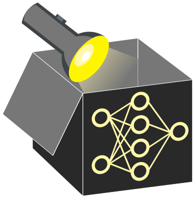
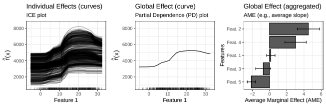
Interpretable Machine Learning



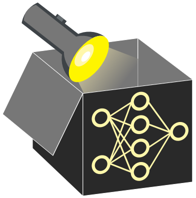
Individual Conditional Expectation (ICE) Plot



Learning goals

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

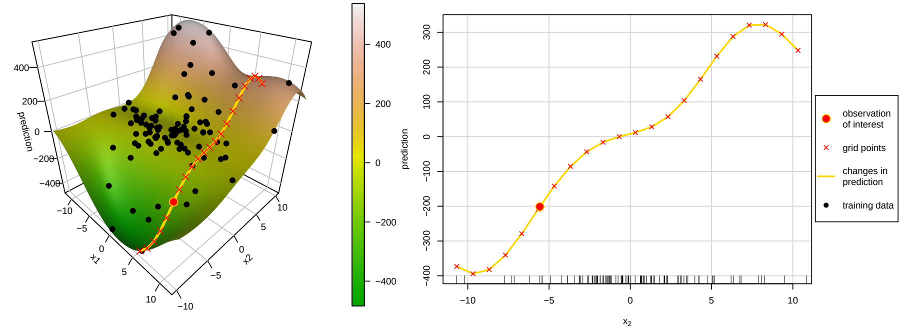
MOTIVATION



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

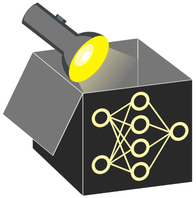
Idea: Change values of observation and feature of interest, and visualize how prediction changes

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



INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

▸ Goldstein et. al (2013)



$\underbrace{\mathbf{x}_S}$ $\underbrace{\mathbf{x}_{-S}}$

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)

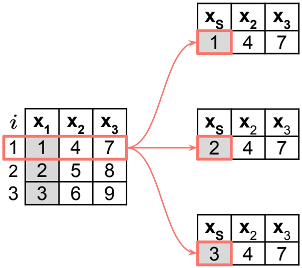
⇝ In practice, $\mathbf{x}_S$ consists of one or two features (i.e., $|S| \leq 2$ and $-S = S^{\complement}$).

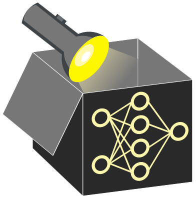
Formal definition of ICE curves:

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

⇝ 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 $\mathbf{x}_S^{*^{(1)}}, \dots, \mathbf{x}_S^{*^{(g)}}$ along feature of interest $\mathbf{x}_S$ and replace vector $\mathbf{x}^{(i)}$ in data with grid
$\Rightarrow$ Creates new artificial points for i-th observation (here: $\mathbf{x}_S^* = x_1^* \in \{1, 2, 3\}$ scalar)

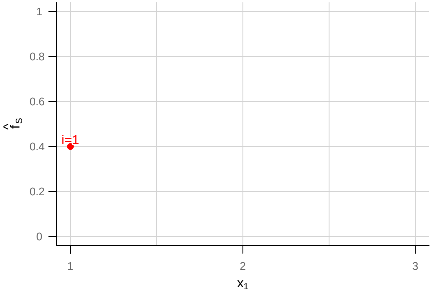
ICE CURVES - ILLUSTRATION

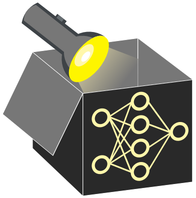
i	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	1	4	7
2	2	5	8
3	3	6	9

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
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 $\mathbf{x}^*_S$:

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

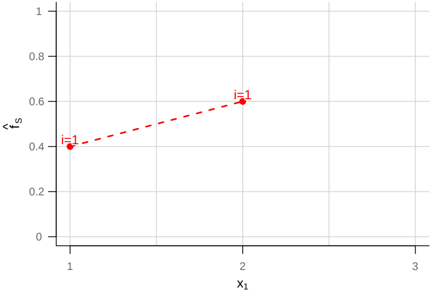
ICE CURVES - ILLUSTRATION

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

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

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

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



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

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

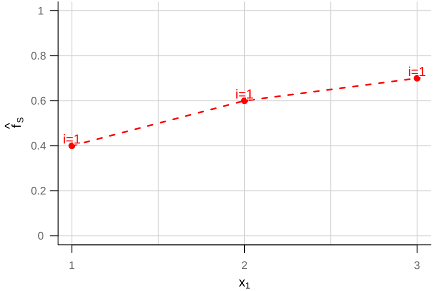
ICE CURVES - ILLUSTRATION

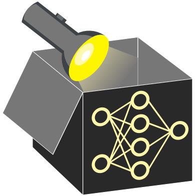
i	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	1	4	7
2	2	5	8
3	3	6	9

