Review Class 7

A Unified Approach to
Interpreting Model Predictions

Paper
link: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

jchen357@g.syr.edu Switch account

Resubmit to save

* Indicates required question

Email *

jchen357@syr.edu

Your Name *

Jingyuan Chen

Your NetID *

jchen357

You're editing your response. Sharing this URL allows others to also edit your response.

FILL OUT A NEW RESPONSE

Write a short summary of the paper *

The goal of this article is to solve the interpretability problem of machine learning models. Complex models are generally difficult to interpret, which creates a tension between accuracy and interpretability. The paper introduce a unified framework for interpreting predictions, SHAP which unified six prior method (LIME, DeepLIFT, Layer-Wise Relevance Propagation, Shapley Regression Values, Shapley sampling values, Quantitative input influence). The paper make assume that feature independence when approximating conditional expectations, the perspective of viewing any explanation of a model's prediction as a model itself and, game theory results guaranteeing a unique solution apply to the entire class of additive feature attribution methods. The approach is using additional feature attribution methods that explain the predictions of complex models using simple, interpretable approximate models. The paper evaluates the benefits of SHAP values by using Kernel SHAP and Deep SHAP approximation methods. And it designed user studies to compare SHAP values with alternative feature importance allocations represented by DeepLIFT and LIME. Finally, it uses MNIST digit image classification to compare SHAP with DeepLIFT and LIME. The result shous SHAP are more consistent with human intuition and provide more accurate and effective feature importance estimates than DeepLIFT and LIME. The SHAP framework identifies a class of additive feature importance methods, showing that there is a unique solution in this class that satisfies a set of desirable properties.

Write about (1) the problem statement of the paper, (2)motivations and/or context and (3) assumptions of the paper. *

Problem statement: The goal of this article is to solve the interpretability problem of machine learning models. That is, how to provide a unified and accurate way to explain the prediction results of different models, especially when facing complex models.

Motivation: Complex models are generally difficult to interpret, which creates a tension between accuracy and interpretability. Therefore, proposing an effective method to explain model predictions, especially predictions from complex models, became the main focus of this study. motivation.

Assumptions:
Feature independence when approximating conditional expectations

The perspective of viewing any explanation of a model's prediction as a model itself

Game theory results guaranteeing a unique solution apply to the entire class of additive feature attribution methods

Good model explanations should be consistent with explanations from humans who

You're editing your response. Sharing this URL allows others to also edit your response.

FILL OUT A NEW RESPONSE

What are the contributions of the paper? What are the novelties? *

The paper introduce a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. It unifies six existing methods (LIME, DeepLIFT, Layer-Wise Relevance Propagation, Shapley Regression Values, Shapley sampling values, Quantitative input influence). Provide Approaches such as Kernel SHAP, Deep SHAP.

Novelties: unifies six existing methods, new SHAP value estimation methods, additive feature attribution methods. Introduce the perspective of viewing any explanation of a model's prediction as a model itself. Kernel SHAP , Deep SHAP.

You're editing your response. Sharing this URL allows others to also edit your response.

FILL OUT A NEW RESPONSE

Write a summary about the paper approach. Is the approach novel compared to the prior works? *

You're editing your response. Sharing this URL allows others to also edit your response.

