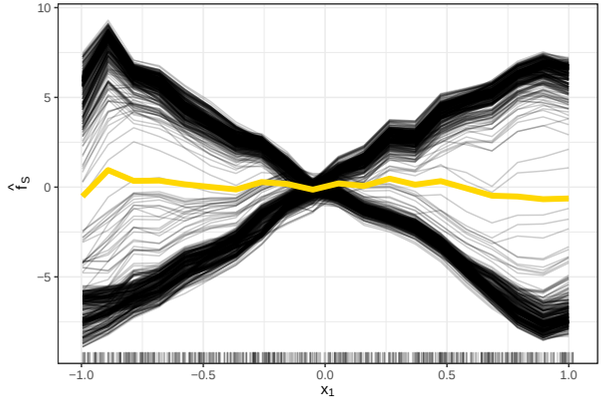
- **Example:** Features x_1 and x_2 are strongly correlated
- **Black points:** Observed points of the original data
- **Red:** Grid points to calculate ICE/PD (many unrealistic x_1, x_2 combinations)
  $\Rightarrow$ **PD at** $x_1 = 0$**:** Averages predictions over *full* marginal distribution of x_2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

COMMENTS ON INTERACTIONS

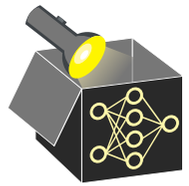


PD plots average ICE curves $\Rightarrow$ May **obscure heterogeneous effects** (interactions)

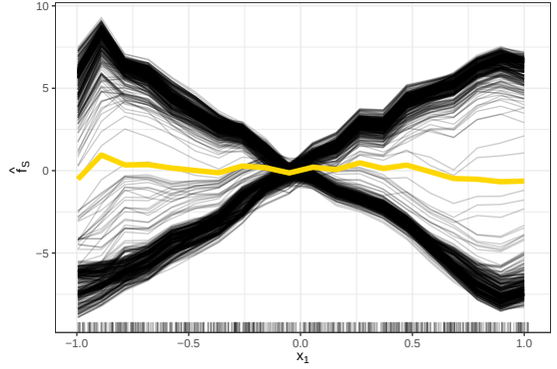
- **Example:** Feature x_1 = treatment dosage; x_2 = gender
  $\Rightarrow$ Males (increasing) and females (decreasing) respond differently to dosage
  $\Rightarrow$ PD curve (yellow) hides this divergence
- Plotting ICE and PD together helps detect interaction
- Diverse ICE shapes suggest interaction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



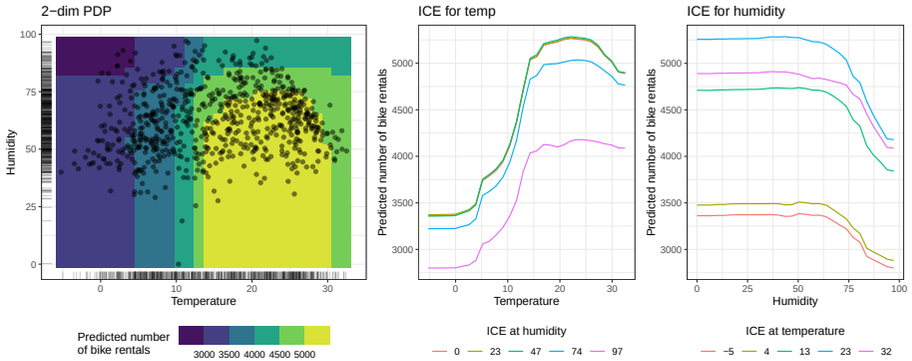
COMMENTS ON INTERACTIONS

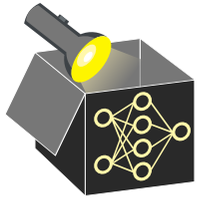


PD plots average ICE curves
$\rightsquigarrow$ May **obscure heterogeneous effects** (interactions)
- **Example:** Feature x_1 = treatment dosage; x_2 = gender
  $\Rightarrow$ Males ($\nearrow$) and females ($\searrow$) respond differently to dosage
  $\Rightarrow$ PD curve (yellow) hides this divergence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



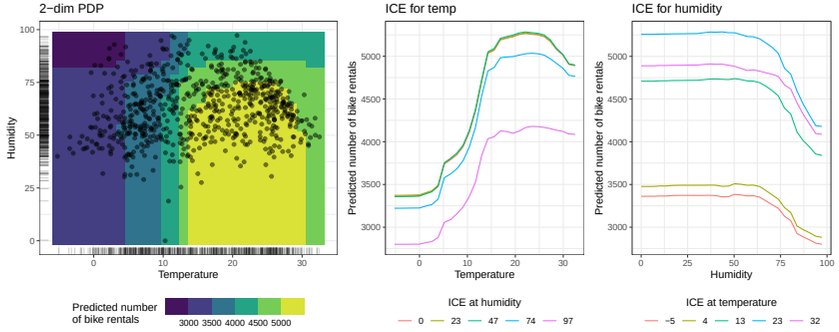
COMMENTS ON INTERACTIONS - 2D PD PLOT

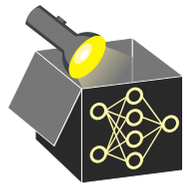




- Humidity and temperature interact at high values (see shape difference)
  ⇝ Shape of ICE curves changes across different (higher) values of other feature
  - ICE (temp): At high humidity, temp effect flattens (pink line)
  - ICE (hum): At high temperature, humidity effect falls steeper (blue/pink)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

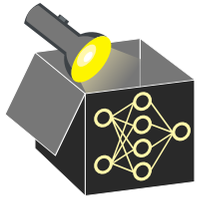
COMMENTS ON INTERACTIONS - 2D PD PLOT





- 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*

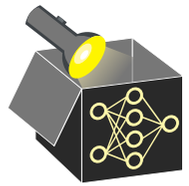
CENTERED ICE PLOT (C-ICE) ▸ Goldstein et al. (2015)



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

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

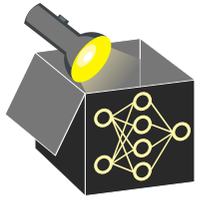
CENTERED ICE PLOT (C-ICE) ▸ GOLDSTEIN_2015



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)

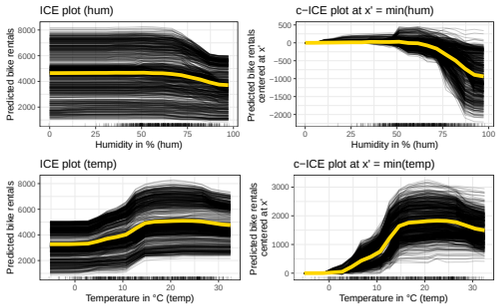


Issue: Varying-intercept (stacked) ICE curves obscure shape heterogeneity
Solution: Center ICE curves at fixed reference value, often $x' = \min(\mathbf{x}_S)$
$\Rightarrow$ Easier to identify heterogeneous shapes with c-ICE curves

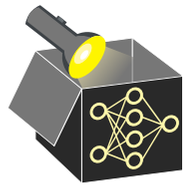
$$\hat{f}^{(i)}_{S,cICE}(\mathbf{x}_S) = \hat{f}(\mathbf{x}_S, \mathbf{x}^{(i)}_{-S}) - \hat{f}(x', \mathbf{x}^{(i)}_{-S}) = \hat{f}^{(i)}_S(\mathbf{x}_S) - \hat{f}^{(i)}_S(x')$$

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

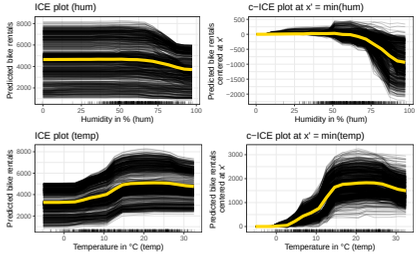


Issue: Varying-intercept (stacked) ICE curves obscure shape heterogeneity
Solution: Center ICE curves at fixed reference value, often $x' = \min(_S)$
$\Rightarrow$ Easier to identify heterogeneous shapes with c-ICE curves

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

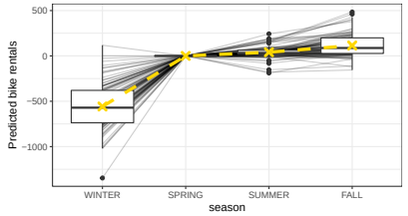
Interpretation

- 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



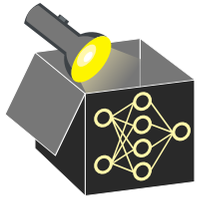
CENTERED ICE PLOT (C-ICE)

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



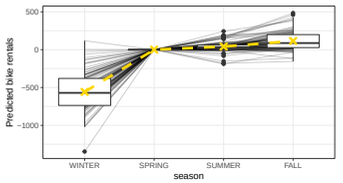
Interpretation:

- 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`



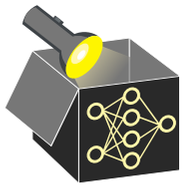
CENTERED ICE PLOT (C-ICE)

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

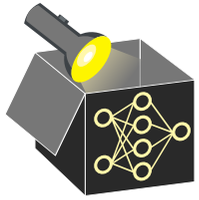


Interpretation:

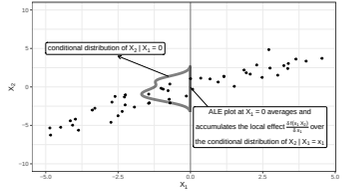
- The reference category is $x' =$ SPRING
- Yellow crosses: Average rentals if we jump from SPRING to any other season $\Rightarrow$ Number of bike rentals drops by $\sim$ 560 in `WINTER` and is slightly higher in `SUMMER` and `FALL` compared to `SPRING`



Interpretable Machine Learning



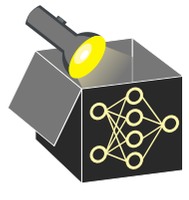
Accumulated Local Effect (ALE): Introduction



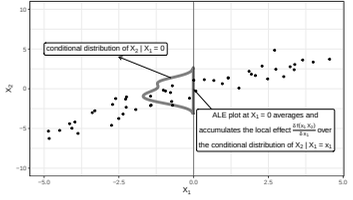
Learning goals

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

Interpretable Machine Learning



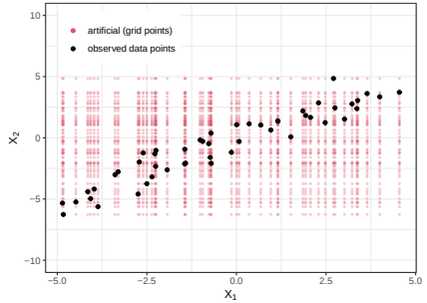
Feature Effects
Accumulated Local Effect (ALE): Intro

