Fairwashing SHAP

or

Interventional and Observational Shapley Values

Motivation

Someone walks into a bank, applied for a loan, gets rejected.
Why?

2 Kinds of explanations:

Recourse: What could they do to get the loan?

R\AI Developer: Why did the model say no? It is bigoted?

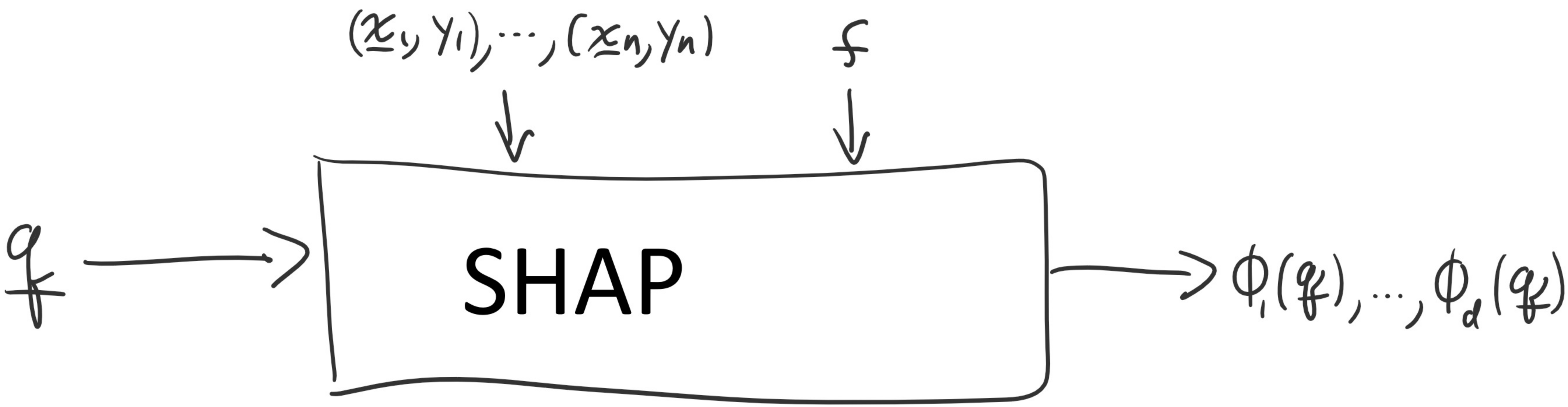
Focus on this

Local Feature Importance

vectors are underlined

Given: dataset $(\underline{x}_1, y_1), \cdots, (\underline{x}_n, y_n) \overset{iid}{\sim} \mathcal{D}$, ML model $f(\underline{x})$, query point $\underline{q} \in \mathbb{R}^d$

Return: Importance of each feature $\phi_1(\underline{q}), \dots, \phi_d(\underline{q})$



Shap is well used, based on nice game theory -- good about proxy vars!

Fairwashing

Suppose f is very racist.

Then $SHAP(f, \mathbf{x})$ should show that race is an important feature

But, we can make ML model $\tilde{f}$ such that

1. For almost all $\mathbf{x}$, $\tilde{f}(\mathbf{x}) = f(\mathbf{x})$
2. $SHAP(\tilde{f}, \mathbf{x})$ shows almost no important for race

$\tilde{f}$ is a "Fairwashed" version of f. It looks fair, but it ain't!

- At least 2 papers: Umang's paper, [Slack et al. 2020]
- But, how is this possible? What about game theory?

Takeaway

Q: How is fairwashing possible? What about game theory?

A: "Interventional" vs "Observational" Shapley Values

What SHAP and QII do
Bad for proxys

Other papers do this.
Good for proxys.

