Causal Shapley values

Anonymous Author(s)
Affiliation
Address
email

Abstract

Explainability. Causal interpretation of Shapley values.

1 Introduction

Explainability. Attribution. Shapley. Conditioning or no conditioning.

2 A causal interpretation of Shapley values

Need more smooth talk.

We assume that we are given a set of features $\mathbf{x}$ with corresponding model output $f(\mathbf{x})$. We compare this output with the average output

$$f_0 = \mathbb{E} f(\mathbf{X}) = \int d\mathbf{X} \, P(\mathbf{X}) f(\mathbf{X}) \, ,$$

with expectation taken over some (for now assumed to be known) probability distribution $P(\mathbf{X})$. We would like to attribute the difference between $f(\mathbf{x})$ and f_0 in a sensible way to the different features $i \in N$ with $N = \{1, \ldots, n\}$ and n the number of features. That is, we would like to write

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} \phi_i \, , \tag{1}$$

where we will refer to ϕ_i as the contribution of feature i to the output $f(\mathbf{x})$. Equation (1) is referred to as the efficiency property, which appears to be a sensible desideratum for any attribution method and we here take as our starting point.

We can think of (at least) two different interpretations that, as we will see, will lead to two different approaches for computing the contributions.

Passive. We interpret the feature vector $\mathbf{x}$ as a passive observation. Feature values come in one after the other and the contribution of feature i should reflect the difference in expected value of $f(\mathbf{X})$ after and before *observing* its feature value x_i.

Active. We interpret the feature vector $\mathbf{x}$ as the result of an action. Feature values are imposed one after the other and the contribution of feature i relates to the difference in expected value of $f(\mathbf{X})$ after and before *setting* its value to x_i.

Following the above reasoning, the contribution of each feature depends on the order π in which the feature values arrive or are imposed. We write the contribution of feature i given the permutation π as

$$\phi_i(\pi) = v(\{j : j \preceq_\pi i\}) - v(\{j : j \prec_\pi i\}) \, , \tag{2}$$

with $j \prec_\pi i$ if j precedes i in the permutation π, i.e., if $\pi(j) < \pi(i)$, and where we define the value function

$$v(S) = \mathbb{E}\left[f(\mathbf{X})|op(\mathbf{x}_S)\right] = \int d\mathbf{X}_{\bar{S}}\, P(\mathbf{X}_{\bar{S}}|op(\mathbf{X}_S = \mathbf{x}_S))f(\mathbf{X}_{\bar{S}}, \mathbf{x}_S)\,. \tag{3}$$

Here S is the subset of indices of features with known "in-coalition" feature values $\mathbf{x}_S$. To compute the expectation, we still need to average over the "out-of-coalition" feature values $\mathbf{X}_{\bar{S}}$ with $\bar{S} = N\backslash S$, the complement of S. The operator $op()$ specifies how the distribution of the "out-of-coalition" features $\mathbf{X}_{\bar{S}}$ depends on the "in-coalition" feature values $\mathbf{x}_S$. To arrive at the passive interpretation, we set $op()$ to conventional conditioning by observation, yielding $P(\mathbf{X}_{\bar{S}}|\mathbf{x}_S)$. For the active interpretation, we need to condition by intervention, for which we resort to Pearl's *do*-calculus [6] and write $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S))$. Do we need a longer introduction/explanation of *do*-calculus? A third option is to ignore the feature values $\mathbf{x}_S$ and just take the unconditional, marginal distribution $P(\mathbf{X}_{\bar{S}})$. We will refer to the corresponding Shapley values as conditional, causal, and marginal, respectively.

It is easy to check that the efficiency property (1) holds for any permutation π. So, for any distribution over permutations $w(\pi)$ with $\sum_\pi w(\pi) = 1$, the contribution

$$\phi_i = \sum_\pi w(\pi)\phi_i(\pi)$$

still satisfies (1). An obvious choice would be to take a uniform distribution $w(\pi) = 1/n!$. We then arrive at the standard definition of Shapley values:

$$\phi_i = \sum_{S\subseteq N\backslash i} \frac{|S|!(n-|S|-1)!}{n!}\left[v(S\cup i) - v(S)\right]\ ,$$

where we use shorthand i for the singleton $\{i\}$. Besides efficiency, these Shapley values uniquely satisfy three other desirable properties [9].

