1. Why let models be explainable?

Since all machine learning models perform in high dimensions, which cannot be interpreted by humans, these models can be referred to as black-box models. Especially in the field of natural language processing (NLP), the dimensions of features are tremendous, which illustrates that the complexity of features becomes crucial. Therefore, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) assist users in explaining the principles of NLP models.

2. What do explainable models mean?

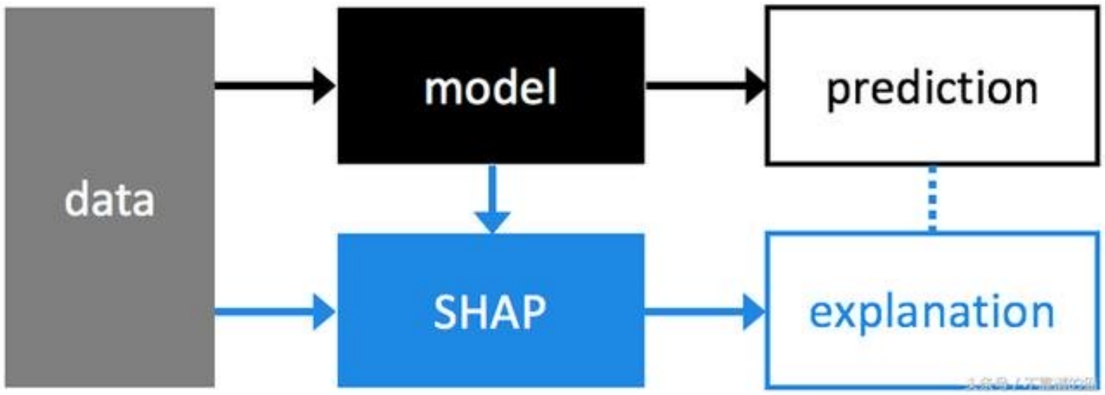
The ability to correctly explain the meanings represented by the outputs of models and the relationship between the inputs and the outputs is essential. This ability convinces users of predictive models and processes of constructing models. This ability also offers advice on improving the performance of models, which makes models evolve better.

In fact, in terms of simple predictive models, the best explanation comes from the model itself. For example, in linear models, one can explain the relationship between X and y through the positiveness and the negativeness of the parameter w. However, when structures of models become complex (e.g. deep learning, continual learning), one may hard to explain the correlation between the inputs and the outputs. Hence, some scholars propose new predictive models which are explainable, such as SHAP or LIME.

3. Concepts of SHAP

Applications of SHAP values is introduced in the paper "A Unified Approach to Interpreting Model Predictions" to evaluate the explainable potential of models. SHAP (Shapley Additive exPlanations) values analyze the predictions of models to discover the influences of each feature on the outputs of models. SHAP values are classified as a technology employed in game theory to assure the degrees of each player's contribution to the success of the coordination of game. In other words, each SHAP value assesses the extents of positive or negative contribution from each feature to the predictions of models.

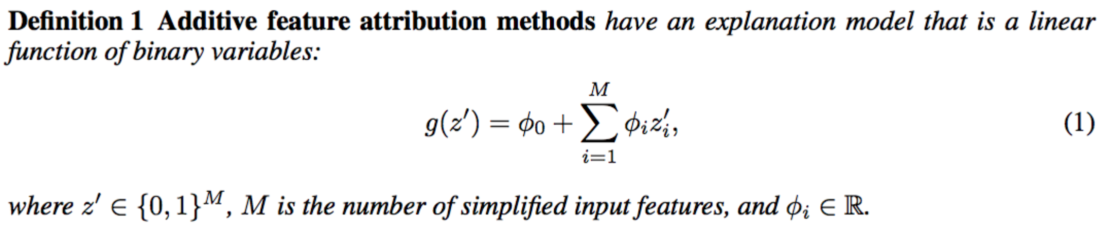
As the following figure:



The paper "A Unified Approach to Interpreting Model Predictions" suggests vital properties as the following.

(1) Additive feature attribution method

SHAP value is an additive feature attribution method. Additive feature attribution method is interpreted as the fact of assuming the impact of each feature is additive. One can observe this property from Definition 1.



