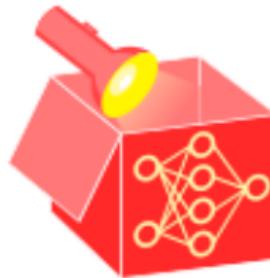
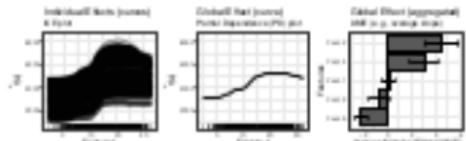


Interpretable Machine Learning



Individual Conditional Expectation (ICE) Plot

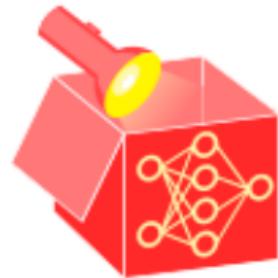


Learning goals

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

MOTIVATION

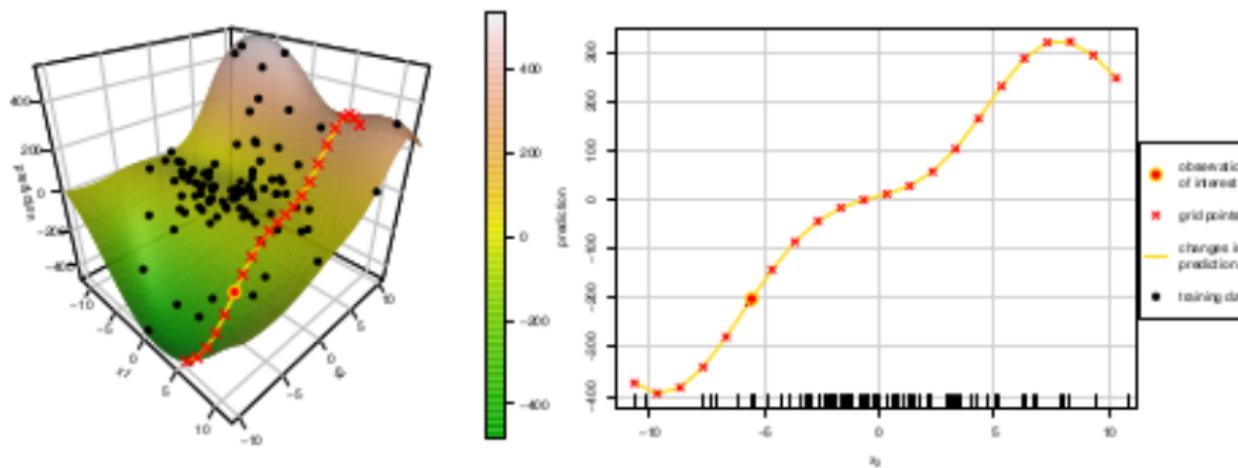
Question: How does changing values of a single feature of an observation affect model prediction?



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

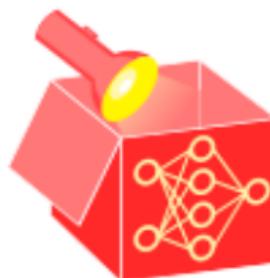
Example: Prediction surface of a model (left), select observation and visualize changes in prediction for different values of x_2 while keeping x_1 fixed

⇒ local interpretation



INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

► Goldstein et al (2013)



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

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

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

Formal definition of ICE curves:

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

- ~ 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

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

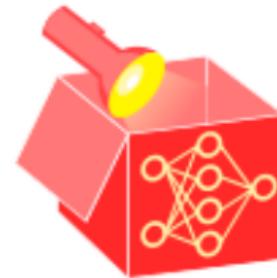
1. Step - Grid points:

