SHAP Summary

Chuhan Jin
cj1436

Introduction:

According to Lundberg and Su-In's work, they provide a basic info about SHAP:
"
Understanding why a model makes a certain prediction can be as crucial as theprediction's accuracy in many applications. However, the highest accuracy for largemodern datasets is often achieved by complex models that even experts struggle tointerpret, such as ensemble or deep learning models, creating a tension betweenaccuracyandinterpretability. In response, various methods have recently beenproposed to help users interpret the predictions of complex models, but it is oftenunclear how these methods are related and when one method is preferable overanother. To address this problem, we present a unified framework for interpretingpredictions, SHAP (SHapleyAdditive exPlanations). SHAP assigns each featurean importance value for a particular prediction. Its novel components include: (1)the identification of a new class of additive feature importance measures, and (2)theoretical results showing there is a unique solution in this class with a set ofdesirable properties. The new class unifies six existing methods, notable becauseseveral recent methods in the class lack the proposed desirable properties. Basedon insights from this unification, we present new methods that show improvedcomputational performance and/or better consistency with human intuition thanprevious approaches.

"

Relate Work:

As AI and data science become more and more popular, the need to interpret those complex models is also rising. Currently, there are few available methods to interpret AI classifiers. One

example is Local Interpretable Model-Agnostic Explanations, which is also known as LIME. The basic idea of LIME is, try to pre-interpret inputs to some level of human understanding. When applying LIME to images classifying, the common tricks LIME play are making super pixels. Super pixels normally represent some critical parts which contribute to the model a lot. And the equation of LIME is:

$$\xi = \arg\min_{g} \in GL(f,g,\pi x') + \Omega(g).$$

According to Lundberg's work.

There is also another approach, which is DeepLift. The basic idea of DeepLift, it gives users the ability to adjust inputs. Like replacing some uninformative inputs to assigned value, this method can help the model take some low noise inputs. Or it can contribute to final outputs, and help users to analyze the different impact of their input settings. The equation provided by Lundberg's article is :

$$n\sum i=1C\Delta xi\Delta o = \Delta o$$

And also, many explanation models integrated the feature of Shapley regression values. The main idea of this feature is, retraining subset of all features. And adjust the importance of each feature. By applying these rules, a model can quickly determine which feature takes more place to contribute to the final predictions.

SHAP

SHAP, which stands for SHapley Additive exPlanation Values, are proposed by Lundberg and Su-In, as a candidate for a unified explanation model.
There are two model-agnosticapproximation methods authors used in SHAP, one is Shapley regression values and another is Kernal Shap. In related work part, I have talked about Shapley regression values, now the new one is Kernal Shap.
Kernal Shap = liner LIME + SHAP values. The idea behind this concept is, LIME will automatically subtract some inputs aka features. However, on the other hand, SHAP will find a solution and better weights in scenarios in which missingness appeared. Thus combining these

two methods can result in a good balance. And Lundberg in his article has also proved mathematically that Knernal Shap is a better version of LIME and Shapley regression values.

Conclusion

SHAP provided a new idea on interpreting model outputs. The SHAP framework identifies the class of additive feature importance methods, include all previous methods. Based on mathematical provement and some experiments, SHAP has shown it's potential ability as an unified explanation model.

Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *ArXiv.org*, 25 Nov. 2017, arxiv.org/abs/1705.07874. Accessed 3 May 2020