Explaining probabilistic prediction on the simplex with Shapley compositions

Anonymous Authors[1]

Abstract

The concept of Shapley value has been widely use for measuring the contribution of each feature on a machine learning model's prediction. However, this has been designed for one-dimensional function's codomain. For multi-class probabilistic classifier, where the output is a discrete probability distribution over the set of more than two possible classes, the output lives on a multidimensional simplex. In this case, people have been applying the concept of Shapley value on each output dimension one-by-one, in an implicit one-vs-rest setting, ignoring the compositional nature of the output distribution where the relative information between probabilities matter. Using the Aitchison geometry of the simplex, coming from the field of compositional data analysis, this paper present a first initiative for a multidimensional extention of the concept of Shapley value, named Shapley composition, for explaining probabilistic predictions on the simplex in machine learning.

1. Introduction

Modern machine learning approaches like the one based on deep learning are often regarded as black-boxes making them not reliable for real-life application where the machine learnbing prediction has to be understood. These last years, the number of contribution to make models more explainable has therefore increased in the machine learning literature. One way to better understand a prediction would be to measure the contribution of each input features on the computation of the model output. The concept of Shapley value is now widely used for this purpose (Štrumbelj & Kononenko, 2014; Datta et al., 2016) especially since the release of the SHAP toolkit (Lundberg & Lee, 2017)[1]. The Shapley value came from cooperative game theory...

explain shapley in game theory,

How it is applied to ML,

Limitation,

...

1.1. Contributions

...

2. The Shapley value in machine learning

In this section, we recall the theoretical formulation of the Shapley value for measuring the contribution of each feature on a machine learning prediction.

Let $f : \mathcal{X} \to \mathbb{R}$ be a learned model one want to locally explain where $f(\boldsymbol{x})$ is the prediction on the instance $\boldsymbol{x} \in \mathcal{X}$. Let Pr be the probability distribution over $\mathcal{X}$ of the data[2]. Let $S \subseteq \{1, 2, \dots d\}$, where d is the number of features that composes an instance $\boldsymbol{x} \in \mathcal{X}$, be a subset of indices. $\boldsymbol{x}_S$ refers to an instance $\boldsymbol{x}$ restricted to the features indicated by the indices in S.

When an instance $\boldsymbol{x}$ is observed, the expected value of the prediction is simply $\mathbb{E}[f(\boldsymbol{x}) \mid \boldsymbol{x}] = f(\boldsymbol{x})$. However, when only $\boldsymbol{x}_S$ is given, i.e. part of the features, there is uncertainty about the other features and we therefore compute the expected prediction given $\boldsymbol{x}_S$: $\mathbb{E}_{\Pr}[f(\boldsymbol{x}) \mid \boldsymbol{x}_S] = \int_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \Pr(\boldsymbol{x} \mid \boldsymbol{x}_S) d\boldsymbol{x}$. The contribution of the feature indexed by $i \notin S$ in the prediction $f(\boldsymbol{x})$ given the known features indexed by S is given by:

$$c_{f,\boldsymbol{x},\Pr}(i, \boldsymbol{X}_S) = v_{f,\boldsymbol{x},\Pr}(\boldsymbol{X}_{S \cup \{i\}}) - v_{f,\boldsymbol{x},\Pr}(\boldsymbol{X}_S), \quad (1)$$

where v is known as the value function:

$$\begin{aligned} v_{f,\boldsymbol{x},\Pr} : 2^{\mathcal{X}} &\to \mathbb{R}, \\ S &\mapsto \mathbb{E}_{\Pr}[f(\boldsymbol{x}) \mid \boldsymbol{x}_S], \end{aligned} \quad (2)$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]https://github.com/shap/shap

[2]In practice, this is usually unknown but the expectation will be replaced by empirical samplings.

where $2^{\mathcal{X}}$ is the set of all subsets of S. This measure the contribution of the ith features with a particular coalition of features indexed by S. The whole contribution of the ith feature is computing by averaging this quantity over all possible coalitions as follow:

$$\phi_{f,\boldsymbol{x},\mathrm{Pr}}(i) = \frac{1}{d!}\sum_{\pi} c_{f,\boldsymbol{x},\mathrm{Pr}}(i, \pi^{<i}_{\boldsymbol{X}}), \tag{3}$$

where π is a permutation of the set S of indexes and $\pi^{<i}_{\boldsymbol{X}}$ is the features of $\boldsymbol{X}$ coming before the ith feature in the ordering given by π. For better clarity, the subscript $_{f,\boldsymbol{x},\mathrm{Pr}}$ wil be dropt from the equations.