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

x_1	x_2	x_3
2	4	7
3	6	9

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

x_1	x_2	x_3
3	4	7



1. Step - Grid points:

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

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

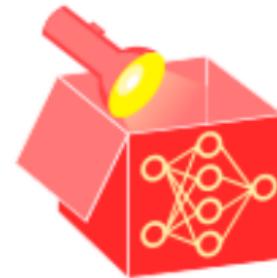
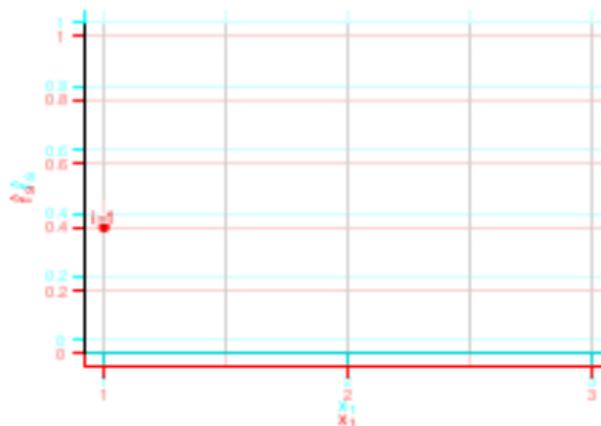
ICE CURVES - ILLUSTRATION

The diagram illustrates the construction of Individual Conditional Expectation (ICE) curves. It starts with a base model f represented by a table:

	x_1	x_2	x_3	f
1	1	4	7	0.4

Three artificial data points are then created by changing one feature at a time while keeping others constant:

i	x_1	x_2	x_3	f
1	1	4	7	0.4
2	2	5	8	0.6
3	3	6	9	0.7



2. Step - Predict and visualize:
2. Step - Predict and visualize:

For each artificially created data point of i -th observation, plot prediction $\hat{f}_i^{(0)}(x_i^*)$ vs.
For each artificially created data point of i -th observation, plot prediction $\hat{f}_{S,ICE}^{(0)}(x_S^*)$ vs.
grid values x_i^* :
grid values x_S^* :

$$\hat{f}_{1,ICE}^{(0)}(x_1^*) = \hat{f}(x_1^*, x_{2,3}^{(0)}) \text{ vs. } x_1^* \in \{1, 2, 3\}$$

ICE CURVES - ILLUSTRATION

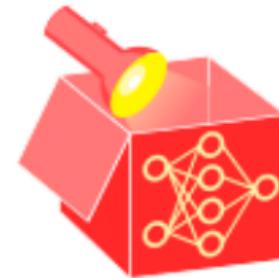
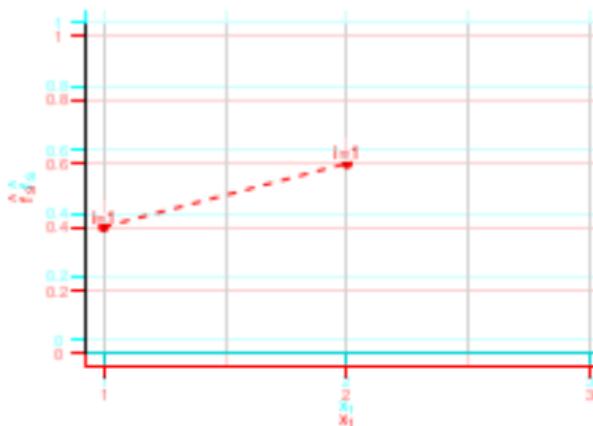
x_1	x_2	x_3	\hat{f}
1	4	7	0.4

