Feature relevance quantification in explainable AI: A causal problem

Dominik Janzing **Lenon Minorics** **Patrick Blöbaum**
Amazon Research Tübingen, Germany

Abstract

We discuss promising recent contributions on quantifying feature relevance using Shapley values, where we observed some confusion on which probability distribution is the right one for dropped features. We argue that the confusion is based on not carefully distinguishing between *observational* and *interventional* conditional probabilities and try a clarification based on Pearl's seminal work on causality. We conclude that *unconditional* rather than *conditional* expectations provide the right notion of *dropping* features in contradiction to the theoretical justification of the software package SHAP. Parts of SHAP are unaffected because unconditional expectations (which we argue to be conceptually right) are used as *approximation* for the conditional ones, which encouraged others to 'improve' SHAP in a way that we believe to be flawed.

1 Motivation

Despite several impressive success stories of deep learning, not only researchers in the field have been shocked more recently about lack of robustness for algorithms that were actually believed to be powerful. Image classifiers, for instance, fail spectacularly once the images are subjected to adversarial changes that appear minor to humans, see e.g. Goodfellow et al. (2015); Sharif et al. (2016); Kurakin et al. (2018); Eykholt et al. (2018); Brown et al. (2018). Understanding these failures is challenging since it is hard to analyze which features were decisive for the classification in a particular case. However, lack of robustness is only one of several different motivations for getting artificial intelligence *interpretable*. Also the demand for getting *fair* decisions, e.g., Dwork et al. (2012); Kilbertus et al. (2017); Barocas et al. (2018), requires understanding of algorithms. In this case, it may even be subject of legal and ethical discussions *why* an algorithm came to a certain conclusion.

To formalize the problem, we describe the input / output behaviour as a function $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \rightarrow \mathbb{R}$ where $\mathcal{X}_1, \ldots, \mathcal{X}_n$ denote the ranges of some input variables $(X_1, \ldots, X_n) =: \mathbf{X}$ (discrete or continuous), while we assume the target variable Y to be real valued for reasons that will become clear later. Given one particular input $\mathbf{x} := (x_1, \ldots, x_n)$ we want to quantify to what extent each x_j is 'responsible' for the output $f(x_1, \ldots, x_n)$. This question makes only sense after specifying what should one input *instead*. Let us first consider the case where $\mathbf{x}$ is compared to some 'baseline' element $\mathbf{x}'$, which has been studied in the literature mostly for the case of real-valued inputs. Based on a hypothetical scenario where only some of the baseline values x'_j are replaced with x_j while others are kept, one wants to quantify to what extent each component j contributes to the difference $f(\mathbf{x}) - f(\mathbf{x}')$. The focus of the present paper, however, is a scenario where the baseline is defined by the expectation $\mathbb{E}[f(\mathbf{X})]$ over the underlying distribution $P_{\mathbf{X}}$. To explain the relevance of each j for the difference $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$ one considers a scenario where only some values are kept and the remaining ones are *averaged over some probability distribution*. The main contribution of this paper is to discuss which distribution is the right one. Recalling the difference between *interventional* and *observational* conditional distributions in the field of causality, we explain why we disagree with the interesting proposal of Lundberg and Lee (2017) in this regard. Further we argue that our criticism is irrelevant for any software that 'approximates' the conditional expectation (which we consider conceptually wrong) by the unconditional expectation, as proposed by Lundberg and Lee (2017). The paper is structured as follows. Section 2 summarizes results from the literature regarding axioms for feature attribution for the case where there is a unique baseline reference input. Here integrated gradients and Shapley values (as the generalization to discrete input) are the unique attribution functions for the stated set of axioms. Section 3 discusses the attribution problem for the case where one averages over unused features as in Lundberg and Lee (2017), and then

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

we present our criticism. We think that the big overlap of the present paper with existing literature is justified by aiming at this clarification only, while keeping this clarification as self-consistent as possible. In particular, the very general discussion of Datta et al. (2016) contains all the ideas of this work at least implicitly, but since it appeared before Lundberg and Lee (2017) it could not explicitly discuss the conceptual problems raised by the latter. Our view on marginalization over unused features is supported by Datta et al. (2016) for similar reasons. In Section 4 we present different experiments that show the practical consequences of our arguments.

2 Prior Work

The growth of deep neural networks recently motivated many researchers to investigate feature attribution, see e.g. Shrikumar et al. (2016) for DeepLIFT, Binder et al. (2016) for Layer-wise Relevance Propagation (LRP), Ribeiro and Singh (2016) for Local Interpretable Model-agnostic Explanations (LIME), and for gradient based methods Chattopadhyay et al. (2019). For a summary of common architecture agnostic methods, see Molnar (2019). We first discuss two closely related concepts that arise from an axiomatic approach.

