

Interpretable Machine Learning

SHAP (SHapley Additive exPlanation) Values



Learning goals

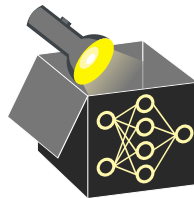
- Get an intuition of additive feature attributions
- Understand the concept of Kernel SHAP
- Ability to interpret SHAP plots
- Global SHAP methods



SHAPLEY VALUES IN ML - A SHORT RECAP

Question: How much does a feat. j contribute to the prediction of a single obs.

Idea: Use Shapley values from cooperative game theory



SHAPLEY VALUES IN ML - A SHORT RECAP

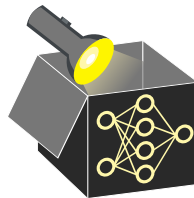
Question: How much does a feat. j contribute to the prediction of a single obs.

Idea: Use Shapley values from cooperative game theory

Procedure:

- Compare “reduced prediction function” of feature coalition S with $S \cup \{j\}$
- Iterate over possible coalitions to calculate marginal contribution of feature j to sample \mathbf{x}

$$\phi_j = \frac{1}{p!} \sum_{\tau \in \Pi} \underbrace{\hat{f}_{S_j^\tau \cup \{j\}}(\mathbf{x}_{S_j^\tau \cup \{j\}}) - \hat{f}_{S_j^\tau}(\mathbf{x}_{S_j^\tau})}_{\text{marginal contribution of feature } j}$$



SHAPLEY VALUES IN ML - A SHORT RECAP

Question: How much does a feat. j contribute to the prediction of a single obs.

Idea: Use Shapley values from cooperative game theory

Procedure:

- Compare “reduced prediction function” of feature coalition S with $S \cup \{j\}$
- Iterate over possible coalitions to calculate marginal contribution of feature j to sample \mathbf{x}

$$\phi_j = \frac{1}{p!} \sum_{\tau \in \Pi} \underbrace{\hat{f}_{S_j^\tau \cup \{j\}}(\mathbf{x}_{S_j^\tau \cup \{j\}}) - \hat{f}_{S_j^\tau}(\mathbf{x}_{S_j^\tau})}_{\text{marginal contribution of feature } j}$$

Remember:

- \hat{f} is the prediction function, p denotes the number of features
- Non-existent feat. in a coalition are replaced by values of random feat. values
- Recall S_j^τ defines the coalition as the set of players before player j in order

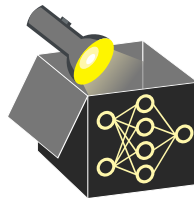
$$\tau = (\tau^{(1)}, \dots, \tau^{(p)})$$

$\tau^{(1)}$...	$\tau^{(S)}$	$\tau^{(S +1)}$	$\tau^{(S +2)}$...	$\tau^{(p)}$
--------------	-----	----------------	------------------	------------------	-----	--------------

S_j^τ : Players before player j

player j

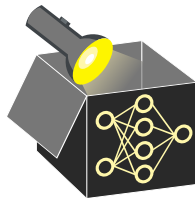
Players after player j



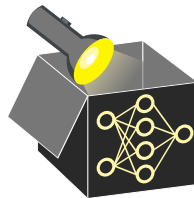
SHAPLEY VALUES IN ML - A SHORT RECAP

Example:

- Train a random forest on bike sharing data only using features humidity (hum), temperature (temp) and windspeed (ws)
- Calculate Shapley value for an observation \mathbf{x} with $\hat{f}(\mathbf{x}) = 2573$
- Mean prediction is $\mathbb{E}(\hat{f}) = 4515$



SHAPLEY VALUES IN ML - A SHORT RECAP



Example:

- Train a random forest on bike sharing data only using features humidity (hum), temperature (temp) and windspeed (ws)
- Calculate Shapley value for an observation \mathbf{x} with $\hat{f}(\mathbf{x}) = 2573$
- Mean prediction is $\mathbb{E}(\hat{f}) = 4515$

Exact Shapley calculation for humidity:

S	$S \cup \{j\}$	\hat{f}_S	$\hat{f}_{S \cup \{j\}}$	weight
\emptyset	hum	4515	4635	2/6
temp	temp, hum	3087	3060	1/6
ws	ws, hum	4359	4450	1/6
temp, ws	hum, temp, ws	2623	2573	2/6

$$\phi_{hum} = \frac{2}{6}(4635 - 4515) + \frac{1}{6}(3060 - 3087) + \frac{1}{6}(4450 - 4359) + \frac{2}{6}(2573 - 2623) = 34$$

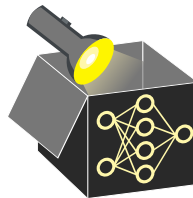
FROM SHAPLEY TO SHAP

Example continued: Same calculation can be done for temperature and windspeed:

- $\phi_{temp} = \dots = -1654$
- $\phi_{ws} = \dots = -323$

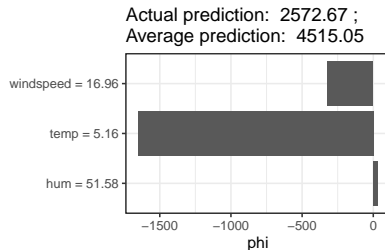
Remember: Shapley values explain difference between actual and average pred.:

$$\begin{aligned} 2573 - 4515 &= 34 - 1654 - 323 = -1942 \\ \hat{f}(\mathbf{x}) - \mathbb{E}(\hat{f}) &= \phi_{hum} + \phi_{temp} + \phi_{ws} \end{aligned}$$



↪ can be rewritten to

$$\hat{f}(\mathbf{x}) = \underbrace{\mathbb{E}(\hat{f})}_{\phi_0} + \phi_{hum} + \phi_{temp} + \phi_{ws}$$



SHAP DEFINITION

► Lundberg et al. 2017

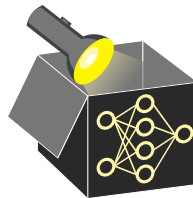
Aim: Find an additive combination that explains the prediction of an observation \mathbf{x} by computing the contribution of each feature to the prediction using a (more efficient) estimation procedure.

Definition

- Simplified (binary) coalition feat. space $\mathbf{Z}' \in \{0, 1\}^{K \times p}$ with K rows and p cols.
- Rows are referred to as $\mathbf{z}'^{(k)} = \{z_1'^{(k)}, \dots, z_p'^{(k)}\}$ with $k \in \{1, \dots, K\}$ (indexes k -th coalition)
- Cols are referred to as \mathbf{z}_j with $j \in \{1, \dots, p\}$ being the index of the original feat.

Example:

Coalition	$\mathbf{z}'^{(k)}$	hum	temp	ws
\emptyset	$\mathbf{z}'^{(1)}$	0	0	0
hum	$\mathbf{z}'^{(2)}$	1	0	0
temp	$\mathbf{z}'^{(3)}$	0	1	0
ws	$\mathbf{z}'^{(4)}$	0	0	1
hum, temp	$\mathbf{z}'^{(5)}$	1	1	0
temp, ws	$\mathbf{z}'^{(6)}$	0	1	1
hum, ws	$\mathbf{z}'^{(7)}$	1	0	1
hum, temp, ws	$\mathbf{z}'^{(8)}$	1	1	1



SHAP DEFINITION


► Lundberg et al. 2017

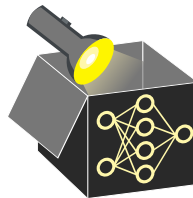
Aim: Find an additive combination that explains the prediction of an observation \mathbf{x} by computing the contribution of each feature to the prediction using a (more efficient) estimation procedure.

Definition

- Simplified (binary) coalition feat. space $\mathbf{Z}' \in \{0, 1\}^{K \times p}$ with K rows and p cols.
- Rows are referred to as $\mathbf{z}'^{(k)} = \{z_1'^{(k)}, \dots, z_p'^{(k)}\}$ with $k \in \{1, \dots, K\}$ (indexes k -th coalition)
- Cols are referred to as \mathbf{z}_j with $j \in \{1, \dots, p\}$ being the index of the original feat.

$\mathbf{z}'^{(k)}$: **Coalition**
simplified features


$$g(\mathbf{z}'^{(k)}) = \phi_0 + \sum_{j=1}^p \phi_j z_j'^{(k)}$$



SHAP DEFINITION

► Lundberg et al. 2017

Aim: Find an additive combination that explains the prediction of an observation \mathbf{x} by computing the contribution of each feature to the prediction using a (more efficient) estimation procedure.

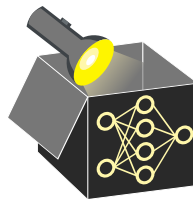
Definition

- Simplified (binary) coalition feat. space $\mathbf{Z}' \in \{0, 1\}^{K \times p}$ with K rows and p cols.
- Rows are referred to as $\mathbf{z}'^{(k)} = \{z_1'^{(k)}, \dots, z_p'^{(k)}\}$ with $k \in \{1, \dots, K\}$ (indexes k -th coalition)
- Cols are referred to as \mathbf{z}_j with $j \in \{1, \dots, p\}$ being the index of the original feat.

$\mathbf{z}'^{(k)}$: **Coalition**
simplified features

$$g(\mathbf{z}'^{(k)}) = \phi_0 + \sum_{j=1}^p \phi_j z_j'^{(k)}$$

ϕ_0 : **Null Output**
Average Model
Baseline ($\mathbb{E}(\hat{f})$)



SHAP DEFINITION

► Lundberg et al. 2017

Aim: Find an additive combination that explains the prediction of an observation \mathbf{x} by computing the contribution of each feature to the prediction using a (more efficient) estimation procedure.

