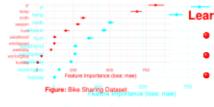
Interpretable Machine Learning

Conditional Feature Importance (CFI)





Learning goals

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

Figure: Bike Sharing Dataset

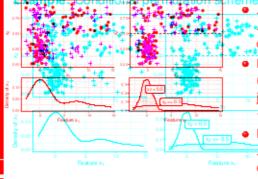
• Permutation Feature Importance Idea: Replace the feat for interest we with an with an independent independent from the marginal district (x/), e.g.eby randomly permember the yobservations in

- Permutation Feature Importance Idea: Replace the feat of interest xicwith an with an independent indep sample from the marginal district (xy), e.g. by randomly permuops rin xyobservations in x
- Problem: Under dependent features, permutation leads to extrapolation
 Problem: Under dependent features, permutation leads to extrapolation

- Permutation Feature Importance Idea: Replace the feat for interest yewith an with an independent indeposample from the marginal dist; R(x), e.g. by randomly permoobs; tin x/observations in x
- Problem: Under dependent features, permutation leads to extrapolation
 Problem: Under dependent features, permutation leads to extrapolation
 Problem: Under dependent features, permutation leads to extrapolation.
- Conditional Feature Importance Idea: Resample x, from the cond. dist.
- P(x, |x_{-i}|), s.t. the joint distribution preserved, i.e., $\mathbb{P}(x_j|x_{-j})\mathbb{P}(x_{-j}) = \mathbb{P}(x_j|x_{-j})$ and distribution $\mathbb{P}(x_j|x_{-j})$, such that the joint distribution is preserved, i.e., $\mathbb{P}(x_j|x_{-j})\mathbb{P}(x_{-j}) = \mathbb{P}(x_j,x_{-j})$

- Permutation Feature Importance Idea: Replace the feat not interest wwith an with an independent indep. sample from the marginal dist; $P(x_i)$, e.g. by randomly permeabs, tin x observations in
- Problem: Under dependent features, permutation leads to extrapolation
- Conditional Feature Importance Idea: Resample x_i from the cond. dist.

 $\mathbb{P}(x_j|X_{-j}^i)$ on the joint distribution preserved, i.e., $\mathbb{P}(x_j|X_{-j}^i)$ $\mathbb{P}(x_j|X_{-j}^i)$ in $\mathbb{P}(x_j|X_{-j}^i)$ $\mathbb{P}(x_j|X_{-j}^i)$ $\mathbb{P}(x_j|X_{-j}^i)$ Example: Conditional permutation scheme

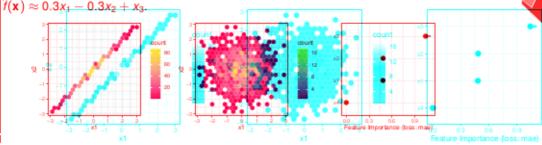


- X₂ ~ U(0, 1) and X₁ ~ N(0, 1) if
 - $X_2 < 0.5$, else $X_1 \cup \{0,0\}$ (4, 24) (black N(0,1) if $X_2 < 0.5$,
 - else $X_1 \sim N(4,4)$ (black dots) dots)
 - Left: For X₂ et 0.5; rpermuting Xpermuting X₁ (crosses) (crosses) preserves marginal (but not of joint) distribution ioint) distribution ttom: Marginal density of X,
 - → Bottom: Marginal density of Xwithin subgroups
 - Right: Permuting X within subgroupsces extrapolation $X < 0.5 & X_0 > 10.5 \text{ reduces y of } X_1 \text{ conditional on groups}$ extrapolation
 - → Bottom: Density of X₁ conditional on groups

Interpretable Machine Learning - 1/6

RECALL: EXTRAPOLATION IN PFI

Example be Let $\mathbf{y} \neq \mathbf{x}_3 + \mathbf{\epsilon}_{\mathbf{y}}$ with $\mathbf{x}_{\mathbf{y}} \sim \mathcal{N}(0,0.1)$ where $\mathbf{x}_{\mathbf{y}} = \mathbf{x}_{\mathbf{y}} = \mathbf{x}_{\mathbf{y}} + \mathbf{\epsilon}_{\mathbf{y}}$ are $\mathbf{\epsilon}_{\mathbf{y}}$ are highly correlated highly correlated $(\mathbf{\epsilon}_{\mathbf{y}} \sim \mathcal{N}(0,0.1), \mathbf{\epsilon}_{\mathbf{y}}) \sim \mathcal{N}(0,0.01)$, and $\mathbf{x}_{\mathbf{y}} := \mathbf{\epsilon}_{\mathbf{y}} + \mathbf{k}_{\mathbf{y}} = \mathbf{k}_{\mathbf{y}} + \mathbf{k}_{\mathbf{y}} + \mathbf{k}_{\mathbf{y}} + \mathbf{k}_{\mathbf{y}} = \mathbf{k}_{\mathbf{y}} + \mathbf{k}_{\mathbf{y}} + \mathbf{k}_{\mathbf{y}} + \mathbf{k}_{\mathbf{y}} = \mathbf{k}_{\mathbf{y}} + \mathbf{$



