Carnegie Mellon University

Faith-Shap: the Faithful Shapley Interaction Index

Che-Ping Tsai, Chih-Kuan Yeh, Pradeep Ravikumar

Machine Learning Department, Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

TL;DR

We propose the *faithful Shapley interaction index* which is the unique interaction index satisfying the original Shapley axioms while being faithful polynomial approximation to the model.

Introduction

- Quantifying feature interaction is useful in some tasks.
  - Question-context interaction in question answering.
  - Medicine-medicine interaction in healthcare.
- Existing works extend the Shapley values to the interaction context by introducing **less suitable axioms to XAI** to ensure uniqueness.
  - The recursive axiom in Shapley interaction indices has almost no physical meaning.
  - The interaction distribution axiom in Shapley-Taylor interaction indices causes impoverished lower-order interaction attribution.
- We extend the "faithful linear approximation" property of the singleton Shapley value to **"faithful polynomial approximation"** for interaction indices.
  - Faithfulness is an important concept in XAI.
  - Together with original Shapley axioms, we prove Faith-Shap is the **unique** interaction index satisfying all these axiomatic properties.
- Similar theoretical results also apply to the Banzhaf

Preliminary: Shapley axioms

- Notations: number of input features $d \in N$, target set function $v: 2^d \to R$, maximum interaction order: $\ell \in N$, number of interactions: $d_\ell = \sum_{j=0}^{\ell} \binom{\ell}{j}$, the vector of interaction indices $\mathcal{E}(v,\ell) \in R^{d_\ell}$, the set of subsets of size $\le \ell$: S_ℓ.

Axiom 3. *(Interaction Linearity): For any maximum interaction order $\ell \in [d]$, and for any two set functions v_1 and v_2, and any two scalars $\alpha_1, \alpha_2 \in \mathbb{R}$, the interaction index satisfies: $\mathcal{E}(\alpha_1 v_1 + \alpha_2 v_2, \ell) = \alpha_1 \mathcal{E}(v_1, \ell) + \alpha_2 \mathcal{E}(v_2, \ell)$.*

Axiom 4. *(Interaction Symmetry): For any maximum interaction order $\ell \in [d]$, and for any set function $v : 2^d \mapsto \mathbb{R}$ that is symmetric with respect to elements $i, j \in [d]$, so that $v(S \cup i) = v(S \cup j)$ for any $S \subseteq [d] \backslash \{i, j\}$, the interaction index satisfies: $\mathcal{E}_{T \cup i}(v, \ell) = \mathcal{E}_{T \cup j}(v, \ell)$ for any $T \subseteq [d] \backslash \{i, j\}$ with $|T| < \ell$.*

Axiom 5. *(Interaction Dummy): For any maximum interaction order $\ell \in [d]$, and for any set function $v : 2^d \mapsto \mathbb{R}$ such that $v(S \cup i) = v(S)$ for some $i \in [d]$ and for all $S \subseteq [d] \backslash \{i\}$, the interaction index satisfies: $\mathcal{E}_T(v, \ell) = 0$ for all $T \in \mathcal{S}_\ell$ with $i \in T$.*

Axiom 6. *(Interaction Efficiency): For any maximum interaction order $\ell \in [d]$, and for any set function $v : 2^d \to \mathbb{R}$, the interaction index satisfies: $\sum_{S \in \mathcal{S}_\ell \backslash \emptyset} \mathcal{E}_S(v, \ell) = v([d]) - v(\emptyset)$ and $\mathcal{E}_\emptyset(v, \ell) = v(\emptyset)$.*

(Faithfulness property) The singleton Shpley values are the minimizer of the following weighted linear approximation:

$$\min_{\mathcal{E} \in \mathbb{R}^{d+1}} \sum_{S \subseteq [d] : \mu(S) < \infty} \mu(S) \left(v(S) - \sum_{i \in S} \mathcal{E}_i \right)^2 \text{ s.t. } v(S) = \sum_{i \in S} \mathcal{E}_i, \ \forall S : \mu(S) = \infty.$$

It has been shown that we can recover the singleton Shapley values as the solution of the weighted regression problem above by setting $\mu(S) \propto \frac{d-1}{\binom{d}{|S|} |S| (d - |S|)}$ and $\mu(\emptyset) = \mu([d]) = \infty$ [2].