Definition

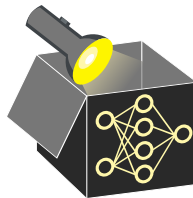
- Simplified (binary) coalition feat. space $\mathbf{Z}' \in \{0, 1\}^{K \times p}$ with K rows and p cols.
- Rows are referred to as $\mathbf{z}'^{(k)} = \{z_1'^{(k)}, \dots, z_p'^{(k)}\}$ with $k \in \{1, \dots, K\}$ (indexes k -th coalition)
- Cols are referred to as \mathbf{z}_j with $j \in \{1, \dots, p\}$ being the index of the original feat.

$\mathbf{z}'^{(k)}$: **Coalition**
simplified features

$$g(\mathbf{z}'^{(k)}) = \phi_0 + \sum_{j=1}^p \phi_j z_j'^{(k)}$$

ϕ_0 : **Null Output**
Average Model
Baseline ($\mathbb{E}(\hat{f})$)

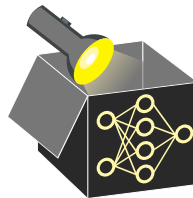
ϕ_j : **Attribution**
How much does
feature j change the
output for coalition k



SHAP DEFINITION

► Lundberg et al. 2017

Aim: Find an additive combination that explains the prediction of an observation \mathbf{x} by computing the contribution of each feature to the prediction using a (more efficient) estimation procedure.



$g(\mathbf{z}'^{(k)})$: **Marginal Contribution**

Contribution of coalition $\mathbf{z}'^{(k)}$ to the prediction

$$g(\mathbf{z}'^{(k)})$$

$$= \underbrace{\phi_0 + \sum_{j=1}^p \phi_j z_j'^{(k)}}_{\text{Additive Feature Attribution}}$$

ϕ_j : **Shapley Values**

Problem

How do we estimate the Shapley values ϕ_j ?

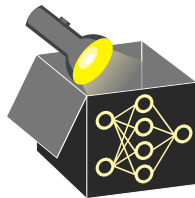
PROPERTIES

Local Accuracy

$$f(\mathbf{x}) = g(\mathbf{x}') = \phi_0 + \sum_{j=1}^p \phi_j x'_j$$

Intuition: If the coalition includes all features ($\mathbf{x}' \in \{1\}^p$), the attributions ϕ_j and the null output ϕ_0 sum up to the original model output $f(\mathbf{x})$

Local accuracy corresponds to the **axiom of efficiency** in Shapley game theory



PROPERTIES

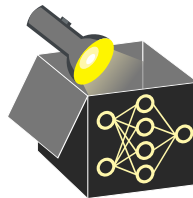
Local Accuracy

$$f(\mathbf{x}) = g(\mathbf{x}') = \phi_0 + \sum_{j=1}^p \phi_j x'_j$$

Missingness

$$x'_j = 0 \implies \phi_j = 0$$

Intuition: A missing feature gets an attribution of zero



PROPERTIES

Local Accuracy

$$f(\mathbf{x}) = g(\mathbf{x}') = \phi_0 + \sum_{j=1}^p \phi_j x'_j$$

Missingness

$$x'_j = 0 \implies \phi_j = 0$$

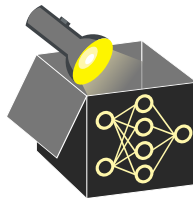
Consistency

$\hat{f}_x(\mathbf{z}'^{(k)}) = \hat{f}(h_x(\mathbf{z}'^{(k)}))$ and $\mathbf{z}'_{-j}^{(k)}$ denote setting $z_j'^{(k)} = 0$. For any two models \hat{f} and \hat{f}' , if

$$\hat{f}'_x(\mathbf{z}'^{(k)}) - \hat{f}'_x(\mathbf{z}'_{-j}^{(k)}) \geq \hat{f}_x(\mathbf{z}'^{(k)}) - \hat{f}_x(\mathbf{z}'_{-j}^{(k)})$$

for all inputs $\mathbf{z}'^{(k)} \in \{0, 1\}^p$, then

$$\phi_j(\hat{f}', \mathbf{x}) \geq \phi_j(\hat{f}, \mathbf{x})$$



PROPERTIES

Local Accuracy

$$f(\mathbf{x}) = g(\mathbf{x}') = \phi_0 + \sum_{j=1}^p \phi_j x'_j$$

Missingness

$$x'_j = 0 \implies \phi_j = 0$$

Consistency

$$\hat{f}'_x(\mathbf{z}'^{(k)}) - \hat{f}'_x(\mathbf{z}'^{(k)}_{-j}) \geq \hat{f}_x(\mathbf{z}'^{(k)}) - \hat{f}_x(\mathbf{z}'^{(k)}_{-j}) \implies \phi_j(\hat{f}', \mathbf{x}) \geq \phi_j(\hat{f}, \mathbf{x})$$

Intuition: If a model changes so that the marginal contribution of a feature value increases or stays the same, the Shapley value also increases or stays the same

From **consistency** the Shapley **axioms of additivity, dummy and symmetry** follow

