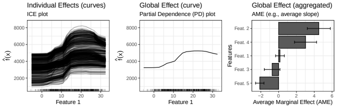
Interpretable Machine Learning

Feature Effects

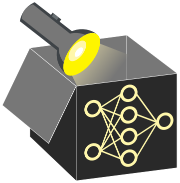
Individual Conditional Expectation (ICE) Plot



Learning goals

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

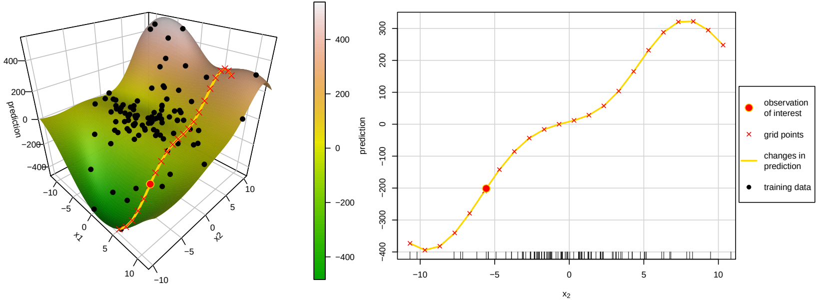
MOTIVATION



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

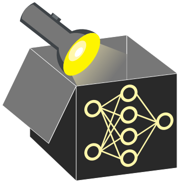
Idea: For a given observation, change the value of the feature of interest, and visualize how prediction changes

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



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

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

$\rightsquigarrow$ 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 $\mathbf{x}_S^* = \mathbf{x}_S^{*^{(1)}}, \dots, \mathbf{x}_S^{*^{(g)}}$ 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

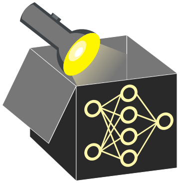
ICE CURVES - ILLUSTRATION

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

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	4	7

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
2	4	7

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$
3	4	7



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)

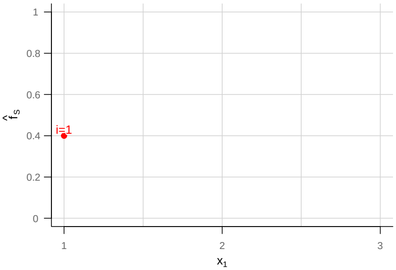
ICE CURVES - ILLUSTRATION

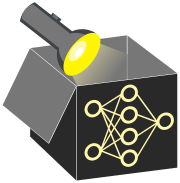
i	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$
1	1	4	7
2	2	5	8
3	3	6	9

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

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

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





2. Step - Predict and visualize:

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

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

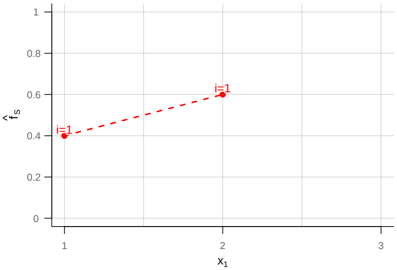
ICE CURVES - ILLUSTRATION

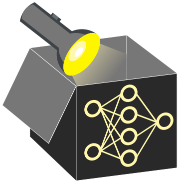
i	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$
1	1	4	7
2	2	5	8
3	3	6	9