Linearity: for two value functions v_1 and v_2, we have $\phi_i(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1\phi_i(v_1) + \alpha_2\phi_i(v_2)$. This guarantees that the Shapley value of a linear ensemble of models is a linear combination of the Shapley values of the individual models.

Null player (dummy): if $v(S \cup i) = v(S)$ for all $S \subseteq N \setminus i$, then $\phi_i = 0$. A feature that never contributes to the model output receives zero Shapley value.

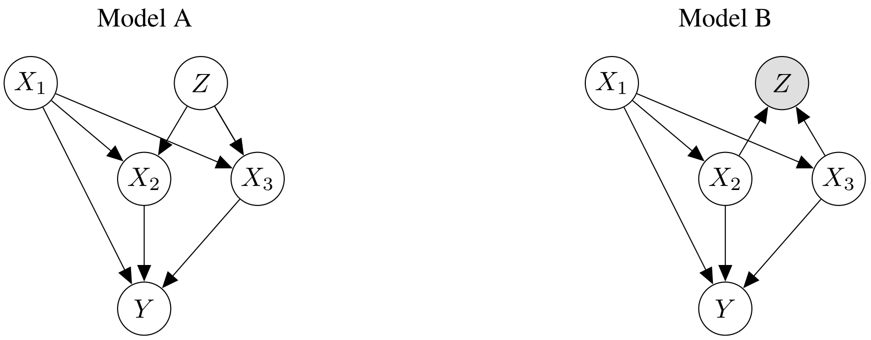
Symmetry: if $v(S \cup i) = v(S \cup j)$ for all $S \subseteq N \setminus \{i, j\}$, then $\phi_i = \phi_j$. Symmetry holds for any of the three different ways (conventional conditioning, *d*o-calculus, ignoring) to average over the "out-of-coalition" features.

Efficiency, linearity, and null player still hold for a non-uniform distribution of permutations, but symmetry is then typically lost.

Part of the next two paragraphs probably to introduction.

Janzing et al. [4] also argued for an active, interventional interpretation of Shapley values. They make a case for using the marginal instead of the (observational) conditional distribution to compute the Shapley values when dependencies between features are due to confounding. This follows directly from our reasoning, since in models with no causal links between the features and any dependencies only due to confounding, conditioning by intervention reduces to the marginal distribution: $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = P(\mathbf{X}_{\bar{S}})$ for any subset S.

Frye et al. [3] introduce asymmetric Shapley values as a way to incorporate causal information. Instead of taking a uniform distribution over all possible permutations, these asymmetric Shapley values only consider those permutations that are consistent with the causal structure between the features, i.e., are such that a known causal ancestor always precedes its descendants. Frye et al. apply conventional conditioning by observation to make sure that the resulting explanations respect the data manifold. This makes the approach somewhat of a mix between an active (incorporating causal structure) and a passive approach (conditioning by observation). In this paper, we suggest a more direct approach to incorporate causality, by replacing conditioning by observation with conditioning by intervention. This can, but does not have to be combined with the idea to make the permutations match the causal structure. We will illustrate the differences in the example below.


Model A

Model B

Figure 1: Two causal models. In both, X_1 causes X_2 and X_3. In Model A the excess correlation between X_2 and X_3 is induced by a common confounder Z, in Model B by selection bias.

3 Illustration

For illustration, we consider two causal models in Figure 1. They have a different causal structure, but the same dependency structure (all features are dependent) and we assume that the probability distribution $P(\mathbf{X})$ is exactly the same for Model A and Model B. Our estimate of the output Y is a linear function of the features:

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{3} \beta_i x_i \, .$$

Too much detail in this section: move parts to supplement? If so, which parts? Combining (2) and (3), we obtain, after some rewriting

$$\phi_i(\pi) = \beta_i \left(x_i - \mathbb{E}[X_i | op(\mathbf{x}_{j:j \prec_\pi i})] \right) + \sum_{k \succ_\pi i} \beta_k \left(\mathbb{E}[X_k | op(\mathbf{x}_{j:j \preceq_\pi i})] - \mathbb{E}[X_k | op(\mathbf{x}_{j:j \prec_\pi i})] \right) \, .$$

For marginal Shapley values only the first term before the sum remains, yielding

$$\phi_i = \phi_i(\pi) = \beta_i (x_i - \mathbb{E}[X_i]) \, ,$$

as also derived in [1].

