Supplementary Material

A. Marginal Shapley values

Marginal Shapley values are heavily influenced by the distribution of the data.

A.1. Shifted distributions

First, let us consider the linear model $f(x) = x_1 + x_2$ with $x_1 \sim \mathrm{Uniform}(-1, 2)$ and $x_2 \sim \mathrm{Uniform}(0, 3)$ and local observation $x = (0, 0)$. We show that although they play symmetric roles in the algebraic formulation of the black box model, their marginal values aren't equal: $\phi_1 = -0.5$ and $\phi_2 = -1.5$.

$$
\begin{aligned}
\phi_1 = & \sum_{S \subseteq \{\emptyset, x_2\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \\
& \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)\right] \\
& = \frac{1}{2}(f_1(x_1) - f_\emptyset(\emptyset)) + \frac{1}{2}(f_{1,2}(x_1, x_2) - f_2(x_2)) \\
& \overset{*}{=} \frac{1}{2}E[f(x_1, X_2)] - \frac{1}{2}E[f(X_1, X_2)] \\
& + \frac{1}{2}E[f(x_1, x_2)] - \frac{1}{2}E[f(X_1, x_2)] \\
& = \frac{1}{2}P(X_2 = 1) - \frac{1}{2}P(X_2 = 1|X_1 = 1)P(X_1 = 1) \\
& + \frac{1}{2} - \frac{1}{2}P(X_1 = 1) \\
& = \frac{1}{2}p - \frac{1}{2}p \cdot 1 + \frac{1}{2} - \frac{1}{2} \cdot 1 = 0
\end{aligned}
$$

where we used the definition $f(x_S) = E[f(x)|do\ x_S]$ from (Janzing et al., 2020) for marginal Shapley values in equation $*$.

$$
\begin{aligned}
2\phi_1 = & \frac{1}{3}[\int_0^3 x_2 dx_2] - \frac{1}{9}[\int_0^3 \int_{-1}^2 (x_1 + x_2) dx_2 dx_1] \\
& - \frac{1}{3}[\int_{-1}^2 x_1 dx_1] \\
& = \frac{1}{3}[\int_0^3 x_2 dx_2] - \frac{1}{3}[\int_{-1}^2 x_1 ddx_1] \\
& - \frac{1}{3}[\int_0^3 x_2 dx_2] - \frac{1}{3}[\int_{-1}^2 x_1 dx_1] \\
& = -\frac{2}{3}[\int_{-1}^2 x_1 dx_1] \\
& = -\frac{2}{3}[\int_{-1}^2 x_1 dx_1] \\
& = -\frac{2}{3}[\frac{4}{2} - \frac{(-1)^2}{2}]
\end{aligned}
$$

Thus

$$\phi_1 = \frac{-1}{2}$$

Symmetrically for x_2,

$$2\phi_2 = -\frac{2}{3}[\int_0^3 x_2 dx_2]$$

and therefore

$$\phi_2 = -\frac{3}{2}$$

A.2. Different spreads

Let us consider a black box $f(x) = x_1^2 + x_2^2$ with $x_1 \sim \mathrm{Normal}(0, 1)$ and $x_2 \sim \mathrm{Normal}(0, 10)$ and local observation $x = (0, 0)$. While the first marginal Shapley value is -1, the second one is -100 as the expected change of model outcome is higher when intervening on the common population by setting $x_2 = 0$ compared to setting $x_1 = 0$.

Similarly to the previous section:

$$\begin{aligned}\phi_1 &= -\int_{-\infty}^{+\infty} \frac{x_1^2}{\sqrt{2\pi}} \exp\frac{-x_1^2}{2} dx_1] \\ &= \frac{1}{\sqrt{2\pi}}[-x_1 \mathrm{e}^{-\frac{x_1^2}{2}} - \int_{-\infty}^{+\infty} -\mathrm{e}^{-\frac{x_1^2}{2}}\, \mathrm{d}x_1\end{aligned}$$

Solving separately:

$$\int -\mathrm{e}^{-\frac{x_1^2}{2}}\, \mathrm{d}x$$

We substitute $u = \frac{x_1}{\sqrt{2}} \longrightarrow \frac{\mathrm{d}u}{\mathrm{d}x_1} = \frac{1}{\sqrt{2}} \longrightarrow \mathrm{d}x_1 = \sqrt{2}\, \mathrm{d}u$:

$$= -\frac{\sqrt{\pi}}{\sqrt{2}} \int \frac{2\mathrm{e}^{-u^2}}{\sqrt{\pi}} \mathrm{d}u$$

