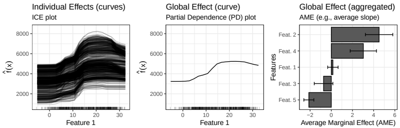
Interpretable Machine Learning

Individual Conditional Expectation (ICE) Plot



Learning goals

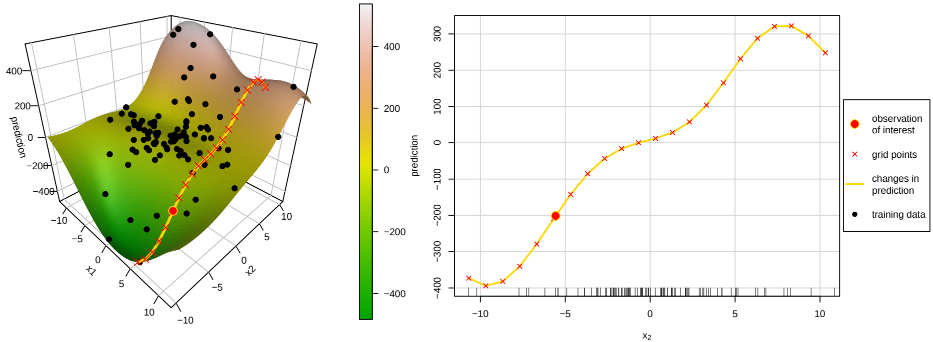
- ICE curves as local effect method
- How to sample grid points for ICE curves

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)

	$\mathbf{x}_S$	$\mathbf{x}_{-S}$	
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 $\mathbf{x}_S$ (features of interest) and remaining features $\mathbf{x}_{-S}$.

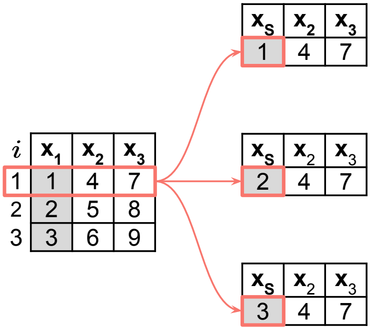
$\rightsquigarrow$ In practice, $\mathbf{x}_S$ consists of one or two features (i.e., $|S| \leq 2$ and $-S = S^\complement$).

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:

$$\hat{f}_S^{(i)}(\mathbf{x}_S^*) \ \text{ vs. } \ \mathbf{x}_S^*$$

where $\hat{f}_S^{(i)}(\mathbf{x}_S^*) = \hat{f}(\mathbf{x}_S^*, \mathbf{x}_{-S}^{(i)})$ is prediction of i-th observation in which original feature value $\mathbf{x}_S^{(i)}$ was replaced by $\mathbf{x}_S^*$.

ICE CURVES - ILLUSTRATION



1. Step - Grid points:

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

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

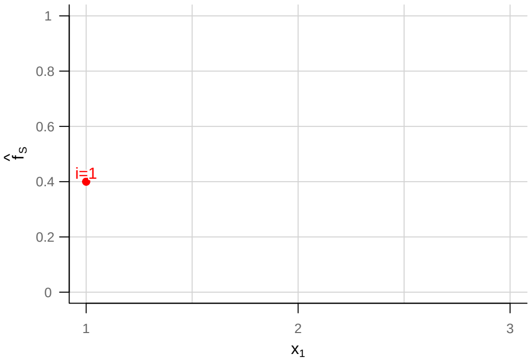
ICE CURVES - ILLUSTRATION

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

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

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

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



2. Step - Predict and visualize:

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

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

ICE CURVES - ILLUSTRATION

x_S	x_2	x_3
1	4	7

$\hat{f}$
0.4

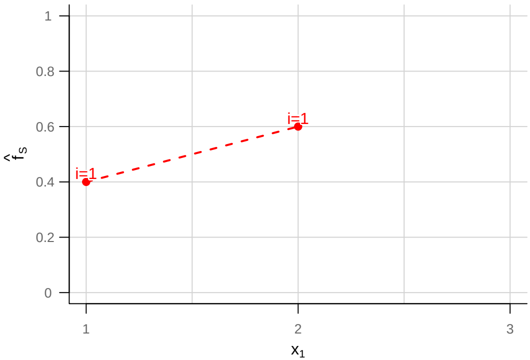
i	x_1	x_2	x_3
1	1	4	7
2	2	5	8
3	3	6	9

x_S	x_2	x_3
2	4	7

$\hat{f}$
0.6

x_S	x_2	x_3
3	4	7

$\hat{f}$
0.7



2. Step - Predict and visualize:

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

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

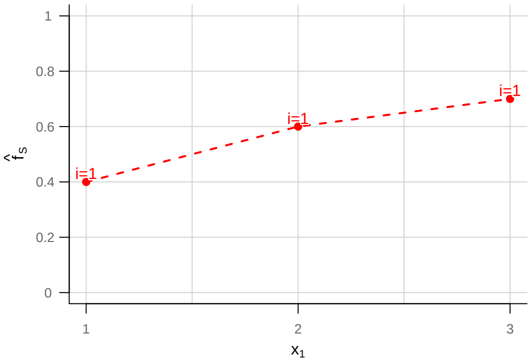
ICE CURVES - ILLUSTRATION

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

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

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

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



2. Step - Predict and visualize:

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

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

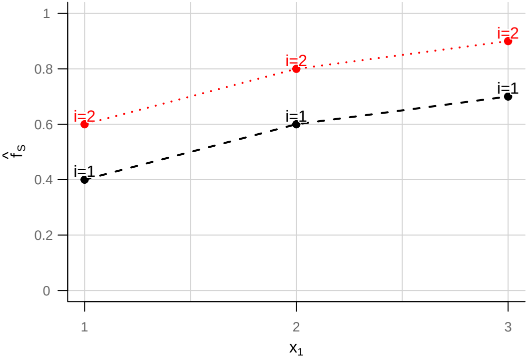
ICE CURVES - ILLUSTRATION

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

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

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

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



3. Step - Repeat for other observations:

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

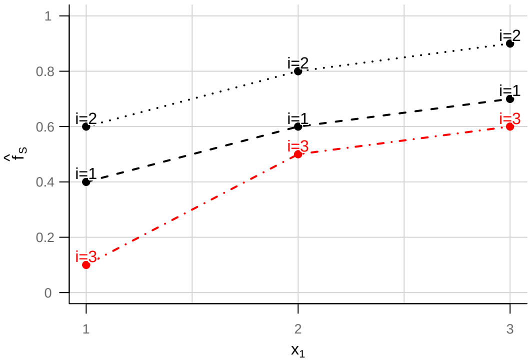
ICE CURVES - ILLUSTRATION

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

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

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

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



3. Step - Repeat for other observations:

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

COMMENTS ON GRID VALUES

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

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