On Locality of Local Explanation Models

Sahra Ghalebikesabi* (sahra.ghalebikesabi@stats.ox.ac.uk), Lucile Ter-Minassian*, Karla Diaz-Ordaz, Chris Holmes

University of Oxford, The London School of Hygiene & Tropical Medicine, and The Alan Turing Institute

Abstract

Shapley values provide model agnostic feature attributions for model outcome at a particular instance by simulating feature absence under a global population distribution. The use of a global population can lead to potentially misleading results when local model behaviour is of interest. Hence we consider the formulation of neighbourhood reference distributions that improve the local interpretability of Shapley values. By doing so, we find that the Nadaraya-Watson estimator, a well-studied kernel regressor, can be expressed as a self-normalised importance sampling estimator. Empirically, we observe that Neighbourhood Shapley values identify meaningful sparse feature relevance attributions that provide insight into local model behaviour, complimenting conventional Shapley analysis. They also increase on-manifold explainability and robustness to the construction of adversarial classifiers.

Definition of Shapley values. For a pre-defined value function $v(T, x)$ that takes a set of features $T \subseteq \{1, ..., m\}$ as input, the Shapley value $\phi_v(j, x)$ of feature j measures the expected change in the value function from including feature j into a random subset of features $S \subseteq \{1, ..., m\} \setminus j$ (without j)

$$\phi_v(j, x) = \mathop{\mathbb{E}}_{S} [v(S \cup j, x) - v(S, x)]$$

where the expectation is taken over the feature coalitions whose distribution is defined by $P(S) = \frac{|S|!(m-|S|)!}{m!}$. This choice of probability distribution ensures that sampling a set of size k has the same probability as sampling one of size l, $P(\{S \mid |S| = k\}) = P(\{S \mid |S| = l\})$ for $k, l \in \{0, ..., m-1\}$.
The value function is typically chosen as the expectation of the black box algorithm f at observation x over the not-included features $\overline{S}$ using a reference distribution $r(X^*_{\overline{S}} \mid x)$ such that

$$v(S, x) = \mathop{\mathbb{E}}_{r(X^*_{\overline{S}} \mid x)} [f(x_S, X^*_{\overline{S}})]$$

for $\overline{S} := \{1, \dots, m\}/S$ and the operation $(x_S, x_{\overline{S}})$ denoting the concatenation of its two arguments. Marginal Shapley values [4, 3] define $r(X^*_{\overline{S}} \mid x) := p(X^*_{\overline{S}})$ where p denotes the marginal data distribution. Conditional Shapley values [1] set the reference distribution equal to the conditional distribution given x_S, $r(X^*_{\overline{S}} \mid x) := p(X^*_{\overline{S}} | X^*_S = x_S)$. All in all, the Shapley value $\phi(j, x)$ is characterised by the expected change in model output, comparing the output when we include j in the model, i.e. integrate out some randomly sampled features $\overline{S} \setminus j$, with the model output where feature j is not included, i.e. we integrated out some randomly sampled features including j, $\overline{S}$

$$\phi(j, x) = \mathop{\mathbb{E}}_{S} \left[\mathop{\mathbb{E}}_{r(X^*_{S \cup \setminus j} \mid x)} [f(x_{S \cup j}, X^*_{\overline{S} \setminus j})] - \mathop{\mathbb{E}}_{r(X^*_{\overline{S}} \mid x)} [f(x_S, X^*_{\overline{S}})] \right].$$

Our contribution. As we see, Shapley values are computed by estimating the change in model outcome when some features are integrated out over the reference distribution $r(X^*_{\overline{S}} \mid x)$, which has so far been defined as either the marginal or conditional *global* population. For marginal Shapley values, the interpretation simplifies: The Shapley value of feature j is the expected change in model outcome when we sample a random individual x^* from the global statistical population and set its feature j equal to x_j (after we already set a random set of features $S \in \{1, \dots, m\} \setminus j$ equal to x_S). This motivates our proposal of neighbourhood distributions where we instead sample a random individual from the immediate neighbourhood of x.

