Proceedings of the DataNinja sAIOnARA 2024 Conference 13-15 DOI: 10.11576/dataninja-1158

Comparing Shapley Value Approximation Methods for Unsupervised Feature Importance

Patrick Kolpaczki PATRICK.KOLPACZKI@UPB.DE
Paderborn University, Germany

Abstract

Assigning importance scores to features is a common approach to gain insights about a prediction model's behavior or even the data itself. Beyond explainability, such scores can also be of utility to conduct feature selection and make unlabeled high-dimensional data manageable. One way to derive scores is by adopting a game-theoretical view in which features are understood as agents that can form groups and cooperate for which they obtain a reward. Splitting the reward among the features appropriately yields the desired scores. The Shapley value is the most popular reward sharing solution. However, its exponential complexity renders it inapplicable for high-dimensional data unless an efficient approximation is available. We empirically compare selected approximation algorithms for quantifying feature importance on unlabeled data.

Keywords: Shapley values, feature importance scores, unsupervised learning

1. Unsupervised Feature Importance

The increasing complexity of machine learning models as well as dimensionality of collected data is calling for a method to make both interpretable to the human user. A universally applicable approach are additive feature explanations which divide an observed numerical effect among the available features. Choosing this effect to be explained appropriately allows to interpret each feature's share as its contribution to the behavior of interest. In particular, the Shapley value [1] has emerged as the most frequently applied scoring rule. Popular examples include the features' contributions to a model's generalization performance [2, 3] and prediction value for a selected instance [4]. In the realm of unlabeled data and absence of a prediction model, Shapley-based feature importance scores have been utilized to perform dimensionality reduction [2]. Balestra et al. [5] refined this approach by proposing a feature ranking based on Shapley values that reduces redundancy among the selected features. Aiming at preserving the information contained in the data while minimizing correlation between the selected feature subset Balestra et al. employ the total correlation of shared by all all available features of the dataset as the numerical effect to be divided. For any subset S it is given by

$$C(S) = \sum_{X \in \mathcal{S}} H(X) - H(S) \tag{1}$$

where $H(X)$ and $H(S)$ denote the Shannon entropy of a single feature X and a set of features S respectively. This is made feasible by viewing the set of all feature values as observed realizations of a random variable.

2. Cooperative Games

A *cooperative game* is formally given by a pair $(\mathcal{N}, \nu)$ containing a finite set of *players* $\mathcal{N} = \{1, \dots, n\}$ and a *value function* $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ that assigns a real-valued *worth* to each *coalition* $S \subseteq \mathcal{N}$. This simple formalism is expressive enough to model feature subsets as coalitions that share some total correlation. The most popular solution to the question of how to divide the achieved worth $\nu(\mathcal{N})$ among all players is the *Shapley value* [1] as it is provably the only solution to fulfill certain axioms [1] that plausibly capture a notion of fairness. It assigns to each $i \in \mathcal{N}$ the share

$$\phi_i = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot [\nu(S \cup \{i\}) - \nu(S)] \tag{2}$$

and can be interpreted as a weighted average of marginal contributions $\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S)$. Given the context of high-dimensional data yielding large player numbers, the computational complexity caused by the exponential number of coalitions renders any attempt to exactly calculate ϕ_i futile.

3. Shapley Value Approximation

The rapid increase of the Shapley value's popularity in recent years, spanning over various machine learning fields [6] and beyond, incentivized the research on how to approximate it, facilitating its practical usage. The approximation problem consists of the task of computing precise estimates $\hat{\phi}_1, \ldots, \hat{\phi}_n$ of all Shapley values with minimal resource consumption.

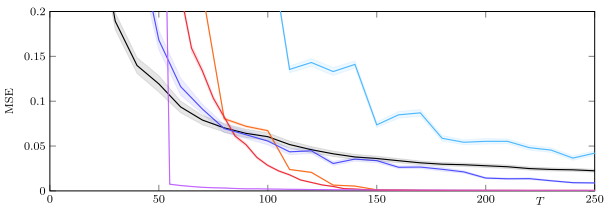
We consider the *fixed-budget setting* in which the number of times an approximation algorithm is allowed to access ν is limited by a *budget* $T \in \mathbb{N}$. This is motivated by the observation that the evaluation of large models or data poses a bottleneck, possibly even causing monetary costs when the access is provided remotely by another party. The quality of the estimates is measured by the mean squared error (MSE) averaged over all players which is to be minimized:

$$\text{MSE} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\phi}_i - \phi_i\right)^2\right] .$$

