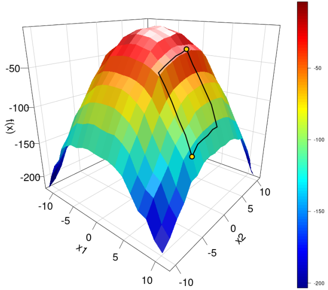
Interpretable Machine Learning

Marginal Effects



Learning goals

- Why parameter-based interpretations are not always possible for parametric models
- How marginal effects can be used in such cases
- Drawbacks of marginal effects
- Model-agnostic applicability

INTERPRETATIONS OF LINEAR MODELS

- The LM can be directly interpreted by evaluating the model coefficients:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \cdots + \epsilon$$

- Increasing x_1 by Δx_1, increases y by $\Delta x_1 \cdot \beta_1$.

- Default interpretations correspond to $\Delta x_1 = 1$, i.e., $\Delta y = \beta$.

- All interpretations are done ceteris paribus, i.e., all remaining features are kept constant.

INTERPRETATIONS OF POLYNOMIAL MODELS

- If higher-order terms or interactions are present, parameter-based interpretations are not possible anymore:

$$y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_{1,2} x_1, x_2 + \epsilon \tag{1}$$

- The isolated main effects of both features vary across different values

- The interaction depends on values of the remaining feature

- The marginal effect (ME) allows us to determine a feature effect nonetheless.

MARGINAL EFFECTS

- The most common definition of the marginal effect (ME) corresponds to the derivative of the prediction function w.r.t. a feature. We refer to this variant as the derivative ME (dME):

$$dME_j(x) = \frac{\partial f(x)}{\partial x_j}$$

- A less commonly known definition corresponds to the change in predicted outcome due to an intervention in the data, e.g., by increasing a feature value by one unit. As this variant corresponds to a forward difference, we refer to it as a forward ME (fME):

$$fME_j(x, h_j) = f(x_1, \ldots, x_j + h_j, \ldots, x_p) - f(x)$$

MARGINAL EFFECTS

- For Eq. (1), the dME and fME of x_1 with step size 1 correspond to:

$$dME_j(x) = 2\beta_1 x_1 + \beta_{1,2} x_2$$
$$fME_j(x, h_j) = 2\beta_1 h_1 + \beta_{1,2} x_2 h_1$$