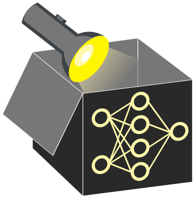
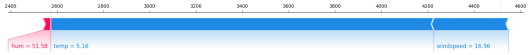
Interpretable Machine Learning

SHAP (SHapley Additive exPlanation) Values



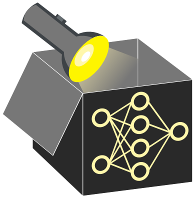


Learning goals

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

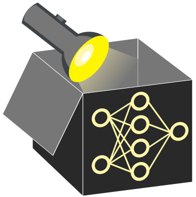
KERNEL SHAP - IN 5 STEPS

Definition: A kernel-based, model-agnostic method to compute Shapley values via local surrogate models (e.g. linear model)



1. Sample coalitions
2. Transfer coalitions into feature space & get predictions by applying ML model
3. Compute weights through kernel
4. Fit a weighted linear model
5. Return Shapley values

KERNEL SHAP - IN 5 STEPS



Step 1: Sample coalitions

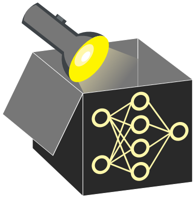
- Sample K coalitions from the simplified feature space

$$\mathbf{z}'^{(k)} \in \{0,1\}^p, \quad k \in \{1,\ldots,K\}$$

- For our simple example, we have in total $2^p = 2^3 = 8$ coalitions (without sampling)

Coalition	$\mathbf{z}'^{(k)}$	hum	temp	ws
$\varnothing$	$\mathbf{z}'^{(1)}$	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

KERNEL SHAP - IN 5 STEPS



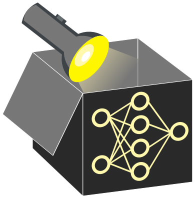
Step 2: Transfer Coalitions into feature space & get predictions by applying ML model

- $\mathbf{z}'^{(k)}$ is 1 if features are are part of the k-th coalition, 0 if they are absent
- To calculate predictions for these coalitions, we need to define a function which maps the binary feature space back to the original feature space

Coalition	$\mathbf{z}'^{(k)}$	hum	temp	ws
$\varnothing$	$\mathbf{z}'^{(1)}$	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

$\mathbf{x}^{coalition}$	hum	temp	ws
$\mathbf{x}^{\{\varnothing\}}$	$\varnothing$	$\varnothing$	$\varnothing$
$\mathbf{x}^{\{hum\}}$	51.6	$\varnothing$	$\varnothing$
$\mathbf{x}^{\{temp\}}$	$\varnothing$	5.1	$\varnothing$
$\mathbf{x}^{\{ws\}}$	$\varnothing$	$\varnothing$	17.0
$\mathbf{x}^{\{hum,temp\}}$	51.6	5.1	$\varnothing$
$\mathbf{x}^{\{temp,ws\}}$	$\varnothing$	5.1	17.0
$\mathbf{x}^{\{hum,ws\}}$	51.6	$\varnothing$	17.0
$\mathbf{x}^{\{hum,temp,ws\}}$	51.6	5.1	17.0

KERNEL SHAP - IN 5 STEPS



Step 2: Transfer Coalitions into feature space & get predictions by applying ML model

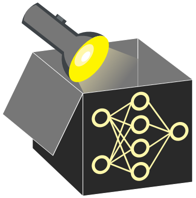
- Define $h_x\left(\mathbf{z}'^{(k)}\right) = \mathbf{z}^{(k)}$ where $h_x : \{0,1\}^p \rightarrow \mathbb{R}^p$ maps 1's to feature values of observation $\mathbf{x}$ for features part of the k-th coalition and 0's to feature values of a randomly sampled observation for features absent in the k-th coalition (feature values are permuted multiple times)
- Predict with ML model on this dataset $\hat{f}$: $\hat{f}\left(h_x\left(\mathbf{z}'^{(k)}\right)\right)$

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

Coalition	$\mathbf{z}'^{(k)}$	hum	temp	ws
$\varnothing$	$\mathbf{z}'^{(1)}$	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