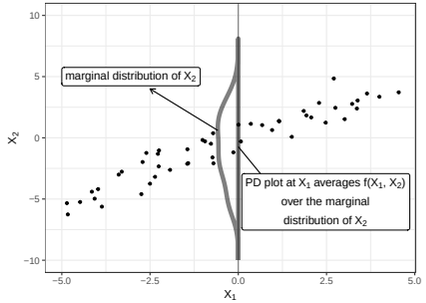


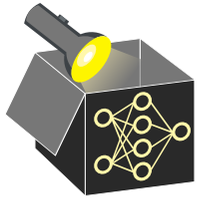
Learning goals

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

MOTIVATION - CORRELATED FEATURES

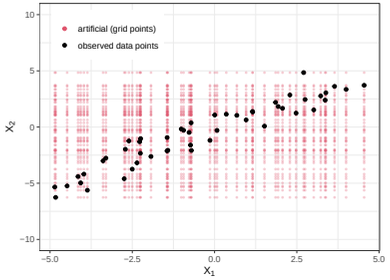


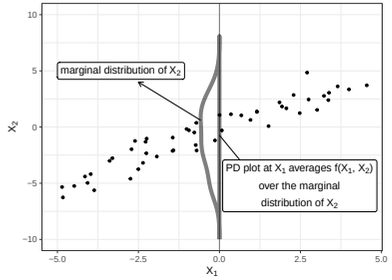


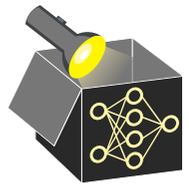


- PD plots **average over predictions** of artificial points that are out of distribution/ unlikely (red)
  $\Rightarrow$ Can lead to misleading / biased interpretations, especially if model also contains interactions
- Not wanted if interest is to interpret effects within data distribution

MOTIVATION - CORRELATED FEATURES

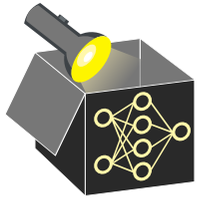






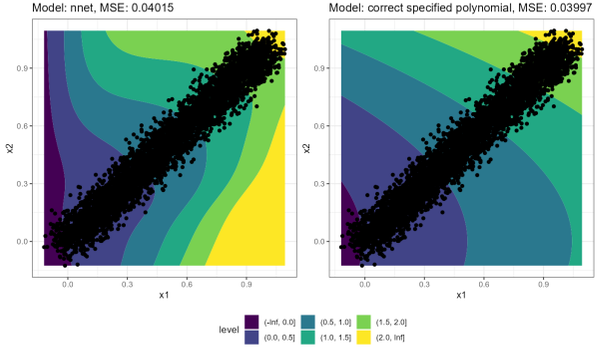
- PD plots **average over predictions** of artificial points that are out of distribution/ unlikely (red)
  $\Rightarrow$ Can lead to misleading / biased interpretations, especially if model also contains interactions
- Not wanted if interest is to interpret effects within data distribution

MOTIVATION - CORRELATED FEATURES



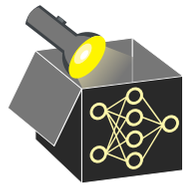
Example: Fit an NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and

$y = x_1 + x_2^2 + \epsilon$, where $x_1 = x + \epsilon_1$, $x_2 = x + \epsilon_2$ and $\epsilon_1, \epsilon_2 \sim N(0, 0.05)$.



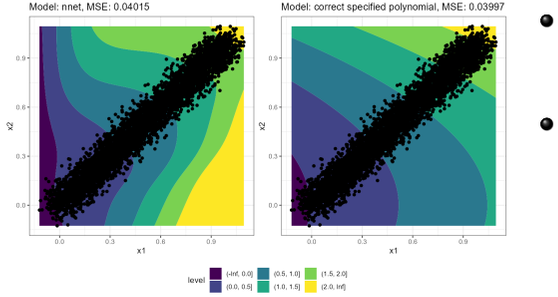
- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)

MOTIVATION - CORRELATED FEATURES



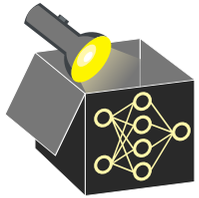
Example: Fit an NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and

$y = x_1 + x_2^2 + \epsilon$, where $x_1 = x + \epsilon_1$, $x_2 = x + \epsilon_2$ and $\epsilon_1, \epsilon_2 \sim N(0, 0.05)$.



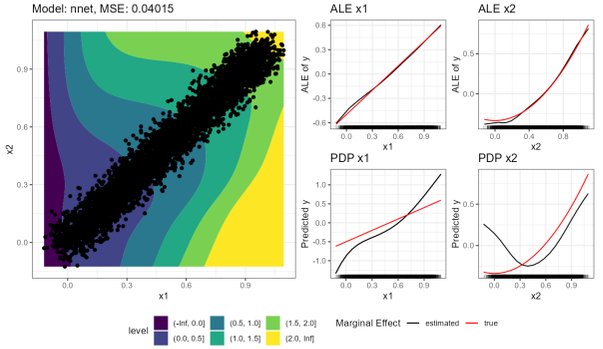
- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)

MOTIVATION - CORRELATED FEATURES



Example: Fit an NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and

$$y = x_1 + x_2^2 + \epsilon, \text{ where } x_1 = x + \epsilon_1, \; x_2 = x + \epsilon_2 \text{ and } \epsilon_1, \epsilon_2 \sim N(0, 0.05).$$

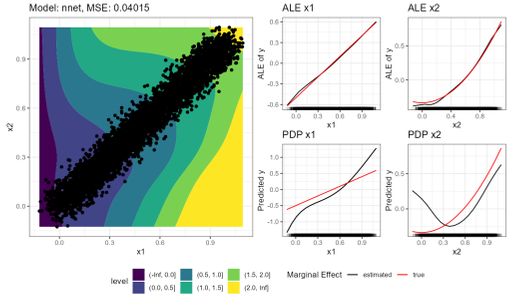


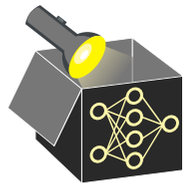
- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)
- ALE in line with ground truth
- PDP does not reflect ground truth effects of DGP well
  $\Rightarrow$ Due to interactions and averaging of points outside data distribution

MOTIVATION - CORRELATED FEATURES

Example: Fit an NN to 5000 simulated data points with $x \sim Unif(0, 1)$, $\epsilon \sim N(0, 0.2)$ and

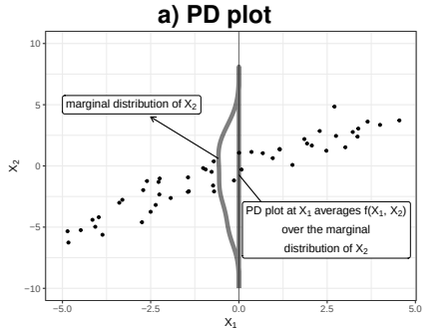
$$y = x_1 + x_2^2 + \epsilon, \text{ where } x_1 = x + \epsilon_1, \; x_2 = x + \epsilon_2 \text{ and } \epsilon_1, \epsilon_2 \sim N(0, 0.05).$$

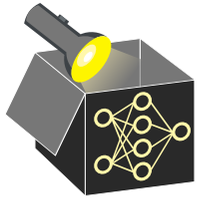




- Test error (MSE) of NN is comparable to other models
- NN contains interactions (see complex pred. surface)
- ALE in line with ground truth
- PDP does not reflect ground truth effects of DGP well
  $\Rightarrow$ Due to interactions and averaging of points outside data distribution

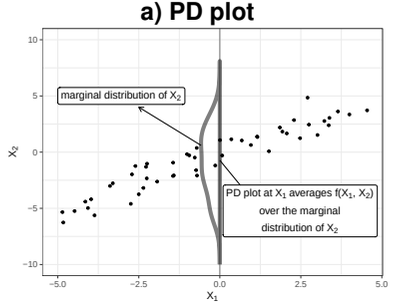
M PLOT VS. PD PLOT

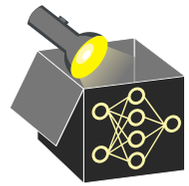




a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1, \mathbf{x}_2^{(i)})$

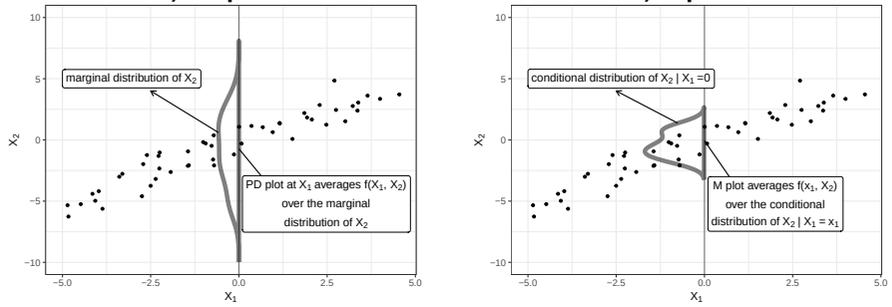
M PLOT VS. PD PLOT

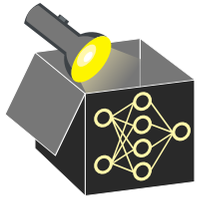




a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1, \mathbf{x}_2^{(i)})$

M PLOT VS. PD PLOT

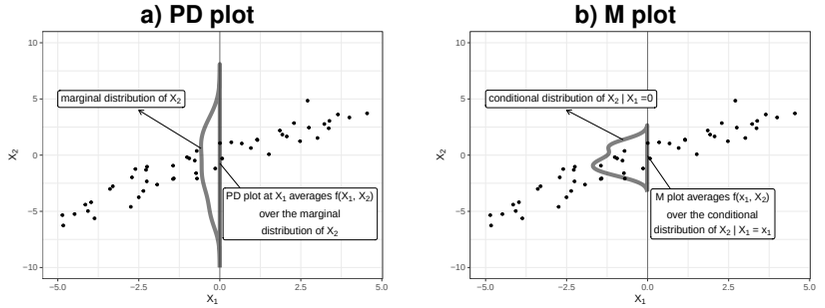


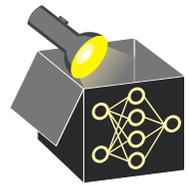


a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1, \mathbf{x}_2^{(i)})$

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

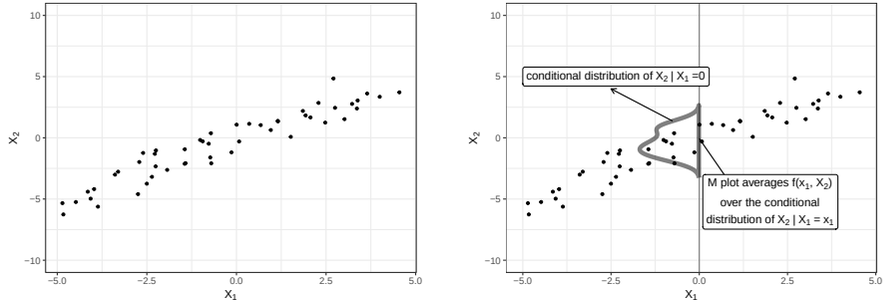


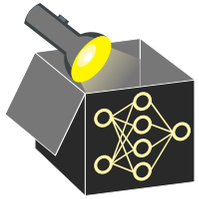


a) PD plot $\mathbb{E}_{\mathbf{x}_2}\left(\hat{f}(x_1, \mathbf{x}_2)\right)$ is estimated by $\hat{f}_{1,PD}(x_1) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_1, \mathbf{x}_2^{(i)})$

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.