Definition 1 **Additive feature attribution methods** *have an explanation model that is a linear function of binary variables:*

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

where $z' \in \{0,1\}^M$, *M is the number of simplified input features, and* $\phi_i \in \mathbb{R}$.

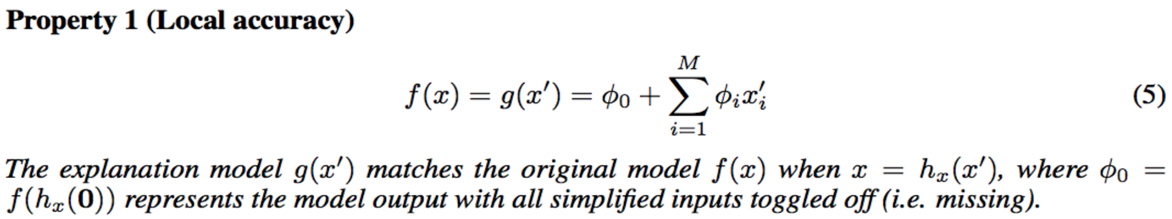
In Definition 1, each ϕ_i represents the magnitude of the effect of the i_{th} feature. Z_i stands for 0 or 1, which determines if a feature should be involved in the predictions of models.

(2) Simple properties uniquely determine additive feature attributions

This attribute elucidates that a unique solution exists in Definition 1 under the restriction of specific characteristics such as local accuracy, missingness or consistency.

- **Local accuracy**

  Property 1 represents that the total of the impacts of all features matches the value of f(x)



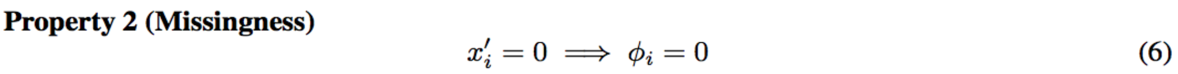
Property 1 (Local accuracy)

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{5}$$

The explanation model $g(x')$ *matches the original model* $f(x)$ *when* $x = h_x(x')$, *where* $\phi_0 = f(h_x(\mathbf{0}))$ *represents the model output with all simplified inputs toggled off (i.e. missing).*

- **Missingness**

  Property 2 represents if the i_{th} feature is not utilized, this feature contributes 0 to the total of the impacts from all features. However, this property is not guaranteed in real scenarios.

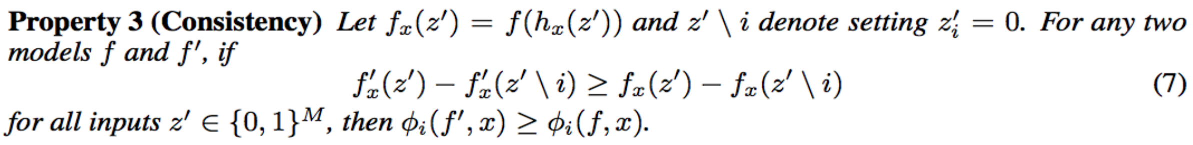


Property 2 (Missingness)

$$x'_i = 0 \implies \phi_i = 0 \tag{6}$$

- **Consistency**

  Property 3 represents if the i_{th} feature applied to the model f' is larger than the i_{th} feature applied to the model f, the i_{th} feature influences the model f' more than the model f.

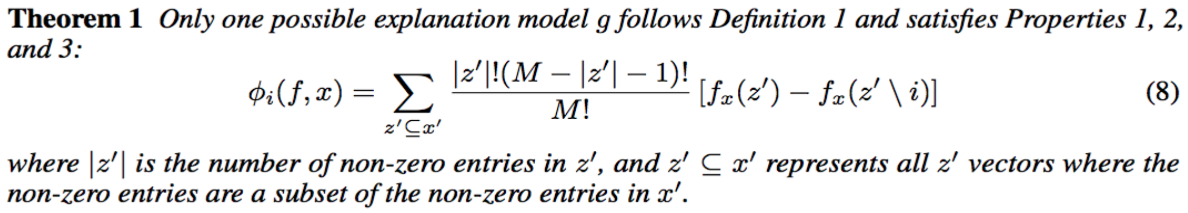


Property 3 (Consistency) *Let* $f_x(z') = f(h_x(z'))$ *and* $z' \setminus i$ *denote setting* $z'_i = 0$. *For any two models f and f', if*

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \tag{7}$$

for all inputs $z' \in \{0,1\}^M$, *then* $\phi_i(f', x) \geq \phi_i(f, x)$.

