# Explaining probabilistic predictions on the simplex with Shapley compositions

**Paul-Gauthier Noé** [1], Miquel Perelló-Nieto [2],
Peter Flach [2], Jean-François Bonastre [1]

[1]Avignon Université

[2]University of Bristol

LIA seminar, Avignon Université
January 18, 2024

# Introduction

**Local explanation in machine learning:**

$$\boldsymbol{x} \in \mathbb{R}^d \longrightarrow \boxed{\text{Model}} \longrightarrow \text{Prediction}$$

*Given one instance* $\boldsymbol{x}$ *with d features, what is the contribution/effect of a feature's value on the prediction?*

$\neq$ Global explanation

# Introduction

Examples of local explanation methods:

- Local Interpretable Model-Agnostic Explanations (LIME)[1],
- Shapley values[2] (SHAP toolkit[3])

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[2] Erik Štrumbelj and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions". In: *Knowledge and information systems* 41 (2014), pp. 647–665.

[3] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774.

# Introduction

**Shapley values in cooperatibe game theory**[4]

- Distributes the total payoff among the players.
- The unique quantity respecting a set of desired axiomatic properties:
  - Linearity:

$$\phi_{\alpha v+(1-\alpha)w}(i) = \alpha\phi_v(i) + (1-\alpha)\phi_w(i), \tag{1}$$

    for a player $i$ and two games $v$ and $w$, and for $\alpha \in [0,1]$
  - Efficiency,

$$\sum_{i\in\mathcal{C}} \phi_v(i) = v(\mathcal{C}), \tag{2}$$

    (the sum of the value is equal to the total payoff)
  - Symmetry

[4] Lloyd S Shapley et al. "A value for n-person games". In: (1953), pp. 307–317.

# Introduction

**Shapley values in machine learning**

# Introduction

**Shapley values in machine learning**

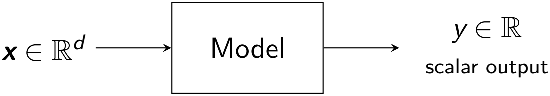
- Features are treated as players, and the scalar output of the model as the payoff,

$$\boldsymbol{x} \in \mathbb{R}^d \longrightarrow \boxed{\text{Model}} \longrightarrow \underset{\text{scalar output}}{y \in \mathbb{R}}$$

# Introduction

**Shapley values in machine learning**

- Features are treated as players, and the scalar output of the model as the payoff,




- Binary classifier, regressor with one-dimensional output ✓

# Introduction

**Shapley values in machine learning**

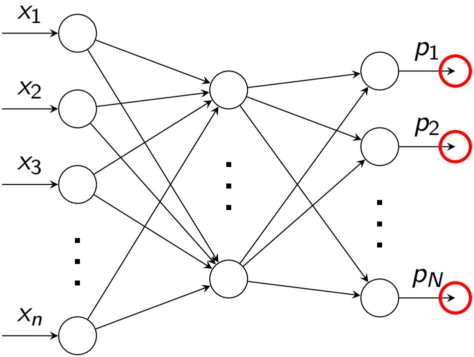
- Features are treated as players, and the scalar output of the model as the payoff,

$$\boldsymbol{x} \in \mathbb{R}^d \longrightarrow \boxed{\text{Model}} \longrightarrow \underset{\text{scalar output}}{y \in \mathbb{R}}$$

- Binary classifier, regressor with one-dimensional output ✓
- Multiclass classifier ✗

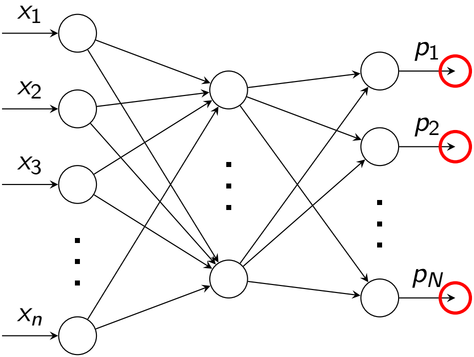
  ex: The output of a softmax layer lives on a multidimensional simplex!

# Introduction




Some explain the output one-by-one,

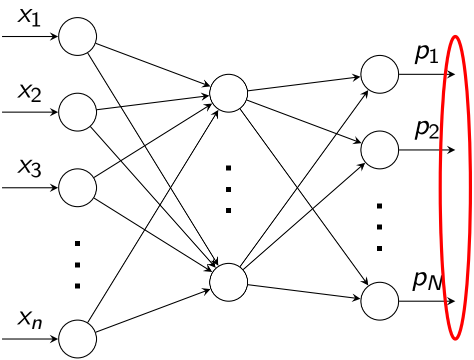
# Introduction




Some explain the output one-by-one,

But a probability distribution lives on a simplex,
**The relative information matter!!**

# Introduction




We will explain the probablities all together using the *Aitchison geometry of the simplex*[5].