Analytically computing the conditional Shapley values is tedious, but conceptually straightforward. To write the equations in a compact form, we define $\bar{x}_{k|S} = \mathbb{E}[X_k | \mathbf{x}_S]$ and combine all expectations in a single matrix $\bar{\mathcal{X}}$:

$$\bar{\mathcal{X}} = \begin{pmatrix} x_1 & x_2 & x_3 \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \\ \bar{x}_{1|2} & \bar{x}_{2|1} & \bar{x}_{3|1} \\ \bar{x}_{1|3} & \bar{x}_{2|3} & \bar{x}_{3|2} \\ \bar{x}_{1|2,3} & \bar{x}_{2|1,3} & \bar{x}_{3|1,2} \end{pmatrix} \, .$$

Taking a uniform distribution over permutations, the conditional Shapley values for any linear model with three variables can be written as

$$\boldsymbol{\phi}^{\text{conditional}} = \mathcal{C}^{\text{conditional}} \times \text{vec}(\bar{\mathcal{X}} \times \text{diag}(\boldsymbol{\beta})) \tag{4}$$

with

$$\mathcal{C}^{\text{conditional}} = \frac{1}{6} \left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -2 & -1 & -1 & -2 & 0 & -2 & 2 & -1 & 1 & 0 & -2 & 2 & -1 & 1 \\ 0 & -2 & 2 & -1 & 1 & 6 & -2 & -1 & -1 & -2 & 0 & -2 & -1 & 2 & 1 \\ 0 & -2 & -1 & 2 & 1 & 0 & -2 & -1 & 2 & 1 & 6 & -2 & -1 & -1 & -2 \end{array} \right) .$$

Skip this explanation? The multiplication with the diagonal matrix of regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ boils down to multiplying the ith column of $\bar{\mathcal{X}}$ with β_i. Vectorization then stacks the columns on top of one another to end up with a 15-dimensional column vector. The vertical bars in the matrix $\mathcal{C}^{\text{conditional}}$ indicate the three blocks, with the first 5 columns in the matrix mapping to the first column of $\bar{\mathcal{X}}$ with expectations of X_1, the next 5 columns to the expectations of X_2, and the final 5 columns to the expectations of X_3.

expectation	model A	model B
$\hat{x}_{1\|2}$	$\bar{x}_1$	
$\hat{x}_{1\|3}$		
$\hat{x}_{1\|2,3}$		
$\hat{x}_{2\|1}$	$\bar{x}_{2\|1}$	
$\hat{x}_{2\|3}$	$\bar{x}_2$	$\bar{x}_{2\|3}$
$\hat{x}_{2\|1,3}$	$\bar{x}_{2\|1}$	$\bar{x}_{2\|1,3}$
$\hat{x}_{3\|1}$	$\bar{x}_{3\|1}$	
$\hat{x}_{3\|2}$	$\bar{x}_3$	$\bar{x}_{3\|2}$
$\hat{x}_{3\|1,2}$	$\bar{x}_{3\|1}$	$\bar{x}_{3\|1,2}$

Table 1: Turning expectations under conditioning by intervention, $\hat{x}_{i|S} = \mathbb{E}[x_i|do(\mathbf{X}_S = \mathbf{x}_S)]$, into expectations under conventional conditioning by observation, $\bar{x}_{i|S} = \mathbb{E}[x_i|\mathbf{X}_S]$, for the two models in Figure 1. Needed? Can put it next to Figure 1 to save space.

Skip this paragraph? It is easy to check that efficiency indeed holds by summing every column of $\mathcal{C}^{\text{conditional}}$. The first column of each block (which relates to the feature values themselves) adds up to 1, the second (corresponding to the marginal expectations) to -1, and the other three columns to zero, as they should. Since Shapley values are constructed by always comparing two (possibly) different expectations, each row within each block sums up to zero.

