Summary and Review of “A Unified Approach to Interpreting Model Predictions”

Huy Truong
Dept. of Computer Science
University of Illinois, Chicago

12/6/2024

Abstract

Lundberg and Lee’s “A Unified Approach to Interpreting Model Predictions” introduces SHAP (SHapley Additive exPlanations), a game-theoretic framework for explaining predictions from any machine learning model [1]. This paper unifies six existing feature attribution methods, demonstrating that they approximate the formalized SHAP values under appropriate assumptions. The authors prove a uniqueness theorem, establishing SHAP as the only additive feature attribution method satisfying desirable properties. They also propose novel estimation algorithms for SHAP values and validate their efficacy through computational experiments and user studies, highlighting improved consistency with human intuition and superior discrimination between prediction classes.

1 Background

The increasing complexity of high-performing machine learning models (deep neural networks, ensembles) often comes at the cost of interpretability. While simpler (linear) models are inherently more transparent, they often lack the predictive power necessary for large, complex, ”untamed” datasets. The trade-off between accuracy and interpretability was addressed by the development of various model explanation methods like LIME and DeepLIFT [2, 3]. However, these methods lacked a unifying framework.

2 Presented Methods

The paper’s core contribution is the SHAP values, which necessitate the formalism of a new class of additive feature attribution (AFA) methods [1]. These methods employ a simplified explanation model g to approximate the original complex model f:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{1}$$

where $z' \in \{0,1\}^M$ represents a simplified input (binary vector indicating feature presence), M is the number of features, and ϕ_i is the attribution for feature i. The paper demonstrates that several existing methods (LIME, DeepLIFT, etc.) fall within this class [2, 3].

The key insight is that there is a unique solution within this class that satisfies three desirable properties: **local accuracy**, **missingness**, and **consistency** (defined formally in the paper). This unique solution is discovered to be the SHAP value, defined as:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{2}$$

where $f_x(z') = E[f(z)|z_S]$ and S is the set of non-zero indices in z'. The SHAP value represents the average marginal contribution of feature i across all possible coalitions of features.

The paper then presents various estimation methods to produce these SHAP values:

1. Model-Agnostic: Kernel SHAP (a modified LIME to recover SHAP values) with the Shapley Sampling procedure.

2. Model-Specific: Linear SHAP (for linear models), Low-Order SHAP, Max SHAP (for max functions), and Deep SHAP (a compositional approach for deep networks leveraging the DeepLIFT design).

3 Empirical Results

The authors conducted experiments on these estimation methods to evaluate their SHAP values:

1. Kernel SHAP demonstrated better sample efficiency compared to LIME and Shapley Sampling values.

2. User studies revealed better alignment between SHAP explanations and human intuition in simple scenarios (sickness prediction, profit allocation) compared to LIME and DeepLIFT [2, 3].

3. Deep SHAP showed improved performance in explaining image classifications compared to DeepLIFT, particularly in capturing changes in class probabilities.

4 Discussion and Critiques

While this SHAP framework offers significant advancements within the explainable machine learning work, some apparent limitations of this body of research remain:

1. Using the binary vector space for the explanation model might oversimplify the input space and remove pivotal information about feature interactions.

2. Computing exact SHAP values is computationally expensive for complex models and large datasets. On the other hand, most SHAP estimation methods rely on strong assumptions (feature independence, model linearity). Therefore, the trade-offs are left to the users to consider.

Future work should focus on training the explanation model g alongside the predictive model f to ensure an automatic tuning of these trade-offs, rather than having the users analyze and employ them post-hoc.

5 Contributions

This paper provides several important contributions:

1. A unifying framework for interpreting model predictions, systematically linking various existing methods under the umbrella of SHAP values.

2. Theoretical justification for using SHAP values based on their unique properties and connections to game theory.

3. Practical algorithms for estimating SHAP values, enabling their application to different types of models.

4. Validation for the effectiveness of SHAP values through both computational and user studies, demonstrating improved consistency with human intuition and better performance in explaining class differences.

References

[1] Scott Lundberg. “A unified approach to interpreting model predictions”. In: *arXiv preprint arXiv:1705.07874* (2017).

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[3] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning important features through propagating activation differences”. In: *International conference on machine learning*. PMlR. 2017, pp. 3145–3153.