[5] John Aitchison. "The Statistical Analysis of Compositional Data". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2 (1982), pp. 139–177, Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

# Outline

# The Shapley values in machine learning

We want to explain a prediction $f(\boldsymbol{x})$ on the instance $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$, where $f \colon \mathcal{X} \to \mathbb{R}$ is the learned model.

$$\boldsymbol{x} \longrightarrow \boxed{f} \longrightarrow \underbrace{f(\boldsymbol{x}) \in \mathbb{R}}_{\text{prediction}}$$

# The Shapley values in machine learning

We want to explain a prediction $f(\boldsymbol{x})$ on the instance $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$, where $f : \mathcal{X} \to \mathbb{R}$ is the learned model.

$$\boldsymbol{x} \longrightarrow \boxed{f} \longrightarrow \underset{\text{prediction}}{f(\boldsymbol{x}) \in \mathbb{R}}$$

Some notation. Let:

- Pr be the probability distribution over $\mathcal{X}$ of the data.

# The Shapley values in machine learning

We want to explain a prediction $f(\boldsymbol{x})$ on the instance $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$, where $f \colon \mathcal{X} \to \mathbb{R}$ is the learned model.

$$\boldsymbol{x} \longrightarrow \boxed{f} \longrightarrow \underset{\text{prediction}}{f(\boldsymbol{x}) \in \mathbb{R}}$$

Some notation. Let:

- Pr be the probability distribution over $\mathcal{X}$ of the data.
- $S \subseteq \mathcal{I} = \{1, 2, \ldots d\}$ be a subset of indices,

# The Shapley values in machine learning

We want to explain a prediction $f(\boldsymbol{x})$ on the instance $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$, where $f\colon \mathcal{X} \to \mathbb{R}$ is the learned model.

$$\boldsymbol{x} \longrightarrow \boxed{f} \longrightarrow \underset{\text{prediction}}{f(\boldsymbol{x}) \in \mathbb{R}}$$

Some notation. Let:

- Pr be the probability distribution over $\mathcal{X}$ of the data.
- $S \subseteq \mathcal{I} = \{1, 2, \ldots d\}$ be a subset of indices,
- $\boldsymbol{x}_S$ refers to an instance $\boldsymbol{x}$ restricted to the features indicated by the indices in $S$.

# The Shapley values in machine learning

The **value function**:

$$\begin{aligned} v_{f,\boldsymbol{x},\mathrm{Pr}} : 2^{\mathcal{I}} &\to \mathbb{R}, \\ S &\mapsto \mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{x}) \mid \boldsymbol{x}_S], \end{aligned} \tag{3}$$

where $\mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{x}) \mid \boldsymbol{x}_S] = \int_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})\mathrm{Pr}(\boldsymbol{x} \mid \boldsymbol{x}_S)d\boldsymbol{x}$.

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 with $\mathcal{S} \neq \mathcal{I}$, there is uncertainty about the non-observed features and we therefore compute the expected prediction given $\boldsymbol{x}_S$.

# The Shapley values in machine learning

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:

$$c_{f,\boldsymbol{x},\mathrm{Pr}}(i, \boldsymbol{X}_S) = v_{f,\boldsymbol{x},\mathrm{Pr}}(\boldsymbol{X}_{S\cup\{i\}}) - v_{f,\boldsymbol{x},\mathrm{Pr}}(\boldsymbol{X}_S), \tag{4}$$

This measures the contribution of the $i$th features with a particular *coalition* of features indexed by $S$.

# The Shapley values in machine learning

The whole contribution of the $i$th feature is computed by averaging this quantity over all possible coalitions of features as follows:

$$\phi_{f,\boldsymbol{x},\mathrm{Pr}}(i) = \frac{1}{d!}\sum_{\pi} c_{f,\boldsymbol{x},\mathrm{Pr}}(i, \pi_{\boldsymbol{X}}^{<i}), \tag{5}$$

where $\pi$ is a permutation of the set $\mathcal{I}$ of indexes and $\pi_{\boldsymbol{X}}^{<i}$ is the features of $\boldsymbol{X}$ coming before the $i$th feature in the ordering given by $\pi$.

# The Shapley values in machine learning

The whole contribution of the $i$th feature is computed by averaging this quantity over all possible coalitions of features as follows:

$$\phi_{f,\boldsymbol{x},\mathrm{Pr}}(i) = \frac{1}{d!} \sum_{\pi} c_{f,\boldsymbol{x},\mathrm{Pr}}(i, \pi_{\boldsymbol{X}}^{<i}), \tag{5}$$

where $\pi$ is a permutation of the set $\mathcal{I}$ of indexes and $\pi_{\boldsymbol{X}}^{<i}$ is the features of $\boldsymbol{X}$ coming before the $i$th feature in the ordering given by $\pi$.