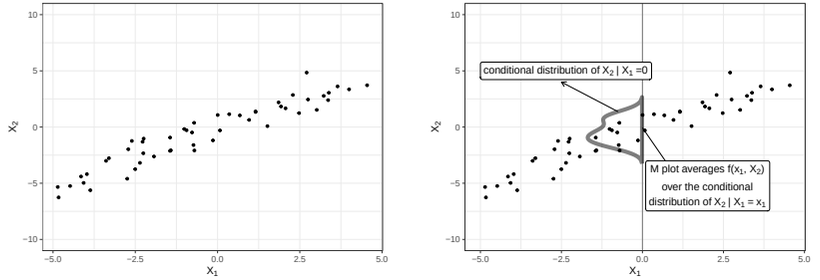
M PLOT VS. PD PLOT

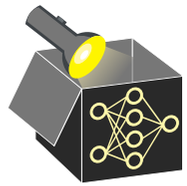




- M plots average predictions over conditional distribution (e.g., $\mathbb{P}(\mathbf{x}_2|x_1)$)
  $\Rightarrow$ Averaging predictions close to data distribution avoid extrapolation issues
- **But:** M plots suffer from omitted-variable bias (OVB)
  - Because of the conditioning M plots contain effects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

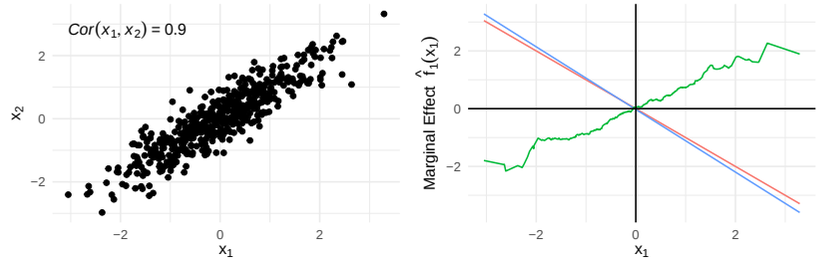
M PLOT VS. PD PLOT





- M plots average predictions over conditional distribution (e.g., $\mathbb{P}(\mathbf{x}_2|x_1)$)
  $\Rightarrow$ Averaging predictions close to data distrib. avoids extrapolation issues
- **But:** M plots suffer from omitted-variable bias (OVB)
  - Because of the conditioning M plots 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



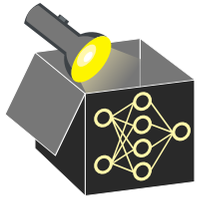
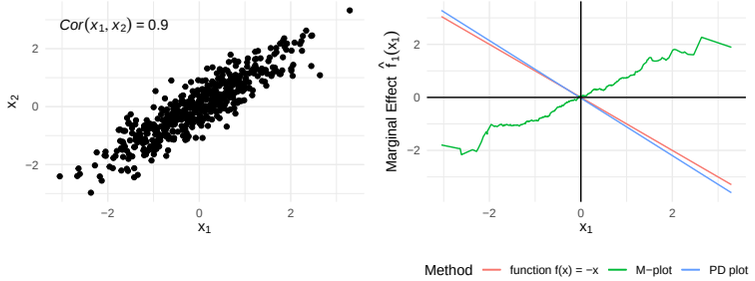


Illustration: Fit LM on 500 i.i.d. observations with features $x_1, x_2 \sim N(0, 1)$, $Cor(x_1, x_2) = 0.9$ and

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

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

M PLOT VS. PD PLOT - OVB EXAMPLE



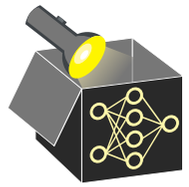
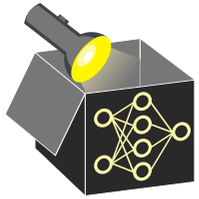


Illustration: Fit LM on 500 i.i.d. observations with features $x_1, x_2 \sim N(0, 1)$, $Cor(x_1, x_2) = 0.9$ and

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

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

IDEA: INTEGRATING PARTIAL DERIVATIVES

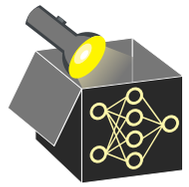


Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ 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

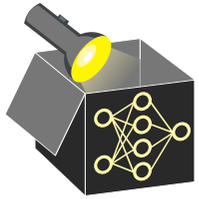


Idea: To remove unwanted effects of other features, take partial derivatives (local effects) of prediction function w.r.t. feature of interest and integrate (accumulate) them w.r.t. the same feature

$\Rightarrow$ 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 $\mathbf{x}_j$

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

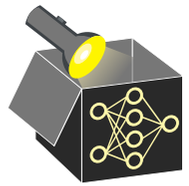
$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

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

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

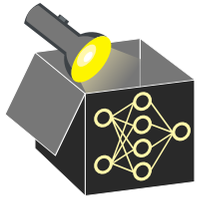
$\Rightarrow$ Integrating again w.r.t. $\mathbf{x}_j$ recovers the original main effect of $\mathbf{x}_j$

Example:

- Consider an additive prediction function:

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

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 $\mathbf{x}_j$

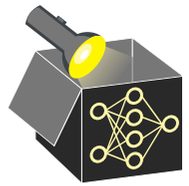
Example:

- 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 $\mathbf{x}_j$

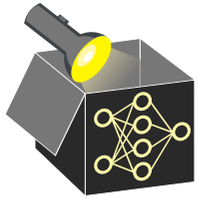
Example:

- 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: $\frac{\partial \hat{f}(x_1, x_2)}{\partial x_1} = 2 - 4x_2$

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:

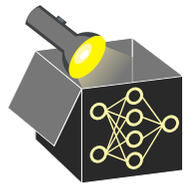
- Consider an additive prediction function:

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

- 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:

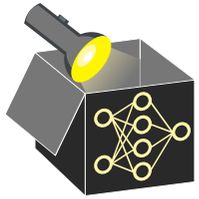
$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- 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 ($z_0 = \min(x_1)$):

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

Interpretable Machine Learning – 6 / 6

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:

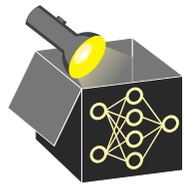
$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2 - 4x_1x_2$$

- 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 ($z_0 = \min(x_1)$):

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

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

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:

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

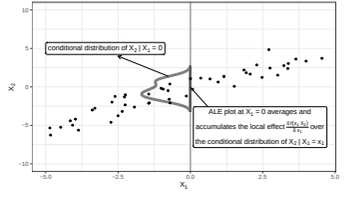
- Partial derivative of $\hat{f}$ w.r.t. x_1: $\frac{\partial \hat{f}(x_1,x_2)}{\partial x_1} = 2 - 4x_2$
- Integral of partial derivative ($z_0 = \min(x_1)$):

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

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

Interpretable Machine Learning

Accumulated Local Effect (ALE) plot

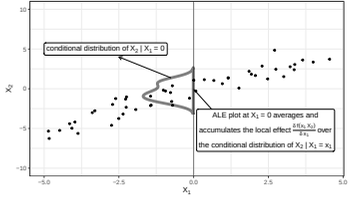


Learning goals

- Understand ALE plots
- Difference between ALE and PD plots

Interpretable Machine Learning

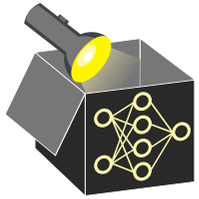
Feature Effects
Accumulated Local Effect (ALE) plot



Learning goals

- Understand ALE plots
- Difference between ALE and PD plots

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

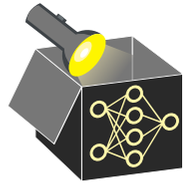


ALE plots estimate the marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)

ACCUMULATED LOCAL EFFECTS (ALE) ZHU_2020



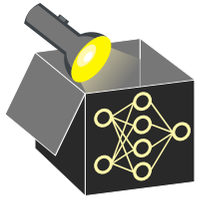
ALE plots estimate marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)

Interpretable Machine Learning – 1 / 8

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

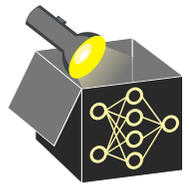


ALE plots estimate the marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)
2. **Average local effects** over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation (unlike PD plots)

ACCUMULATED LOCAL EFFECTS (ALE) ▸ ZHU_2020



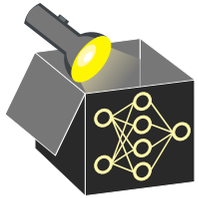
ALE plots estimate marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)
2. **Average local effects** over conditional distr. $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation (unlike PD plots)

Interpretable Machine Learning – 1 / 8

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

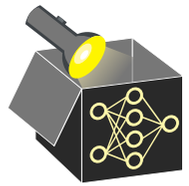


ALE plots estimate the marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)
2. **Average local effects** over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation (unlike PD plots)
3. **Accumulate:** Integrate averaged local effects up to a specific value $x \in \mathcal{X}_S$
   $\Rightarrow$ Reconstructs main effect of x_S

ACCUMULATED LOCAL EFFECTS (ALE) ▸ ZHU_2020

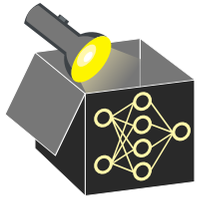


ALE plots estimate marginal effect of a feature by accumulating its local effects (integrating partial derivatives), evaluated in regions supported by the data.

