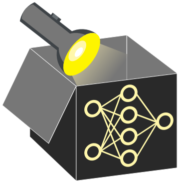
Interpretable Machine Learning



Feature Importance

Conditional Feature Importance (CFI)

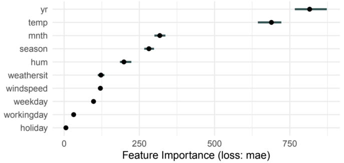
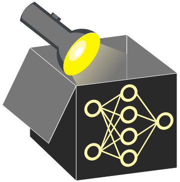


Figure: Bike Sharing Dataset

Learning goals

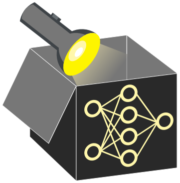
- Extrapolation and Conditional Sampling
- Conditional Feature Importance (CFI)
- Interpretation of CFI and difference to PFI

CFI MOTIVATION



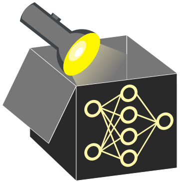
- **PFI Idea:** Replace feature(s) X_S with perturbed $\tilde{X}_S$ to preserve marginal distr. $\mathbb{P}(X_S)$ so that $\tilde{X}_S \perp\!\!\!\perp Y$ (indep.), e.g., by random permutations

CFI MOTIVATION



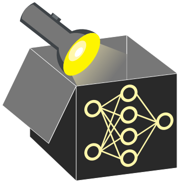
- **PFI Idea:** Replace feature(s) X_S with perturbed $\tilde{X}_S$ to preserve marginal distr. $\mathbb{P}(X_S)$ so that $\tilde{X}_S \perp\!\!\!\perp Y$ (indep.), e.g., by random permutations
- **Problem:** Breaks not only association between X_S and Y (what we want) but also between X_S, $X_{-S} \Rightarrow \mathbb{P}(X_S, X_{-S}) \neq \mathbb{P}(\tilde{X}_S, X_{-S})$ (extrapolation)

CFI MOTIVATION



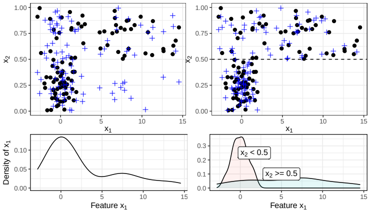
- **PFI Idea:** Replace feature(s) X_S with perturbed $\tilde{X}_S$ to preserve marginal distr. $\mathbb{P}(X_S)$ so that $\tilde{X}_S \perp\!\!\!\perp Y$ (indep.), e.g., by random permutations
- **Problem:** Breaks not only association between X_S and Y (what we want) but also between X_S, $X_{-S} \Rightarrow \mathbb{P}(X_S, X_{-S}) \neq \mathbb{P}(\tilde{X}_S, X_{-S})$ (extrapolation)
- **CFI Idea:** Replace feature(s) X_S with perturbed $\tilde{X}_S$ to preserve joint distr. so that $\mathbb{P}(X_S, X_{-S}) = \mathbb{P}(\tilde{X}_S, X_{-S})$ (no extrapolation) while still $\tilde{X}_S \perp\!\!\!\perp Y$

CFI MOTIVATION



- **PFI Idea:** Replace feature(s) X_S with perturbed $\tilde{X}_S$ to preserve marginal distr. $\mathbb{P}(X_S)$ so that $\tilde{X}_S \perp\!\!\!\perp Y$ (indep.), e.g., by random permutations
- **Problem:** Breaks not only association between X_S and Y (what we want) but also between X_S, $X_{-S} \Rightarrow \mathbb{P}(X_S, X_{-S}) \neq \mathbb{P}(\tilde{X}_S, X_{-S})$ (extrapolation)
- **CFI Idea:** Replace feature(s) X_S with perturbed $\tilde{X}_S$ to preserve joint distr. so that $\mathbb{P}(X_S, X_{-S}) = \mathbb{P}(\tilde{X}_S, X_{-S})$ (no extrapolation) while still $\tilde{X}_S \perp\!\!\!\perp Y$

Example: Conditional permutation scheme
Black dots: $X_2 \sim \mathcal{U}(0, 1)$ and $X_1 \sim \mathcal{N}(0, 1)$ (if $X_2 < 0.5$) or $\mathcal{N}(4, 4)$ (if $X_2 \geq 0.5$)



Left: For $X_2 < 0.5$, permuting X_1 (crosses) preserves marginal (but not joint) distrib.
$\rightsquigarrow$ Bottom: Marginal density of X_1

