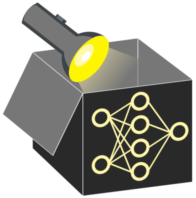
Interpretable Machine Learning

SHAP (SHapley Additive exPlanation) Values

Learning goals

- Understand KernelSHAP as weighted least-squares regression over coalitions
- Grasp how background samples impute "absent" features
- Observational vs. interventional SHAP

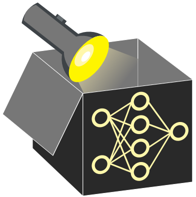
KERNEL SHAP - IN 5 STEPS



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

1. Sample coalition vectors $\mathbf{z}' \in \{0, 1\}^p$
2. Map coalition vectors to original feature space and predict
3. Compute kernel weights for surrogate model
4. Fit a weighted linear model
5. Return Shapley values

KERNEL SHAP - IN 5 STEPS



Step 1: Sample coalition vectors

- Sample K coalitions from the simplified (binary) feature space

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

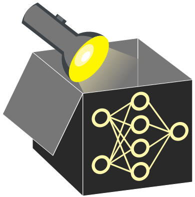
- $\mathbf{z}'^{(k)} \in \{0,1\}^p$ indicates which features are present in k-th coalition
- To evaluate the model on each coalition, we must map $\mathbf{z}'^{(k)}$ to original space
- Example ($\mathbf{x} = (51.6, 5.1, 17.0)) \Rightarrow 2^p = 2^3 = 8$ coalitions (without sampling)

Map to 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

$\mathbf{z}^{(k)}$	hum	temp	ws
$\mathbf{z}^{(1)}$	?	?	?
$\mathbf{z}^{(2)}$	51.6	?	?
$\mathbf{z}^{(3)}$	?	5.1	?
$\mathbf{z}^{(4)}$	?	?	17.0
$\mathbf{z}^{(5)}$	51.6	5.1	?
$\mathbf{z}^{(6)}$	?	5.1	17.0
$\mathbf{z}^{(7)}$	51.6	?	17.0
$\mathbf{z}^{(8)}$	51.6	5.1	17.0

KERNEL SHAP – IN 5 STEPS



Step 2: Map coalition vectors to original feature space and predict

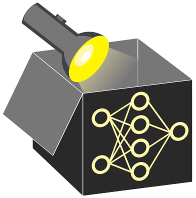
- Define mapping $h_{\mathbf{x},\mathbf{x}'} : \{0,1\}^p \rightarrow \mathbb{R}^p$, where: $(h_{\mathbf{x},\mathbf{x}'}(\mathbf{z}'))_j = \begin{cases} x_j & \text{if } z'_j = 1 \\ x'_j & \text{if } z'_j = 0 \end{cases}$
- Construct $\mathbf{z} = h_{\mathbf{x},\mathbf{x}'}(\mathbf{z}')$ where present features take their values from $\mathbf{x}$ and absent features are imputed with values from a random background sample $\mathbf{x}' = (64.3, 28.0, 14.5)$
- Evaluate the model on each constructed vector: $\hat{f} = \hat{f}(h_{\mathbf{x},\mathbf{x}'}(\mathbf{z}'^{(k)}))$

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

$\xrightarrow{h_{\mathbf{x},\mathbf{x}'}(\mathbf{z}'^{(k)})}$

$\mathbf{z}^{(k)}$	hum	temp	ws	$\hat{f}(h_{\mathbf{x}}(\mathbf{z}'^{(k)}))$
$\mathbf{z}^{(1)}$	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 2: Map coalition vectors to original feature space and predict

Fix coalition vector $\mathbf{z}' = (1, 0, 0)$; draw multiple background samples $\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(B)}$
$\Rightarrow$ keep **hum**, replace **temp** and **ws** by draws from the background data.

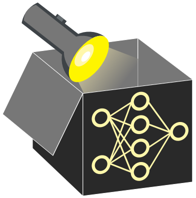
Sample b	hum (from $\mathbf{x}$)	temp (from $\mathbf{x}'^{(b)}$)	ws (from $\mathbf{x}'^{(b)}$)	$\hat{f}(h_{\mathbf{x},\mathbf{x}'^{(b)}}(\mathbf{z}'))$
1	51.6	28.0	14.5	4635
2	51.6	5.1	14.5	3295
3	51.6	28.0	17.0	5586
$\vdots$		$\ldots$		$\ldots$

- Typically, many background samples $\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(B)}$ are used to approximate the marginal expectation required for KernelSHAP via Monte-Carlo average:

$$\mathbb{E}_{\mathbf{X}_{-S}}\left[f(\mathbf{x}_S, \mathbf{X}_{-S})\right] \approx \tfrac{1}{B} \textstyle\sum_{b=1}^{B} \hat{f}(h_{\mathbf{x},\mathbf{x}'^{(b)}}(\mathbf{z}'))$$

- Background samples $\mathbf{x}'^{(b)}$ are drawn from:
  - Conditional distribution $\mathbf{x}'^{(b)} \sim P_{\mathbf{X}|\mathbf{X}_S=\mathbf{x}_S} \rightsquigarrow$ **Observational SHAP**
  - Marginal distribution $\mathbf{x}'^{(b)} \sim P_{\mathbf{X}} \rightsquigarrow$ **Interventional SHAP**
- The same procedure applies to every other coalition vector $\mathbf{z}'^{(k)}$.