Computation Steps:

1. **Estimate local effects** $\frac{\partial \hat{f}(x_S, \mathbf{x}_{-S})}{\partial x_S}$ (via finite differences)
   $\Rightarrow$ Removes unwanted main effects of other features $\mathbf{x}_{-S}$ (unlike M plots)
2. **Average local effects** over conditional distr. $\mathbb{P}(\mathbf{x}_{-S}|x_S)$ similar to M plots
   $\Rightarrow$ Avoids extrapolation (unlike PD plots)
3. **Accumulate:** Integrate averaged local effects up to a specific $x \in \mathcal{X}_S$
   $\Rightarrow$ Reconstructs main effect of x_S

FIRST ORDER ALE FUNCTION

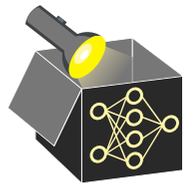


Uncentered ALE Function evaluated at $x \in \mathcal{X}_S$ (domain of feature x_S):

$$\tilde{f}_{S,\text{ALE}}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}}_{\substack{(2)\text{ average}\\ \text{locally}}} \left(\underbrace{\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}}_{(1)\text{ local effect}} \right) dz_S = \int_{z_0}^{x} \int \frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S} d\mathbb{P}(\mathbf{x}_{-S} \mid z_S) dz_S$$

- x_S is feature of interest, with minimum value $z_0 = \min(x_S)$
- z_S is integration variable ranging over $\mathcal{X}_S$, used to evaluate local effects
- $\mathbf{x}_{-S}$ denotes all other features (complement of S)

FIRST ORDER ALE FUNCTION

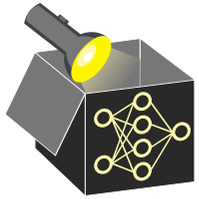


Uncentered ALE Function evaluated at $x \in \mathcal{X}_S$ (domain of feature x_S):

$$\tilde{f}_{S,\text{ALE}}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}}_{\substack{(2)\text{ average}\\ \text{locally}}} \left(\underbrace{\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}}_{(1)\text{ local effect}} \right) dz_S = \int_{z_0}^{x} \int \frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S} d\mathbb{P}(\mathbf{x}_{-S} \mid z_S) dz_S$$

- x_S is feature of interest, with minimum value $z_0 = \min(x_S)$
- z_S is integration variable ranging over $\mathcal{X}_S$, used to evaluate local effects
- $\mathbf{x}_{-S}$ denotes all other features (complement of S)

FIRST ORDER ALE FUNCTION



Uncentered ALE Function evaluated at $x \in \mathcal{X}_S$ (domain of feature x_S):

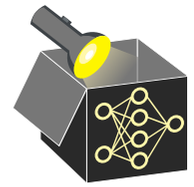
$$\tilde{f}_{S,\text{ALE}}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}}_{\substack{(2)\text{ average}\\ \text{locally}}} \left(\underbrace{\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}}_{(1)\text{ local effect}} \right) dz_S = \int_{z_0}^{x} \int \frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S} d\mathbb{P}(\mathbf{x}_{-S} \mid z_S) dz_S$$

- x_S is feature of interest, with minimum value $z_0 = \min(x_S)$
- z_S is integration variable ranging over $\mathcal{X}_S$, used to evaluate local effects
- $\mathbf{x}_{-S}$ denotes all other features (complement of S)

Centering (to ensure identifiability):

$$f_{S,\text{ALE}}(x) = \tilde{f}_{S,\text{ALE}}(x) - \underbrace{\int \tilde{f}_{S,\text{ALE}}(x_S)\, d\mathbb{P}(x_S)}_{\text{constant shift to mean zero}}$$

FIRST ORDER ALE FUNCTION



Uncentered ALE Function evaluated at $x \in \mathcal{X}_S$ (domain of feature x_S):

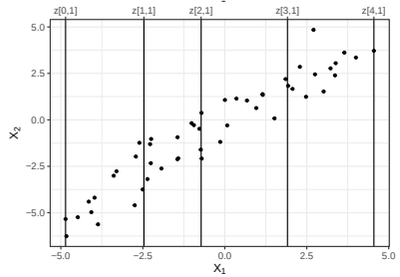
$$\tilde{f}_{S,\text{ALE}}(x) = \underbrace{\int_{z_0}^{x}}_{(3)} \underbrace{\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}}_{\substack{(2)\text{ average}\\ \text{locally}}} \left(\underbrace{\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}}_{(1)\text{ local effect}} \right) dz_S = \int_{z_0}^{x} \int \frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S} d\mathbb{P}(\mathbf{x}_{-S} \mid z_S) dz_S$$

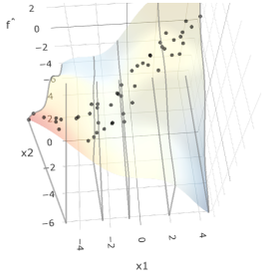
- x_S is feature of interest, with minimum value $z_0 = \min(x_S)$
- z_S is integration variable ranging over $\mathcal{X}_S$, used to evaluate local effects
- $\mathbf{x}_{-S}$ denotes all other features (complement of S)

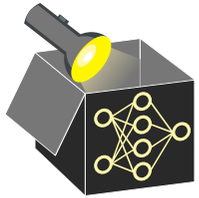
Centering (to ensure identifiability):

$$f_{S,\text{ALE}}(x) = \tilde{f}_{S,\text{ALE}}(x) - \underbrace{\int \tilde{f}_{S,\text{ALE}}(x_S)\, d\mathbb{P}(x_S)}_{\text{constant shift to mean zero}}$$

ALE ESTIMATION: ILLUSTRATION





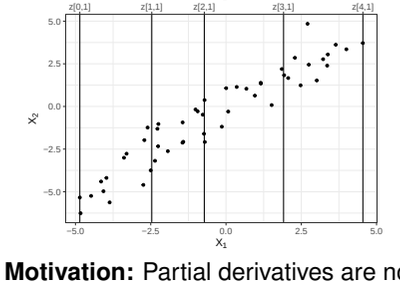


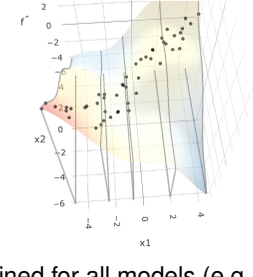
- **Motivation:** Partial derivatives are not well-defined for all models (e.g., tree-based methods). $\Rightarrow$ Use finite differences within intervals instead.
- Partition the feature range of x_S into K intervals (vertical lines)
  - Define intervals:

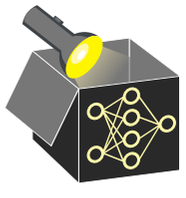
    $x_S \in [\min(x_S), \max(x_S)] \Rightarrow x_S \in [z_0, z_{1,S}] \cup [z_{1,S}, z_{2,S}] \cup \cdots \cup [z_{K-1,S}, z_{K,S}]$

  - *Equidistant*: preserves resolution
  - *Quantile-based*: balances sample size per interval

ALE ESTIMATION: ILLUSTRATION





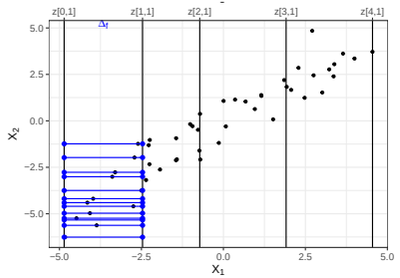


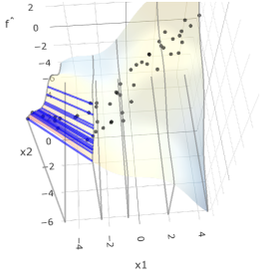
- **Motivation:** Partial derivatives are not well-defined for all models (e.g., tree-based methods). $\Rightarrow$ Use finite differences within intervals instead.
- Partition the feature range of x_S into K intervals (vertical lines)
  - Define intervals:

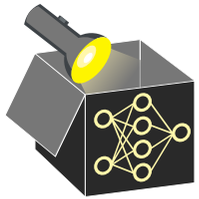
    $x_S \in [\min(x_S), \max(x_S)] \Rightarrow x_S \in [z_0, z_{1,S}] \cup [z_{1,S}, z_{2,S}] \cup \cdots \cup [z_{K-1,S}, z_{K,S}]$

  - *Equidistant*: preserves resolution
  - *Quantile-based*: balances sample size per interval

ALE ESTIMATION: ILLUSTRATION

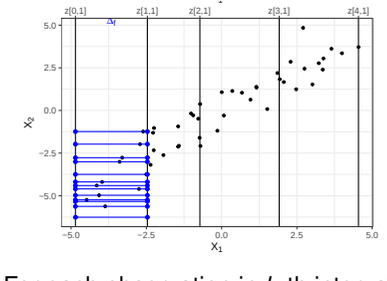


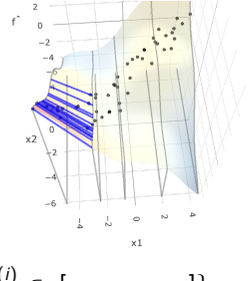


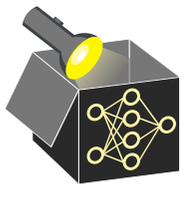


- For each observation in k-th interval, i.e., $\{i : x_S^{(i)} \in [z_{k-1,S}, z_{k,S}]\}$:
  - Replace $x_S^{(i)}$ with upper/lower interval bounds, keeping $\mathbf{x}_{-S}^{(i)}$ fixed
  - Compute observation-wise finite difference of i-th obs. in k-th interval
    $\rightsquigarrow \hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$ (approximates local effect)

ALE ESTIMATION: ILLUSTRATION

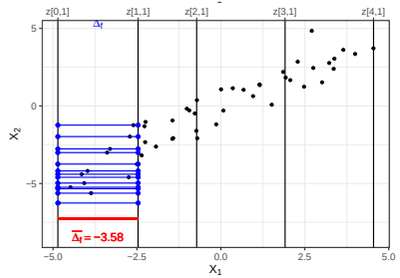


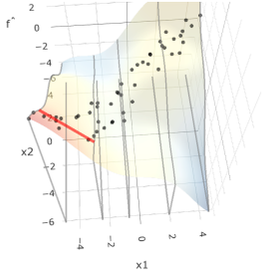


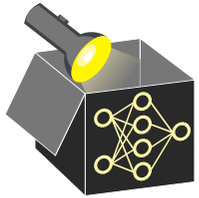


- For each observation in k-th interval, i.e., $\{i : x_S^{(i)} \in [z_{k-1,S}, z_{k,S}]\}$:
  - Replace $x_S^{(i)}$ with upper/lower interval bounds, keeping $\mathbf{x}_{-S}^{(i)}$ fixed
  - Compute obs.-wise finite difference of i-th obs. in k-th interval
    $\rightsquigarrow \hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$ (approximates local effect)

ALE ESTIMATION: ILLUSTRATION

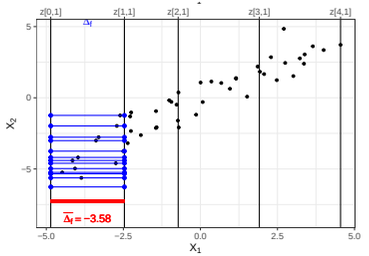


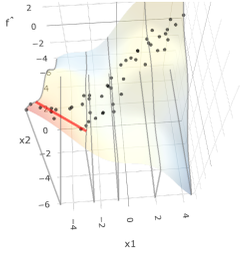


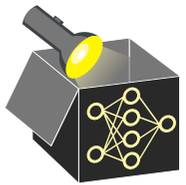


- For each observation in k-th interval, i.e., $\{i : x_S^{(i)} \in [z_{k-1,S}, z_{k,S}]\}$:
  - Replace $x_S^{(i)}$ with upper/lower interval bounds, keeping $\mathbf{x}_{-S}^{(i)}$ fixed
  - Compute observation-wise finite difference of i-th obs. in k-th interval
    $\rightsquigarrow \hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$ (approximates local effect)
- Average these finite differences over all observations in each interval
  $\rightsquigarrow$ Approximates **inner integral** $\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left[\partial\hat{f}/\partial z_S\right]$
- Accumulate these averages from z_0 to the point of interest $x \in \mathcal{X}_S$
  $\rightsquigarrow$ Approximates **outer integral** over $z_S \in [z_0, x]$ $\Rightarrow$ uncentered ALE function

