University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Shapley Additive Explanations

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

WORLD CHANGING GLASGOW

Model Agnostic Approaches

- Permutation Feature Importance
- Local Interpretable Model-agnostic Explanations
- **Shapley Additive Explanations**

Shapley Additive Explanations (SHAP)

- SHAP identifies a new way of estimating importance scores
- Each explanation is treated as a model itself (surrogate model)
- There is a unique solution in this class with a set of desirable properties
- Exploiting game theory guarantees a unique solution

Feature Attribution Methods

- SHAP is a member of the **additive feature attribution** methods class:

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

Feature Importance - Multicollinearity

- The importance of a feature might be underestimated under the presence of multicollinearity
- Estimate every possible combination of features subsets
- The differences between the original model and every possible subset are estimated
- The total score **(shapley sampling value)** of a feature is a weighted average of all possible combinations with this feature included

Uniqueness of Additive Feature Attributions

- Local accuracy
- Missingness
- Consistency

Kernel SHARP

- Linear LIME & Shapley values.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

Explanation

Model to explain

Proximity Measure

Kernel SHARP

- Linear LIME & Shapley values.

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

Explanation **Model to explain** **Proximity Measure**

$$\Omega(g) = 0,$$

$$\pi_{x'}(z') = \frac{(M-1)}{(M \; choose \; |z'|)|z'|(M-|z'|)},$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f(h_x^{-1}(z')) - g(z')\right]^2 \pi_{x'}(z'),$$

Drawbacks of SHAP

- Shapley sampling values require estimating importance score for every possible subset combination of input features
- Even with KernelSHAP computational complexity is high
- With local approximators, we may still have problems to understand the model behavior
- Time-series dependencies are not taken into consideration

Summary

- SHAP is an additive feature attribution method
- SHAP handles well multicollinearity
- SHAP improves over LIME because it finds a unique solution which satisfies the properties of local accuracy, consistency and missingness

References

- Ribeiro et al. ‘Model-Agnostic Interpretability of Machine Learning’, ICML Workshop on Human Interpretability in Machine Learning, 2016.
- Lundberg et al. ‘A Unified Approach to Interpreting Model Predictions’, NIPS 2017.
- Neves et al. ‘Interpretable heartbeat classification using local model-agnostic explanations on ECGs’, Computers in Biology and Medicine, 2021.