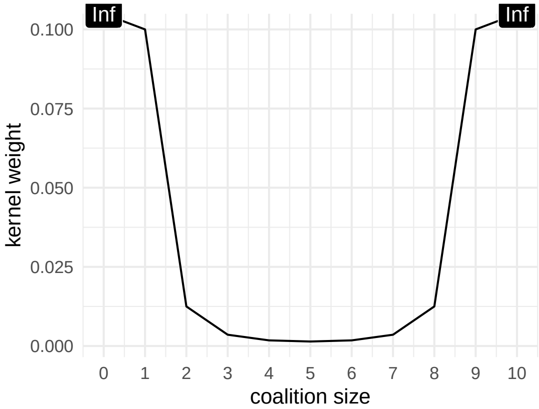
$\mathbf{z}^{(k)}$	hum	temp	ws	$\hat{f}\left(h_x\left(\mathbf{z}'^{(k)}\right)\right)$
$\mathbf{z}^{(1)}$	64.3	28.0	14.5	6211
$\mathbf{z}^{(2)}$	51.6	28.0	14.5	5586
$\mathbf{z}^{(3)}$	64.3	5.1	14.5	3295
$\mathbf{z}^{(4)}$	64.3	28.0	17.0	5762
$\mathbf{z}^{(5)}$	51.6	5.1	14.5	2616
$\mathbf{z}^{(6)}$	64.3	5.1	17.0	2900
$\mathbf{z}^{(7)}$	51.6	28.0	17.0	5411
$\mathbf{z}^{(8)}$	51.6	5.1	17.0	2573

KERNEL SHAP - IN 5 STEPS

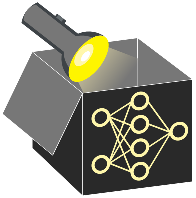


Step 3: Compute weights through Kernel

Intuition: We learn most about individual features if we can study their effects in isolation or at maximal interaction: Small coalitions (few 1's) and large coalitions (i.e. many 1's) get the largest weights



KERNEL SHAP - IN 5 STEPS



Step 3: Compute weights through Kernel ▸ see shapley_kernel_proof.pdf

Intuition: We learn most about individual features if we can study their effects in isolation or at maximal interaction: Small coalitions (few 1's) and large coalitions (i.e. many 1's) get the largest weights

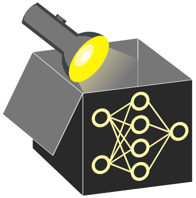
$\pi_x(\mathbf{z}'^{(k)})$: kernel weight for coalition $\mathbf{z}'^{(k)}$

p: Number of features in $\mathbf{x}$

$$\pi_x\left(\mathbf{z}'^{(k)}\right) = \frac{(p-1)}{\binom{p}{\left|\mathbf{z}'^{(k)}\right|}\left|\mathbf{z}'^{(k)}\right|\left(p - \left|\mathbf{z}'^{(k)}\right|\right)}$$

$|\ \mathbf{z}'^{(k)}\ |$: coalition size / sum of 1s in $\mathbf{z}'^{(k)}$

KERNEL SHAP - IN 5 STEPS



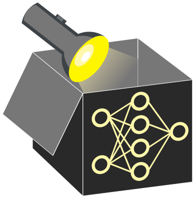
Step 3: Compute weights through Kernel

Purpose: to include this knowledge in the local surrogate model (linear regression), we calculate weights for each coalition which are the observations of the linear regression

$$\pi_x\left(\mathbf{z}'\right) = \frac{(p-1)}{\binom{p}{|\mathbf{z}'|}|\mathbf{z}'|(p-|\mathbf{z}'|)} \rightsquigarrow \pi_x\left(\mathbf{z}' = (1,0,0)\right) = \frac{(3-1)}{\binom{3}{1}1\,(3-1)} = \frac{1}{3}$$

Coalition	$\mathbf{z}'^{(k)}$	hum	temp	ws	weight
$\varnothing$	$\mathbf{z}'^{(1)}$	0	0	0	∞
hum	$\mathbf{z}'^{(2)}$	1	0	0	0.33
temp	$\mathbf{z}'^{(3)}$	0	1	0	0.33
ws	$\mathbf{z}'^{(4)}$	0	0	1	0.33
hum, temp	$\mathbf{z}'^{(5)}$	1	1	0	0.33
temp, ws	$\mathbf{z}'^{(6)}$	0	1	1	0.33
hum, ws	$\mathbf{z}'^{(7)}$	1	0	1	0.33
hum, temp, ws	$\mathbf{z}'^{(8)}$	1	1	1	∞