ALE ESTIMATION: ILLUSTRATION

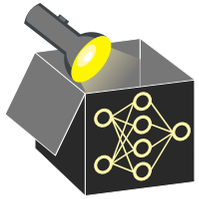






- For each observation in k-th interval, i.e., $\{i : x_S^{(i)} \in [z_{k-1,S}, z_{k,S}]\}$:
  - Replace $x_S^{(i)}$ with upper/lower interval bounds, keeping $\mathbf{x}_{-S}^{(i)}$ fixed
  - Compute obs.-wise finite difference of i-th obs. in k-th interval
    $\rightsquigarrow \hat{f}(z_{k,S}, \mathbf{x}_{-S}^{(i)}) - \hat{f}(z_{k-1,S}, \mathbf{x}_{-S}^{(i)})$ (approximates local effect)
- Average these finite differences over all observations in each interval
  $\rightsquigarrow$ Approximates **inner integral** $\mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left[\partial\hat{f}/\partial z_S\right]$
- Accumulate these averages from z_0 to the point of interest $x \in \mathcal{X}_S$
  $\rightsquigarrow$ Approximates **outer integral** over $z_S \in [z_0, x]$
  $\Rightarrow$ uncentered ALE function

ALE ESTIMATION: FORMULA

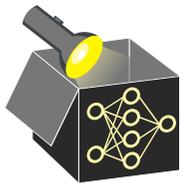


Estimated uncentered ALE: For a point $x \in \mathcal{X}_S$, define:

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

- $[z_{k-1,S}, z_{k,S}]$: k-th interval of feature x_S with interval bounds $z_{k-1,S}$ and $z_{k,S}$
- $k_S(x)$: index of the interval in which x lies
- $n_S(k)$: number of observations in interval k
- $\mathbf{x}_{-S}^{(i)}$: all other features held fixed for i-th observation

ALE ESTIMATION: FORMULA

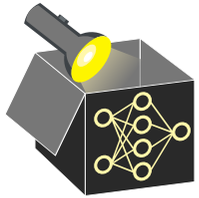


Estimated uncentered ALE: For a point $x \in \mathcal{X}_S$, define:

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

- $[z_{k-1,S}, z_{k,S}]$: k-th interval of feat. x_S with interval bounds $z_{k-1,S}$ and $z_{k,S}$
- $k_S(x)$: index of the interval in which x lies
- $n_S(k)$: number of observations in interval k
- $\mathbf{x}_{-S}^{(i)}$: all other features held fixed for i-th observation

ALE ESTIMATION: FORMULA



Estimated uncentered ALE: For a point $x \in \mathcal{X}_S$, define:

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

- $[z_{k-1,S}, z_{k,S}]$: k-th interval of feature x_S with interval bounds $z_{k-1,S}$ and $z_{k,S}$
- $k_S(x)$: index of the interval in which x lies
- $n_S(k)$: number of observations in interval k
- $\mathbf{x}_{-S}^{(i)}$: all other features held fixed for i-th observation

Centering: Ensure identifiability by subtracting mean uncentered ALE (constant c):

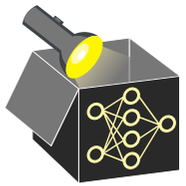
$$\hat{f}_{S,\text{ALE}}(x) = \hat{\tilde{f}}_{S,\text{ALE}}(x) \; - \; c, \qquad c = \tfrac{1}{n} \sum_{i=1}^{n} \hat{\tilde{f}}_{S,\text{ALE}}\left(x_S^{(i)}\right).$$

Efficient centering (used in implementations): Use weighted trapezoidal averaging of interval-wise boundary values (avoids redundant re-evaluation at all n points):

$$c = \sum_{k=1}^{K} \tfrac{1}{2} \cdot \left(\hat{\tilde{f}}_{S,\text{ALE}}(z_{k-1,S}) + \hat{\tilde{f}}_{S,\text{ALE}}(z_{k,S})\right) \cdot \tfrac{n_S(k)}{n}$$

Plotting ALE: Visualize the pairs $\left(z_{k,S}, \hat{f}_{S,\text{ALE}}(z_{k,S})\right)$ for all interval boundaries $z_{k,S}$.

ALE ESTIMATION: FORMULA



Estimated uncentered ALE: For a point $x \in \mathcal{X}_S$, define:

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

- $[z_{k-1,S}, z_{k,S}]$: k-th interval of feat. x_S with interval bounds $z_{k-1,S}$ and $z_{k,S}$
- $k_S(x)$: index of the interval in which x lies
- $n_S(k)$: number of observations in interval k
- $\mathbf{x}_{-S}^{(i)}$: all other features held fixed for i-th observation

Centering: Ensure identifiability by subtracting mean uncentered ALE (c):

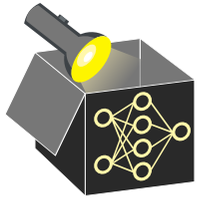
$$\hat{f}_{S,\text{ALE}}(x) = \hat{\tilde{f}}_{S,\text{ALE}}(x) \; - \; c, \qquad c = \tfrac{1}{n} \sum_{i=1}^{n} \hat{\tilde{f}}_{S,\text{ALE}}\left(x_S^{(i)}\right).$$

Efficient centering (used in implementations): Use weighted trapezoidal averaging of interval-wise boundary values (avoids redundant re-evaluation at all n points):

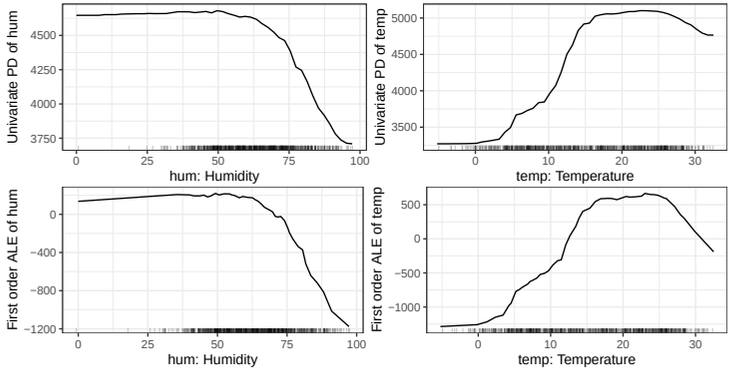
$$c = \sum_{k=1}^{K} \tfrac{1}{2} \cdot \left(\hat{\tilde{f}}_{S,\text{ALE}}(z_{k-1,S}) + \hat{\tilde{f}}_{S,\text{ALE}}(z_{k,S})\right) \cdot \tfrac{n_S(k)}{n}$$

Plotting: Visualize pairs $\left(z_{k,S}, \hat{f}_{S,\text{ALE}}(z_{k,S})\right)$ for all interval boundaries $z_{k,S}$.

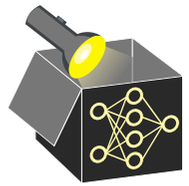
BIKE SHARING DATASET: FIRST ORDER ALE



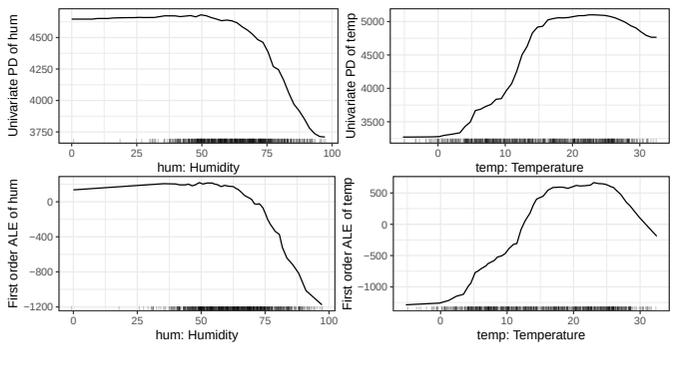
- **Visual comparison:** PD plot (top) vs. First-order ALE plot (bottom)
- **Shape:** Both plots show similar trends, but differ in y-axis scale due to centering
- **Interpretation:** ALE accounts for feature dependencies and avoids extrapolation into unsupported regions
  - $\rightsquigarrow$ PD reflects model behavior in entire feature space ("true to the model")
  - $\rightsquigarrow$ ALE focuses on effects in data-supported regions ("true to the data")



BIKE SHARING DATASET: FIRST ORDER ALE

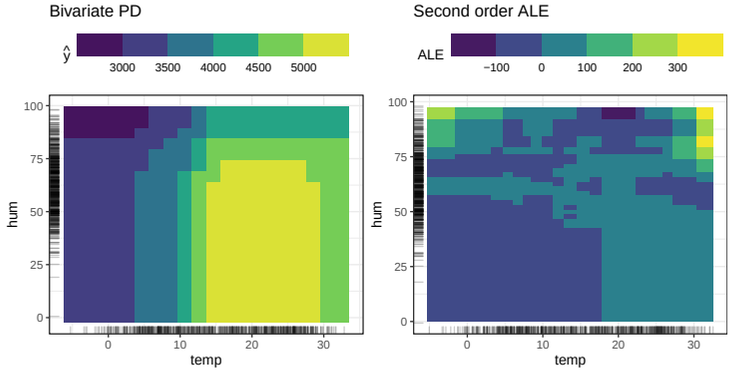


- **Visual comparison:** PD plot (top) vs. First-order ALE plot (bottom)
- **Shape:** Similar trends in both plots; y-axis scale differs due to centering
- **Interpretation:** ALE accounts for feature dependencies and avoids extrapolation into unsupported regions
  - $\rightsquigarrow$ PD reflects model behavior in entire feature space ("true to the model")
  - $\rightsquigarrow$ ALE focuses on effects in data-supported regions ("true to the data")



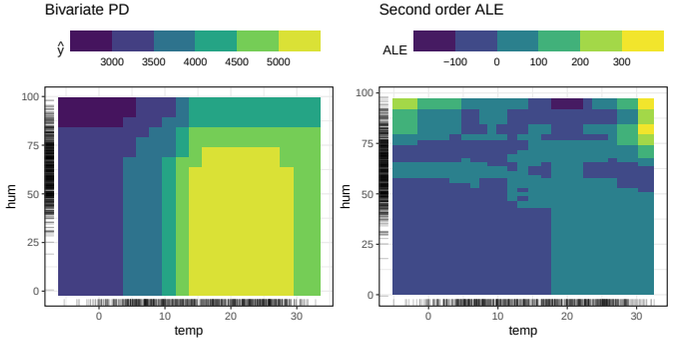
BIKE SHARING DATASET: SECOND ORDER ALE

Unlike bivariate PD plots, 2nd-order ALE plots only estimate pure interaction between two features (1st-order effects are not included).

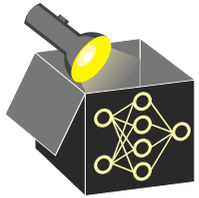


BIKE SHARING DATASET: SECOND ORDER ALE

Unlike bivariate PD plots, 2nd-order ALE plots only estimate pure interaction between two features (1st-order effects are not included).