Putting X_1 before X_2 and X_3, and X_2 and X_3 on equal footing, asymmetric Shapley values only consider the two permutations where x_1 is observed before x_2 and x_3, leading to (we divide by 6 to make it easier to compare with the other Shapley values)

$$\mathcal{C}^{\text{asymmetric}} = \frac{1}{6}\left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -6 & 0 & 0 & 0 & 0 & -6 & 6 & 0 & 0 & 0 & -6 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & -3 & 0 & -3 & 0 & 0 & -3 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3 & 0 & 3 & 6 & 0 & -3 & 0 & -3 \end{array}\right).$$

Using the standard rules for *do*-calculus [6], we show in Table 1 how the expectations under conditioning by intervention reduce to expectations under conditioning by observation. Since in Model A the correlation between X_2 and X_3 is due to confounding, the interventional expectations and thus Shapley values simplify considerably:

$$\mathcal{C}^{\text{causal,A}} = \frac{1}{6}\left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -6 & 0 & 0 & 0 & 0 & -3 & 3 & 0 & 0 & 0 & -3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & -3 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & -3 & 3 & 0 & 0 \end{array}\right).$$

For Model B, on the other hand, features X_2 and X_3 do affect each other when intervened upon, which makes that compared to the conditional Shapley values only the first block changes:

$$\mathcal{C}^{\text{causal,B}} = \frac{1}{6}\left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -6 & 0 & 0 & 0 & 0 & -2 & 2 & -1 & 1 & 0 & -2 & 2 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 6 & -2 & -1 & -1 & -2 & 0 & -2 & -1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -1 & 2 & 1 & 6 & -2 & -1 & -1 & -2 \end{array}\right).$$

To make this more concrete, let us assume that $P(\mathbf{X})$ follows a multivariate normal distribution corresponding to the causal model

$$X_1 \sim \mathcal{N}(0;1) \;\text{ and }\; (X_2, X_3|X_1) \sim \mathcal{N}\left((\alpha_2 X_1, \alpha_3 X_1); \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

where for notational convenience, we chose zero means and unit variance for all noise variables. The first feature drives the second and third feature with coefficients α_2 and α_3. The confounding in Model A or selection bias in Model B leads to correlation ρ on top of the correlation induced by the joint dependence on the first feature. Straightforward calculations yield

$$\bar{\mathcal{X}} = \begin{pmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 0 \\ \frac{\alpha_2}{\sigma_2^2} x_2 & \alpha_2 x_1 & \alpha_3 x_1 \\ \frac{\alpha_3}{\sigma_3^2} x_3 & \frac{\gamma}{\sigma_3^2} x_3 & \frac{\gamma}{\sigma_2^2} x_2 \\ \frac{\delta_{23} x_2 + \delta_{32} x_3}{1 - \rho^2 + \delta_{23}\alpha_2 + \delta_{32}\alpha_3} & (\alpha_2 - \rho\alpha_3) x_1 + \rho x_3 & (\alpha_3 - \rho\alpha_2) x_1 + \rho x_2 \end{pmatrix}.$$

with $\delta_{ij} = \alpha_i - \rho\alpha_j$, $\sigma_i^2 = 1 + \alpha_i^2$ (the marginal correlation for feature i), and $\gamma = \rho + \alpha_2\alpha_3$ (the total correlation between the second and third feature). Plugging this and the expressions for the different $\mathcal{C}$ matrices into (4), we obtain the asymmetric Shapley values

$$\begin{aligned}
\phi_1^{\text{asymmetric}} &= \beta_1 x_1 + (\alpha_2\beta_2 + \alpha_3\beta_3)x_1 \\
\phi_2^{\text{asymmetric}} &= \beta_2 x_2 - \alpha_2\beta_2 x_1 + \frac{\rho}{2}(\alpha_3\beta_2 - \alpha_2\beta_3)x_1 + \frac{\rho}{2}(\beta_3 x_2 - \beta_2 x_3) \\
\phi_3^{\text{asymmetric}} &= \beta_3 x_3 - \alpha_3\beta_3 x_1 + \frac{\rho}{2}(\alpha_2\beta_3 - \alpha_3\beta_2)x_1 + \frac{\rho}{2}(\beta_2 x_3 - \beta_3 x_2)\,,
\end{aligned}$$