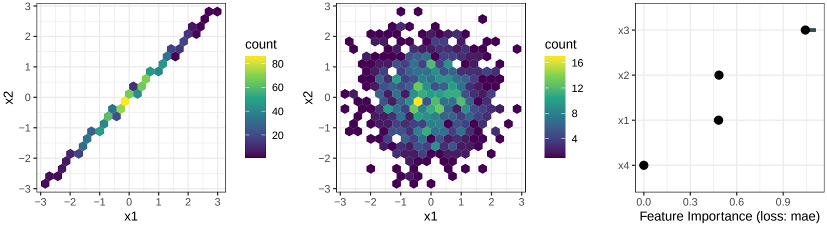
Right: Permuting X_1 within subgroups $X_2 < 0.5$ & $X_2 \geq 0.5$ reduces extrapolation
$\rightsquigarrow$ Bottom: X_1-density cond. on groups

▶ "Molnar et. al" 2020

RECALL: EXTRAPOLATION IN PFI

Recall: Let $y = x_3 + \epsilon_y$, with $\epsilon_y \sim \mathcal{N}(0, 0.1)$.

- $x_1 := \epsilon_1$, $x_2 := x_1 + \epsilon_2$; highly correlated ($\epsilon_1 \sim \mathcal{N}(0,1)$, $\epsilon_2 \sim \mathcal{N}(0, 0.01)$)
- $x_3 := \epsilon_3$, $x_4 := \epsilon_4$, with $\epsilon_3, \epsilon_4 \sim \mathcal{N}(0,1)$; all noise terms ϵ_j are indep.
- Fitting a linear model yields $\hat{f}(\mathbf{x}) \approx 0.3x_1 - 0.3x_2 + x_3$



Hexbin plot of (x_1, x_2) before (left) and after (center) permuting x_1;
PFI scores (right).

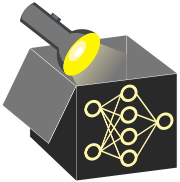
$\Rightarrow$ x_1, x_2 cancel in $\hat{f}$ and should be irrelevant

$\Rightarrow$ But PFI evaluates model on unrealistic inputs (caused by permutation)

$\rightsquigarrow$ *PFI* > 0 for x_1, x_2 due to extrapolation

$\rightsquigarrow$ x_1, x_2 are misleadingly considered relevant

CFI ▸ "Strobl et al." 2008 ▸ "Hooker et al." 2021



CFI for X_S using test data $\mathcal{D}$:

- Measure the error **with unperturbed features** x_S.
- Measure the error **with perturbed feature values** $\tilde{x}_S \sim \mathbb{P}(X_S | X_{-S})$
- Repeat perturbing X_S (e.g., m times) and avg. difference of both errors:

$$\widehat{CFI}_S = \frac{1}{m} \sum_{k=1}^{m} \mathcal{R}_{\text{emp}}(\hat{f}, \tilde{\mathcal{D}}_{(k)}^{S|-S}) - \mathcal{R}_{\text{emp}}(\hat{f}, \mathcal{D})$$

Here, $\tilde{\mathcal{D}}^{S|-S}$ denotes data, where x_S values are conditionally resampled given x_{-S}.

Illustrative example: Conditional permutation when X_{-S} is categorical:

Original Data

ID	X_{-S}	X_S
1	A	3.1
2	A	2.7
3	A	3.4
4	B	6.0
5	B	5.4
6	B	6.2

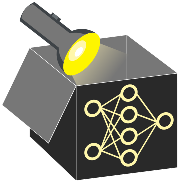
Permuted Conditionally on X_{-S}

ID	X_{-S}	X_S
1	A	2.7
2	A	3.1
3	A	3.4
4	B	6.2
5	B	6.0
6	B	5.4

Here, X_S is permuted *within* each group of X_{-S} to preserve $\mathbb{P}(X_S, X_{-S})$.

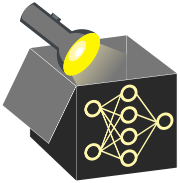
IMPLICATIONS OF CFI

▸ "König et al." 2020



Interpretation: Due to the conditional sampling w.r.t. all other features, CFI quantifies a feature's unique contribution to the model performance.

IMPLICATIONS OF CFI ▸ "König et al." 2020

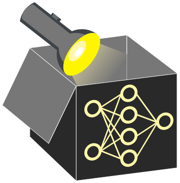


Interpretation: Due to the conditional sampling w.r.t. all other features, CFI quantifies a feature's unique contribution to the model performance.

Entanglement with data:

- If feat x_S does not contrib. unique information about y, i.e., $x_S \perp\!\!\!\perp y | x_{-S}$ $\Rightarrow$ CFI $= 0$
- Why? Under the conditional indep. $\mathbb{P}(\tilde{X}_S, X_{-S}, Y) = \mathbb{P}(X_S, X_{-S}, Y)$ $\rightsquigarrow$ no prediction-relevant information is destroyed by permutation of x_S conditional on x_{-S}

IMPLICATIONS OF CFI ▸ "König et al." 2020



Interpretation: Due to the conditional sampling w.r.t. all other features, CFI quantifies a feature's unique contribution to the model performance.