KERNEL SHAP - IN 5 STEPS



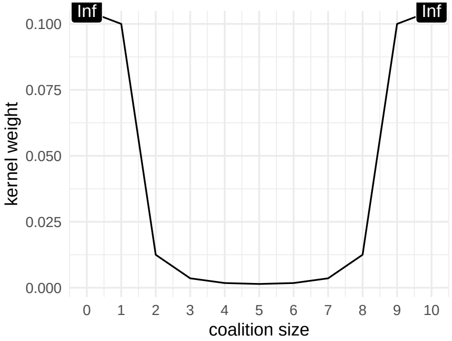
Step 3: Compute kernel weights for surrogate model

Intuition: We learn most about a feature's effect when (recall multinomial coefficient in Shapley value's set definition):

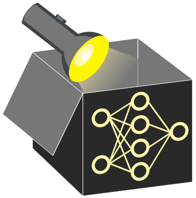
- it appears **in isolation** (small coalition), or
- in **near-complete context** (large coalition).

$\Rightarrow$ SHAP assigns highest weights to very small and very large coalitions.

Note: The figure below is illustrative and not tied to the running example.



KERNEL SHAP - IN 5 STEPS



Step 3: Compute kernel weights for surrogate model

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

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

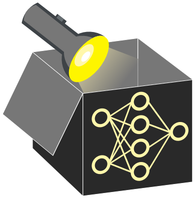
$$\pi_x\left(\mathbf{z}'^{(k)}\right) = \frac{(p-1)}{\binom{p}{\left|\mathbf{z}'^{(k)}\right|}\left|\mathbf{z}'^{(k)}\right|\left(p - \left|\mathbf{z}'^{(k)}\right|\right)}$$

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

Note: Weights differ from multinomial coefficient in the Shapley value set-definiton but are constructed to yield the same Shapley values via weighted linear regression.

▸ see shapley_kernel_proof.pdf

KERNEL SHAP - IN 5 STEPS



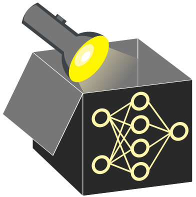
Step 3: Compute kernel weights for surrogate model

Purpose: Assign observation weights $\pi_x(\mathbf{z}')$ to each coalition vector $\mathbf{z}'$ when solving the local surrogate (weighted linear regression), e.g.:

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

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

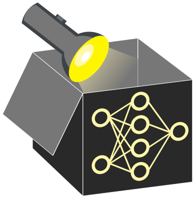


Step 3: Compute kernel weights for surrogate model

- For $p > 3$ features, the finite weights are all 0.33 as every shown coalition has the same size ($|S| = 1$ and $|-S| = 2$ and vice versa for $p = 3$).
- In general (when $p > 3$), weights vary with coalition size.
- Empty and full coalitions receive weight ∞ (division-by-zero term)
  $\rightsquigarrow$ These coalition vectors are not used as observations for the linear regression
  $\rightsquigarrow$ Instead constraints are used to ensure *local accuracy* and *missingness*

Coalition	$\mathbf{z}'^{(k)}$	hum	temp	ws	weight $\pi_x(\mathbf{z}')$
$\emptyset$	$\mathbf{z}'^{(1)}$	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 4: Fit a weighted linear model

Goal Estimate Shapley values ϕ_j as coefficients of a local, weighted linear surrogate.

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

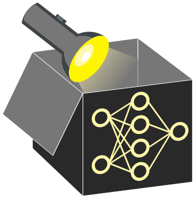
Weighted least-squares objective

$$\min_{\phi} \; \sum_{k=1}^{K} \pi_x(\mathbf{z}'^{(k)}) \left[\hat{f}\left(h_{\mathbf{x}}(\mathbf{z}'^{(k)})\right) - g(\mathbf{z}'^{(k)})\right]^2$$

Boundary coalitions ($\mathbf{z}' = \mathbf{1}$ and $\mathbf{z}' = \mathbf{0}$) enforce constraints on coefficients

$$\phi_0 = \mathbb{E}[\hat{f}(\mathbf{X})], \qquad \sum_{j=1}^{p} \phi_j = \hat{f}(\mathbf{x}) - \phi_0.$$

KERNEL SHAP - IN 5 STEPS



Step 4: Fit a weighted linear model

Goal Estimate Shapley values ϕ_j as coefficients of a local, weighted linear surrogate.

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

Numeric illustration ($p = 3$)

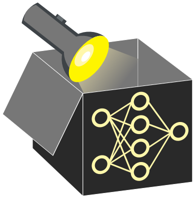
$$g(\mathbf{z}') = 4515 + 34\, z'_1 - 1654\, z'_2 - 323\, z'_3$$

$\mathbf{z}'$	hum	temp	ws	weight $\pi_x(\mathbf{z}')$	$\hat{f}(h_{\mathbf{x}}(\mathbf{z}'))$	$g(\mathbf{z}')$
$(1, 0, 0)$	1	0	0	0.33	4635	4549
$(0, 1, 0)$	0	1	0	0.33	3087	2861
$(0, 0, 1)$	0	0	1	0.33	4359	4192
$(1, 1, 0)$	1	1	0	0.33	3060	2895
$(0, 1, 1)$	0	1	1	0.33	2623	2538
$(1, 0, 1)$	1	0	1	0.33	4450	4226

inputs outputs

The inputs and outputs are used to learn the weighted linear regression model.

KERNEL SHAP - IN 5 STEPS



Step 5: Return SHAP values

Intuition: Estimated Kernel SHAP values are equivalent to Shapley values

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