The paper “A Unified Approach to Interpreting Model Predictions” prove that the model g follows Definition 1 and satisfies Properties 1, 2 and 3. Also, the solution is SHAP values.



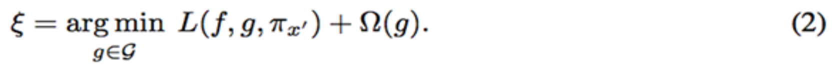
Theorem 1 *Only one possible explanation model g follows Definition 1 and satisfies Properties 1, 2, and 3:*

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}\left[f_x(z') - f_x(z' \setminus i)\right] \tag{8}$$

where $|z'|$ is the number of non-zero entries in z', and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x'.

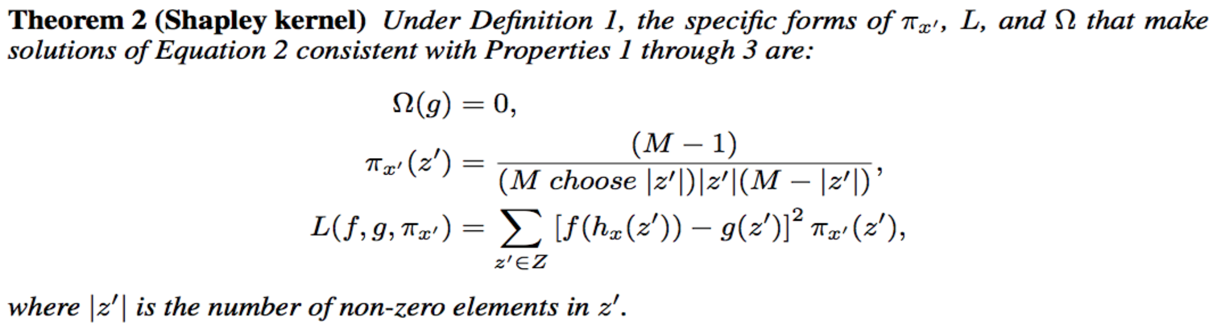
(3) Model-Agnostic Approximations

In LIME, parameters L, π and regularization are required. However, from theorem 1, only a unique model g exists in Shapley kernel to follow additive feature attribution method and to satisfy local accuracy, missingness and consistency.



$$\xi = \underset{g \in \mathcal{G}}{\arg\min}\; L(f,g,\pi_{x'}) + \Omega(g). \tag{2}$$

Also, the mode of penalization or of choosing weights may affect model g, which may cause local accuracy or consistency to disappear. Thus, applying Theorem 2 can tackle this issue and can retrieve the solution derived from Theorem 1.



Theorem 2 (Shapley kernel) *Under Definition 1, the specific forms of $\pi_{x'}$, L, and Ω that make solutions of Equation 2 consistent with Properties 1 through 3 are:*

$$\begin{aligned} \Omega(g) &= 0, \\ \pi_{x'}(z') &= \frac{(M-1)}{(M\ choose\ |z'|)|z'|(M-|z'|)}, \\ L(f,g,\pi_{x'}) &= \sum_{z' \in Z} \left[f(h_x(z')) - g(z')\right]^2 \pi_{x'}(z'), \end{aligned}$$

where $|z'|$ is the number of non-zero elements in z'.

From Theorem 2, one can observe that the loss function is transformed into the weighted least square loss one. Hence, applying Theorem 2, one can transform a problem into that regarding weighted linear regression in advance to eschew intricate calculations of SHAP values.

When employing SHAP value, one can transform the predictions of models into a weighted average value or the total of the degrees of the impacts from each feature. Hence, convoluted models are partially explainable. The paper “A Unified Approach to Interpreting Model Predictions” provides both theories and practices concerning optimizing calculations.

Works Cited

Lundberg, Scott, and Su-In Lee. *Arxiv.Org*, 2020, https://arxiv.org/pdf/1705.07874.pdf. Accessed 3 May 2020.