KERNEL SHAP - IN 5 STEPS



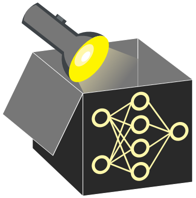
Step 3: Compute weights through Kernel

Purpose: to include this knowledge in the local surrogate model (linear regression), we calculate weights for each coalition which are the observations of the linear regression

Coalition	$\mathbf{z}'^{(k)}$	hum	temp	ws	weight
$\varnothing$	$\mathbf{z}'^{(1)}$	0	0	0	∞
hum	$\mathbf{z}'^{(2)}$	1	0	0	0.33
temp	$\mathbf{z}'^{(3)}$	0	1	0	0.33
ws	$\mathbf{z}'^{(4)}$	0	0	1	0.33
hum, temp	$\mathbf{z}'^{(5)}$	1	1	0	0.33
temp, ws	$\mathbf{z}'^{(6)}$	0	1	1	0.33
hum, ws	$\mathbf{z}'^{(7)}$	1	0	1	0.33
hum, temp, ws	$\mathbf{z}'^{(8)}$	1	1	1	∞

$\rightsquigarrow$ weights for empty and full set are infinity and not used as observations for the linear regression
$\rightsquigarrow$ instead constraints are used such that properties (local accuracy and missingness) are satisfied

KERNEL SHAP - IN 5 STEPS



Step 4: Fit a weighted linear model

Aim: Estimate a weighted linear model with Shapley values being the coefficients ϕ_j

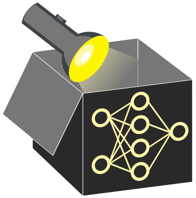
$$g\left(\mathbf{z}'^{(k)}\right) = \phi_0 + \sum_{j=1}^{p} \phi_j z_j'^{(k)}$$

and minimize by WLS using the weights π_x of step 3

$$L\left(\hat{f}, g, \pi_x\right) = \sum_{k=1}^{K} \left[\hat{f}\left(h_x\left(\mathbf{z}'^{(k)}\right)\right) - g\left(\mathbf{z}'^{(k)}\right)\right]^2 \pi_x\left(\mathbf{z}'^{(k)}\right)$$

with $\phi_0 = \mathbb{E}(\hat{f})$ and $\phi_p = \hat{f}(x) - \sum_{j=0}^{p-1} \phi_j$ we receive a $p-1$ dimensional linear regression problem

KERNEL SHAP - IN 5 STEPS



Step 4: Fit a weighted linear model

Aim: Estimate a weighted linear model with Shapley values being the coefficients ϕ_j

$$g\left(\mathbf{z}'^{(k)}\right) = \phi_0 + \sum_{j=1}^{p} \phi_j z_j'^{(k)} \rightsquigarrow g\left(\mathbf{z}'^{(k)}\right) = 4515 + 34 \cdot z_1'^{(k)} - 1654 \cdot z_2'^{(k)} - 323 \cdot z_3'^{(k)}$$

$\mathbf{z}'^{(k)}$	hum	temp	ws	weight	$\hat{f}$
$\mathbf{z}'^{(2)}$	1	0	0	0.33	4635
$\mathbf{z}'^{(3)}$	0	1	0	0.33	3087
$\mathbf{z}'^{(4)}$	0	0	1	0.33	4359
$\mathbf{z}'^{(5)}$	1	1	0	0.33	3060
$\mathbf{z}'^{(6)}$	0	1	1	0.33	2623
$\mathbf{z}'^{(7)}$	1	0	1	0.33	4450

input *output*

KERNEL SHAP - IN 5 STEPS

Step 5: Return SHAP values

Intuition: Estimated Kernel SHAP values are equivalent to Shapley values

$$
\begin{aligned}
g(\mathbf{z}'^{(8)}) &= \hat{f}(h_x(\mathbf{z}'^{(8)})) = 4515 + 34 \cdot 1 - 1654 \cdot 1 - 323 \cdot 1 \\
&= \underbrace{\mathbb{E}(\hat{f})}_{\phi_0} + \phi_{hum} + \phi_{temp} + \phi_{ws} = \hat{f}(\mathbf{x}) = 2573
\end{aligned}
$$