PD VS. ALE



PD:

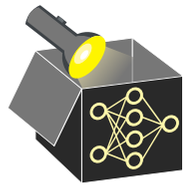
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

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

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: Needs $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$

PD VS. ALE



PD:

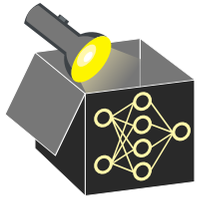
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

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

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$

PD VS. ALE



PD:

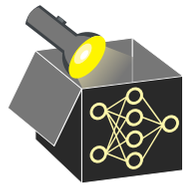
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

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

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: Needs $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages the
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates these partial derivatives over $z_S \in [z_0, x] \subseteq \mathcal{X}_S$
  $\rightsquigarrow$ isolates effect of x_S and removes main effect of other dependent features

PD VS. ALE



PD:

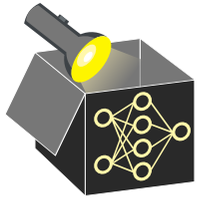
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

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

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages the
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates these partial deriv. over $z_S \in [z_0, x] \subseteq \mathcal{X}_S$
  $\rightsquigarrow$ isolates effect of x_S and removes main effect of other dependent feat.

PD VS. ALE



PD:

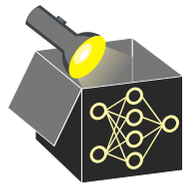
$$f_{S,PD}(x_S) = \mathbb{E}_{\mathbf{x}_{-S}}\left(\hat{f}(x_S, \mathbf{x}_{-S})\right)$$

ALE:

$$f_{S,\text{ALE}}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left(\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}\right) dz - \int \tilde{f}_{S,\text{ALE}}(x_S)\, d\mathbb{P}(x_S)$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: Needs $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages the
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates these partial derivatives over $z_S \in [z_0, x] \subseteq \mathcal{X}_S$
  $\rightsquigarrow$ isolates effect of x_S and removes main effect of other dependent features
- Difference 3: ALE is centered so that $\mathbb{E}_{x_S}\left(f_{S,ALE}(x)\right) = 0$

PD VS. ALE



PD:

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

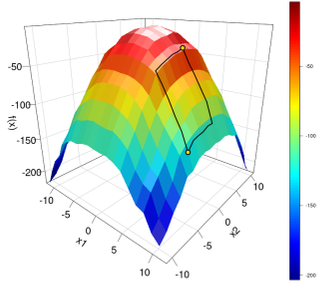
ALE:

$$f_{S,\text{ALE}}(x) = \int_{z_0}^{x} \mathbb{E}_{\mathbf{x}_{-S}|x_S=z_S}\left(\frac{\partial \hat{f}(z_S, \mathbf{x}_{-S})}{\partial z_S}\right) dz - \int \tilde{f}_{S,\text{ALE}}(x_S)\, d\mathbb{P}(x_S)$$

- Recall: PD directly averages predictions over marginal distribution of $\mathbf{x}_{-S}$
- ALE is faster: $O(2 \cdot n)$ model calls vs. $O(n \cdot g)$ for PD with g grid points
- Difference 1: ALE averages the
  - prediction changes (via partial derivatives, estimated by finite differences)
  - over conditional distribution $\mathbb{P}(\mathbf{x}_{-S}|x_S = z_S)$
- Difference 2: ALE integrates these partial deriv. over $z_S \in [z_0, x] \subseteq \mathcal{X}_S$
  $\rightsquigarrow$ isolates effect of x_S and removes main effect of other dependent feat.
- Difference 3: ALE is centered so that $\mathbb{E}_{x_S}\left(f_{S,ALE}(x)\right) = 0$

Interpretable Machine Learning

Marginal Effects

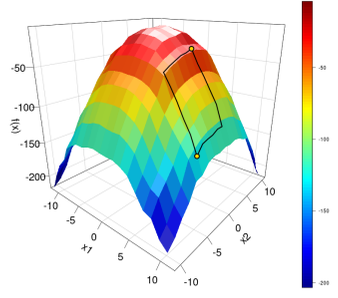


Learning goals

- Why parameter-based interpretations are not always possible for parametric models
- How marginal effects can be used in such cases
- Drawbacks of marginal effects
- Model-agnostic applicability

Interpretable Machine Learning

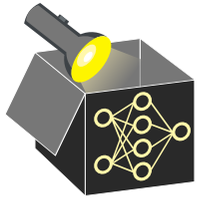
Feature Effects
Marginal Effects



Learning goals

- Why parameter-based interpretations are not always possible for parametric models
- How marginal effects can be used in such cases
- Drawbacks of marginal effects
- Model-agnostic applicability

INTERPRETATION OF SIMPLE MODELS

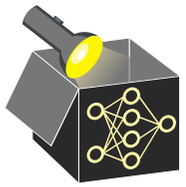


- **Linear Models**:
  - Change in x_j by Δx_j results in change in y by $\Delta y = \Delta x_j \cdot \theta_j$
  - Model equation:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

  - Default interpretations correspond to $\Delta x_j = 1$, i.e., $\Delta y = \theta_j$
  - Assumes "ceteris paribus" (all other features held constant)

INTERPRETATION OF SIMPLE MODELS



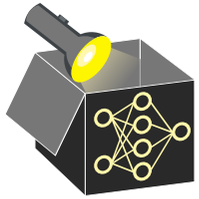
- **Linear Models**:
  - Change in x_j by Δx_j results in change in y by $\Delta y = \Delta x_j \cdot \theta_j$
  - Model equation:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

  - Default interpretations correspond to $\Delta x_j = 1$, i.e., $\Delta y = \theta_j$
  - Assumes "ceteris paribus" (all other features held constant)

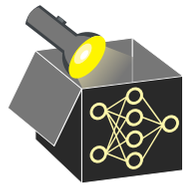
Interpretable Machine Learning – 1 / 9

INTERPRETATION OF SIMPLE MODELS



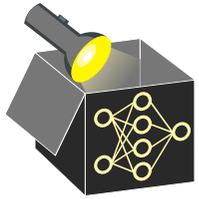
- **Linear Models**:
  - Change in x_j by Δx_j results in change in y by $\Delta y = \Delta x_j \cdot \theta_j$
  - Model equation:
$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$
  - Default interpretations correspond to $\Delta x_j = 1$, i.e., $\Delta y = \theta_j$
  - Assumes "ceteris paribus" (all other features held constant)
- **Non-Linear Models with Interactions**:
  - For models with higher-order or interaction terms, single coefficients are not sufficient:
$$y = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2 + \theta_{1,2} x_1 x_2 + \epsilon$$
  - Marginal effect of x_1 varies with different values of x_2 (and vice versa)
  - Interactions depend on the values of other features

INTERPRETATION OF SIMPLE MODELS



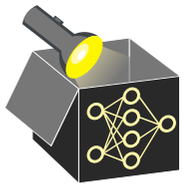
- **Linear Models**:
  - Change in x_j by Δx_j results in change in y by $\Delta y = \Delta x_j \cdot \theta_j$
  - Model equation:
$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$
  - Default interpretations correspond to $\Delta x_j = 1$, i.e., $\Delta y = \theta_j$
  - Assumes "ceteris paribus" (all other features held constant)
- **Non-Linear Models with Interactions**:
  - For models with higher-order or interaction terms, single coefficients are not sufficient:
$$y = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2 + \theta_{1,2} x_1 x_2 + \epsilon$$
  - Marginal effect of x_1 varies with different values of x_2 (and vice versa)
  - Interactions depend on the values of other features

MARGINAL EFFECTS (ME) ▸ Bartus, 2005 ▸ Scholbeck, 2024



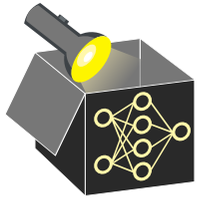
- MEs measure changes in predictions due to changes in *one/several* features.
- **How to compute it?**
  1. **Derivative Marginal Effects (dMEs)**: *numeric derivative* (slope of tangent)
     $\rightsquigarrow$ needs differentiability, fails for step-wise models.
  2. **Forward Marginal Effects (fMEs)**: *forward difference* $\hat{f}(\boldsymbol{x} + \boldsymbol{h}) - \hat{f}(\boldsymbol{x})$
     $\rightsquigarrow$ works for *any* model, any feature type.
- **Caveat**: dMEs can mislead whenever the prediction surface is non-smooth (e.g., decision trees); fMEs remain well-defined (due to finite differences).

MARGINAL EFFECTS (ME) ▸ BARTUS_2005 ▸ SCHOLBECK_2024



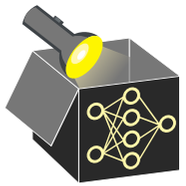
- MEs measures prediction changes due to varying *one/several* features.
- **How to compute it?**
  1. **Derivative MEs (dMEs)**: *numeric deriv.* (slope of tangent)
     $\rightsquigarrow$ needs differentiability, fails for step-wise models.
  2. **Forward MEs (fMEs)**: *forward difference* $\hat{f}(\boldsymbol{x} + \boldsymbol{h}) - \hat{f}(\boldsymbol{x})$
     $\rightsquigarrow$ works for *any* model, any feature type.
- **Caveat**: dMEs can mislead when the prediction surface is non-smooth (e.g., decision trees); fMEs remain well-defined (due to finite differences).

MARGINAL EFFECTS (ME) ▸ Bartus, 2005 ▸ Scholbeck, 2024



- MEs measure changes in predictions due to changes in *one/several* features.
- **How to compute it?**
  1. **Derivative Marginal Effects (dMEs)**: *numeric derivative* (slope of tangent)
     $\rightsquigarrow$ needs differentiability, fails for step-wise models.
  2. **Forward Marginal Effects (fMEs)**: *forward difference* $\hat{f}(\boldsymbol{x} + \boldsymbol{h}) - \hat{f}(\boldsymbol{x})$
     $\rightsquigarrow$ works for *any* model, any feature type.
- **Caveat**: dMEs can mislead whenever the prediction surface is non-smooth (e.g., decision trees); fMEs remain well-defined (due to finite differences).
- **Local instantiations (one number per data point)**
  - **ME** (at observed point $\boldsymbol{x}^{(i)}$): Individual, observation-specific "what-if" effect.
  - **MEM** (at mean $\bar{\boldsymbol{x}}$): Effect at artificial profile ("average obs.").
  - **MER** (at representative value $\boldsymbol{x}^*$): Effect at a user-defined profile.
- **Global summary – Average Marginal Effect (AME)**:
  Expectation of the (d/f)MEs; captures the *global overall* effect.