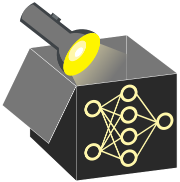
Entanglement with data:

- If feat x_S does not contrib. unique information about y, i.e., $x_S \perp\!\!\!\perp y | x_{-S}$ $\Rightarrow$ CFI $= 0$
- Why? Under the conditional indep. $\mathbb{P}(\tilde{X}_S, X_{-S}, Y) = \mathbb{P}(X_S, X_{-S}, Y)$ $\rightsquigarrow$ no prediction-relevant information is destroyed by permutation of x_S conditional on x_{-S}

Entanglement with model:

- If the model does not use a feature $\Rightarrow$ CFI $= 0$
- Why? Then the prediction is not affected by any perturbation of the feat $\rightsquigarrow$ model performance does not change after conditional permutation

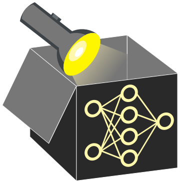
IMPLICATIONS OF CFI



Can we gain insight into whether ...

1. the feature x_j is causal for the prediction?
   - $CFI_j \neq 0 \Rightarrow$ model relies on x_j
     (converse does not hold, see next slide)

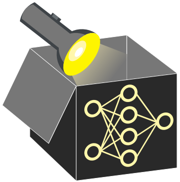
IMPLICATIONS OF CFI



Can we gain insight into whether ...

1. the feature x_j is causal for the prediction?
   - $CFI_j \neq 0 \Rightarrow$ model relies on x_j
     (converse does not hold, see next slide)
2. the variable x_j contains prediction-relevant information?
   - If $x_j \not\perp\!\!\!\perp y$ but $x_j \perp\!\!\!\perp y | x_{-j}$ (e.g., x_j and x_{-j} share information)
     $\Rightarrow CFI_j = 0$
   - x_j is not exploited by model (regardless of its usefulness for y)
     $\Rightarrow CFI_j = 0$

IMPLICATIONS OF CFI



Can we gain insight into whether ...

1. the feature x_j is causal for the prediction?
   - $CFI_j \neq 0 \Rightarrow$ model relies on x_j
     (converse does not hold, see next slide)
2. the variable x_j contains prediction-relevant information?
   - If $x_j \not\perp\!\!\!\perp y$ but $x_j \perp\!\!\!\perp y | x_{-j}$ (e.g., x_j and x_{-j} share information)
     $\Rightarrow CFI_j = 0$
   - x_j is not exploited by model (regardless of its usefulness for y)
     $\Rightarrow CFI_j = 0$
3. Does the model need access to x_j to achieve its prediction performance?
   - $CFI_j \neq 0 \Rightarrow x_j$ contributes unique information (meaning $x_j \not\perp\!\!\!\perp y | x_{-j}$)
   - Only uncovers the relationships that were exploited by the model

EXTRAPOLATION: COMPARE PFI AND CFI

Recall: Let $y = x_3 + \epsilon_y$, with $\epsilon_y \sim \mathcal{N}(0, 0.1)$.

- $x_1 := \epsilon_1$, $x_2 := x_1 + \epsilon_2$; highly correlated ($\epsilon_1 \sim \mathcal{N}(0, 1)$, $\epsilon_2 \sim \mathcal{N}(0, 0.01)$)
- $x_3 := \epsilon_3$, $x_4 := \epsilon_4$, with $\epsilon_3, \epsilon_4 \sim \mathcal{N}(0, 1)$; all noise terms ϵ_j are indep.
- Fitting a linear model yields $\hat{f}(\mathbf{x}) \approx 0.3x_1 - 0.3x_2 + x_3$

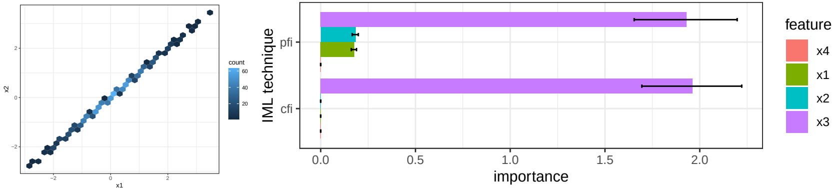


Figure: Density plot for x_1, x_2 before permuting x_1 (left). PFI and CFI (right).

- x_1 and x_2 cancel in $\hat{f}(\mathbf{x})$ and should be irrelevant for the prediction
- PFI evaluates model on unrealistic obs.
  $\rightsquigarrow$ x_1, x_2 appear relevant (PFI > 0)
- CFI evaluates model on realistic obs. (due to conditional sampling)
  $\rightsquigarrow$ x_1, x_2 appear irrelevant (CFI $= 0$)