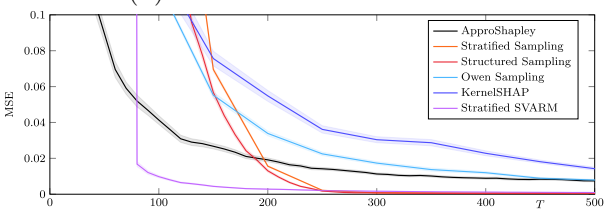
We shortly describe selected algorithms that we use for our experiments in Section 4. The first and simplest class of approximation methods leverages the fact that ϕ_i can be interpreted as player i's expected marginal contribution. This allows to obtain a mean estimate by randomly sampling marginal contributions. Castro et al. [7] propose with *ApproShapley* an algorithm that draws random permutations of $\mathcal{N}$. It extracts a marginal contribution of each player by iterating through a permutation. Following the spirit, *Stratified Sampling* [8] partitions the population of a player's marginal contributions into strata, each containing marginal contributions to coalitions S of the same size. This technique can increase estimation quality if $|S|$ has an influence on $\Delta_i(S)$. Closely related, *Structured Sampling* [9] modifies sampled permutations such that the marginal contributions to coalitions of different sizes appear in the same frequency. Departing from the discrete sum, *Owen Sampling* [10] updates an integral representation of the Shapley value [11]. Introducing another representation, Kolpaczki et al. [12] sample with *Stratified SVARM* single coalitions instead of marginal contributions. In combination with stratification it reaches higher sample efficiency as all players' estimates are updated with each coalition. Adopting a different view, *KernelSHAP* [4] solves a weighted least squares problem, filled by randomly drawn coalitions, of which the Shapley values are the solution.

4. Empirical Evaluation

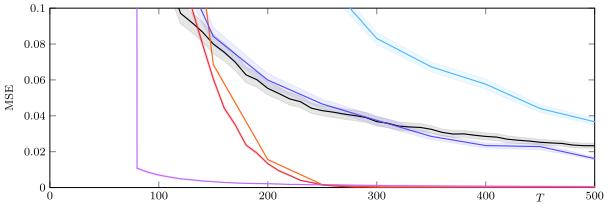
We compare the approximation quality of selected algorithms depending on the available budget T for unsupervised feature importance. In particular we use three real-world datasets: Breast Cancer, Big Five Personality Test, and FIFA 21 prepared as in [5]. A cooperative game is built from each dataset by interpreting the features as players and applying the total correlation as the corresponding coalition's worth. The approximation algorithms are run for a range of different budget values for multiple repetitions. In order to track the MSE, we calculate the Shapley values exhaustively beforehand. From Figure 1, *Stratified SVARM* emerges as significantly superior once it completes its warmup. *Stratified Sampling* and *Structured Sampling* perform on par or marginally better for higher budget ranges. The advantage of stratifying methods is likely to be caused by the impact of the feature subset size on the total correlation. In contrast, other methods including *KernelSHAP* perform clearly worse, except for *ApproShapley* displaying the lowest MSE given exteremely small budget.



(*a*) Breast Cancer: 9 features



(*b*) Big Five: 12 features



(*c*) FIFA 21: 12 features

Figure 1: Averaged MSE and std. error over 50 repetitions depending on available budget T.

Acknowledgments

This research was supported by the research training group Dataninja, funded by the German federal state of North Rhine-Westphalia.

References

[1] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, 1953.

[2] Shay B. Cohen, Eytan Ruppin, and Gideon Dror. Feature selection based on the shapley value. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 665–670, 2005.

[3] Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Proceeedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceeedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4768–4777, 2017.

[5] Chiara Balestra, Florian Huber, Andreas Mayr, and Emmanuel Müller. Unsupervised features ranking via coalitional game theory for categorical data. In *Proceedings of Big Data Analytics and Knowledge Discovery (DaWaK)*, pages 97–111, 2022.

[6] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Oliver Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5572–5579, 2022.

[7] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[8] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. *CoRR*, abs/1306.4265, 2013.

[9] Tjeerd van Campen, Herbert Hamers, Bart Husslage, and Roy Lindelauf. A new approximation method for the shapley value applied to the wtc 9/11 terrorist attack. *Social Network Analysis and Mining*, 8(3):1–12, 2018.

[10] Ramin Okhrati and Aldo Lipani. A multilinear sampling algorithm to estimate shapley values. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 7992–7999, 2020.

[11] Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):64–79, 1972. ISSN 00251909, 15265501.

[12] Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the shapley value without marginal contributions. In *Proceeedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 13246–13255, 2024.