Efficient Shapley value explanation through feature groups

Martin Jullum[1], **Annabelle Redelmeier**[2] **and Kjersti Aas**[3]

[1] *Norwegian Computing Center, Norway, jullum@nr.no*
[2] *Norwegian Computing Center, Norway, anr@nr.no*
[3] *Norwegian Computing Center, Norway, kjersti@nr.no*

Shapley values has established itself as one of the most appropriate and theoretically sound frameworks for explaining predictions from complex regression/machine learning models. The popularity of Shapley values in the explanation setting is probably due to its unique theoretical properties. The main drawback with Shapley values, however, is that its computational complexity grows exponentially in the number of input features (covariates), making it unfeasible in many real world situations where there could be hundreds or thousands of features. Furthermore, with many (dependent) features, presenting/visualizing and interpreting the computed Shapley values also becomes challenging, see e.g. [1].

I hereby present *groupShapley*: a conceptually simple approach for dealing with the aforementioned bottlenecks. The idea is to group the features, for example by type or dependence, and then compute and present Shapley values for these groups instead of for all individual features. Reducing hundreds or thousands of features to half a dozen or so, makes precise computations practically feasible and the presentation and knowledge extraction greatly simplified.

It may be shown that under certain conditions, groupShapley is equivalent to summing the feature-wise Shapley values within each feature group, but simulations indicate that they may differ significantly outside the conditions. The groupShapley method is implemented in a development version of the R-package `shapr` [2].

I will introduce the method, and talk about it from both theoretical and practical sides. The practical simplifications will be showcased through a real world car insurance example.

References

[1] Aas, K., Jullum, M., Løland, A (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, **298**, https://doi.org/10.1016/j.artint.2021.103502

[2] Sellereite, N. and Jullum, M. (2020). shapr: An R-package for explaining machine learning models with dependence-aware Shapley values. *Journal of Open Source Software*, **5**(46), 2027, https://doi.org/10.21105/joss.02027