2.1 Integrated gradient

Sundararajan et al. (2017), investigated the attribution of x_i to the difference

$$f(\mathbf{x}) - f(\mathbf{x}'), \tag{1}$$

where $\mathbf{x}'$ is a given baseline. Under the assumption that f is differentiable almost everywhere[1], they defined the attribution of x_i to (1) as

$$\text{IntegratedGrads}_i(\mathbf{x}; f) := (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i}\, d\alpha.$$

Contrary to LIME, DeepLIFT and LRP, this attribution method has the advantage that all of the following 5 properties are satisfied (see Sundararajan et al. (2017) and Aas et al. (2019)):

1. *Completeness:* If $atr_i(\mathbf{x}; f)$ denotes the attribution of x_i to (1), then
$$\sum_i atr_i(\mathbf{x}; f) = f(\mathbf{x}) - f(\mathbf{x}').$$
2. *Sensitivity:* If f does not depend on x_i, then $atr_i(\mathbf{x}; f) = 0$.
3. *Implementation Invariance:*[2] If f and f' are equal for all inputs, then
$$atr_i(\mathbf{x}; f) = atr_i(\mathbf{x}; f') \quad \text{for all} \quad i.$$
4. *Linearity:* For $a, b \in \mathbb{R}$ holds
$$atr_i(\mathbf{x}; af_1 + bf_2) = a \cdot atr_i(\mathbf{x}; f_1) + b \cdot atr_i(\mathbf{x}; f_2).$$
5. *Symmetry-Preserving:* If f is symmetric in component i and j and $x_i = x_j$ and $x_i' = x_j'$, then
$$atr_i(\mathbf{x}; f) = atr_j(\mathbf{x}; f).$$

Integrated gradients can be generalized by integrating over an arbitrary path γ instead of the straight line. This attribution method is called *path method* and the following theorem holds.

Theorem 1. *((Friedman, 2004, Theorem 1) and (Sundararajan et al., 2017, Theorem 1)) If an attribution method satisfies the properties Completeness, Sensitivity, Implementation Invariance and Linearity, then the attribution method is a convex combination of path methods. Furthermore, integrated gradients is the only path method that is symmetry preserving.*

Notice that convex combinations of path methods can also be symmetry preserving even if the attribution method is not given by integrated gradients.

2.2 Shapley values

To assess feature relevance *relative to the average*, Lundberg and Lee (2017) use a concept that relies on first defining an attribution for binary functions, or, equivalently, functions with subset as input ('set functions'). We first explain this concept and describe in Section 3 how it solves the attribution relative to the expectation. Assume we are given a set with n elements, say $U := \{1, \ldots, n\}$ and a function

$$g : 2^U \to \mathbb{R} \quad \text{ with } g(U) \neq 0,\ g(\emptyset) = 0.$$

We then ask to what extent each single $j \in U$ contributes to $g(U)$. A priori, the contribution of each j depends on the order in which more elements are included. We can thus define the contribution of j, given $T \subseteq U \setminus \{j\}$ by

$$C(j|T) := g(T \cup \{j\}) - g(T)$$

[1] see Sundararajan et al. (2017, Proposition 1)

[2] Note that this axiom is pointless if it refers to properties of *functions* rather than properties of *algorithms*. We have listed it for completeness and for consistency with the literature.

(note that it can be negative and also exceed $g(U)$). With

$$\phi_i := \sum_{T \subseteq U \setminus \{i\}} \frac{1}{n\binom{n-1}{|T|}} C(i|T). \quad (2)$$

it then holds

$$g(U) = \sum_{i=1}^{n} \phi_i.$$

The quantity ϕ_i is called the *Shapley value* (Shapley, 1953) of i, which can be considered the average contribution of i to $g(U)$. At first glance, Shapley values only solve the attribution problem for binary inputs by canonically identifying subsets T with binary words $\{0,1\}^n$. To show that Shapley values also solve the above attribution problem, one can simply define a set function by

$$g(T) := f_T(\mathbf{x}_T) - f(\mathbf{x}'),$$

for any subset $T \subseteq \{1,\ldots,n\}$. Here, f_T is the 'simplified' function with the reduced input $\mathbf{x}_T$ obtained from f when all remaining features are taken from the baseline input $\mathbf{x}'$, that is, $f_\emptyset(\mathbf{x}_\emptyset) = f(\mathbf{x}')$.

Since Shapley Values also satisfy Completeness, Sensitivity, Implementation Invariance and Linearity (Aas et al., 2019) with respect to the binary function defined by the set function g, they are given by a convex combination of path methods. Furthermore, Shapley Values with respect to g are symmetry-preserving, but don't coincide with integrated gradients.