and the causal Shapley values, for Model A,

$$\begin{aligned}
\phi_1^{\text{causal,A}} &= \beta_1 x_1 + \frac{1}{2}(\alpha_2\beta_2 + \alpha_3\beta_3)x_1 \\
\phi_2^{\text{causal,A}} &= \beta_2 x_2 - \frac{1}{2}\alpha_2\beta_2 x_1 \\
\phi_3^{\text{causal,A}} &= \beta_3 x_3 - \frac{1}{2}\alpha_3\beta_3 x_1\,.
\end{aligned}$$

and, for Model B, (really ugly. . . can we prevent this?)

$$\begin{aligned}
\phi_1^{\text{causal, B}} &= \beta_1 x_1 + \frac{1}{2}(\alpha_2\beta_2 + \alpha_3\beta_3)x_1 - \frac{\rho}{6}(\alpha_3\beta_2 + \alpha_2\beta_3)x_1 - \frac{1}{6}\left(\frac{\alpha_3\delta_{23}}{\sigma_3^2}\beta_2 x_3 + \frac{\alpha_2\delta_{32}}{\sigma_2^2}\beta_3 x_2\right) \\
\phi_2^{\text{causal, B}} &= \beta_2 x_2 - \frac{1}{2}\alpha_2\beta_2 x_1 + \frac{\rho}{6}(2\alpha_3\beta_2 - \alpha_2\beta_3)x_1 + \\
&\quad \frac{1}{6}\left(2\frac{\alpha_2\delta_{32}}{\sigma_2^2}\beta_3 x_2 - \frac{\alpha_3\delta_{23}}{\sigma_3^2}\beta_2 x_3\right) + \frac{\rho}{2}(\beta_3 x_2 - \beta_2 x_3) \\
\phi_3^{\text{causal, B}} &= \beta_3 x_3 - \frac{1}{2}\alpha_3\beta_3 x_1 + \frac{\rho}{6}(2\alpha_2\beta_3 - \alpha_3\beta_2)x_1 + \\
&\quad \frac{1}{6}\left(2\frac{\alpha_3\delta_{23}}{\sigma_3^2}\beta_2 x_3 - \frac{\alpha_2\delta_{32}}{\sigma_2^2}\beta_3 x_2\right) + \frac{\rho}{2}(\beta_2 x_3 - \beta_3 x_2)\,.
\end{aligned}$$

In this linear model, the asymmetric Shapley value for the first feature adds its indirect causal effects on the output through the second and third feature, $\alpha_2\beta_2 x_1 + \alpha_3\beta_3 x_1$, to its direct effect, $\beta_1 x_1$. The causal Shapley values for the first feature are somewhat more conservative: they essentially claim only half of the indirect effects through the other two features. Move the rest of this paragraph elsewhere? This is a direct consequence of taking a uniform distribution over all permutations: for any pair of features i and j, the feature value x_i is set before and after x_j for exactly half of the number of permutations. Which distribution over permutations to prefer, a uniform one or one that respects the causal structure, depends on the question the practitioner tries to answer and possibly on the application. For example, when a causal link represents a temporal relationship, it may make no sense to set a feature value before the values of all features preceding it in time have been set. In that case, it would be wise to consider a non-uniform distribution over permutations as in [3]. On the other hand, for causal models without temporal interpretation, e.g., describing presumed causal relationships between personal and biomedical variables related to Alzheimer [10] or between social and economic characteristics in census data [2], deviating from a uniform distribution over permutations (and hence sacrificing the symmetry property) seems unnecessary. With or without uniform distribution over permutations, applying *do*-calculus instead of conditioning by observation is a natural way to incorporate causal information.

The Shapley values for Model A are different from those for Model B, even though the observable probability distribution $P(\mathbf{X})$ is exactly the same. Those for Model A simplify a lot, because in this model any excess correlation between X_2 and X_3 beyond the correlation resulting from the common parent X_1 results from a confounder. This correlation vanishes when we intervene on either X_2 or X_3. The contributions of the second and third feature are therefore just their direct effect, minus half of the indirect effect, which already has been attributed to the first feature. In Model B, on the other hand, for most expectations conditioning by intervention reduces to conditioning by observation on the same variables and does not further simplify to conditioning on less or even no variables as for Model A. Since the causal Shapley values consider all six permutations, in contrast to the asymmetric Shapley values which only consider two of them, expectations such as $\mathbb{E}[X_2|x_3]$ now also enter the equation, which considerably complicates the analytical expressions.