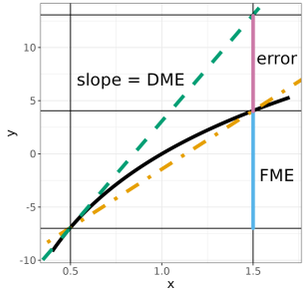
MARGINAL EFFECTS (ME) ▸ BARTUS_2005 ▸ SCHOLBECK_2024



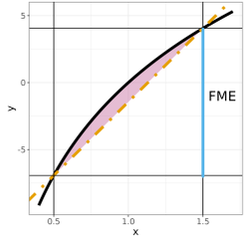
- MEs measures prediction changes due to varying *one/several* features.
- **How to compute it?**
  1. **Derivative MEs (dMEs)**: *numeric deriv.* (slope of tangent)
     $\rightsquigarrow$ needs differentiability, fails for step-wise models.
  2. **Forward MEs (fMEs)**: *forward difference* $\hat{f}(\boldsymbol{x} + \boldsymbol{h}) - \hat{f}(\boldsymbol{x})$
     $\rightsquigarrow$ works for *any* model, any feature type.
- **Caveat**: dMEs can mislead when the prediction surface is non-smooth (e.g., decision trees); fMEs remain well-defined (due to finite differences).
- **Local instantiations (one number per data point)**
  - **ME** (at observed point $\boldsymbol{x}^{(i)}$): Individual, obs.-specific "what-if" effect.
  - **MEM** (at mean $\bar{\boldsymbol{x}}$): Effect at artificial profile ("average obs.").
  - **MER** (at representative value $\boldsymbol{x}^*$): Effect at a user-defined profile.
- **Global summary – Average Marginal Effect (AME)**:
  Expectation of the (d/f)MEs; captures the *global overall* effect.

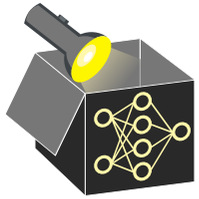
DERIVATIVE VS. FORWARD DIFFERENCE

- **dME (tangent, green)**
  - slope of the tangent at x;
  - delivers a *rate* of change $\frac{\partial \widehat{f}}{\partial x}$.
- **fME (secant, orange)**
  - vertical gap between two model evaluations;
  - always *exact* change in predicted outcome.
  - Non-linearity measure (pink band, bottom): quantifies deviation of secant and true curve
- **When the two differ**
  - Curvature makes the tangent overshoot or undershoot ⇒ dME may be badly biased.
  - fME is robust to kinks, plateaus, trees, . . .



black = non-lin. function
blue = fME; pink = dME error



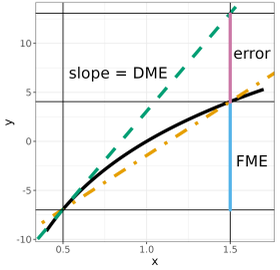


DERIVATIVE VS. FORWARD DIFFERENCE

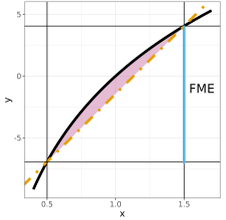
dME (tangent, green)
- slope of the tangent at x;
- delivers a *rate* of change $\frac{\partial \widehat{f}}{\partial x}$.

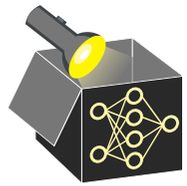
fME (secant, orange)
- vertical gap between two model evaluations;
- always *exact* change in predicted outcome.
- Non-linearity measure (pink band, bottom): quantifies deviation of secant and true curve

When the two differ
- Curvature makes the tangent overshoot or undershoot ⇒ dME may be badly biased.
- fME is robust to kinks, plateaus, trees, . . .



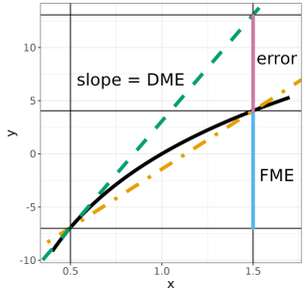
black = non-lin. function
blue = fME; pink = dME error



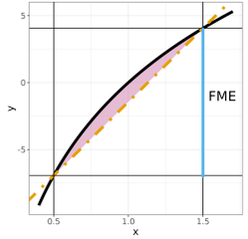


DERIVATIVE VS. FORWARD DIFFERENCE

- **dME (tangent, green)**
  - slope of the tangent at x;
  - delivers a *rate* of change $\frac{\partial \widehat{f}}{\partial x}$.
- **fME (secant, orange)**
  - vertical gap between two model evaluations;
  - always *exact* change in predicted outcome.
  - Non-linearity measure (pink band, bottom): quantifies deviation of secant and true curve
- **When the two differ**
  - Curvature makes the tangent overshoot or undershoot $\Rightarrow$ dME may be badly biased.
  - fME is robust to kinks, plateaus, trees, ...
- Use fME for any non-linear or non-smooth model
- Use dME for lin. functions or analytic convenience



black = non-lin. function
blue = fME; pink = dME error



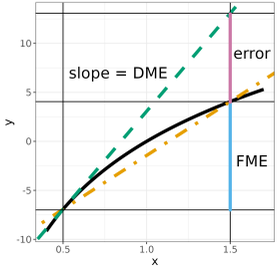
DERIVATIVE VS. FORWARD DIFFERENCE

dME (tangent, green)
- slope of the tangent at x;
- delivers a *rate* of change $\frac{\partial \widehat{f}}{\partial x}$.

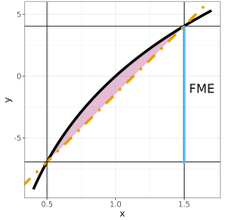
fME (secant, orange)
- vertical gap between two model evaluations;
- always *exact* change in predicted outcome.
- Non-linearity measure (pink band, bottom): quantifies deviation of secant and true curve

When the two differ
- Curvature makes the tangent overshoot or undershoot $\Rightarrow$ dME may be badly biased.
- fME is robust to kinks, plateaus, trees, ...

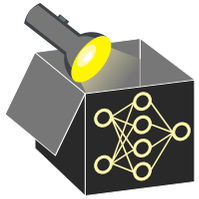
Recommendations
- Use fME for any non-linear / non-smooth model
- Use dME for lin. func.-s or analytic convenience



black = non-lin. function
blue = fME; pink = dME error



MARGINAL EFFECTS FOR CONTINUOUS FEATURES



- **Derivative Marginal Effect (dME)**:

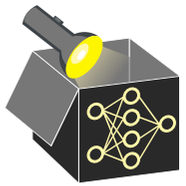
$$\text{dME}_j(\mathbf{x}) = \frac{\partial \hat{f}(\mathbf{x})}{\partial x_j} \approx \frac{\hat{f}(x_1, \ldots, x_j + h_j, \ldots, x_p) - \hat{f}(x_1, \ldots, x_j - h_j, \ldots, x_p)}{2h_j}$$

- **Forward Marginal Effect (fME)**:

$$\text{fME}_j(\mathbf{x}, h_j) = \hat{f}(x_1, \ldots, x_j + h_j, \ldots, x_p) - \hat{f}(\mathbf{x})$$

- **Note:** fME is not scale-invariant – halving the step size does not halve the effect.
- **Additive Recovery:** dME and fME isolate terms involving the target feature.
  - **Example:** For $\widehat{f}(\mathbf{x}) = ax_1 + bx_2$: $\text{dME}_1(\mathbf{x}) = a, \quad \text{fME}_1(\mathbf{x}, h_1) = ah_1$
  - Effects from additively linked features (e.g., x_2) are canceled.
  - Enables focus on direct feature-specific influence in $\widehat{f}$.

ME FOR CONTINUOUS FEATURES



- **Derivative Marginal Effect (dME)**:

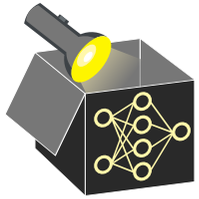
$$\text{dME}_j(\mathbf{x}) = \frac{\partial \hat{f}(\mathbf{x})}{\partial x_j} \approx \frac{\hat{f}(x_1, \ldots, x_j + h_j, \ldots, x_p) - \hat{f}(x_1, \ldots, x_j - h_j, \ldots, x_p)}{2h_j}$$

- **Forward Marginal Effect (fME)**:

$$\text{fME}_j(\mathbf{x}, h_j) = \hat{f}(x_1, \ldots, x_j + h_j, \ldots, x_p) - \hat{f}(\mathbf{x})$$

- **Note:** fME is not scale-invariant – halving the step size does not halve the effect.
- **Additive Recovery:** dME and fME isolate terms involving the target feat.
  - **Example:** For $\widehat{f}(\mathbf{x}) = ax_1 + bx_2$: $\text{dME}_1(\mathbf{x}) = a, \quad \text{fME}_1(\mathbf{x}, h_1) = ah_1$
  - Effects from additively linked features (e.g., x_2) are canceled.
  - Enables focus on direct feature-specific influence in $\widehat{f}$.

MARGINAL EFFECTS FOR CATEGORICAL FEATURES

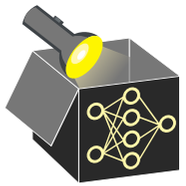


- **Traditional Approach**:
  - Choose a baseline category for the categorical feature x_j
    $\rightsquigarrow$ Either the observed value x_j or a fixed reference x_j^{ref}
  - Replace x_j with an alternative category x_j^{new}
  - Compute the change in prediction, keeping all other features $\mathbf{x}_{-j}$ fixed
- **fME Definition for Categorical Features**:

$$\text{fME}_j(\mathbf{x}; x_j^{\text{new}}) = \widehat{f}(x_j^{\text{new}}, \mathbf{x}_{-j}) - \widehat{f}(x_j, \mathbf{x}_{-j})$$

  - x_j: original category of feature j in obs. $\mathbf{x}$ (or reference category x_j^{ref})
  - x_j^{new}: new category to evaluate
  - $\mathbf{x}_{-j}$: all other features held fixed
- **Advantages**:
  - Mirrors continuous feature fME: measures discrete change in prediction.
  - Any level can act as baseline - no fixed reference needed.

ME FOR CATEGORICAL FEATURES

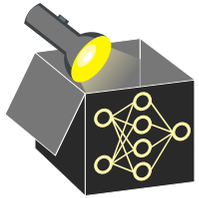


- **Traditional Approach**:
  - Choose a baseline category for the categorical feature x_j
    $\rightsquigarrow$ Either the observed value x_j or a fixed reference x_j^{ref}
  - Replace x_j with an alternative category x_j^{new}
  - Compute the change in prediction, keeping all other feat. $\mathbf{x}_{-j}$ fixed
- **fME Definition for Categorical Features**:

$$\text{fME}_j(\mathbf{x}; x_j^{\text{new}}) = \widehat{f}(x_j^{\text{new}}, \mathbf{x}_{-j}) - \widehat{f}(x_j, \mathbf{x}_{-j})$$

  - x_j: original category of feature j in obs. $\mathbf{x}$ (or reference category x_j^{ref})
  - x_j^{new}: new category to evaluate
  - $\mathbf{x}_{-j}$: all other features held fixed
- **Advantages**:
  - Mirrors continuous feature fME: measures discrete change in pred.
  - Any level can act as baseline - no fixed reference needed.

AVERAGE MARGINAL EFFECTS



Definition (based on fMEs with step h_S, can also be based on dMEs):

$$\text{AME}_S = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{f}(\boldsymbol{x}_S^{(i)} + \boldsymbol{h}_S, \boldsymbol{x}_{-S}^{(i)}) - \hat{f}(\boldsymbol{x}^{(i)}) \right]$$

Why they work in GLMs:

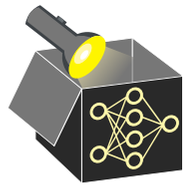
- Link function is monotonic $\Rightarrow$ direction of effect stable.
- Averaging gives sensible results (e.g., logit, probit).