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

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

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



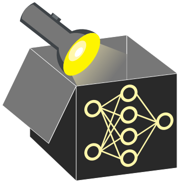


2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION

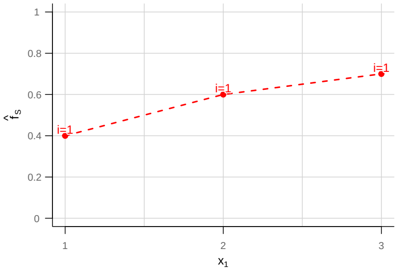


i	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	1	4	7
2	2	5	8
3	3	6	9

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
1	4	7	0.4

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
2	4	7	0.6

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
3	4	7	0.7

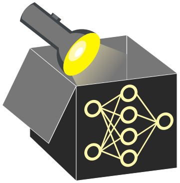


2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION

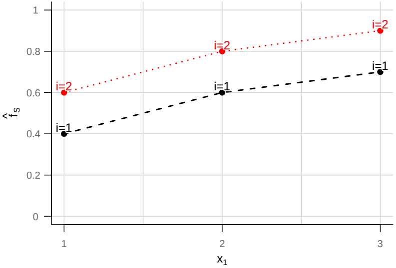


i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

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

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

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



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

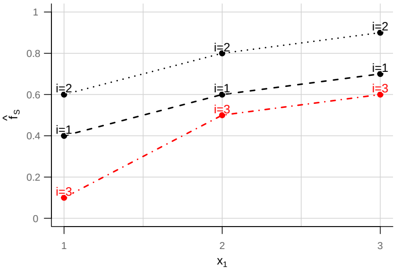
ICE CURVES - ILLUSTRATION

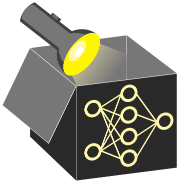
i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

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

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

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

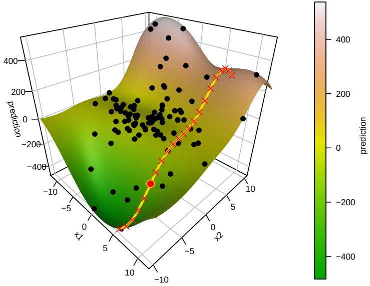


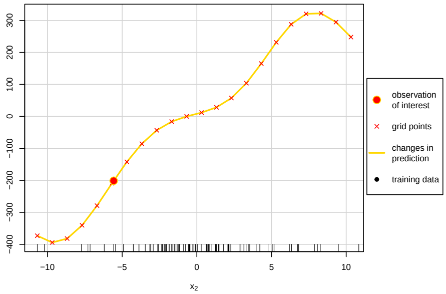


3. Step - Repeat for other observations:
ICE curve for $i = 3$ connects all predictions at grid values associated to the i-th observation.

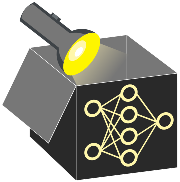
ICE CURVES - INTERPRETATION

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**

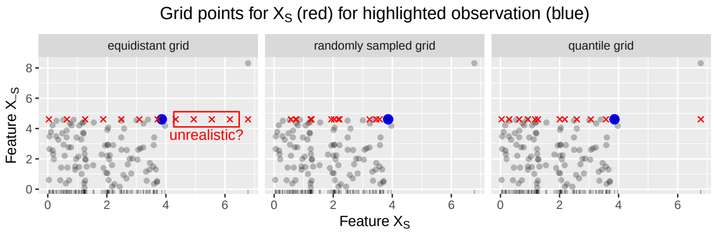




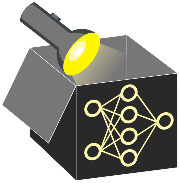
COMMENTS ON GRID VALUES



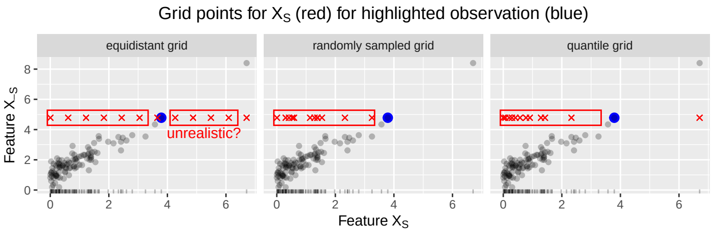
- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$; shown on x-axis
- **Three common strategies** for grid definition:
  - Equidistant grid values within feature range
  - Random samples from observed feature values
  - Quantiles of observed feature values
- **Marginal realism:**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



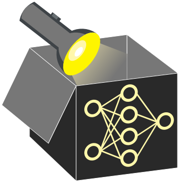
COMMENTS ON GRID VALUES



- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$; 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)

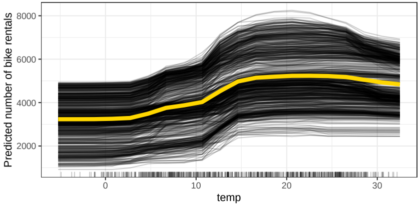
- Too coarse $\Rightarrow$ may miss sharp nonlinearities or discontinuities
- Too fine $\Rightarrow$ high runtime (without gaining much)
- Fix: cap at $\approx 50 - 100$ grid points; vectorize predictions by feeding the model a single data frame containing all grid-modified instances

ICE curves (number of instances/curves visualized)

- Too few $\Rightarrow$ hides 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**)