Interpretable Machine Learning

Shapley Additive Global Importance (SAGE)

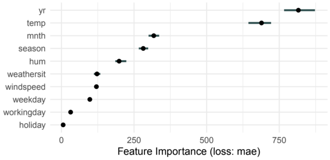
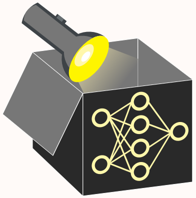


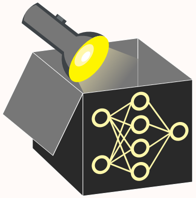
Figure: Bike Sharing Dataset

Learning goals

- How SAGE fairly distributes importance
- Definition of SAGE value function
- Difference SAGE value function and SAGE values
- Marginal and Conditional SAGE

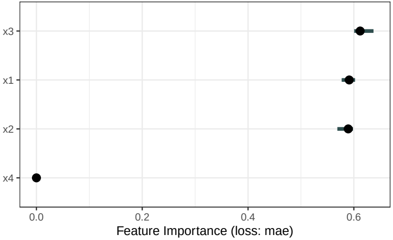


CHALLENGE: FAIR ATTRIBUTION OF IMPORTANCE



Recap:

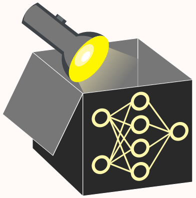
- Data: $x_1, \dots, x_4$ uniformly sampled from $[-1, 1]$
- DGP: $y := x_1 x_2 + x_3 + \epsilon_Y$ with $\epsilon_Y \sim N(0, 1)$
- Model: $\hat{f}(x) \approx x_1 x_2 + x_3$



Although x_3 alone contributes as much to the prediction as x_1 and x_2 jointly, all three are considered equally relevant by PFI.

Reason: PFI assesses importance given that all remaining features are preserved. If we first permute x_1 and then x_2, permutation of x_2 would have no effect on the performance (and vice versa).

SAGE IDEA ▸ Covert et al. (2020)



SAGE: Use Shapley values to compute a fair attribution of importance (via model performance)

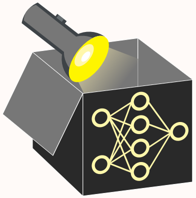
Idea:

- Feature importance attribution can be regarded as cooperative game
  $\rightsquigarrow$ features jointly contribute to achieve a certain model performance
- Players: features
- Payoff to be fairly distributed: model performance
- Surplus contribution of a feature depends on the coalition of features that are already accessible by the model

Note:

- Same idea (called SFIMP) was proposed in ▸ Casalicchio et al. (2018)
- Definition based on model refits was proposed in context of feature selection in ▸ Cohen et al. (2007)

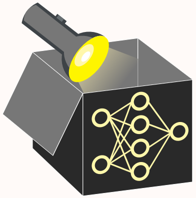
SAGE - VALUE FUNCTION



Removal Idea: To deprive information of the non-coalition features $-S$ from the model, marginalize the prediction function over the features $-S$ to be "dropped".

$$\hat{f}_S(x_S) = \mathbb{E}[\hat{f}(x)|X_S = x_S]$$

SAGE - VALUE FUNCTION



Removal Idea: To deprive information of the non-coalition features $-S$ from the model, marginalize the prediction function over the features $-S$ to be "dropped".

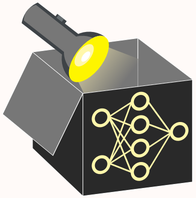
$$\hat{f}_S(x_S) = \mathbb{E}[\hat{f}(x)|X_S = x_S]$$

SAGE value function:

$$v_{\hat{f}}(S) = \mathcal{R}(\hat{f}_\emptyset) - \mathcal{R}(\hat{f}_S), \text{ where } \mathcal{R}(\hat{f}_S) = \mathbb{E}_{Y,X_S}[L(y, \hat{f}_S(x_S))]$$

$\rightsquigarrow$ Quantify the predictive power of a coalition S in terms of reduction in risk
$\rightsquigarrow$ Risk of predictor $\hat{f}_S(x_S)$ is compared to the risk of the mean prediction $\hat{f}_\emptyset$

