SHAP (SHapley Additive exPlanations) Summary

David Chu
dfc296

May 3, 2020

1 Introduction

This is a summary of the feature attribution method SHapley Additive exPlanations, referred in the remainder of the paper as SHAP. SHAP is a collection of models which attempt to estimate feature importance.

2 Explanations in the context of AI

In the context of AI, we seek to explain why P happened instead of some other event Q (termed as our foil). In the case of binary classification of positive or negative for covid-19, our foil is implicit. This generally means trying to understand the importance of each input feature to determining the prediction of the model.[Google, 2020, p.5] However, this can be difficult to do directly as complex models are too difficult to understand. Therefore, a more interpretable version of the model, termed an explanation model, must be constructed. This explanations of this explanation model must, of course, also translate to explanations about the original prediction model.[Lundberg and Lee, 2017, p.2]

3 Classic Shapley Value

The classic Shapley value, proposed by Lloyed Shapley in 1951, is a solution in cooperative game theory that attempts to calculate the importance of each

player. It does so by averaging a player's contribution across all permutations of n players, where each permutation represents an order at which the player contributed. Mathematically it can be more formally stated in the next section.

4 Shapley Regression Values

Shapley regression values represent the impact that each input feature had on the outcome. It is calculated by the below equation:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{1}$$

This equation states that to compute the feature importance of feature i (ϕ_i), we must do several steps. First we train a model $f_{S \cup \{i\}}$ which is a trained with feature i present, and a model f_S which is a model trained without feature i. Then we find the difference in predictions by finding the difference $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$. We do this for all possible subsets of $S \subseteq F$, and get the weighted average of all the differences.

In a more concrete example, if we have a feature set $F = \{x_1, x_2, x_3\}$ and we want to know the importance of x_1, we would calculate the shapely regression value ϕ_1 using the subsets $F \setminus \{i\} = \{0, x_2, x_3\}, \{0, 0, x_3\}, \{0, 0, 0\}, \{0, x_2, 0\}$, where 0 represents some baseline value such as a median or mean value, or presence.

5 Additive Feature Attribution Methods

Additive feature attribution methods are methods which construct an explanation model $g(z')$, where z' is a simplified version of the original inputs, that is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2}$$

As you can see from the above equation, the explanation model is a sum of features weighted by their importance (ϕ). As proven by Lundberg et all, all additive feature attribution methods have several desirable properties, and only one formula fulfills all the properties. These properties are:

- Local Accuracy - given x and a simplified version x', and a model f and an explanation model g, $f(x) = g(x')$. In other words, the explanation model such make the same prediction as the actual model when using the simplified input.
- Missingness - If the simplified inputs represent feature presence, then the features that are missing must have no impact on the prediction.
- Consistency - If a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, the input's attribution should not decrease. In other words, an input's importance should never decrease as a result of any other feature's change in importance.

It concludes that any method not based on Shapley values violates local accuracy and/or consistency. The authors then propose modifications to other methods to use SHAP values.

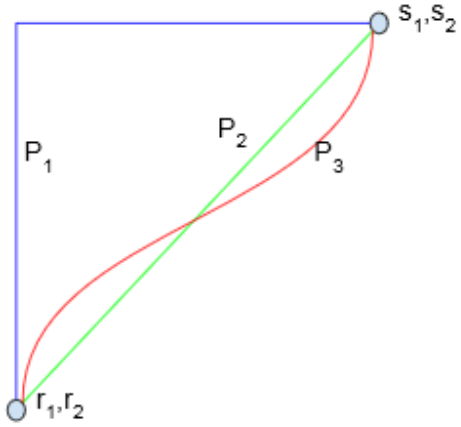
6 SHAP (SHapley Additive exPlanations)

SHAP are modifications to existing Additive Feature Attribution methods to ensure that they use values derived from Shapley values. For our purposes, we will use SHAP GradientExplainer to explain our COVID-19 model, which is a modification of Integrated Gradients (described below) to attribute feature importance.

7 Integrated Gradients

Integrated Gradients (IG) uses a generalization of the Aumann-Shapley method. Classical Shapley values are based on binary values (present/not present). However, for deep neural networks, features are almost always continuous. The classical Shapley value takes an average over several discrete paths; In each step of a discrete path, a variable is either 'off' or 'on'. IG traverses a single, smooth path (in R^n space) between the baseline and the explicand, and computes the gradients at all points along the path. Integrated gradients are obtained by accumulating these gradients.[Sundararajan et al., 2017, p.3] Integrated gradients are good for neural networks because it offers computational advantages for large input spaces.[Google, 2020, p.10]

Figure 1: Example of different paths through R^2 space. P1 is classic Shapely, while P2 represents IG



The proof that Integrated Gradients does in fact derive from the binary Shapley values can be found in Mukund Sundararajan and Amir Najmi's paper.[Sundararajan and Najmi, 2019, p.11]

In Tensorflow, to calculate we compute the gradient in a for loop by calling tf.gradients in a loop over the set of inputs[Sundararajan et al., 2017, p.5]:

$$x' + \frac{k}{m} * (x = x') \text{ for } k = 1, ..., m \tag{3}$$

where m is the number of steps in our Riemman sum.

References

Google. Ai explainability whitepaper, 2020. URL `https://storage.googleapis.com/cloud-ai-whitepapers/AI Explainability Whitepaper.pdf`.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. URL `http://arxiv.org/abs/1705.07874`.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. *CoRR*, abs/1908.08474, 2019. URL `http://arxiv.org/abs/1908.08474`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.