Hexbin plot of x_1, x_2 before permuting x_1 (left), after permuting x_1 (center), and PFI scores (right) $x_1 = x_1 + x_2 + x_3 + x_4 + x_4 + x_5 +$

 $\{\mathbf{x} \geq \mathbb{P}(\mathbf{x}) \leq 0\}$ as $0.3x_1 = 0.3x_2 \approx 0$ as $0.3x_1 = 0.3x_2 \approx 0$ $\Rightarrow PFI$ evaluates model on unrealistic observation $\hat{f}(\mathbf{x})$ for $\{\mathbf{x} : \mathbb{P}(\mathbf{x}) > 0\}$ as $0.3x_1 = 0.3x_2 \approx 0$.

relevant

CONDITIONAL FEATURE IMPORTANCE Stroblet al. (2003) + Hooker et al. (2021)

Hooker et al. (2021)

Conditional feature importance (CFI) for features x_S using test data \mathcal{D} :

Conditional feature importance (CFI) for features x_s using test data \mathcal{D} :

- Measure the error with unperturbed features. • Measure the error with unperturbed features. $\tilde{x}^{S|-S}$, where $\tilde{x}^{S|-S}_S \sim \mathbb{P}(x_S|x_{-S})$ • Measure the error with perturbed feature values $\tilde{x}^{S|-S}$, where $\tilde{x}^{S|-S}_S$ where
- $\tilde{x}_e^{S|ES} \sim \mathbb{P}(x_S|x_{-S})$
- Repeat permuting the feature (e.g., m times) and average the difference of both errors:

Here, $\tilde{\mathcal{D}}^{S|-S}$ denote $\widehat{\mathcal{GFl}}$ to \overline{d} and $\widehat{\mathcal{C}}_{S}^{m}$ and $\widehat{\mathcal{C}}_{S}^{m}$ denote $\widehat{\mathcal{GFl}}$ to \overline{d} and $\widehat{\mathcal{C}}_{S}^{m}$ denote $\widehat{\mathcal{C}}_{S}^{m}$ denote $\widehat{\mathcal{C}}_{S}^{m}$ and $\widehat{\mathcal{C}}_{S}^{m}$ denote $\widehat{\mathcal{C}}_{S}^{m}$ d

Here. \mathcal{D}^{Sp} denotes the dataset where features x_{S} where sampled conditional on the remaining features x_{-S}.

IMPLICATIONS OF CF (König et al. (2020)

Interpretation: Due to the conditional sampling live, twall other features. CFI, CFI quantifies a feature's quantifies a feature's unique contribution to the model performance.

IMPLICATIONS OF CFI (E König et al. (2020))

Interpretation: Due to the conditional sampling liver twall other features. CFI, CFI quantifies a feature's quantifies a feature's unique contribution to the model performance.

Entanglement with data:ta:

- If Ifeature x_S does not contribute unique information tabout y Lite, $x_S \perp |x_y| x_{-S} \Rightarrow CFI = 0$ CFU \overline{y} ? Under the conditional independence $\mathbb{P}(\tilde{x}^{S|-S}, y) = \mathbb{P}(x, y)$
- Why? Under the conditional independence ⊕(x̄S ȳS y y) → ⊕(x,y) ion of x_S conditional on x_S
 → no prediction-relevant information is destroyed by permutation of x_S
 conditional on x_S

IMPLICATIONS OF CF (F König et al. (2020)

Interpretation: Due to the conditional sampling liver twall other features. CFI. CFI quantifies a features quantifies a feature's unique contribution to the model performance.