4 Causal chain graphs

This section may need some more “fluff” with Theorems and an algorithm or ??? All we use/need is in fact in [5]... Computing causal Shapley values not only requires knowledge of the probability distribution $P(\mathbf{X})$, but also of the underlying causal structure. And even then, there is no guarantee that any causal query is identifiable (see e.g., [7]). For example, if Model A or B also includes a causal link from X_2 to X_3, even knowing the probability distribution $P(\mathbf{X})$ and the causal structure is insufficient: it is impossible to express, for example, $P(X_3|do(X_2 = x_2))$ in terms of $P(\mathbf{X})$, essentially because, without knowing the parameters of the causal model, there is no way to tell which part of the observed dependence between X_2 and X_3 is due to the causal link and which due to the confounding or selection bias.

Furthermore, and perhaps more importantly, requiring a practitioner to specify a complete causal structure, possibly even including some of its parameters, would be detrimental to the method’s generic applicability. We therefore follow the same line of reasoning as in [3] and assume that a practitioner may be able to specify a causal ordering, but not much more.

In the special case that a complete causal ordering of the features can be given and that all causal relationships are unconfounded, $P(\mathbf{X})$ satisfies the Markov properties associated with a directed acyclic graph (DAG) and can be written in the form

$$P(\mathbf{X}) = \prod_{j \in N} P(X_j|\mathbf{X}_{pa(j)}) \; ,$$

with $pa(j)$ the parents of node j. If no further conditional independences are assumed, the parents of j are all nodes that precede j in the causal ordering. For causal DAGs, we have the interventional formula [5]:

$$P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{j \in \bar{S}} P(X_j|\mathbf{X}_{pa(j)\cap\bar{S}}, \mathbf{x}_{pa(j)\cap S}) \; , \tag{5}$$

with $pa(j) \cap T$ the parents of j that are also part of subset T. The interventional formula can be used to answer any causal query of interest. We will often approximate the expectations needed to compute the Shapley values through sampling, which is particularly straightforward for causal DAGs under conditioning by intervention. Variables are sampled consecutively by following the causal ordering. The probability distribution for a feature then only depends on the values of its parents, which by then is either sampled or fixed. Since the intervention blocks the influence of all descendants, there is no need for an MCMC approach such as Gibbs sampling: the values of all features can be sampled in a single pass through the graph.

If we only consider permutations that follow the causal ordering, as with asymmetric Shapley values, conditioning by intervention reduces to conditioning by observation.

We may not always be willing or able to give a complete ordering between the individual variables, but rather a partial ordering as, for example, in Figure 1 where we have the partial ordering $(\{1\}, \{2, 3\})$: the first feature precedes the second and third feature in the causal ordering, with the second and third feature on equal footing, i.e., without specifying whether the second causes the third or vice versa. Here causal chain graphs [5] come to the rescue. A causal chain graph has directed and undirected edges. All features that are treated on an equal footing are linked together with undirected edges and become part of the same chain component. Edges between chain components are directed and represent causal relationships. The probability distribution $P(\mathbf{X})$ now factorizes as a “DAG of chain components”:

$$P(\mathbf{X}) = \prod_{\tau} P(\mathbf{X}_\tau|\mathbf{X}_{pa(\tau)}) \; ,$$

with each τ corresponding to a chain component, consisting of all features that are treated on an equal footing.

How to compute the effect of an intervention now depends on the interpretation of the generative process leading to the undirected part of the probability distribution. One possible interpretation is that the undirected part corresponds to the equilibrium distribution of a dynamic process resulting from interactions between the variables within a component [5]. In this case, setting the value of a feature does affect the distribution of the variables within the same component and the interventional formula reads

$$P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{\tau} P(\mathbf{X}_{\tau \cap \bar{S}}|\mathbf{X}_{pa(\tau)\cap \bar{S}}, \mathbf{x}_{pa(\tau)\cap S}, \mathbf{x}_{\tau \cap S}) \, .$$

That is, in the observational distribution we need to condition on intervened features that are parents and those that are within the same component. The same interventional formula applies when the dependencies within a component are the result of a selection bias as in Model B in Figure 1.