Different ways of feature attribution based on Shapley Values were recently investigated by Sundararajan and Najmi (2019). Their main consideration is feature relevance relative to an auxiliary baseline, but feature attribution relative to the expectation (according to an arbitrary distribution) is also mentioned. Furthermore, Sundararajan and Najmi (2019) already discussed that Shapley Values based on conditional distributions can assign unimportant features nonzero attribution. However, Sundararajan and Najmi (2019) didn't consider the problem from a causal perspective.

3 How should we sample the *dropped* features?

We now want to attribute the difference between $f(\mathbf{x})$ and the expectation $\mathbb{E}[f(\mathbf{X})]$ to individual features. Explaining *why* the output for one particular input $\mathbf{x}$ deviates strongly from the average output is particularly interesting for understanding 'outliers'. Let us introduce some notation first. For any $T \subseteq U$ let $\mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})|\mathbf{X}_T = \mathbf{x}_T]$ denote the conditional expectation of f, given $\mathbf{X}_T = \mathbf{x}_T$. By $\mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})]$ we denote the expectation of $f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})$ with respect to the distribution of $\mathbf{X}_{\bar{T}}$ *without conditioning* on $\mathbf{X}_T = \mathbf{x}_T$. Let us call this expression *'marginal expectation'* henceforth.

Accordingly, we now discuss two different options for defining 'simplified functions' f_T where all features from $\bar{T}$ are dropped:

$$f_T(\mathbf{x}) := \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})|\mathbf{X}_T = \mathbf{x}_T] \quad (3)$$

$$\textbf{vs} \quad f_T(\mathbf{x}) := \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})] \quad \textbf{?} \quad (4)$$

Lundberg and Lee (2017) propose (3), but since it is difficult to compute they *approximate* it by (4), which they justify by the simplifying assumption of feature independence. Using the set function $g(T) := f_T(\mathbf{x}) - f_\emptyset(\mathbf{x})$, they compute Shapley values ϕ_i according to (2). We will argue that using (4) rather than (3) is conceptually the right thing in the first place. Our clarification is supposed to prevent others from 'improving' SHAP by finding an approximation for the conditional expectation that is better than the marginal expectation, like, for instance Aas et al. (2019) and Lundberg et al. (2018).[3]

To explain our arguments, let us first discuss why marginal expectations occur naturally in the field of causal inference.

Observational versus interventional conditional distributions The main ideas of this paragraph can already be found in Datta et al. (2016) in more general and abstract form, see also Friedman (2001) and Zhao and Hastie (2019), but we want to rephrase them in a way that optimally prepares the reader to the below discussion. Assume we are given the causal structure shown in Figure 1.

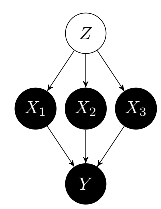

Figure 1: A simple causal structure where the observational conditional $p(y|x_1)$ does not correctly describe how Y changes after *intervening* on X_1 because the common cause Z 'confounds' the relation between X_1 and Y. Z is drawn in white color because it may be latent.

Further, assume we are interested in how the expectation of Y changes when we manually set X_1 to some value x_1. This is *not* given by $\mathbb{E}[Y|X_1 = x_1]$ because observing $X_1 = x_1$ changes also the distribution of X_2, X_3 due to the dependences between X_1 and X_2, X_3 (which are generated by the common cause Z). This way, the difference between $\mathbb{E}[Y]$ and $\mathbb{E}[Y|X_1 = x_1]$ is not only due to the

[3]Note that TreeExplainer in SHAP has meanwhile been changed accordingly.

influence of X_1, but can also be caused by the influence of X_2, X_3. The impact of setting X_1 to x_1 is captured by Pearl's do-operator Pearl (2000) instead, which yields

$$\begin{aligned} &\mathbb{E}[Y|do(X_1 = x_1)] \\ &= \int \mathbb{E}[Y|x_1, x_2, x_3] p(x_2, x_3) dx_2 dx_3. \end{aligned} \tag{5}$$

This can be easily verified using the backdoor criterion Pearl (2000) since (phrased in Pearl's language) the variables X_2, X_3 'block the backdoor path' $X_1 \leftarrow Z \rightarrow Y$. Observations from Z are not needed, we may therefore assume Z to be latent.