This quantity is known as the **Shapley value** for the $i$th feature.

# The Shapley values in machine learning

It comes from cooperative game theory and is known to be the only quantity respecting a set of desired axiomatic properties[6].

[6] Lloyd S Shapley et al. "A value for n-person games". In: (1953), pp. 307–317.

# The Shapley values in machine learning

It comes from cooperative game theory and is known to be the only quantity respecting a set of desired axiomatic properties[6].

- Linearity with respect to the model ($\alpha, \beta \in \mathbb{R}$): $\phi_{\alpha f+\beta g}(i) = \alpha\phi_f(i) + \beta\phi_g(i)$,

[6] Lloyd S Shapley et al. "A value for n-person games". In: (1953), pp. 307–317.

# The Shapley values in machine learning

It comes from cooperative game theory and is known to be the only quantity respecting a set of desired axiomatic properties[6].

- Linearity with respect to the model ($\alpha, \beta \in \mathbb{R}$): $\phi_{\alpha f+\beta g}(i) = \alpha\phi_f(i) + \beta\phi_g(i)$,
- The "centered" learned model is additively separable with respect to the Shapley values:

$$f(\boldsymbol{x}) - \mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{X})] = \sum_{i=1}^{d} \phi_f(i), \tag{6}$$

which is known as the *efficiency* property.

[6] Lloyd S Shapley et al. "A value for n-person games". In: (1953), pp. 307–317.

# The Shapley values in machine learning

It comes from cooperative game theory and is known to be the only quantity respecting a set of desired axiomatic properties[6].

- Linearity with respect to the model ($\alpha, \beta \in \mathbb{R}$): $\phi_{\alpha f+\beta g}(i) = \alpha\phi_f(i) + \beta\phi_g(i)$,
- The "centered" learned model is additively separable with respect to the Shapley values:

$$f(\boldsymbol{x}) - \mathbb{E}_{\mathrm{Pr}}[f(\boldsymbol{X})] = \sum_{i=1}^{d} \phi_f(i), \tag{6}$$

  which is known as the *efficiency* property.
- Symmetry

[6] Lloyd S Shapley et al. "A value for n-person games". In: (1953), pp. 307–317.

# The Shapley values in machine learning

Example of explanation:

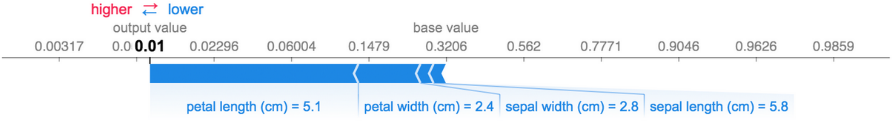
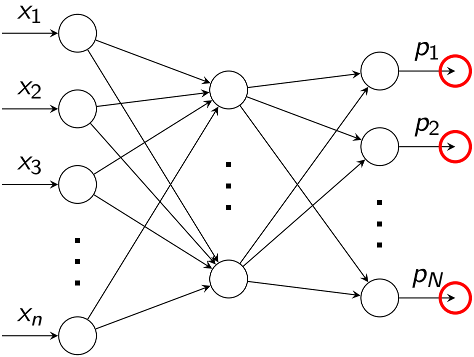



Figure: Explanation of the probability for the class Setosa for a flower from the Iris dataset. The classifier is an SVM with radial basis function and pairwise coupling. Image from `https://github.com/shap/shap/tree/master`.

Note that the Shapley explanation is ran in the *logit* domain!

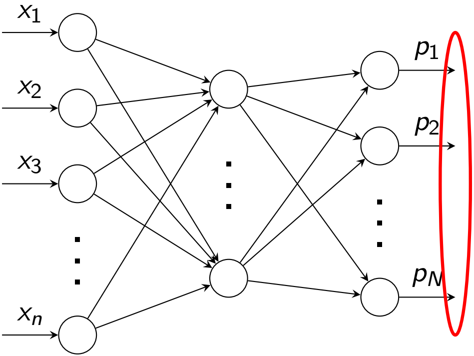
# The Shapley values in machine learning




Some explain the output one-by-one,

But a probability distribution lives on a simplex,
**The relative information matter!!**

# The Shapley values in machine learning




We will explain the probablities all together using the *Aitchison geometry of the simplex*.

# Compositional data analysis

# Shapley composition on the simplex

# Explaining a prediction with Shapley compositions

# Discussion and conclusion

# References I

[1] John Aitchison. "The Statistical Analysis of Compositional Data". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2 (1982), pp. 139–177.

[2] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774.

[3] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. """ Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

# References II


[5] Lloyd S Shapley et al. “A value for n-person games”. In: (1953), pp. 307–317.

[6] Erik Štrumbelj and Igor Kononenko. “Explaining prediction models and individual predictions with feature contributions”. In: *Knowledge and information systems* 41 (2014), pp. 647–665.

# Thank you!!