This quantity is known as the Shapley value for the ith feature. It comes from cooperative game theory and is known to be only quantity respecting a set of desired axiomatic properties (Shapley et al., 1953). It is linear as a function of the model ($\alpha, \beta \in \mathbb{R}$): $\phi_{\alpha f+\beta g}(i) = \alpha\phi_f(i) + \beta\phi_g(i)$, and the "centered" learned model is additively separable with respect to the Shapley values: $(\boldsymbol{x}) - \mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{X})] = \sum_{i=1}^{d} \phi_f(i)$, which is known as the efficiency property.

Like originally developed in game theory, the Shapley value is designed for one-dimensinal codomain of the function f. For explaining machine learning model with a multidimensional output like for multiclass classifiers, people have been explaining each output dimension one-by-one ignoring the relative information between them. Indeed, for application with more than two classes, the probabilistic output of a classifier lives on a multidimensional simplex. The latter is the sample space of data refered as compositional data we briefly review in the next section.

3. Compositional data

Compositional data carries relative information. Each element of a composition describes a part of some whole (Pawlowsky-Glahn et al., 2015) like vectors of proportions, concentrations, and discrete probability distributions. A N-part composition is a vector of N non-zero positive real numbers that sum to a constant k. Each element of the vector is a part of the whole k. The sample space of compositional data is known as the simplex: $\mathcal{S}^N = \left\{\boldsymbol{x} = [x_1, x_2, \dots x_N]^T \in \mathbb{R}_+^{*N} \middle| \sum_{i=1}^{N} x_i = k\right\}$. In a composition, only the relative information between parts matters and John Aitchison introduced the use of log-ratios of components to handle this (Aitchison, 1982). He defined several operations on the simplex which leads to what is called the Aitchison geometry of the simplex.

3.1. The Aitchison geometry of the simplex

John Aitchison defined an internal operation called perturbation, an external one called powering and an inner product (Aitchison, 2001):

- a perturbation: $\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}\left([x_1y_1, \dots x_Ny_N]\right)$ seen as an addition between two compositions $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^N$,
- a powering: $\alpha \odot \boldsymbol{x} = \mathcal{C}\left([x_1^\alpha, \dots x_N^\alpha]\right)$ seen as a multiplication by a scalar $\alpha \in \mathbb{R}$,
- an inner product: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} \log\frac{x_i}{x_j}\log\frac{y_i}{y_j}$.

$\mathcal{C}(\cdot)$ is the closure operator. Since only the relative information matter, scaling factors are irrelevant and a composition $\boldsymbol{x}$ is equivalent to $\lambda\boldsymbol{x} = [\lambda x_1, \lambda x_2, \dots \lambda x_N]$ for all $\lambda > 0$. This equivalence is materialized by the closure operator defined for $k > 0$ as: $\mathcal{C}(\boldsymbol{x}) = \left[\frac{kx_1}{\|\boldsymbol{x}\|_1}, \frac{kx_2}{\|\boldsymbol{x}\|_1}, \dots \frac{kx_N}{\|\boldsymbol{x}\|_1}\right]^T$, where $\boldsymbol{x} \in \mathbb{R}_+^{*N}$ and $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{N}|x_i|$.

This give to the simplex a $(N-1)$-dimensional Euclidean vector space structure called Aitchison geometry of the simplex. In this paper, since we are interested in classifiers' outputs as discrete probability distributions, we restrict ourselves to the probability simplex where $k = 1$.

3.2. The isometric log-ratio transformation

...

4. Shapley compositions on the simplex

5. Explaining probabilistic prediction

6. Discussion and conclusion

References

Aitchison, J. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological), 44(2):139–177, 1982.

Aitchison, J. Simplicial inference. In Marlos A. G. Viana, D. S. P. R. e. A. S. S. o. A. M. i. S. (ed.), Algebraic Methods in Statistics and Probability, Contemporary Mathematics 287. American Mathematical Society, 2001.

Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE Symposium on Security and Privacy (SP), pp. 598–617, 2016. doi: 10.1109/SP.2016.42.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc., 2017.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. Modeling and Analysis of Compositional Data. John Wiley & Sons, 2015.

Shapley, L. S. et al. A value for n-person games. pp. 307–317, 1953.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41:647–665, 2014.

A. You can have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.