SAGE - VALUE FUNCTION



Removal Idea: To deprive information of the non-coalition features $-S$ from the model, marginalize the prediction function over the features $-S$ to be "dropped".

$$\hat{f}_S(x_S) = \mathbb{E}[\hat{f}(x)|X_S = x_S]$$

SAGE value function:

$$v_{\hat{f}}(S) = \mathcal{R}(\hat{f}_\emptyset) - \mathcal{R}(\hat{f}_S), \text{ where } \mathcal{R}(\hat{f}_S) = \mathbb{E}_{Y,X_S}[L(y, \hat{f}_S(x_S))]$$

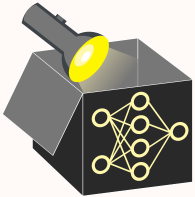
$\leadsto$ Quantify the predictive power of a coalition S in terms of reduction in risk
$\leadsto$ Risk of predictor $\hat{f}_S(x_S)$ is compared to the risk of the mean prediction $\hat{f}_\emptyset$

Surplus contribution of feature x_j over coalition x_S:

$$v_{\hat{f}}(S \cup \{j\}) - v_{\hat{f}}(S) = \mathcal{R}(\hat{f}_S) - \mathcal{R}(\hat{f}_{S\cup\{j\}})$$

$\leadsto$ Quantifies the added value of feature j when it is added to coalition S

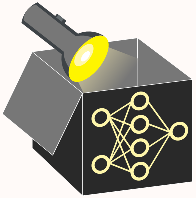
SAGE - MARGINAL AND CONDITIONAL SAMPLING



When computing the marginalized prediction $\hat{f}_S(x_S)$, the “dropped” features can be sampled from

- the marginal distribution $\mathbb{P}(x_{-S}) \Rightarrow$ marginal SAGE
- the conditional distribution $\mathbb{P}(x_{-S}|x_S) \Rightarrow$ conditional SAGE

SAGE - MARGINAL AND CONDITIONAL SAMPLING



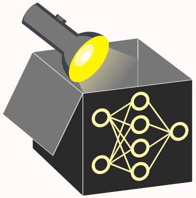
When computing the marginalized prediction $\hat{f}_S(x_S)$, the “dropped” features can be sampled from

- the marginal distribution $\mathbb{P}(x_{-S}) \Rightarrow$ marginal SAGE
- the conditional distribution $\mathbb{P}(x_{-S}|x_S) \Rightarrow$ conditional SAGE

Interpretation marginal sampling: $v(S)$ quantifies the reliance of the model on features x_S

- features x_S not being causal for the prediction $\Rightarrow v(S) = 0$

SAGE - MARGINAL AND CONDITIONAL SAMPLING



When computing the marginalized prediction $\hat{f}_S(x_S)$, the “dropped” features can be sampled from

- the marginal distribution $\mathbb{P}(x_{-S}) \Rightarrow$ marginal SAGE
- the conditional distribution $\mathbb{P}(x_{-S}|x_S) \Rightarrow$ conditional SAGE

Interpretation marginal sampling: $v(S)$ quantifies the reliance of the model on features x_S

- features x_S not being causal for the prediction $\Rightarrow v(S) = 0$

Interpretation conditional sampling: $v(S)$ quantifies whether variables x_S contains prediction-relevant information (e.g. $y \not\perp x_S$) that is (directly or indirectly) exploited by the model

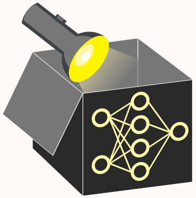
- features x_S not being causal for the prediction $\not\Rightarrow v(S) = 0$
  - e.g., if x_1 and x_2 are perfectly correlated, even if only x_1 has a nonzero coefficient, both are considered equally important
- under model optimality, links to mutual information or the conditional variance exist

SAGE - MARGINAL AND CONDITIONAL SAMPLING

Example:

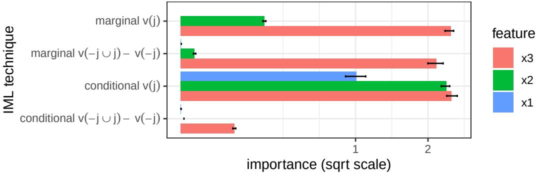
- $y = x_3 + \epsilon_y$
  $x_1 = \epsilon_1$
  $x_2 = x_1 + \epsilon_2$
  $x_3 = x_2 + \epsilon_3$ (all ϵ_j i.i.d.)
- Causal DAG:
  $x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow y$
- Fitted LM:
  $\hat{f} \approx 0.95 x_3 + 0.05 x_2$

SAGE - MARGINAL AND CONDITIONAL SAMPLING



Example:

- $y = x_3 + \epsilon_y$
  $x_1 = \epsilon_1$
  $x_2 = x_1 + \epsilon_2$
  $x_3 = x_2 + \epsilon_3$ (all ϵ_j i.i.d.)
- Causal DAG:
  $x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow y$
- Fitted LM:
  $\hat{f} \approx 0.95x_3 + 0.05x_2$

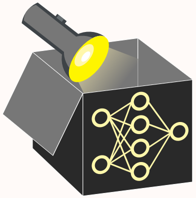


- Marginal $v(j)$ are only nonzero for features that are used by $\hat{f}$
- Conditional $v(j)$ are also nonzero for features that are not used by $\hat{f}$ (e.g., due to correlation)
- For conditional value function v, the difference $v(-j \cup j) - v(-j)$ quantifies the unique contribution of x_j over remaining features x_{-j}
  $\Rightarrow$ Since $y \perp x_1, x_2 | x_3$, only $v(\{1,2,3\}) - v(\{1,2\})$ is nonzero (i.e., for feature $j = 3$)

SAGE VALUE FUNCTIONS VERSUS SAGE VALUES

SAGE value function $v(S)$: measure contribution of a specific feature set over the empty coalition

SAGE VALUE FUNCTIONS VERSUS SAGE VALUES

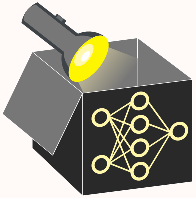


SAGE value function $v(S)$: measure contribution of a specific feature set over the empty coalition

SAGE values ϕ_j: fair attribution of importance

- can be computed by averaging the contribution of x_j over all feature orderings
- for feature permutation τ, the contribution of j in the set S_j^τ is given as
  $v(S_j^\tau \cup \{j\}) - v(S_j^\tau)$
  Note: S_j^τ is the set of features preceding j in permutation τ

SAGE VALUE FUNCTIONS VERSUS SAGE VALUES



SAGE value function $v(S)$: measure contribution of a specific feature set over the empty coalition

SAGE values ϕ_j: fair attribution of importance

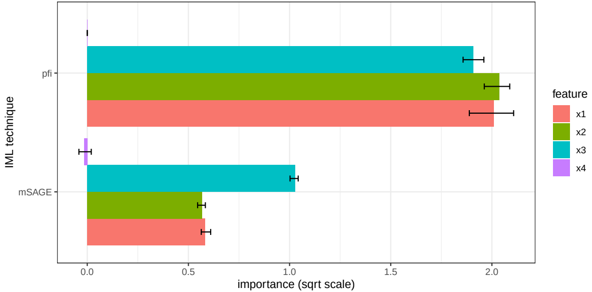
- can be computed by averaging the contribution of x_j over all feature orderings
- for feature permutation τ, the contribution of j in the set S_j^τ is given as
  $v(S_j^\tau \cup \{j\}) - v(S_j^\tau)$
  Note: S_j^τ is the set of features preceding j in permutation τ

SAGE value approximation: Average over the contributions for M randomly sampled permutations

$$\phi_j = \frac{1}{M} \sum_{m=1}^{M} v(S_j^\tau \cup \{j\}) - v(S_j^\tau)$$

INTERACTION EXAMPLE REVISITED

Recap: Data: $x_1, \ldots, x_4$ uniformly sampled from $\{-1, 1\}$ and $y := x_1 x_2 + x_3 + \epsilon_Y$ with $\epsilon_Y \sim N(0, 1)$. Model: $\hat{f}(x) \approx x_1 x_2 + x_3$.

