UNIVERSITÉ DE LA RÉUNION

Shapley Values: Sobol' versus Hoeffding

Thierry A. Mara

University of Reunion

12^{th} SAMO Summer School
June 24-28 2024
University of Parma

Outline

Notations & Assumptions

- $y = f(\boldsymbol{x})$ the scalar QoI
- $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \sim p_x = \dfrac{\partial^d F_x}{\partial x_1 \partial \ldots \partial x_d}$
- $x_i \sim p_i = \dfrac{\mathrm{d}F_i}{\mathrm{d}x_i}$, the marginal pdf ($F_i$ cdf) of x_i
- $p_{i|j} = \dfrac{\mathrm{d}F_{i|j}}{\mathrm{d}x_i}$, the conditional pdf (cdf) of x_i over x_j, ...
- $\mathcal{D} = \{0, 1, \ldots, d\}$, $\mathcal{D}_{-i} \bigcap \{i\} = \emptyset$, $\mathcal{D}_{+i} \bigcap \{i\} \neq \emptyset$
- $\boldsymbol{x} = (\boldsymbol{x}_\alpha, \boldsymbol{x}_{-\alpha})$ with $\boldsymbol{x}_\alpha \bigcap \boldsymbol{x}_{-\alpha} = \emptyset$
- $\alpha = \{i_1, \ldots, i_k\} \subseteq \mathcal{D} \Leftrightarrow \boldsymbol{x}_\alpha = \boldsymbol{x}_{i_1, \ldots, i_k} = (x_{i_1}, \ldots, x_{i_k})$
- $f_\alpha(\boldsymbol{x}_\alpha) = \mathbb{E}\left[f | \boldsymbol{x}_\alpha\right] = \int_{\mathbb{R}^{d-|\alpha|}} f(\boldsymbol{x}) p_{-\alpha|\alpha} \mathrm{d}\boldsymbol{x}_{-\alpha}$

Assumption: $\mathbb{E}\left[f^2\right] = \int_{\mathbb{R}^d} f^2(\boldsymbol{x}) p_x \mathrm{d}\boldsymbol{x} < \infty$

Hoeffding HDMR

Hoeffding's High-Dimensional Model Representation:

$$f(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}} f_{\alpha}^{H}(\boldsymbol{x}_{\alpha}) \tag{1}$$

with $f_0^H = \mathbb{E}\left[f(\boldsymbol{x})\right]$, and,

$$f_{\alpha}^{H}(\boldsymbol{x}_{\alpha}) = f_{\alpha}(\boldsymbol{x}_{\alpha}) - \sum_{\beta \subsetneq \alpha} f_{\beta}^{H}(\boldsymbol{x}_{\beta}) = \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} f_{\beta}(\boldsymbol{x}_{\beta}) \tag{2}$$

Hoeffding HDMR

Hoeffding's High-Dimensional Model Representation:

$$f(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}} f_\alpha^H(\boldsymbol{x}_\alpha) \tag{1}$$

with $f_0^H = \mathbb{E}[f(\boldsymbol{x})]$, and,

$$f_\alpha^H(\boldsymbol{x}_\alpha) = f_\alpha(\boldsymbol{x}_\alpha) - \sum_{\beta \subsetneq \alpha} f_\beta^H(\boldsymbol{x}_\beta) = \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} f_\beta(\boldsymbol{x}_\beta) \tag{2}$$

As a consequence, Eq.(1) can be rewritten:

$$f(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}} f_\alpha^H(\boldsymbol{x}_\alpha) = \sum_{\alpha \subseteq \mathcal{D}} \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} f_\beta(\boldsymbol{x}_\beta) \tag{3}$$

Hoeffding HDMR

Hoeffding's High-Dimensional Model Representation:

$$f(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}} f_{\alpha}^{H}(\boldsymbol{x}_{\alpha}) \tag{1}$$

with $f_0^H = \mathbb{E}\left[f(\boldsymbol{x})\right]$, and,

$$f_{\alpha}^{H}(\boldsymbol{x}_{\alpha}) = f_{\alpha}(\boldsymbol{x}_{\alpha}) - \sum_{\beta \subsetneq \alpha} f_{\beta}^{H}(\boldsymbol{x}_{\beta}) = \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} f_{\beta}(\boldsymbol{x}_{\beta}) \tag{2}$$

As a consequence, Eq.(1) can be rewritten:

$$f(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}} f_{\alpha}^{H}(\boldsymbol{x}_{\alpha}) = \sum_{\alpha \subseteq \mathcal{D}} \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} f_{\beta}(\boldsymbol{x}_{\beta}) \tag{3}$$

As a result:

- **Hoeffding's hdmr is unique (even when $p_x \neq \prod_{i=1}^{d} p_i$)**
- **The $f_{\alpha}^{H}(\boldsymbol{x}_{\alpha})$'s are pairwise $\perp$ only if $p_x = \prod_{i=1}^{d} p_i$**