For our purpose, two observations are important: first, (5) does not contain the conditional distribution, given $X_1 = x_1$. Replacing $p(x_2, x_3)$ with $p(x_2, x_3|x_1)$ in (5) would yield the *observational* conditional expectation $\mathbb{E}[Y|X_1 = x_1]$, which we are not interested in. In other words, the intervention on X_1 breaks the dependences to X_2, X_3. The second observation that is crucial for us is that the dependences between X_2, X_3 are kept, they are unaffected by the intervention on X_1.

Why observational conditionals are flawed Let us start with a simple example.

Example 1 (irrelevant feature)**.** *Assume we have*

$$f(x_1, x_2) = x_1.$$

Obviously, the feature X_2 is irrelevant. Let both X_1, X_2 be binaries and

$$p(x_1, x_2) = \begin{cases} 1/2 & \textit{for } x_1 = x_2 \\ 0 & \textit{otherwise} \end{cases}.$$

(1) with conditional expectations:

$$\begin{aligned} f_\emptyset(\mathbf{x}) &= \mathbb{E}[f(X_1, X_2)] = 1/2 & (6) \\ f_{\{1\}}(\mathbf{x}) &= \mathbb{E}[f(x_1, X_2)|x_1] = x_1 & (7) \\ f_{\{2\}}(\mathbf{x}) &= \mathbb{E}[f(X_1, x_2)|x_2] = x_2 & (8) \\ f_{\{1,2\}}(\mathbf{x}) &= f(x_1, x_2) = x_1 & (9) \end{aligned}$$

Therefore,

$$\begin{aligned} C(2|\emptyset) &= f_{\{2\}}(\mathbf{x}) - f_\emptyset(\mathbf{x}) = x_1 - 1/2 \\ C(2|\{1\}) &= f_{\{1,2\}}(\mathbf{x}) - f_{\{1\}}(\mathbf{x}) = x_1 - x_1. \end{aligned}$$

Hence, the Shapley value for X_2 reads:

$$\phi_2 = \frac{1}{2}(x_1 - 1/2 + x_1 - x_1) = x_1/2 - 1/4 \neq 0.$$

(2) with marginal expectations:

$$\begin{aligned} f_\emptyset(\mathbf{x}) &= \mathbb{E}[f(X_1, X_2)] = 1/2 & (10) \\ f_{\{1\}}(\mathbf{x}) &= \mathbb{E}[f(x_1, X_2)] = x_1 & (11) \\ f_{\{2\}}(\mathbf{x}) &= \mathbb{E}[f(X_1, x_2)] = 1/2 & (12) \\ f_{\{1,2\}}(\mathbf{x}) &= f(x_1, x_2) = x_1. & (13) \end{aligned}$$

We then obtain

$$\begin{aligned} C(2|\emptyset) &= f_{\{2\}}(\mathbf{x}) - f_\emptyset(\mathbf{x}) = 0 \\ C(2|\{1\}) &= f_{\{1,2\}}(\mathbf{x}) - f_{\{1\}}(\mathbf{x}) = 0, \end{aligned}$$

which yields $\phi_2 = 0$.

The example proves the follow result, which were already discussed in Sundararajan and Najmi (2019):

Lemma 1 (failure of Sensitivity)**.** *When the relevance of ϕ_i is defined by defining 'simplified' functions f_T via conditional expectations*

$$f_T(\mathbf{x}_T) := \mathbb{E}[f(\mathbf{x})|\mathbf{X}_{\bar{T}} = \mathbf{x}_{\bar{T}}],$$

then $\phi_i \neq 0$ does not imply that f depends on x_i.

The example is particularly worrisome because we mentioned earlier that Shapley values satisfy the axiom of sensitivity, while Lemma 1 seems to claim the opposite. The resolve this paradox, note that the Shapley values refer to binary functions (or set functions) and reading (6) to (8) as the values of a binary function $\tilde{g}$ with inputs $(z_1, z_2) = 00, 10, 01, 11$ we clearly observe that $\tilde{g}$ depends also on the second bit. This way, the Shapley values do not violate sensitivity for $\tilde{g}$, but we certainly care about 'sensitivity for f'. Note that this distinction between the binary function $\tilde{g}$ and f is crucial although in our example f is binary itself. Fortunately, the second bit is irrelevant for the binary function $\tilde{g}$ defined by (10) and (13) and we do not obtain the above paradox.

To assess the impact of changing the inputs of f, we now switch to a more causal language and state that we consider the inputs of an algorithm as *causes* of the output. Although this remark seems trivial it is necessary to emphasize that we are not talking about the causal relation between any features in the real world outside the computer (where the attribute predicted by Y may be the cause of the features), but only about causality of this technical input / output system[4]. To facilitate this view, we formally distinguish between the true features $\tilde{X}_1, \dots, \tilde{X}_n$ obtained from the objects and the corresponding features $X_1, \dots, X_n$ plugged into the algorithm. This way, we are able to talk about a hypothetical scenario where the inputs are changed compared to the true features. Let us first consider the causal structure in figure 2, top, where the inputs are determined by the true features. In contrast, figure 2, bottom, shows the causal structure after an intervention on X_1, X_2 has adjusted these variables to fixed values x_1, x_2.