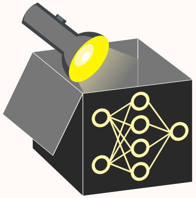


- PFI regards x_1, x_2 to be equally important as x_3
- Marginal SAGE fairly divides the contribution of the interaction x_1 and x_2

SAGE LOSS FUNCTIONS

When the loss-optimal model f^* is inspected using *conditional-sampling* based SAGE value functions, interesting links exist.

SAGE LOSS FUNCTIONS



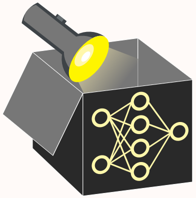
When the loss-optimal model f^* is inspected using *conditional-sampling* based SAGE value functions, interesting links exist.

For cross-entropy loss:

- value function is the mutual information: $v_{f^*}(S) = I(y; x_S)$
- surplus contribution of a feature x_j is the conditional mutual information: $v_{f^*}(S \cup \{j\}) - v_{f^*}(S) = I(y, x_i | x_S)$

SAGE LOSS FUNCTIONS

When the loss-optimal model f^* is inspected using *conditional-sampling* based SAGE value functions, interesting links exist.



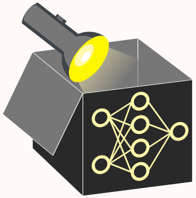
For cross-entropy loss:

- value function is the mutual information: $v_{f^*}(S) = I(y; x_S)$
- surplus contribution of a feature x_j is the conditional mutual information: $v_{f^*}(S \cup \{j\}) - v_{f^*}(S) = I(y, x_i | x_S)$

For MSE loss:

- value function is the expected reduction in variance given knowledge of the features x_S: $v_{f^*}(S) = Var(y) - \mathbb{E}[Var(y|x_S)]$
- surplus contribution is the respective reduction over x_S: $v_{f^*}(S \cup \{j\}) - v_{f^*}(S) = \mathbb{E}[Var(y|x_S)] - \mathbb{E}[Var(y|x_{S \cup j})]$

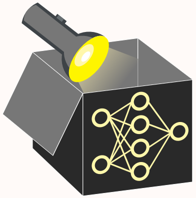
IMPLICATIONS MARGINAL SAGE VALUES



Can we gain insight into whether the ...

1. feature x_j is causal for the prediction?
   - for all coalitions S, $v(j \cup S) - v(S)$ can only be nonzero if $x_j \rightarrow \hat{f}(x)$ (as for PFI)
     $\rightsquigarrow$ ϕ_j is only nonzero if x_j is causal for the prediction
   - $v(j \cup S) - v(S)$ may be zero due to independence $x_j \perp y | x_S$ (as for PFI)
     $\rightsquigarrow$ ϕ_j may be zero although the feature is causal for the prediction

IMPLICATIONS MARGINAL SAGE VALUES

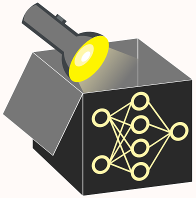


Can we gain insight into whether the ...

2. feature x_j contains prediction-relevant information about y?
   - value functions may be nonzero despite independence due to extrapolation (as for PFI)
     $\rightsquigarrow$ ϕ_j may be nonzero without x_j being dependent with y
   - value functions may be zero despite x_j containing prediction-relevant information due to underfitting (as for PFI)
     $\rightsquigarrow$ ϕ_j may be zero although prediction-relevant information contained

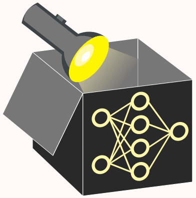
IMPLICATIONS MARGINAL SAGE VALUES

Can we gain insight into whether the ...



3. model requires access to x_j to achieve it's prediction performance?
   - like PFI, in general marginal value functions do not allow insight into unique contribution $\rightsquigarrow$ no insight from ϕ_j

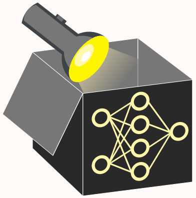
IMPLICATIONS CONDITIONAL SAGE VALUES



Can we gain insight into whether the ...