"True to the Model or True to the Data" (on arXiv)

Great jumping off point, easy read

SHAPLEY VALUES

Some notation

- We have d features: $\underline{x}_1, \dots, \underline{x}_n, \underline{q} \in \mathbb{R}^d$
- Let $S \subseteq \{1, 2, 3, \dots, d\}$ be a subset of features
- Let $\bar{S}$ complement of S
- Let $(\underline{q})_S$ = entries of $\underline{q}$ indexed by S
- Notice: $\underline{q} = (\underline{q})_S + (\underline{q})_{\bar{S}}$

Frankenstein Point: $(\underline{q})_S + (\underline{x}_j)_{\bar{S}}$

$d = 5$

$\underline{q} = [9 \;\; 7 \;\; 5 \;\; 3 \;\; 1]$

$S = \{1, 2\}$

$\bar{S} = \{3, 4, 5\}$

$(\underline{q})_S = [9 \;\; 7 \;\; 0 \;\; 0 \;\; 0]$

$(\underline{q})_{\bar{S}} = [0 \;\; 0 \;\; 5 \;\; 3 \;\; 1]$

$\underline{x}_1 = [2 \;\; 2 \;\; 2 \;\; 2 \;\; 2]$

$(\underline{q})_S + (\underline{x}_1)_{\bar{S}} = [9 \;\; 7 \;\; 2 \;\; 2 \;\; 2]$

Shapley Values

Not "Quantity of Interest"

Let $v_S(\mathbf{x})$ be the Value of $\mathbf{x}$ with respect to S.

The average marginal contribution of feature i.

$$\phi_i(\mathbf{x}) := \underbrace{\sum_{\substack{S \subseteq \{1,\dots,d\} \\ i \notin S}} \frac{1}{d} \cdot \frac{1}{\binom{d-1}{|S|}}}_{\text{Average over } S} \underbrace{\left(v_{S \cup \{i\}}(\mathbf{x}) - v_S(\mathbf{x}) \right)}_{\text{How much } i \text{ increases value, starting from } S}$$

① Interventional $\quad v_S(\mathbf{x}) = \mathop{\mathbb{E}}_{\underline{x} \sim \mathcal{D}} \left[f\left((\mathbf{x})_S + (\underline{x})_{\bar{S}} \right) \right] \approx \frac{1}{n} \sum_{j=1}^{n} f\left((\mathbf{x})_S + (\underline{x}_j)_{\bar{S}} \right)$

- Used by SHAP (by default), QII
- Creates very fake-looking data points (ignores dependencies)

① Interventional $v_S(\underline{x}) = \mathbb{E}_{\underline{x} \sim \mathcal{D}}\left[f\left((\underline{x})_S + (\underline{x})_{\bar{S}} \right)\right] \approx \frac{1}{n} \sum_{j=1}^{n} f\left((\underline{x})_S + (\underline{x}_j)_{\bar{S}} \right)$

- Used by SHAP (by default), QII
- Creates very fake-looking data points (ignores dependencies)

② Observational $v_S(\underline{x}) = \mathbb{E}_{\underline{x} \sim \mathcal{D}}\left[f\left((\underline{x})_S + (\underline{x})_{\bar{S}} \right) \;\middle|\; (\underline{x})_S = (\underline{x})_S \right]$

$\approx$ Average of $\left((\underline{x})_S + (\underline{x}_j)_{\bar{S}} \right)$ for all $(\underline{x}_j)_S = (\underline{x})_S$?

- Less used, but is used
- Creates real-looking data points
- Hard to compute (few points to average)

How can we compare these 2 approaches?

Simple Linear Model

Suppose $f(\underline{x}) = \langle \underline{w}, \underline{x} \rangle + b$ for some $\underline{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$.

How do interventional & observational $\phi_i(\underline{x})$ look?

Lemma

Interventionally,

average of feature i

$$\phi_i(\underline{x}) = w_i (x_i - \mu_i)$$

Takeaway: Completely ignores proxy variables

Takeaway: Completely ignores proxy variables

$\underline{x} = (\text{Kindness}, \text{salary}, \text{age})$ where $\text{salary} = 1000 \cdot \text{age}$

$\underline{w}_{BAD} = [1, 0, 1000]$ is unfair: uses age a lot

$\phi_{AGE}(\underline{q}) = 1000(q_i - \mu_i) \neq 0$, great!