We now consider the impact of an hypothetical intervention, which leaves the remaining components *unaffected*. They are therefore sampled from their natural joint distribution *without* conditioning. Similar to the above paragraph,

[4] Accordingly, Y is the output of the system and not a property of the external world.

we then obtain

$$\mathbb{E}[Y|do(\mathbf{X}_T = \mathbf{x}_T)] = \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})]. \tag{14}$$

Our formal separation between the *true* values of the features $\tilde{X}_j$ of some object and the corresponding *inputs* X_j of the algorithms allows us to be agnostic about the causal relations between the true features in the real world, the fact that the inputs $X_1, \dots, X_n$ *cause* the output Y is the only causal knowledge needed to compute (14). Since the interventional expectations coincide with the marginal expectations, we have thus justified the use of marginal expectations for the Shapley values from the causal perspective.

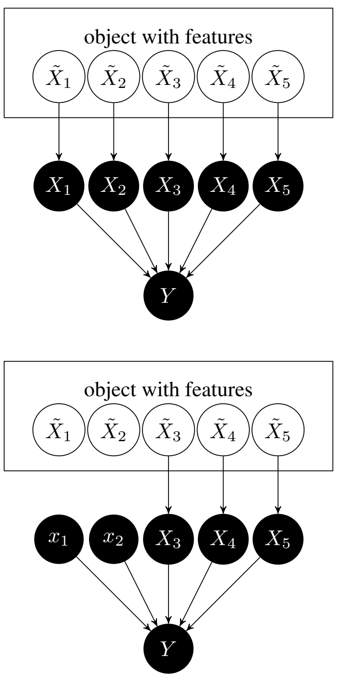

Figure 2: Top: Causal structure of our prediction scenario: The output Y is determined by the inputs $X_1, \dots, X_n$. In the usual learning scenario these inputs coincide with features $\tilde{X}_1, \dots, \tilde{X}_n$ of some object, that is $X_j = \tilde{X}_j$. Bottom: To evaluate the impact of some inputs, say X_1, X_2, for the output Y we consider a hypothetical scenario where we adjust these inputs to some fixed values x_1, x_2 and sample the remaining inputs from the usual joint distribution $P_{X_3,\dots,X_n}$.

The problem with the symmetry axiom We briefly rephrase Example 4.9 of Sundararajan and Najmi (2019) showing that the symmetry axiom is violated when Shapley values are used for quantifying the influence relative to conditional or marginal expectations. Figure 3 shows values and probabilities of two random variables X_1 and X_2 and the values of the function $f(X_1, X_2) = X_1 + X_2$. As explained by Sundararajan and Najmi (2019), for the input $(x_1, x_2) = (2, 2)$ the value x_1 gets attribution $(1-p)$ and x_2 gets attribution $(1-q)$. Therefore, if $p \neq q$, x_1 and x_2 get different attribution, although f is symmetric. They conclude that this is a violation of symmetry. Since X_1 and X_2 are independent, this problem occurs regardless of whether one defines the simplified function f_T with respect to marginal or conditional expectations. One can argue, however, that this result makes intuitively sense because the value x_j that is farther from its mean contributes *more* to the fact that $f(x_1, x_2)$ deviates from its mean. If we have even $x_1 = \mathbb{E}[X_1]$, we would certainly say that x_1 does not contribute to the deviation from the mean at all. For this reason we do not follow Sundararajan and Najmi (2019) in regarding this phenomenon as a problem of this kind of attribution analysis. Recall furthermore that we have already mentioned that the symmetry axiom does hold for the corresponding binary function defined by including or not certain features (simply because symmetry holds for Shapley values). For the above example this binary function is indeed asymmetric. To check this, define

Probability	X_1	X_2	$f = X_1 + X_2$
$(1-p)\cdot(1-q)$	1	1	2
$(1-p)\cdot q$	1	2	3
$(1-q)\cdot p$	2	1	3
$p\cdot q$	2	2	4

Figure 3: Table 3 from Sundararajan and Najmi (2019) which shows an example for alleged lack of symmetry of Shapley Values with respect to the marginal expectation.

$$\tilde{g}(z_1, z_2) := \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})],$$

where T is the set of all j for which $z_j = 1$. This function is not symmetric in Z_1 and Z_2, since we have, for instance, $\tilde{g}(1,0) = x_1 + \mathbb{E}[X_2] \neq \tilde{g}(0,1) = x_2 + \mathbb{E}[X_1]$.