Observation $i = 1$: $x_1 = 1, x_2 = 4, x_3 = 7$

i	x_1	x_2	x_3	\hat{f}
1	1	4	7	0.4
2	2	5	8	
3	3	6	9	

Observation $i = 2$: $x_1 = 2, x_2 = 5, x_3 = 8$

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



2. Step - Predict and visualize:

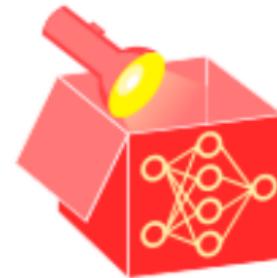
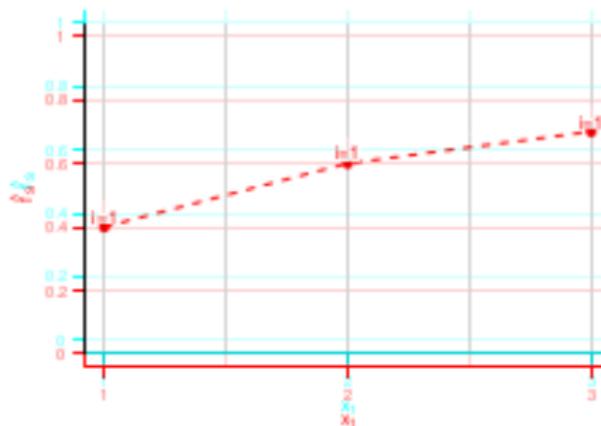
For each artificially created data point of i -th observation, plot prediction $\hat{f}_{1,ICE}^{(i)}(x_S^*)$ vs. grid values x_S^* :

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

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

ICE CURVES - ILLUSTRATION

	x_1	x_2	x_3	f^*
1	1	4	7	0.4
2	2	5	8	0.6
3	3	6	9	0.7
	x_S	x_2	x_3	f^*
	3	4	7	0.7



2. Step - Predict and visualize:

2. Step - Predict and visualize:

For each artificially created data point of i -th observation, plot prediction $\hat{f}_{1,ICE}^{(i)}(x_S^*)$ vs. grid values x_S^* :

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

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

ICE CURVES - ILLUSTRATION

The diagram illustrates the construction of three Individual Conditional Expectation (ICE) curves. It starts with a table of data points:

i	x_1	x_2	x_3	f_i
1	1	4	7	0.4
2	2	5	8	0.6
3	3	6	9	0.8

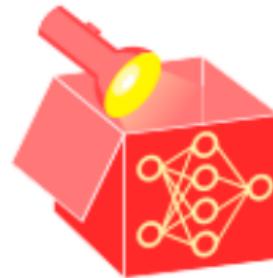
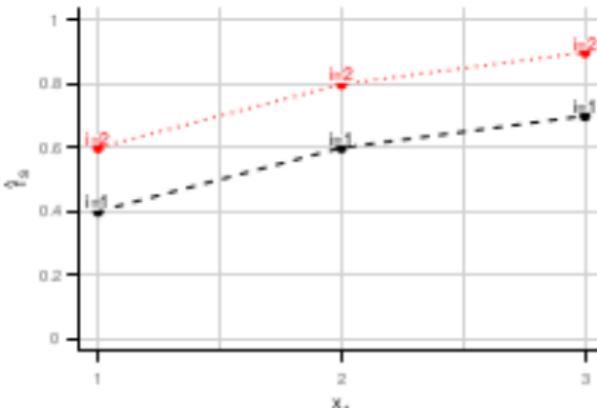
Three additional tables show the predictions for each observation at different grid values of x_1 :

- For $i=1$:

x_8	x_2	x_3	f_i
1	4	7	0.4
2	5	8	0.6
- For $i=2$:

x_8	x_2	x_3	f_i
2	4	7	0.6
2	5	8	0.8
- For $i=3$:

x_8	x_2	x_3	f_i
3	4	7	0.7
3	5	8	0.9



3. Step - Repeat for other observations:

3. Step - Repeat for other observations:

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

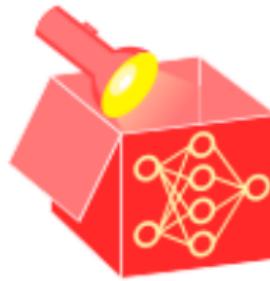
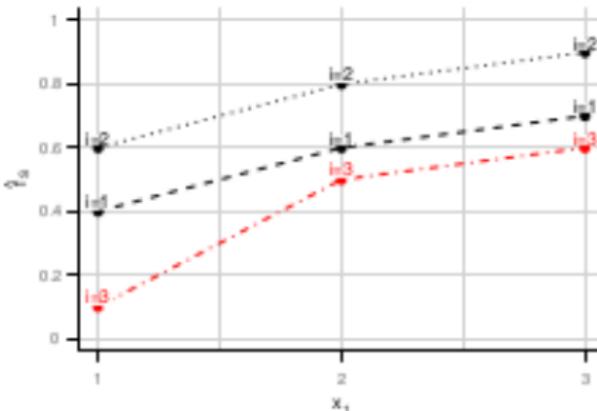
ICE CURVES - ILLUSTRATION

