Research Ideas

Vasilis Gkolemis

April 22, 2023

1 Evaluation framework for interaction methods

1.1 Idea 1

We may split every $f : \mathbb{R}^D \to \mathbb{R}$ into a model without interaction between $\mathbf{x_c}$ and x_s, i.e., $f_{ni}(\mathbf{x}) = f^{(x_s)}(x_s) + f^{(\mathbf{x_c})}(\mathbf{x_c})$, and the interaction term $\kappa(\mathbf{x_c}, x_s)$:

$$f(\mathbf{x}) = \underbrace{f^{(x_s)}(x_s) + f^{(\mathbf{x_c})}(\mathbf{x_c})}_{f_{ni}(\mathbf{x})} + \kappa(\mathbf{x_c}, x_s)$$

A simple example is setting f to be a Neural Network and f_{ni} a Neural Additive Model without interaction between x_s and $\mathbf{x_c}$. Then $\kappa(\mathbf{x_c}, x_s) = f(\mathbf{x}) - f_{ni}(\mathbf{x})$ and we quantify the importance of κ as $\mathbb{E}_{X_c, X_s}\left[|\kappa(X_c, X_s)|\right] \approx \sqrt{\frac{1}{N}\sum_i \kappa^2(\mathbf{x_c}, x_s)}$

1.2 Idea 2