4 Numerical Evidence

In this section, we show numerically that the marginal expectation $\mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})]$ is a better choice than $\mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})|\mathbf{X}_T = \mathbf{x}_T]$ to quantify the attribution of each observation x_j of a particular input $\mathbf{x} = (x_1, \dots, x_n)$ to $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$. Admittedly, one can argue that the experiments are pointless because simple equations prove this claim. However, some readers get a better intuition of conceptual problems when simple experiments show a significant difference between the two different averages.

4.1 Computation of Shapley Values

As explained by Aas et al. (2019, Section 2.3), the implementation of KernelSHAP (Lundberg and Lee, 2017) consists of two parts:

1. Using a representation of Shapley Values as the solution of a weighted least square problem for a computationally tractable approximation.

2. Approximation of $g(T)$.

4.1.1 Shapley Values as solution of weighted least square problem

By Charnes et al. (1988), the Shapley Values to the set function g are given as the solution $(\phi_1, \ldots, \phi_n)$ of

$$\min_{\phi_1,\ldots,\phi_n} \left\{ \sum_{T \subseteq U} \left[g(T) - \Big(\sum_{j \in T} \phi_j \Big) \right]^2 k(U,T) \right\}, \quad (15)$$

where $k(U,T) = (|U|-1)/(\binom{|U|}{|T|}|T|(|U|-|T|))$ are the *Shapley kernel weights*. Since $k(U,U) = \infty$, we use the constraint $\sum_j \phi_j = g(U)$, or, for numerical calculation, we set $k(U,U)$ to a large number.

Since the power set of U consists of 2^n elements, the computation time of the Sharpley Values increases exponentially. KernelSHAP therefore samples subsets of U according to the probability distribution induced by the Shapley kernel weights.

4.1.2 Approximation of the set function

As discussed in the previous sections, Lundberg and Lee (2017) define

$$f_T(\mathbf{x}) = \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}}) | \mathbf{X}_T = \mathbf{x}_T].$$

To evaluate the conditional expectation, they assume feature independence (or weak dependence) to obtain $\mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}}) | \mathbf{X}_T = \mathbf{x}_T] \approx \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})]$ and use the approximation

$$f_{T,\text{KernelSHAP}}(\mathbf{x}) \approx \frac{1}{K} \sum_k f(\mathbf{x}_T, \mathbf{x}_{\bar{T}}^k), \quad (16)$$

where $\mathbf{x}_{\bar{T}}^k$, $k = 1, \ldots, K$ are our samples from $\mathbf{X}_{\bar{T}}$.

4.2 Experiments

To show in an experimental setup that the marginal expectation is a better choice, we consider functions f for which we can calculate analytically the attribution of x_j. This is possible for linear functions

$$f(\mathbf{x}) = \alpha_0 + \sum_i \alpha_i x_i, \quad \alpha_i \in \mathbb{R}$$

since

$$f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})] = \sum_i \alpha_i (x_i - \mathbb{E}X_i)$$

and hence, the attribution of x_j is $\alpha_j(x_j - \mathbb{E}[X_j])$. Our experiments are divided into the following setups:

1. We assume that the feature vector $\mathbf{X}$ follows a multivariate Gaussian distribution.
2. We use a kernel estimation to approximate the conditional expectation.

For the experiments, we use the KernelExplainer class of the python SHAP package from Lundberg and Lee (2017) to calculate Shapley Values with respect to the marginal expectation and the R package SHAPR, in which the methodology of Aas et al. (2019) is implemented, to calculate Shapley Values with respect to the conditional distribution.

Notice that calculating Shapley Values is also possible for non-linear functions. Further, approximating the marginal expectation is computationally inexpensive compared to the approximation of the conditional expectation with kernel estimation.

4.2.1 Multivariate Gaussian distribution

If $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with some mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, it holds that

$$\mathbb{P}(\mathbf{X}_{\bar{T}} | \mathbf{X}_T = \mathbf{x}_T) = N(\boldsymbol{\mu}_{\bar{T}|T}, \boldsymbol{\Sigma}_{\bar{T}|T}),$$

(see (Aas et al., 2019, Section 3.1)), where

$$\begin{aligned} \boldsymbol{\mu}_{\bar{T}|T} &= \boldsymbol{\mu}_{\bar{T}} + \boldsymbol{\Sigma}_{\bar{T}T}\boldsymbol{\Sigma}_{TT}^{-1}(\mathbf{x}_T - \boldsymbol{\mu}_T) \\ \boldsymbol{\Sigma}_{\bar{T}|T} &= \boldsymbol{\Sigma}_{\bar{T}\bar{T}} - \boldsymbol{\Sigma}_{\bar{T}T}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{T\bar{T}}, \end{aligned}$$

with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_T \\ \boldsymbol{\mu}_{\bar{T}} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{TT} & \boldsymbol{\Sigma}_{T\bar{T}} \\ \boldsymbol{\Sigma}_{\bar{T}T} & \boldsymbol{\Sigma}_{\bar{T}\bar{T}} \end{pmatrix}.$$

Hence, we can approximate the conditional expectation by sampling $X_{\bar{T}}$ directly from its distribution.