Entanglement with data: a:

- If feature x_S does not contribute unique information about y_Y i.e., $x_S \perp \!\!\!\perp \!\!\!\perp y \mid x_{-S} \Rightarrow \mathsf{CFI} = 0$ ${}^{\circ}$ CFII \overline{y} ?Ounder the conditional independence $\mathbb{P}(\tilde{x}^{S|-S},y) = \mathbb{P}(x,y)$
- Why? Under the conditional independence $\mathbb{P}(\bar{x}^S \cup \bar{y}^S \cup \bar{y}) = \mathbb{P}(x_* y_*)$ ion of x_S conditional on x_{-S} \rightarrow no prediction-relevant information is destroyed by permutation of x_S

Entanglement with snodel:

If the model does not use a feature ⇒ CFI = 0.

Entanglement with model:

Why? I hen the prediction is not affected by any perturbation of the feature

- If the model does not use a feature chaffer after conditional permutation
- Why? Then the prediction is not affected by any perturbation of the feature → model performance does not change after conditional permutation

IMPLICATIONS OF CFI

Can we gain insight into whether ...

- the feature x_j is causal for the prediction? n?
 - • CEI; ≠ 0 ⇒ model:relies on x₁ (converse does not hold; see next slide): lide)



IMPLICATIONS OF CFI

Can we gain insight into whether ...

- the feature x_j is causal for the prediction? n?
 - • CEI; ≠ 0 ⇒ modetireliesion x₁ (converse does not hold; see next slide): lide)
- (a) the variable x_j contains prediction-relevant information? on?
 - olf lxjx,/LL/y jbut xjx, LLL y jxx, (e.g.g xj, and xLy share information) ⇒ CFIC= 0
 - • x_j is not exploited by model (regardless of whether it is useful for y for not) not) $\Rightarrow CFI_j = 0$ $\Rightarrow CFI_j = 0$



IMPLICATIONS OF CFI

Can we gain insight into whether ...

- the feature x_j is causal for the prediction? n?
 - CFI_I ≠ 0 ⇒ model relies on x_I (converse does not hold, see next slide): lide)
- Ontains prediction relevant information?
 - • If $|\hat{x}_j x_j L / y_j|$ but $|\hat{x}_j x_j L L y_j|$ $|\hat{x}_j x_j L L y_j|$
 - •• x_j is not exploited by model (regardless of whether it is useful for y for not) not) $\Rightarrow CFI_j = 0$
 - O Doest f_i hodel require access to x_i to achieve it's prediction performance?
- Does the model require access to x_i to achieve its prediction performance?
 - CFI_jI ≠ .01⇔ x_j: contributes unique information (meaning lx) #hey |xx_j|e|
 - . Only uncovers the relationships that were exploited by the model



COMPARISON: PFI AND CFI

Example: Let $y \neq x_3 + \epsilon_y$ with with $\sim N(0,0.1)$ where $x_1 = x_2 = x_1 + \epsilon_z$ are ϵ_z are highly correlated highly correlated $(\epsilon_1 \sim N(0,0.1),\epsilon_2)$ and $x_3 := \epsilon_z$, $\epsilon_z = \epsilon_z$, with (0,1). All noise terms are $\epsilon_z = \epsilon_z = \epsilon_z$

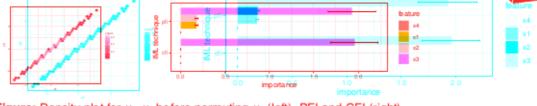


Figure: Density plot for x_1 , x_2 before permuting x_1 (left). PFI and CFI (right). Figure: Density plot for x_1 , x_2 before permuting x_1 (left). PFI and CFI (right).

- \Rightarrow x_1 and x_2 are irrelevant for the prediction $\hat{f}(\mathbf{x})$ for $\{\mathbf{x}: \mathbb{P}(\mathbf{x})>0\}$ as $0.3x_1^{\chi_1} + 30.9x_2^{\chi_2} \approx 0$ irrelevant for the prediction $\hat{f}(\mathbf{x})$ for $\{\mathbf{x}: \mathbb{P}(\mathbf{x})>0\}$ as $0.3x_1-0.3x_2\approx0$
- \Rightarrow PFF evaluates mode for unrealistic obs. but side $\mathbb{P}(\mathbf{x}) \hookrightarrow x_1, x_2$ are considered event (PFI > 0) relevant (PFI > 0) be reconstructed from x_2 and vice versa, CFI considers x_1 and x_2 to be irrelevant
- \Rightarrow Since x_1 can be reconstructed from x_2 and vice versa, CFI considers x_1 and x_2 to

Interpretable Machine Learning - 6 / 6