$\mathbf{x}_S$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\hat{f}$
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

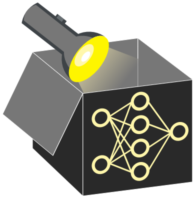




2. Step - Predict and visualize:
For each artificially created data point of i-th observation, plot prediction $\hat{f}^{(i)}_{S,ICE}(\mathbf{x}^*_S)$ vs. grid values $\mathbf{x}^*_S$:

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

ICE CURVES - ILLUSTRATION

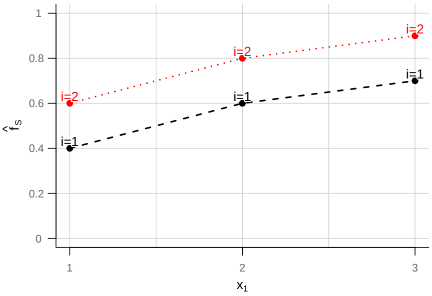


i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

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

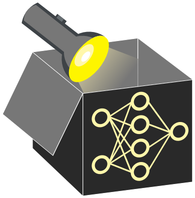
x_S	x_2	x_3	$\hat{f}$
3	4	7	0.7
3	5	8	0.9



3. Step - Repeat for other observations:

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

ICE CURVES - ILLUSTRATION

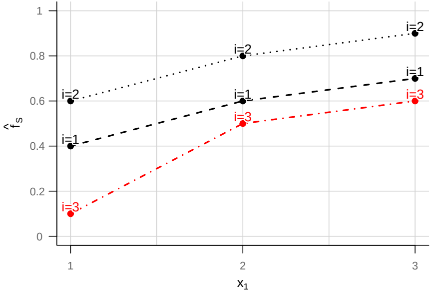


i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

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

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

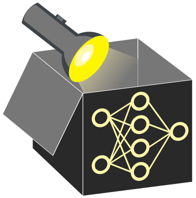
x_S	x_2	x_3	$\hat{f}$
3	4	7	0.7
3	5	8	0.9
3	6	9	0.6



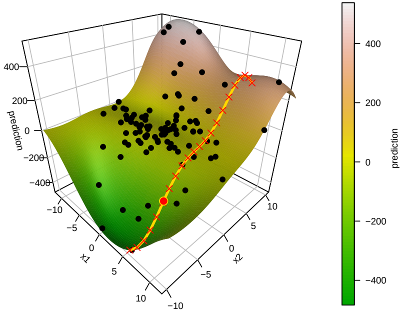
3. Step - Repeat for other observations:

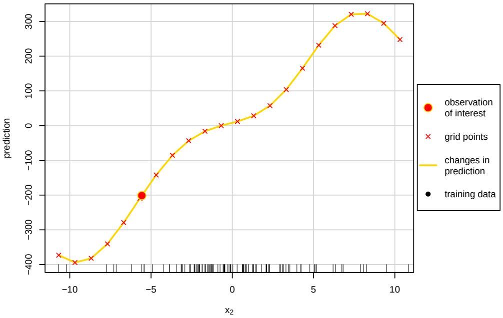
ICE curve for $i = 3$ connects all predictions at grid values associated to i-th observation.

ICE CURVES - INTERPRETATION

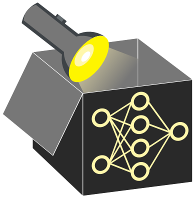


Example: Prediction surface of a model (left), select observation and visualize changes in prediction for different values of x_2 while keeping x_1 fixed
$\Rightarrow$ **local interpretation**

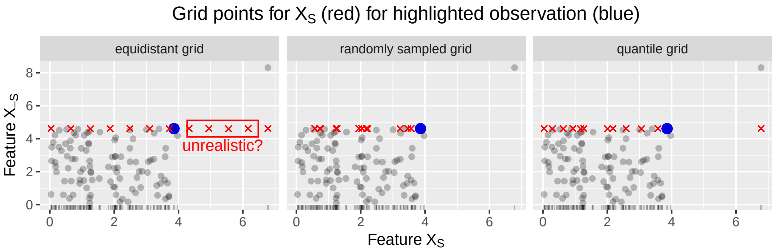




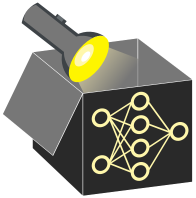
COMMENTS ON GRID VALUES



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



COMMENTS ON GRID VALUES



- Plotting ICE curves involves generating grid values $\mathbf{x}_S^*$; visualized on x-axis
- Common choices for grid values are
  - equidistant grid values within feature range
  - randomly sampled values or quantile values of observed feature values
- Except equidistant grid, the other two options preserve (approximately) the marginal distribution of feature of interest
- Correlations/interactions $\rightsquigarrow$ unrealistic values in all three methods