We simulate Gaussian data and run the experiment for different number of features. For every experiment with multivariate Gaussian distribution, we set the intercept to 0, i.e. $\alpha_0 = 0$.

Dimension n=3. In the first 3-dimensional experiment, we let $\alpha_1 = 0$ and choose in every run α_1 and α_2 independently from the standard normal distribution. Further, we let $\boldsymbol{\mu} = (0,0,0)^T$ and $\boldsymbol{\Sigma} = cc^T$, where we choose the entries of c in every run independently from the standard normal distribution and $\mathbf{x}$ also randomly in every run. The number of runs and the sample size of $\mathbf{X}$ is 1000. Figure 4 shows the errors $\phi_j - \text{contr}_j(\mathbf{x})$ of the Shapley Values ϕ_j with respect to the set function $g(T) = \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}})] - \mathbb{E}f(\mathbf{X})$ (blue) and the set function $g(T) = \mathbb{E}[f(\mathbf{x}_T, \mathbf{X}_{\bar{T}}) | \mathbf{X}_T = \mathbf{x}_T] - \mathbb{E}f(\mathbf{X})$ (red). The very precise results for the marginal expectation are mainly from feature 1.

Dimension n=10. In 10-dimensions, we take almost the same setting with the difference that we set the first 3 coefficients to zero, i.e. $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Again, the very precise results for the marginal expectation are from the features whose coefficients we set to 0.

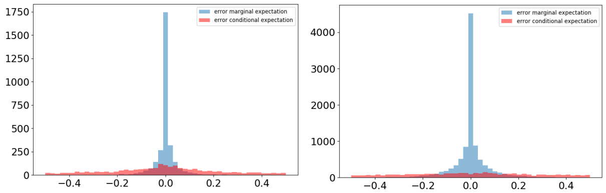

Figure 4: Histogram showing the error of the Shapley Values for multivariate Gaussian distribution in the 3-dimensional (left) and 10-dimensional (right) setting with $\alpha_1 = 0$. Blue: error using marginal expectation, Red: error using conditional expectation.

4.2.2 Approximation via kernel estimation

If we have no information about the underlying distribution, it is hard to approximate the conditional distribution sufficiently. However, in low dimensions kernel estimates can provide a good approximation. We take the kernel estimation method from Aas et al. (2019) to show how strongly the Shapley Values w.r.t. conditional expectation deviate from $\alpha_j(x_j - \mathbb{E}[X_j])$. Their approximation is as follows:

1. Let Σ_T be the covariance matrix of our samples from $\mathbf{X}_T$. To each point $\mathbf{x}^i$, calculate the Mahalanobis distance (see Mahalanobis (1936))

   $$\mathrm{dist}_T(\mathbf{x},\mathbf{x}^i) := \sqrt{\frac{(\mathbf{x}_T - \mathbf{x}_T^i)'\Sigma_T^{-1}(\mathbf{x}_T - \mathbf{x}_T^i)}{|T|}},$$

   where $(\mathbf{x}_T - \mathbf{x}_T^i)'$ denotes the transpose of $(\mathbf{x}_T - \mathbf{x}_T^i)$.

2. Calculate the *Kernel weights*

   $$w_T(\mathbf{x},\mathbf{x}^i) := \exp\left(-\frac{\mathrm{dist}_T(\mathbf{x},\mathbf{x}^i)^2}{2\sigma^2}\right).$$

   Hereby, $\sigma^2 > 0$ is a bandwidth which has to be specified.