We notice the Gaussian error function below:

$$\int \frac{2\mathrm{e}^{-u^2}}{\sqrt{\pi}} \mathrm{d}u = \mathrm{erf}(u)$$

We plug in solved integrals:

$$\begin{aligned}&-\frac{\sqrt{\pi}}{\sqrt{2}} \int \frac{2\mathrm{e}^{-u^2}}{\sqrt{\pi}} \mathrm{d}u \\ &= -\frac{\sqrt{\pi}\,\mathrm{erf}(u)}{\sqrt{2}}\end{aligned}$$

We undo the substitution $u = \frac{x}{\sqrt{2}}$:

$$= -\frac{\sqrt{\pi}\,\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)}{\sqrt{2}}$$

Ultimately:

$$\int_{-\infty}^{+\infty} \frac{x^2 \mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2}\sqrt{\pi}} \mathrm{d}x = [\frac{\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)}{2} - \frac{x\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2}\sqrt{\pi}}]_{-\infty}^{+\infty} = 1$$

Symmetrically for x_2,

$$\begin{aligned}\phi_2 &= -\int_{-\infty}^{+\infty} \frac{x_1^2}{10\sqrt{2\pi}} \exp\frac{-x_1^2}{200} dx_1 \\ &= -100\end{aligned}$$

B. Counterfactual Fairness

Because of the dummy property of marginal Shapley values we have that

$$\text{counterfactual fairness} \rightarrow \text{marginal Shapley value of } 0$$

for deterministic models. The back direction as we will see does not hold: Let our feature space comprise two binary variables $x_{canLift}$ and x_{male} with

$$\begin{aligned} P(x_{male} = 1) &= 1 \\ P(x_{canLift} = 1) &= p \\ P(x_{canLift} = 1 | x_{male} = 1) &= p \end{aligned}$$

where p is an arbitrary probability.

Our black box algorithm is $f(x) = x_{male} \cdot x_{canLift}$ and the feature attribution of x_{male} for $x_{male} = x_{canLift} = 1$ can be computed as follows

$$\begin{aligned} &\phi_{male} = \\ &\sum_{S \subseteq \{\emptyset, x_{canLift}\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}} \left(x_{S \cup \{i\}}\right) \\ &- f_S\left(x_S\right)] \\ &= \frac{1}{2}(f_{male}(x_{male}) - f_{\emptyset}(\emptyset)) \\ &+ \frac{1}{2}(f_{male,canLift}(x_{male}, x_{canLift}) - f_{canLift}(x_{canLift})) \\ &\stackrel{*}{=} \frac{1}{2}E[f(x_{male}, X_{canLift})] - \frac{1}{2}E[f(X_{male}, X_{canLift})] \\ &+ \frac{1}{2}E[f(x_{male}, x_{canLift})] - \frac{1}{2}E[f(X_{male}, x_{canLift})] \\ &= \frac{1}{2}P(X_{canLift} = 1) + \frac{1}{2} - \frac{1}{2}P(X_{male} = 1) \\ &- \frac{1}{2}P(X_{canLift} = 1 | X_{male} = 1)P(X_{male} = 1) \\ &= \frac{1}{2}p - \frac{1}{2}p \cdot 1 + \frac{1}{2} - \frac{1}{2} \cdot 1 = 0 \end{aligned}$$

where we used the definition $f(x_S) = E[f(x) | do\ x_S]$ from (Janzing et al., 2020) for marginal Shapley values in equation $*$.

C. Feature Selection

Let our feature space comprise two independent binary variables $X_1, X_2 \sim \mathrm{Normal}(1, 1)$. Our black box algorithm is defined by $f(x) = \mathbb{I}(x_1 > 1) \cdot 3x_2 - \mathbb{I}(x_1 \leq 1) \cdot x_2$ and the conditional feature attribution of feature 2 at $x = (0.5, 0.5)$

can be computed as follows

$$\begin{aligned}
\phi_2 = &\sum_{S \subseteq \{\emptyset, 1\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) \\
&- f_S\left(x_S\right)] \\
&= \frac{1}{2}(f_2(x_2) - f_\emptyset(\emptyset)) + \frac{1}{2}(f_{1,2}(x_1, x_2) - f_1(x_1)) \\
&\stackrel{*}{=} \frac{1}{2}E[f(X_1, x_2 = 0.5)] - \frac{1}{2}E[f(X_1, X_2)] \\
&+ \frac{1}{2}E[f(x_1 = 0.5, x_2 = 0.5)] - \frac{1}{2}E[f(x_1 = 0.5, X_2)] \\
&= \frac{1}{2}(0.5 \cdot 0.5 \cdot 3 - 0.5 \cdot 0.5) - \frac{1}{2}E[0.5 \cdot 3X_2 - 0.5X_2] \\
&- \frac{1}{2}0.5 - \frac{1}{2}E[-X_2] \\
&= \frac{1}{4} - \frac{1}{2} - \frac{1}{4} + \frac{1}{2} = 0
\end{aligned}$$

where we used $f_S(x_S) = E[f(x_S, X_{\bar{S}})]$ in equation $*$.