Rosenblatt Transformations

Transforms $\boldsymbol{x} \sim p_x$ into $\boldsymbol{u} \sim \mathcal{U}(0,1)^d$

$$\begin{cases} u_{i_1} = F_{i_1}(x_{i_1}) \\ u_{i_2} = F_{i_2|i_1}(x_{i_2}|x_{i_1}) \\ \vdots \\ u_{i_d} = F_{i_d|i_1,\ldots,i_d}(x_{i_d}|\boldsymbol{x}_{-i_d}) \end{cases} \tag{4}$$

- the u-variables are independent of each other by definition
- RT is unique only if $p_x = \prod_{i=1}^{d} p_i$
- RT requires the knowledge of the conditional cdf's

Sobol' HDMR

RT turn $\boldsymbol{x} \sim p_x$ into $\boldsymbol{u} \sim \mathcal{U}(0,1)^d$ (i.e. $p_u = 1$). As a consequence it also turns $f(\boldsymbol{x})$ into $g(\boldsymbol{u})$. Sobol' hdmr is as follows,

$$g(\boldsymbol{u}) = \sum_{\alpha \subseteq \mathcal{D}} g_\alpha(\boldsymbol{u}_\alpha) \tag{5}$$

with $g_0 = \mathbb{E}\left[g(\boldsymbol{u})\right]$, and, $\int_0^1 g_\alpha(\boldsymbol{u}_\alpha)\mathrm{d}u_k = 0, \forall k \in \alpha$

Sobol' HDMR

RT turn $\boldsymbol{x} \sim p_x$ into $\boldsymbol{u} \sim \mathcal{U}(0,1)^d$ (i.e. $p_u = 1$). As a consequence it also turns $f(\boldsymbol{x})$ into $g(\boldsymbol{u})$. Sobol' hdmr is as follows,

$$g(\boldsymbol{u}) = \sum_{\alpha \subseteq \mathcal{D}} g_\alpha(\boldsymbol{u}_\alpha) \tag{5}$$

with $g_0 = \mathbb{E}[g(\boldsymbol{u})]$, and, $\int_0^1 g_\alpha(\boldsymbol{u}_\alpha)\mathrm{d}u_k = 0, \forall k \in \alpha$

As a result:

- **Sobol's hdmr is not unique as the RTs are not unique (unless $p_x = \prod_{i=1}^d p_i$)**
- **The g_α's are pairwise $\perp$**

Sobol' HDMR

RT turn $\boldsymbol{x} \sim p_x$ into $\boldsymbol{u} \sim \mathcal{U}(0,1)^d$ (i.e. $p_u = 1$). As a consequence it also turns $f(\boldsymbol{x})$ into $g(\boldsymbol{u})$. Sobol' hdmr is as follows,

$$g(\boldsymbol{u}) = \sum_{\alpha \subseteq \mathcal{D}} g_\alpha(\boldsymbol{u}_\alpha) \tag{5}$$

with $g_0 = \mathbb{E}\left[g(\boldsymbol{u})\right]$, and, $\int_0^1 g_\alpha(\boldsymbol{u}_\alpha)\mathrm{d}u_k = 0, \forall k \in \alpha$

As a result:

- **Sobol's hdmr is not unique as the RTs are not unique (unless $p_x = \prod_{i=1}^d p_i$)**
- **The g_α's are pairwise $\perp$**

Turning back to the original variables, we can write,

$$f(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}} f_\alpha^S(\boldsymbol{x}_\alpha) \tag{6}$$

It is shown that $f_\alpha^S = f_\alpha^H, \forall \alpha \subseteq \mathcal{D}$ if and only if $p_x = \prod_{i=1}^d p_i$

Shapley Value

Shapley values have several fomulations among which,

$$f(\boldsymbol{x}) = f_0^H + \sum_{i=1}^{d} \Phi_i(\boldsymbol{x}) \tag{7}$$

with

$$\Phi_i(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}_{+i}} \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} \frac{f_\beta(\boldsymbol{x}_\beta)}{|\alpha|} \tag{8}$$

Shapley Value

Shapley values have several fomulations among which,

$$f(\boldsymbol{x}) = f_0^H + \sum_{i=1}^{d} \Phi_i(\boldsymbol{x}) \tag{7}$$

with

$$\Phi_i(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}_{+i}} \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} \frac{f_\beta(\boldsymbol{x}_\beta)}{|\alpha|} \tag{8}$$

Let $\boldsymbol{x}^*$ be a given draw, $f(\boldsymbol{x}^*)$ the associated response:

The Shapley value $\Phi_i(\boldsymbol{x}^*)$ is the fair contribution of x_i to $f(\boldsymbol{x}^*)$. Fair = the mutual contributions (correlations+interactions) are fairly shared among the input variables.

