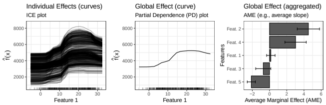
Interpretable Machine Learning

Individual Conditional Expectation (ICE) Plot



Learning goals

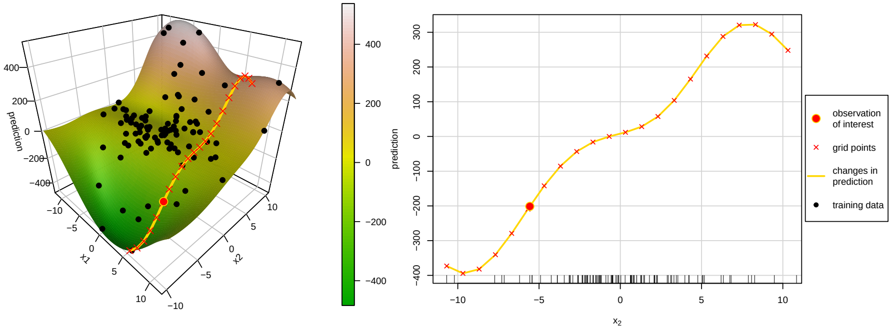
- ICE curves as local effect method
- How to sample grid points for ICE curves

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)

Partition each observation $\mathbf{x}$ into $\mathbf{x}_S$ (features 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:

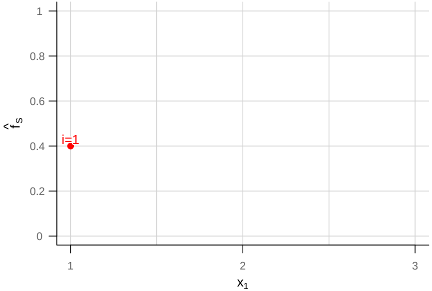
- Choose grid points $\mathbf{x}^*_S = \mathbf{x}^{*^{(1)}}_S, \dots, \mathbf{x}^{*^{(g)}}_S$ to vary $\mathbf{x}_S$
- Plot point pairs $\left\{\left(\mathbf{x}^{*^{(k)}}_S, \hat{f}^{(i)}_{S,ICE}(\mathbf{x}^{*^{(k)}}_S)\right)\right\}_{k=1}^{g}$
  where $\hat{f}^{(i)}_{S,ICE}(\mathbf{x}^*_S) = \hat{f}(\mathbf{x}^*_S,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

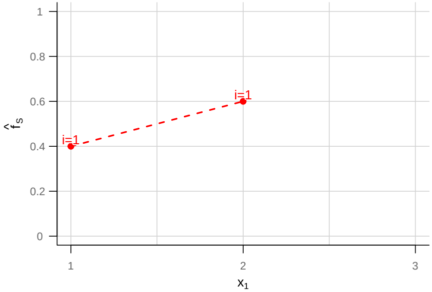


2. Step - Predict and visualize:

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

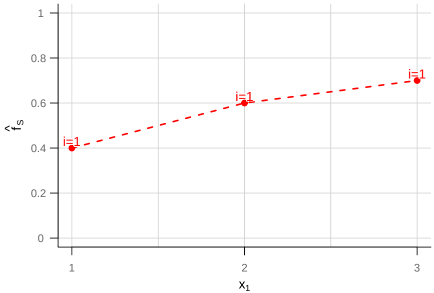


2. Step - Predict and visualize:

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

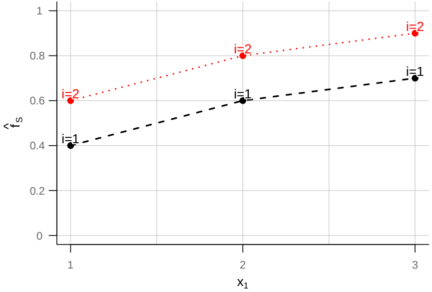


2. Step - Predict and visualize:

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

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

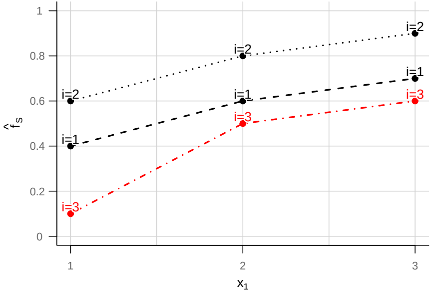
ICE CURVES - ILLUSTRATION



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

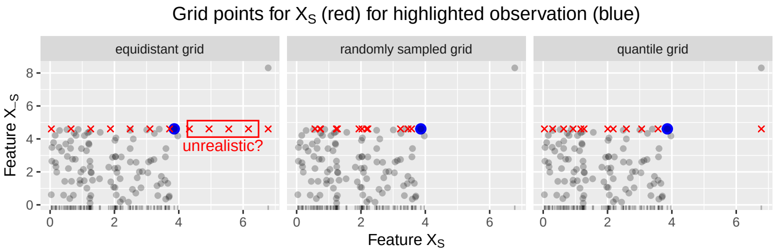


3. Step - Repeat for other observations:

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 $\mathbf{x}_S^*$; 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