$\underline{w}_{WASHED} = [1, 1, 0]$ makes exact same predictions

$\phi_{AGE}(\underline{q}) = 0$, terrible!

Lemma Proof: $\phi_i(q) = w_i(q_i - \mu_i)$ interventionally

$$v_S(q) = \frac{1}{n}\sum_{j=1}^{n} f\big((q)_S + (x_j)_{\bar S}\big) = \frac{1}{n}\sum_{j=1}^{n} \langle w, (q)_S + (x_j)_{\{i\}} + (x_j)_{\bar S \setminus \{i\}} \rangle + \cancel{b}$$

$$v_{S\cup\{i\}}(q) = \frac{1}{n}\sum_{j=1}^{n} f\big((q)_{S\cup\{i\}} + (x_j)_{\bar S\setminus\{i\}}\big) = \frac{1}{n}\sum_{j=1}^{n} \langle w, (q)_S + (q)_{\{i\}} + (x_j)_{\bar S \setminus \{i\}} \rangle + \cancel{b}$$

$$v_{S\cup\{i\}}(q) - v_S(q) = \frac{1}{n}\sum_{j=1}^{n} \langle w, (q)_{\{i\}} - (x_j)_{\{i\}} \rangle$$

$$= \langle w, (q)_{\{i\}} - (\mu)_{\{j\}} \rangle$$

$$= w_i(q_i - \mu_i)$$

$$[0 \cdots 0\ (q_i - \mu_i)\ 0 \cdots 0]$$

$\phi_i(q) =$ Average of $v_{S\cup\{i\}}(q) - v_S(q)$ across all $= w_i(q_i - \mu_i)$ $\blacksquare$

Beyond Intervensional Value

Above proof does not work for observation values

$$v_S(\underline{x}) = \mathop{\mathbb{E}}_{\underline{x} \sim \mathcal{D}}\left[f((\underline{x})_S + (\underline{x})_{\bar{S}}) \mid (\underline{x})_S = (\underline{x})_S \right]$$

Only Frankenstein $\underline{x}$ with x_j's that match $\underline{x}$ on S.

So,

$v_{S\cup\{i\}}(\underline{x}) - v_S(\underline{x})$ depends on S, breaking the proof.

Claim

If $f(\underline{x}) = \langle \underline{w}, \underline{x} \rangle + \beta$ and $x_j \overset{iid}{\sim} N(\mu, \Sigma)$, then we can write $\phi_i(\underline{x})$ exactly (but it's ugly).

But it super depends on correlation in Σ!

Comparison

Let $\underline{x}_j \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & & \\ & 1 & 0.99 \\ & 0.99 & 1 \end{bmatrix} \right)$

Features 2 and 3
are super correlated

$$\underline{w} = [\,1 \quad 2 \quad 3\,], \quad \underline{q} = [\,1 \quad 1 \quad 1\,]$$

Interventional

$\phi_1(\underline{q}) = 1$

$\phi_2(\underline{q}) = 2$

$\phi_3(\underline{q}) = 3$

Treated Very Different

Observational

$\phi_1(\underline{q}) = 1$

$\phi_2(\underline{q}) \approx 2.5$

$\phi_3(\underline{q}) \approx 2.5$

Treated the Same!

Interventional	Observational
$\phi_1(f) = 1$	$\phi_1(f) = 1$
$\phi_2(f) = 2$, $\phi_3(f) = 3$ — Treated Very Different	$\phi_2(f) \approx 2.5$, $\phi_3(f) \approx 2.5$ — Treated the Same!
$w = [1\ 0\ 3]$	
If $w_2 = 0$,	If $w_2 = 0$
Then $\phi_2(f) = 0$	Then $\phi_2(f)$ may be $\neq 0$
	If f is blind to age, it possible that $\phi_{AGE}(f) \neq 0$!
Better recourse IF Feature are independent	Better recourse in my opinion because Features are dependant

Conclusion

Going back to the bank example,

Both recourse and R\AI Engineering prefer observational values

But we use interventional almost always.
Also, it's harder to compute.