3. Sort the weights $w_T(\mathbf{x},\mathbf{x}^i)$ in increasing order and let $\tilde{\mathbf{x}}^i$ be the corresponding ordered sampling instances. Then, approximate $g(T)$ by

   $$g_{\mathrm{cond}}(T) := \frac{\sum_{i=1}^K w_T(\mathbf{x},\tilde{\mathbf{x}}^i) f(\mathbf{x}_{\bar{T}}^i, \mathbf{x}_T)}{\sum_{i=1}^K w_T(\mathbf{x},\tilde{\mathbf{x}}^i)}.$$

For the experiment, we use the data set *Human Activity Recognition Using Smartphones Data Set* (see Anguita et al. (2013)) from the UCI repository. The data set consists of 561 features with a training sample size of 7352 and test sample size of 2948. In this experiment, we merge these samples together and therefore our sample size is 10299. We take randomly 4 features and train a linear model with 3 of these features as inputs and with the 4-th feature as target. We don't consider the label (which is a daily activity performed by the human) of the data set, but the different features have the true label as a common cause. Notice that we are not interested in the quality of the model, but rather in a model for which the ground truth of the attribution is known (because we can certainly look at the linear model obtained).

Afterwards, we calculate the Shapley Values with SHAP and SHAPR (with σ^2 set to 0.1 in SHAPR which is the default value) using the first 1000 samples and approximate the expected value $\mathbb{E}X_j$ using the whole data set. The observation $\mathbf{x}$ is also randomly picked from the data and we run this experiment 1000 times. Figure 5 shows the histogram of the error $\phi_j - \mathrm{contr}_j(\mathbf{x})$ for the marginal expectation (blue) and conditional expectation (red).

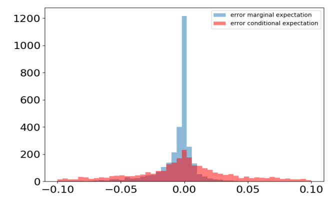

Figure 5: Histogram showing the error of the Shapley Values for the data set *Human Activity Recognition Using Smartphones Data Set*. Blue: error using marginal expectation, Red: error using conditional expectation.

5 Conclusion

In this work we considered the problem of attributing the output from one particular multivariate input to individual features. We argued that there is a misconception also in recent proposals for feature attribution because they use observational conditional distributions rather than interventional distributions. Our arguments are phrased in terms of the causal language introduced by Pearl (2000). We argue that parts of the package SHAP from Lundberg and Lee (2017) are unaffected by this misconception (although the corresponding theory part of the paper suffers from this issue) since it 'approximates' the observational expectations by an expression that would have been the right one in the first place. We think that this clarification is important since other authors tried to 'improve' the SHAP package in a way that we consider conceptually flawed. Moreover, we revisited some properties that were stated as desirable in the context of attribution analysis. If stated in a too vague manner, there is some room for interpretation. We argued, for instance, why we think that our attribution method satisfies a reasonable symmetry property, since attribution via interventional probabilities has been criticised for violating alleged desirable symmetry properties.

Acknowledgements: The authors would like to thank Scott Lundberg and Anders Løland for their valuable feedback and Atalanti Mastakouri for remarks on the presentation.

References

K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. ArXiv:1903.10464, 2019.

D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 24–26, April 2013.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. http://www.fairmlbook.org.

A. Binder, G. Montavon, S. Lapuschkin, K. R. Müller, and W. Samek. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In *Artificial Neural Networks and Machine Learning – ICANN*, volume 9887, 2016.

T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial Patch. ArXiv:1712.09665, 2018.

A. Charnes, B. Golany, M. Keane, and J. Rousseau. Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations. *Econometrics of Planning and Efficiency*, 11:123–133, 1988.

A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. Balasubramanian. Neural network attributions: A causal perspective. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 981–990, Long Beach, California, USA, 09–15 Jun 2019.

A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.

K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1625–1634, 2018.

E. J. Friedman. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4):501–518, 2004.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings from the conference Neural Information Processing Systems (NeurIPS) 2017*, pages 656–666.

A. Kurakin, I. J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018.

S. Lundberg and S. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings from the Conference Neural Information Processing Systems (NeurIPS) 2017*, pages 4765–4774.

S. Lundberg, G. Erion, and S. Lee. Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888, 2018.

P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, April 1936.

C. Molnar. *Interpretable Machine Learning*. Molnar, C., 2019. https://christophm.github.io/interpretable-ml-book/.

J. Pearl. *Causality*. Cambridge University Press, 2000.

M. Ribeiro and C. Singh, S.and Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, NY, USA, 2016.

L. Shapley. A value for n-person games. In H. Kuhn und A. Tucker, editors, *Contributions to the Theory of Games (AM-28)*, 2, Princeton University Press 1953.

M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, 2016.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features Through Propagating Activation Differences. In D. Precup and Y. Teh, editors, *Proceedings of of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153, 2016.

M. Sundararajan and A. Najmi. The many Shapley values for model explanation. *arXiv:1908.08474*, 2019.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of Proceedings of Machine Learning Research, pages 3319–3328, August 2017.

Q. Zhao and T. Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, pages 1–10, 2019.