Why they fail on non-parametric models:

- AMEs assume a consistent effect across the feature space.
- Non-parametric models can model complex, non-linear relationships.
- Averaging effects can obscure important heterogeneities.

Takeaway: AMEs can be useful summaries for smooth, monotonic models. For black-boxes, use **local fMEs** and support them with a non-linearity measure.

AVERAGE MARGINAL EFFECTS



Definition (based on fMEs with step h_S, can also be based on dMEs):

$$\text{AME}_S = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{f}(\boldsymbol{x}_S^{(i)} + \boldsymbol{h}_S, \boldsymbol{x}_{-S}^{(i)}) - \hat{f}(\boldsymbol{x}^{(i)}) \right]$$

Why they work in GLMs:

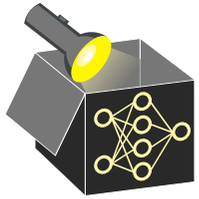
- Link function is monotonic $\Rightarrow$ direction of effect stable.
- Averaging gives sensible results (e.g., logit, probit).

Why they fail on non-parametric models:

- AMEs assume a consistent effect across the feature space.
- Non-parametric models can model complex, non-linear relationships.
- Averaging effects can obscure important heterogeneities.

Takeaway: AMEs can be useful summaries for smooth, monotonic models.
For black-boxes, use **local fMEs** and support them with non-linearity measure.

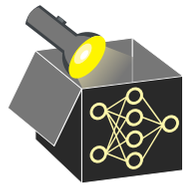
WHY MARGINAL EFFECTS *STILL* MATTER



- **Single, formal number:** One scalar per observation; can be averaged (AME), reported with CIs, audited, stored easily.
- **Multivariate changes** Simultaneously perturb multiple *continuous/categorical* features. Still yields a scalar (unlike PD/ICE, which require multivariate plots).
- **Model-faithful, assumption-light** Measured at the *actual data point*. Captures interactions, no independence or surrogate-model assumptions (LIME).
- **Non-Linearity Measure:** Quantifies how well local linear approximation holds (e.g., via a normalized squared deviation from the secant).
  $\rightsquigarrow$ Local reliability measure, something PD/ICE plots cannot quantify.
- **Computationally cheap** Just two forward passes (or $k-1$ for a k-level factor) per observation vs. grid$\times n$ for PD/ICE.

Conclusion: Plots let you *see* the landscape; ME give numbers you can *use*.

WHY MARGINAL EFFECTS *STILL* MATTER



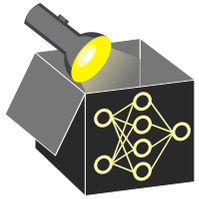
- **Single, formal number:** One scalar per observation; can be averaged (AME), reported with CIs, audited, stored easily.
- **Multivariate changes** Simultaneously perturb multiple *continuous/categ.* feat. Still yields a scalar (unlike PD/ICE, which require multivar. plots).
- **Model-faithful, assumption-light** Measured at the *actual data point*. Captures interactions, no indep. or surrogate-model assumptions (LIME).
- **Non-Linearity Measure:** Quantifies how well local linear approximation holds (e.g., via a normalized squared deviation from the secant).
  $\rightsquigarrow$ Local reliability measure, something PD/ICE plots cannot quantify.
- **Computationally cheap** Just two forward passes (or $k-1$ for a k-level factor) per observation vs. grid$\times n$ for PD/ICE.

Conclusion:

Plots let you *see* the landscape; ME give numbers you can *use*.

Interpretable Machine Learning – 7 / 9

USE-CASE: SCALAR VS. VISUAL ESTIMATION



Setting: A clinical model predicts heart attack risk from patient features, e.g., x_1 : systolic blood pressure (BP), x_2 : LDL cholesterol, x_3 : age, ...

Clinician's questions

- *"What if this patient's systolic BP increases by 10 mmHg?"*
- *"What if BP increases by 10 mmHg & LDL by 15 mg/dL?"*

Route A – ICE / PD

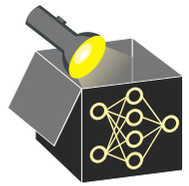
- Plot prediction as a function of BP (1-D) or BP+LDL (2-D) on a grid.
- Manual interpretation of change by looking at curve/surface.

$\rightarrow$ Visual and local; limited to 1–2 features at a time.

Route B – Forward Marginal Effect: $\mathrm{fME} = \widehat{f}(\mathbf{x} + \mathbf{h}) - \widehat{f}(\mathbf{x})$

- **1-D case: h** $= (10, 0, 0, \dots)$ $\Rightarrow$ risk increases by **+3 percentage points**
- **2-D case: h** $= (10, 15, 0, \dots)$ $\Rightarrow$ risk increases by **+4.1 percentage points**
- One scalar answer per query, extensible to higher dimensions.

USE-CASE: SCALAR VS. VISUAL ESTIMATION



Setting: A clinical model predicts heart attack risk from patient features, e.g., x_1 : systolic blood pressure (BP), x_2 : LDL cholesterol, x_3 : age, ...

Clinician's questions

- *"What if this patient's systolic BP increases by 10 mmHg?"*
- *"What if BP increases by 10 mmHg & LDL by 15 mg/dL?"*

Route A – ICE / PD

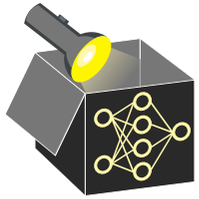
- Plot prediction as a function of BP (1-D) or BP+LDL (2-D) on a grid.
- Manual interpretation of change by looking at curve/surface.

$\rightarrow$ Visual and local; limited to 1–2 features at a time.

Route B – Forward Marginal Effect: $\mathrm{fME} = \widehat{f}(\mathbf{x} + \mathbf{h}) - \widehat{f}(\mathbf{x})$

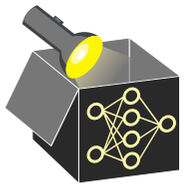
- **1-D case: h** $= (10, 0, 0, \dots)$ $\Rightarrow$ risk increases by **+3 % points**
- **2-D case: h** $= (10, 15, 0, \dots)$ $\Rightarrow$ risk increases by **+4.1 % points**
- One scalar answer per query, extensible to higher dimensions.

RELATION TO ICE AND PD



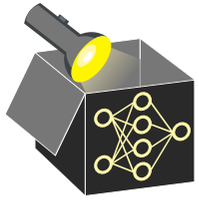
- **Individual Conditional Expectation (ICE)**:
  - Visualizes predictions for an observation across a range of feature values.
  - fME corresponds to vertical differences between points on an ICE curve.
- **Partial Dependence (PD)**:
  - Shows average predictions across a range of feature values.
  - AME is equivalent to vertical differences on PD for linear models.
- **Advantages of fMEs**:
  - Provide exact change in prediction.
  - Applicable to high-dimensional feature changes.
  - Quantifiable and not limited to visual interpretation.

RELATION TO ICE AND PD

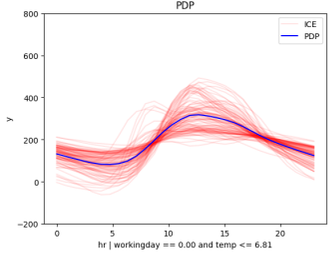


- **Individual Conditional Expectation (ICE)**:
  - Visualizes predictions for an obs. across a range of feature values.
  - fME corresponds to vertical diff. between points on an ICE curve.
- **Partial Dependence (PD)**:
  - Shows average predictions across a range of feature values.
  - AME is equivalent to vertical differences on PD for linear models.
- **Advantages of fMEs**:
  - Provide exact change in prediction.
  - Applicable to high-dimensional feature changes.
  - Quantifiable and not limited to visual interpretation.

Interpretable Machine Learning



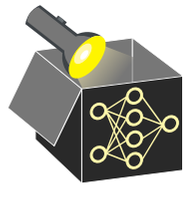
Feature effects: Further Resources and Software



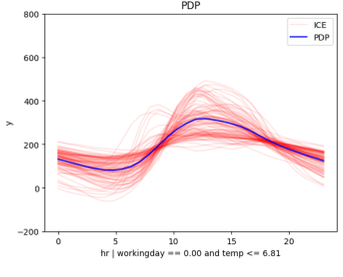
Learning goals

- Learn about further resources
- Get an overview of software packages in R and Python

Interpretable Machine Learning



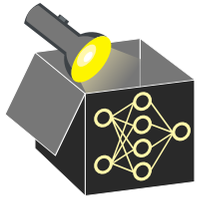
Feature Effects
Further Resources and Software



Learning goals

- Learn about further resources
- Get an overview of software packages in R and Python

RESOURCES AND SOFTWARE:



Individual Conditional Expectation (ICE)

- The paper: ▸ Goldstein et al. 2015
- Chapter in Interpretable Machine Learning book: ▸ Molnar 2025
- R packages: ▸ iml; ▸ pdp; ▸ ICEbox
- Python packages: ▸ PiML; ▸ sklearn

Partial Dependence (PD) plot

- Chapter in Interpretable Machine Learning book: ▸ Molnar 2025
- R packages: ▸ iml; ▸ pdp; ▸ DALEX
- Python packages: ▸ sklearn; ▸ effector; ▸ PDPbox

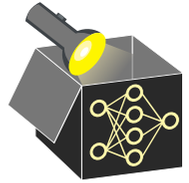
Accumulated Local Effect (ALE)

- The paper: ▸ Apley et al. 2020
- Chapter in Interpretable Machine Learning book: ▸ Molnar 2025
- R packages: ▸ ALEPlot; ▸ iml
- Python packages: ▸ effector; ▸ PiML; ▸ ALEPython; ▸ Alibi

Regional effect plots (REPID)

- The paper: ▸ Herbinger et al. (2022)
- Codebase: ▸ repid

RESOURCES AND SOFTWARE:



Individual Conditional Expectation (ICE)

- The paper: ▸ Goldstein 2015
- Chapter in Interpretable Machine Learning book: ▸ Molnar 2025
- R packages: ▸ iml n.d.; ▸ pdp n.d.; ▸ ICEbox n.d.
- Python packages: ▸ PiML n.d.; ▸ sklearn n.d.

Partial Dependence (PD) plot

- Chapter in Interpretable Machine Learning book: ▸ Molnar 2025
- R packages: ▸ iml n.d.; ▸ pdp n.d.; ▸ DALEX n.d.
- Python packages: ▸ sklearn n.d.; ▸ effector n.d.; ▸ PDPbox n.d.

Accumulated Local Effect (ALE)

- The paper: ▸ Apley 2020
- Chapter in Interpretable Machine Learning book: ▸ Molnar 2025
- R packages: ▸ ALEPlot n.d.; ▸ iml n.d.
- Python packages: ▸ effector n.d.; ▸ PiML n.d.; ▸ ALEPython n.d.; ▸ Alibi n.d.

Regional effect plots (REPID)

- The paper: ▸ Herbinger 2022
- Codebase: ▸ repid n.d.