Main Theoretical Results

- We say $\mathcal{E}(v,\ell)$ are Faith-Interaction Indices if there exists a weighting function $\mu(\cdot)$ such that

$$\mathcal{E}(v,\ell) = \min_{\mathcal{E} \in \mathbb{R}^{d_\ell}} \sum_{S \subseteq [d]} \mu(S) \left(v(S) - \sum_{T \subseteq S, T \le \ell} \mathcal{E}_T \right)^2.$$

- Our theoretical results:
  - (Linearity) Faith-Interaction Indices satisfy the interaction linearity axiom.
  - (Symmetry) Faith-Interaction Indices satisfy the interaction symmetry axiom if and only if the weighting function $\mu(S) = \mu(T)$ for all $|S| = |T|$.
  - (Linearity, Symmetry, Dummy) Faith-Interaction Indices satisfy the interaction linearity, symmetry, and dummy axiom if and only if the weighting function satisfies:

$$\mu(S) \propto \sum_{i=|S|}^{d} \binom{d - |S|}{i - |S|} (-1)^{i - |S|} g(a, b, i), \ \text{where } g(a, b, i) = \begin{cases} 1 & \text{if } i = 0 \\ \prod_{j=0}^{j=i-1} \frac{a(a-b) + j(b - a^2)}{a - b + j(b - a^2)} & \text{if } 1 \le i \le d, \end{cases}$$

  - (Efficiency) Faith-Interaction Indices satisfy the interaction efficiency axiom if and only if $\mu([d]) = \mu(\emptyset) = \infty$.
  - (Linearity, Symmetry, Dummy, Efficiency) **Faith-Shap is the unique interaction index** that satisfies the interaction linearity, symmetry, dummy, and efficiency axiom. The weighting function is given as $\mu(S) \propto \frac{d-1}{\binom{d}{|S|} |S| (d - |S|)}$ *for all $S \subseteq [d]$ with $1 \le |S| \le d - 1$, and $\mu(\emptyset) = \mu([d]) = \infty$.*

*We term this unique interaction index as the **Faithful Shapley Interaction index** (Faith-Shap), which has the form:*

$$\mathcal{E}_S^{\text{F-Shap}}(v, \ell) = a(v, S) + (-1)^{\ell - |S|} \frac{|S|}{\ell + |S|} \binom{\ell}{|S|} \sum_{T \supset S, |T| > \ell} \frac{\binom{|T|-1}{\ell}}{\binom{|T|+\ell-1}{\ell+|S|}} a(v, T), \ \ \forall S \in \mathcal{S}_\ell, \tag{16}$$

where $a(v, \cdot)$ is the Möbius transform of $v(\cdot)$. Moreover, its highest-order interaction terms can be expressed as a weighted average of discrete derivatives:

$$\mathcal{E}_S^{\text{F-Shap}}(v, \ell) = \frac{(2\ell - 1)!}{((\ell - 1)!)^2} \sum_{T \subseteq [d] \backslash S} \frac{(\ell + |T| - 1)!(d - |T| - 1)!}{(d + \ell - 1)!} \Delta_S(v(T)) \ \ \text{for all } S \in \mathcal{S}_\ell \text{ with } |S| = \ell. \tag{17}$$

- Though the computation of the exact Faith-Shap is exponential, in practice, we use the following procedure to approximate it:
  1. Sample each coalition $S_i \subseteq [d]$ with probability $\mu(S_i) \propto \frac{1}{\binom{d}{|S_i|} |S_i| (d - |S_i|)}$ and compute $v(S_i)$.
  2. Solve the following L1-regularized regression problem:

$$\underset{\mathcal{E} \in \mathbb{R}^{d_\ell}}{\operatorname{argmin}} \frac{1}{n} \sum \left(v(S_i) - \sum_{T \subseteq S_i, T \le \ell} \mathcal{E}_T \right)^2 + \underbrace{INFTY * \sum_{S \in \{\emptyset, [d]\}} \left(v(S) - \sum_{T \subseteq S, T \le \ell} \mathcal{E}_T \right)^2}_{\text{Ensure efficiency}} + \underbrace{\lambda |\mathcal{E}|_1}_{\text{Encourage Sparsity}}$$

Qualitative Examples

- Explain a binary sentiment classification model on the IMDB dataset.
  - A BERT model
  - Faith-Shap with the maximum interaction order $\ell = 2$.

Index	Sentences (bold words are the interactions with the highest (absolute) importance values)	Model Prediction	Interaction score
1	I have **Never forgot** this movie. All these years and it has remained in my life.	Positive	0.818
2	TWINS EFFECT is a poor film in so many respects. The **only good** element is that it doesn't take itself seriously..	Negative	-0.375
3	I rented this movie to get an easy, entertained view of the history of Texas. I got a **headache instead**.	Negative	0.396
4	Truly **appalling waste** of space. Me and my friend tried to watch this film to its conclusion but had to switch it off about 30 minutes from the end.	Negative	0.357
5	I still remember watching Satya for the first time. I was completely **blown away**.	Positive	0.283

Table 5: Top interactions of different examples on IMDB. See more results in Appendix B.

- (Example 1) While individual words "Never" and "forgot" are negative, their joint effect is positive.
- (Example 2,3,4) non-complementary interaction effect.
- (Example 5) Complementarity effects: words in a phrase are only meaningful when all words are present.