FILL OUT A NEW RESPONSE

Approach:
Additional feature attribution method: This is a method of interpreting complex model predictions. Interpret models using simpler, interpretable approximate models
LIME: LIME revolves around local approximations of predictive models, using local linear interpretations of the model to understand why the model makes a specific prediction. It reduces the input to an understandable form and then finds the best interpretation within this reduced input space. To find the best explanation, LIME optimizes an objective function that takes into account the accuracy and complexity of the explanatory model. It finds the best explanation by minimizing a loss function that takes into account how accurately the explanatory model predicts the original model, as well as the complexity of the explanatory model itself.
DeepLIFT: In DeepLIFT, the input data is mapped into binary values, where 1 means that the input feature retains its original value, while 0 means that the input feature takes a reference value. The reference value is selected by the user. DeepLIFT assigns a value (C delta x_i delta y) that reflects the impact on the output when an input feature changes compared to a reference value.
Layer-Wise Relevance Propagation: Similar to DeepLIFT, but when the input data is mapped to binary values, , 0 means that the input feature takes 0 as the value.
Shapley regression: used to solve problems when features are highly correlated. It evaluates the impact of each feature by retraining the model by comparing the difference in model predictions with and without features. It then calculates the predicted differences for all possible feature combinations, using a weighted average of these differences to represent the importance of each feature.
Shapley sampling value: First, it approximates the effect of removing a feature by sampling; second, it estimates the effect of the feature by comparing predictions with and without the feature. This approach avoids retraining the model and reduces the number of prediction differences that need to be calculated.
Quantitative input impact: Provides a more comprehensive framework to assess not only feature importance, but also other aspects. This method also proposes an approximate calculation method similar to Shapley sampling values

The SHAP value method provides a unified framework for interpreting model predictions by assigning feature importance to the model predictions. This method is based on the Shapley value of the conditional expectation function and is an innovative way to account for variation in model predictions by averaging over all possible orderings of φi values.

Kernel SHAP method, which combines linear LIME and Shapley values, focuses on how to satisfy the three properties of consistency, accuracy and missingness through specific loss functions, weight kernels and regularization terms.

Deep SHAP: It combines linear DeepLIFT and Shapley values, which utilizes additional knowledge of the compositional properties of deep networks to improve the efficiency of calculating SHAP values by analyzing small components of the network. Deep SHAP does not require heuristic selection of methods to linearize components, but instead derives SHAP values by performing efficient linearization on each component.

Compared with previous work, the novelty of the SHAP method is that it provides a unified method to understand and compare different feature importance measures while

You're editing your response. Sharing this URL allows others to also edit your response.
FILL OUT A NEW RESPONSE

Write a summary about the experimental design. *

Experimental design: The paper evaluates the benefits of SHAP values by using Kernel SHAP and Deep SHAP approximation methods. And it designed user studies to compare SHAP values with alternative feature importance allocations represented by DeepLIFT and LIME. Finally, it uses MNIST digit image classification to compare SHAP with DeepLIFT and LIME.

Comparison: Computational efficiency and accuracy of Kernel SHAP vs. LIME and Shapley sampling values
Results: Comparing Shapley sampling, SHAP, and LIME on dense and sparse decision tree models illustrates that kernel SHAP improves sampling efficiency and that LIME values can differ significantly from SHAP values that satisfy local accuracy and consistency.

Comparison: SHAP values compared to alternative feature importance assignments represented by DeepLIFT and LIME.
Assume Good model explanations should be consistent with explanations from humans who understand that model. LIME, DeepLIFT, and SHAP were compared to human interpretation of two settings. In the first setting disease scores were higher when only one of the two symptoms was present, and in the second setting participants were told a story about how three men behaved according to Short story about their highest score and money making

Result:
The agreement between human interpretation and SHAP is much stronger than other methods.

Compare: MNIST digit image classification to compare SHAP with
DeepLIFT and LIME.
Result: kernel SHAP and LIME interpret the model's output

What are the fallacies of the paper? (If any) Reference: Slide 9 of 'Paper Presentation and Review.ppt' *

The paper may contain too much assumption.
Feature independence when approximating conditional expectations this assumption may not work in real life data.

Participants Due to limited sample size.

The SHAP value method mainly focuses on the additive contribution of features and may not fully capture the complex interactions between features.

You're editing your response. Sharing this URL allows others to also edit your response.

FILL OUT A NEW RESPONSE

Never submit passwords through Google Forms.

This form was created outside of your domain. Report Abuse - Terms of Service - Privacy Policy

Google Forms

You're editing your response. Sharing this URL allows others to also edit your response.

FILL OUT A NEW RESPONSE

You're editing your response. Sharing this URL allows others to also edit your response.

FILL OUT A NEW RESPONSE