Shapley Value & Hoeffding

Shapley values have several fomulations among which,

$$f(\boldsymbol{x}) = f_0^H + \sum_{i=1}^{d} \Phi_i(\boldsymbol{x}) \tag{9}$$

with

$$\Phi_i(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}_{+i}} \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} \frac{f_\beta(\boldsymbol{x}_\beta)}{|\alpha|} \tag{10}$$

By comparing, Eq.(10) to Eq.(3), that is,

$$f(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}} f_\alpha^H(\boldsymbol{x}_\alpha) = \sum_{\alpha \subseteq \mathcal{D}} \sum_{\beta \subseteq \alpha} (-1)^{|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|} f_\beta(\boldsymbol{x}_\beta)$$

We can infer that

$$\Phi_i(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}_{+i}} \frac{f_\alpha^H(\boldsymbol{x}_\alpha)}{|\alpha|} \tag{11}$$

We see that Hoeffding's hdmr is the key for interpreting Machine Learning outcomes.

Shapley Value: Computational Issue

Shapley values,

$$f(\boldsymbol{x}) = f_0^H + \sum_{i=1}^{d} \Phi_i(\boldsymbol{x})$$

with

$$\Phi_i(\boldsymbol{x}) = \sum_{\alpha \subseteq \mathcal{D}_{+i}} \frac{f_\alpha^H(\boldsymbol{x}_\alpha)}{|\alpha|}$$

Computing the Shapley Values requires the computation of the 2^d $f_\alpha(\boldsymbol{x}_\alpha)$ functions from which one can infer the $f_\alpha^H(\boldsymbol{x}_\alpha)$'s.
$f_\alpha(\boldsymbol{x}_\alpha) = \mathbb{E}\left[f(\boldsymbol{x})|\boldsymbol{x}_\alpha\right]$ can be estimated with any regression technique (that avoids over/under fitting)

Shapley Value: Computational Issue

Suppose that Sobol's hdmr has been obtained for the following ordering $(i_1, \ldots, i_d)$, it has been shown that (Mara & Tarantola 2012):

- $f_0 = \mathbb{E}[g(\boldsymbol{u})] = g_0$
- $f_{i_1}(x_{i_1}) = \mathbb{E}[g(\boldsymbol{u})|u_1] = g_0 + g_1(u_1)$
- $f_{i_1,i_2}(x_{i_1}, x_{i_2}) = \mathbb{E}[g(\boldsymbol{u})|u_1, u_2] = g_0 + g_1(u_1) + g_2(u_2) + g_{1,2}(u_1, u_2)$
- $\vdots$

d of the $f_\alpha(\boldsymbol{x}_\alpha)$ functions can be deduced from one single Sobol' hdmr

The remainder can be obtained with the Sobol's hdmr for a different ordering of the indexes $(1, \ldots, d)$

Shapley Value: Computational Issue

Suppose that Sobol's hdmr has been obtained for the following ordering $(i_1, \ldots, i_d)$, it has been shown that (Mara & Tarantola 2012):

- $f_0 = \mathbb{E}\left[g(\boldsymbol{u})\right] = g_0$
- $f_{i_1}(x_{i_1}) = \mathbb{E}\left[g(\boldsymbol{u})|u_1\right] = g_0 + g_1(u_1)$
- $f_{i_1,i_2}(x_{i_1}, x_{i_2}) = \mathbb{E}\left[g(\boldsymbol{u})|u_1, u_2\right] = g_0 + g_1(u_1) + g_2(u_2) + g_{1,2}(u_1, u_2)$
- $\vdots$

d of the $f_\alpha(\boldsymbol{x}_\alpha)$ functions can be deduced from one single Sobol' hdmr

The remainder can be obtained with the Sobol's hdmr for a different ordering of the indexes $(1, \ldots, d)$

Pros: BSPCE is known to provide efficiently the Sobol's hdmr

Cost: The cost to estimate the overall $f_\alpha(\boldsymbol{x}_\alpha)$'s is $\dfrac{d!}{\left(\frac{d}{2}!\right)^2}$

Con: RTs require the knowledge of the overall conditional and marginal cdfs.

Conclusion

Assumption: $p_x = \prod_{i=1}^{d} p_i$

	Hoeffding	Sobol'
Conditional cdfs knowledge	No	No
Uniqueness	Yes	Yes
Orthogonality	Yes	Yes
Shapley Values Estimate	Yes	Yes
Cost	1	1

Conclusion

Assumption: $p_x = \prod_{i=1}^{d} p_i$

	Hoeffding	Sobol'
Conditional cdfs knowledge	No	No
Uniqueness	Yes	Yes
Orthogonality	Yes	Yes
Shapley Values Estimate	Yes	Yes
Cost	1	1

Assumption: $p_x \neq \prod_{i=1}^{d} p_i$

	Hoeffding	Sobol'
Conditional cdfs knowledge	No	Yes
Uniqueness	Yes	No
Orthogonality	No	Yes
Shapley Values Estimate	Yes	Yes
Cost	2^d	$d!/\left(\frac{d}{2}!\right)^2$