The figure shows three tables illustrating feature values and predictions for three observations ($i = 1, 2, 3$). The first table lists the overall feature values. The subsequent two tables show the feature values for each observation, with the last column indicating the prediction (f_i). The arrows point from the first table to the second, and from the second to the third.

	x_1	x_2	x_3	f_i
1	1	4	7	0.4
2	2	5	8	0.6
3	3	6	9	0.1

	x_1	x_2	x_3	f_i
1	1	4	7	0.6
2	2	5	8	0.8
3	2	6	9	0.5

	x_1	x_2	x_3	f_i
1	3	4	7	0.7
2	3	5	8	0.9
3	3	6	9	0.6



3. Step - Repeat for other observations:

3. Step - Repeat for other observations:

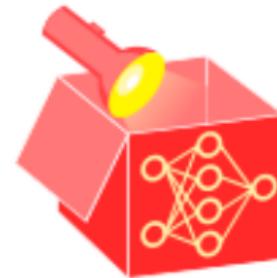
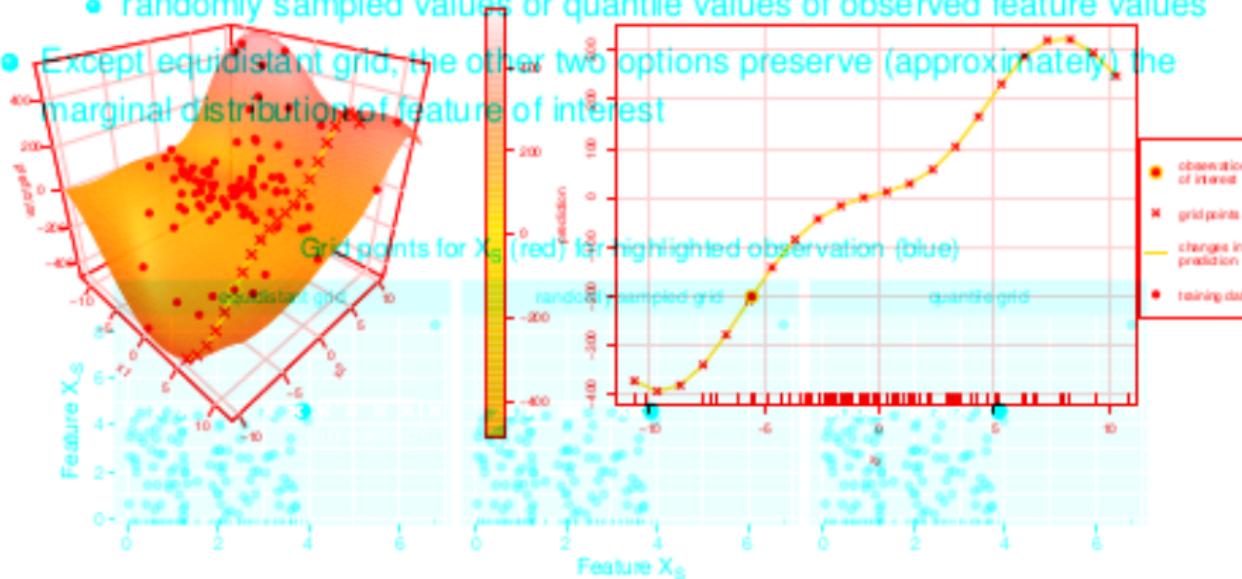
ICE curve for $i = 3$ connects all predictions at grid values associated to i -th observation.
ICE curve for $i = 3$ connects all predictions at grid values associated to i -th observation.

ICE CURVES - INTERPRETATION

- Plotting ICE curves involves generating grid values x_S , visualized on X-axis changes in prediction for different values of x_2 while keeping x_1 fixed
- Common choices for grid values are

⇒ Local interpretation

- 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 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

