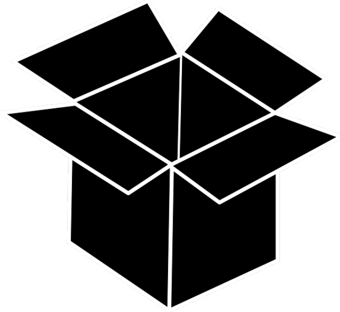
Interpretable Machine Learning

Introduction and Background



Learning goals

- Interpretable Models

HIGH-DIMENSIONAL MODEL REPRESENTATION

- A high-dimensional model representation (HDMR) decomposes the model into a sum of effect terms of increasing order:

$$\hat{f}(x) = g_{\{0\}} + g_{\{1\}}(x_1) + g_{\{2\}}(x_2) + \ \dots \ + g_{\{1,2\}}(x_1, x_2) \\ + \ \dots \ + g_{\{1,\dots,p\}}(x_1, \dots, x_p)$$

- The features need to be independent to make the HDMR unique.
- Different techniques to estimate an additive decomposition exist, e.g., repeated expectations (partial dependence / PD) or accumulated local effects (ALE).

ADDITIVE DECOMPOSITION OF A PREDICTION FUNCTION

Consider the estimation via iterative expectations:

$$
\begin{aligned}
g_{\{0\}} &= \mathbb{E}_X \left[\widehat{f}(x)\right] \\
g_{\{1\}}(x_1) &= \mathbb{E}_{X_{-1}} \left[\widehat{f}(x) \mid X_1\right] - g_{\{0\}} \\
g_{\{2\}}(x_2) &= \mathbb{E}_{X_{-2}} \left[\widehat{f}(x) \mid X_2\right] - g_{\{0\}} \\
g_{\{1,2\}}(x_1, x_2) &= \mathbb{E}_{X_{-\{1,2\}}} \left[\widehat{f}(x) \mid X_1, X_2\right] - g_{\{2\}}(x_2) - g_{\{1\}}(x_1) - g_{\{0\}} \\
&\vdots \\
g_{\{1,\ldots,p\}}(x) &= \widehat{f}(x) - \cdots - g_{\{1,2\}}(x_1, x_2) \\
&\quad - g_{\{2\}}(x_2) - g_{\{1\}}(x_1) - g_{\{0\}}
\end{aligned}
$$

FUNCTIONAL ANOVA

After $\hat{f}$ has been decomposed, we can conduct a functional analysis of variance (functional ANOVA / FANOVA):

- $$\begin{aligned} Var\left[\hat{f}(x)\right] &= Var\left[g_{\{0\}} + g_{\{1\}}(x_1) + g_{\{2\}}(x_2) + \ \dots \ + g_{\{1,2\}}(x_1, x_2) \right. \\ &\quad \left. + \ \dots \ + g_{\{1,\dots,p\}}(x)\right] \end{aligned}$$

- If the features are independent, the variance can be additively decomposed without covariances:

$$\begin{aligned} Var\left[\hat{f}(x)\right] &= Var\left[g_{\{0\}}\right] + Var\left[g_{\{1\}}(x_1)\right] + Var\left[g_{\{2\}}(x_2)\right] \\ &\quad + Var\left[g_{\{1,2\}}(x_1, x_2)\right] + \ \dots \ + Var\left[g_{\{1,\dots,p\}}(x)\right] \end{aligned}$$

FUNCTIONAL ANOVA

- Dividing by the prediction variance results in the fraction of variance explained by each term:

$$1 = \frac{Var\left[g_{\{0\}}\right]}{Var\left[\hat{f}(x)\right]} + \frac{Var\left[g_{\{1\}}(x_1)\right]}{Var\left[\hat{f}(x)\right]} + \frac{Var\left[g_{\{2\}}(x_2)\right]}{Var\left[\hat{f}(x)\right]} + \frac{Var\left[g_{\{1,2\}}(x_1, x_2)\right]}{Var\left[\hat{f}(x)\right]} + \ldots + \frac{Var\left[g_{\{1,\ldots,p\}}(x)\right]}{Var\left[\hat{f}(x)\right]}$$

- The fraction of variance explained by a term is referred to as the Sobol index:

$$S_j = \frac{Var\left[g_{\{j\}}(x_j)\right]}{Var\left[\hat{f}(x)\right]}$$