However, if we assume that the undirected part of the probability distribution is the consequence of marginalizing out a common confounder, we should take

$$P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{\tau} P(\mathbf{X}_{\tau \cap \bar{S}}|\mathbf{X}_{pa(\tau)\cap \bar{S}}, \mathbf{x}_{pa(\tau)\cap S}) \, ,$$

i.e., we should only condition on the parent components, not on the intervened variables within the same component. This interpretation is consistent with [4] and Model A in Figure 1.

So, to be able to compute the expectations in the Shapley equations under an interventional interpretation, we need to specify (1) a partial order and (2) whether any dependencies between features that are treated on an equal footing are most likely the result of mutual interaction or selection, or of a common confounder. Based on this information, any expectation by intervention can be translated to an expectation by observation.

To compute these expectations, we can rely on the various methods that have been proposed to compute conditional Shapley values [1, 3]. Following [1], we will assume a multivariate Gaussian distribution for $P(\mathbf{X})$ that we estimate from the training data. Alternative proposals include assuming a Gaussian copula distribution, estimating from the empirical (conditional) distribution (both from [1]) and a variational autoencoder [3].

5 Experiments

From here on just rough text and ideas.

Show that it works. For now: example on bike rental. Do we predict more bike shares on a warm, but cloudy day in August because of the season or because of the weather? ADNI as another example? Tried German Credit Data, but hard to see differences between causal and conditioning, mainly because the features, such as gender and age, that can be considered causes of some of the others, hardly affect the prediction. Other suggestions? Currently using a relatively straightforward adaptation of the code of [1]. How to describe this? Do we need to publish the code? Do we need to show results for asymmetric Shapley values as well? If so, need to dig deeper into the code. Also: currently no code for handling discrete variables. Could connect to Ruifei's Gaussian copula's for mixed missing data, if needed?

6 Discussion

Which Shapley value to use depends on interpretation: what happens if we set the features to their values compared to what happens to when we observe that features obtain their values. Both interpretations are fine, but in most cases the causal interpretation seems to be the one to prefer/implicitly implied.

Marginal fine when dependencies are purely the result of confounding, as Janzing argued. However, when there actually are causal relationships between the features and/or the dependencies between features results from selection bias or mutual feedback, expectations under conditioning by intervention do not simplify to marginal expectations.

Novel algorithm, based on causal chain graphs. All that a practitioner needs to provide is a partial causal order (as for asymmetric Shapley) and how to interpret dependencies between features that are on the same footing. Conditioning by intervention becomes conditioning by observation, but only on ancestors and possibly, depending on the interpretation, features within the same component. Any existing approach for conditional Shapley values can be easily adapted: if anything the expectations simplify, since no conditioning on descendants.

Provides an interpretation for asymmetric Shapley values: for all permutations that are consistent with the causal ordering, conditioning by intervention boils down to conditioning by observation. However, current definition [3] implicitly assumes that dependencies between features that are on the same footing is due to mutual feedback or selection bias, not common confounding. Whether or not to only consider permutations that match the causal ordering depends on application and possibly taste. This paper shows that it is unnecessary to give away symmetry in order to arrive at a causal interpretation of Shapley values. Roughly speaking, asymmetric (causal/conditioning) Shapley values attribute all indirect effects of a causal variable through a mediator on the output to the causal variable (is this the right term?), subtracting it from the Shapley value of the mediator (since efficiency should hold). Symmetric Shapley values are more conservative and attribute only half to the causal variable.

Discuss non-manipulable causes as in [8]?

Mention that it's easy to combine with any of TreeSHAP, KernelShap, and so on?

Compare with counterfactual explanations?

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] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[3] Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.

[4] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causality problem. *arXiv preprint arXiv:1910.13413*, 2019.

[5] Steffen L Lauritzen and Thomas S Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002.

[6] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[7] Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

[8] Judea Pearl. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference*, 6(2), 2018.

[9] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[10] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. *Scientific reports*, 10(1):1–12, 2020.