1. feature $\mathbf{x}_j$ is causal for the prediction?
    - value functions may be nonzero although feature is not directly used by the model
      $\leadsto$ nonzero ϕ_j does not imply $\mathbf{x}_j \rightarrow \hat{y}$
    - value functions may be zero although feature may be used by the model, e.g. if feature is independent with y and all other features $\leadsto$ zero ϕ_j does not imply $\mathbf{x}_j \not\rightarrow \hat{y}$

IMPLICATIONS CONDITIONAL SAGE VALUES

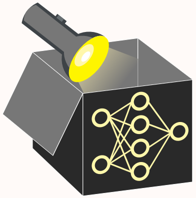


Can we gain insight into whether the ...

2. feature $\mathbf{x}_j$ contains prediction-relevant information about y?
   - e.g. for cross-entropy optimal $\hat{f}$, $v(j)$ measures mutual information $I(y; x_j)$
     $\rightsquigarrow$ prediction-relevance implies nonzero ϕ_j
   - $x_j \perp y$ does not imply $x_j \perp y | x_S$ and consequently does not imply $v(j \cup S) - v(S) = 0$
     $\rightsquigarrow$ ϕ_j may be nonzero although $\mathbf{x}_j \perp y$

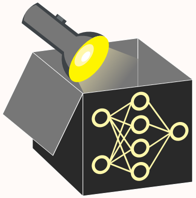
IMPLICATIONS CONDITIONAL SAGE VALUES

Can we gain insight into whether the ...



3. model requires access to x_j to achieve it's prediction performance?
   - e.g. for cross-entropy optimal $\hat{f}$, the surplus contribution $v(j \cup -j) - v(-j)$ captures the conditional mutual information $I(y; x_j | x_{-j})$
     $\rightsquigarrow \phi_j$ is nonzero for features with unique contribution
   - $x_j \perp y | x_{-j}$ does not imply conditional independence w.r.t. to arbitrary coalitions $x_j \perp y | x_S$

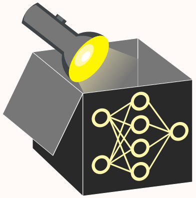
DEEP DIVE: SHAPLEY AXIOMS FOR SAGE



The Shapley axioms can be translated into properties of SAGE. The interpretation depends on whether conditional or marginal sampling is used.

Shapley property $\Longrightarrow$	conditional SAGE property
efficiency	$\sum_{i=1}^{p} \phi_j(v) = \mathcal{R}(\hat{f}_\emptyset) - \mathcal{R}(\hat{f})$
symmetry	$x_j = x_i \implies \phi_i = \phi_j$
linearity	ϕ_j expecation of per-instance conditional SHAP applied to model loss
monotonicity	given models f, f', if $\forall S$: $v_f(S \cup j) - v_f(S) \geq v_{f'}(S \cup j) - v_{f'}(S)$ then $\phi_j(v_f) \geq \phi_j(v_{f'})$
dummy	if $\forall S : \hat{f}(x) \perp x_j \vert x_S \Rightarrow \phi_j = 0$

DEEP DIVE: SHAPLEY AXIOMS FOR SAGE



The Shapley axioms can be translated into properties of SAGE. The interpretation depends on whether conditional or marginal sampling is used.

Shapley property $\Longrightarrow$	marginal SAGE property
efficiency	$\sum_{i=1}^{p} \phi_j(v) = \mathcal{R}(\hat{f}_\emptyset) - \mathcal{R}(\hat{f})$
symmetry	no intelligible implication
linearity	ϕ_j expecation of per-instance
	marginal SHAP applied to model loss
monotonicity	given models f, f', if $\forall S$:
	$v_f(S \cup j) - v_f(S) \geq v_{f'}(S \cup j) - v_{f'}(S)$
	then $\phi_j(v_f) \geq \phi_j(v_{f'})$
dummy	model invariant to $x_j \Rightarrow \phi_j = 0$