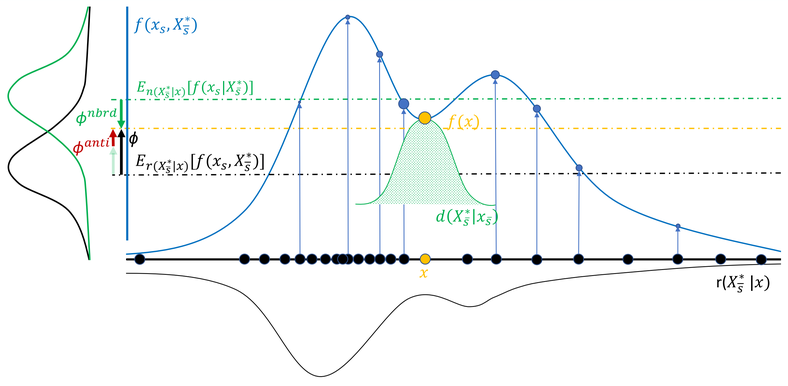

Figure 1: When sampling $\{x^*_i\}_{i=1}^L$ (black dots) from reference distribution $r(X^*_{\overline{S}} \mid x)$ (here $S = \emptyset$), the Shapley value ϕ at x is positive since $f(x)$ is larger than $\mathbb{E}_{r(X^*_{\overline{S}} \mid x)}[f(x_S, X^*_{\overline{S}})]$. In contrast, Neighbourhood SHAP ϕ^{nbrd} is negative since $\mathbb{E}_{n(X^*_{\overline{S}} \mid x)}[f(x_S, X^*_{\overline{S}})]$ is larger than $f(x)$. This difference results from the fact that, first, the model outcome has a local minimum at x, and second, $f(x_S, X^*_{\overline{S}})$ takes its smallest values at the tails of the data distribution (right-skewed density of $f(x_S, X^*_{\overline{S}})$ when $X^*_{\overline{S}} \sim p(X^*_{\overline{S}})$, black line on the left). SHAP only captures that $f(x)$ is higher than the average model outcome but not that $f(\cdot)$ is smaller at x than it is for any other close observation – this is reflected by Neighbourhood SHAP.

Instead of estimating the neighbourhood distribution, we approximate the expectation of the model outcome in the neighbourhood around x using self-normalised importance sampling [2] with proposal distribution $r(X^*_{\overline{S}} \mid x)$: $\mathbb{E}_{n(X^*_{\overline{S}} \mid x)}[f(x_S, X^*_{\overline{S}})] =$

$$\mathop{\mathbb{E}}_{r(X^*_{\overline{S}} \mid x)} \left[n_c \cdot d(X^*_{\overline{S}} \mid x_{\overline{S}}) f(x_S, X^*_{\overline{S}}) \right] \approx \frac{\sum_{i=1}^L d(x^*_{i,\overline{S}} \mid x_{\overline{S}}) f(x_S, x^*_{i,\overline{S}})}{\sum_{i=1}^L d(x^*_{i,\overline{S}} \mid x_{\overline{S}})}.$$

We note that the proposed local neighbourhood sampling scheme has a convenient form which corresponds to the well-known Nadaraya-Watson estimator [5, 7, 6] used for kernel regression.

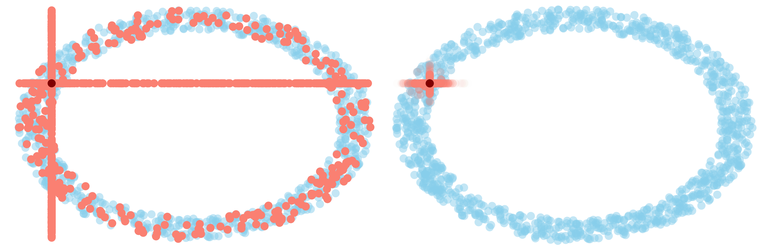

Figure 2: Concatenated data (pink dots) used for model evaluations for the computation of KernelSHAP (left) and Neighbourhood SHAP ($\sigma_{nbrd} = 0.1$, right) at a randomly sampled instance (maroon dots) where the data manifold is a ring in $\mathbb{R}^2$. Even though the background references (blue dots) lie on the data manifold, marginal Shapley values are evaluated at instances that lie off the data manifold.

Simulated experiment. As a simple motivating example as to why this question matters, consider a black box model given by $f(x) = x_1 > 0 2x_2^2 - x_1 \leq 0 x_2^2$ where $\cdot$ denotes the indicator function. When attributing the local feature importance at a test instance $x = (x_1, 2)$, with x_2 fixed at 2, we would expect Feature-1 to receive a higher absolute attribution when x is closer to the decision boundary at $x_1 = 0$. In Figure 3 we report the results on this example from LIME and SHAP as well as for our proposed 'Neighbourhood SHAP' approach.

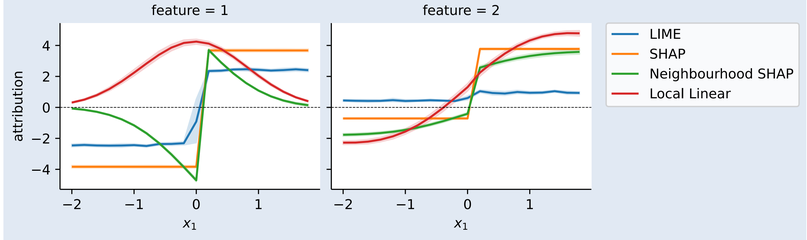

Figure 3: Attributions at $x = (x_1, 2)$ with x_1 varying for a reference distribution of $X \sim (0, 1)$ and black box $f(x) = x_1 > 0 2x_2^2 - x_1 \leq 0 x_2^2$ averaged over 10 runs displayed with 95% confidence intervals (see next section for details). While (Tabular) LIME and SHAP assign the same absolute attribution to Feature-1 no matter how large x_1 is, our neighbourhood approach takes its distance to the decision boundary into consideration. A local linear approximation to the black box trained with a Ridge Regressor gives misleading attributions to Feature-1 for $-0.4 < x_1 < 0$.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv preprint arXiv:1903.10464* (2019).

[2] Arnaud Doucet, Nando De Freitas, and Neil Gordon. "An introduction to sequential Monte Carlo methods". *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 3–14.

[3] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem". *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2907–2916.

[4] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[5] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142.

[6] David Ruppert and Matthew P Wand. Multivariate locally weighted least squares regression. *The annals of statistics* (1994), pp. 1346–1370.

[7] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* (1964), pp. 359–372.