

MASTER THESIS
COMPUTER SCIENCE



RADBOUD UNIVERSITY

Causal Shapley values for interpretable machine learning

Author:

E.M.C. Sijben (Evi)
s4476727

Supervisor & examiner:

Prof. dr. T.M. Heskes (Tom)
T.Heskes@science.ru.nl

Examiner:

Prof. dr. E. Marchiori (Elena)
E.Marchiori@cs.ru.nl

July 3, 2020

Abstract

As the use and impact of machine learning models expands, criticism grows on our lack of deeper understanding on the behaviour of these models. Shapley values are at present presumably the most popular method for interpreting the behaviour of these models. However, clarity lacks on how to take care of available knowledge on causal relationships between input variables when calculating Shapley values. When this is handled inadequately, it can leave us with counter-intuitive explanations for our models' behaviour. This work introduces causal Shapley values which provide a theoretical substantiated view on how causality should be taken into account when calculating the Shapley value.

Contents

1	Introduction	2
2	Background	4
2.1	Shapley values in game theory	4
2.1.1	Shapley equation	5
2.1.2	Permutation-based Shapley equation	6
2.2	Probability distributions and causal structures	7
2.2.1	Conditional probability	7
2.2.2	Conditionally independent	7
2.2.3	Causal structure	8
2.2.4	Causal model	8
2.2.5	D-separation	8
2.2.6	Do-operator: interventional conditional probability . .	9
3	Shapley values for machine learning	12
3.1	Expected value of a model	12
3.2	Properties of Shapley values for machine learning	13
3.3	Approximating Shapley values	14
3.4	Taking dependencies into account	14
3.4.1	Conditional probability distribution	14
3.4.2	Interventional conditional probability distribution . .	15
3.4.3	Asymmetric Shapley values	16
4	Causal Shapley values	18
4.1	Approximating causal Shapley values	18
4.2	Results of different approaches	19
4.2.1	Example 1a: direct relationship	20
4.2.2	Example 1b: direct relationship	20
4.2.3	Example 2: multiple direct relationships	21
4.2.4	Example 3: direct relationship and confounder	22
4.2.5	Conclusion	23
5	Conclusions	24

Chapter 1

Introduction

Many machine learning models have a high black-box degree. Despite the fact that these models may have a good predictive ability, this does not guarantee that these models behave in the way we want them to (think for example about models that discriminate on gender or race [7]). Without any interpretation on the behaviour of these models, undesired behaviour can easily go unnoticed. According to the European commission, a human should be in command on when and how to use an AI system [3]. But this leaves humans with an impossible task. Professionals working with these models are unable to judge the validity of a models' output without a further interpretation on why this output was generated. Furthermore, the GDPR includes that in case of automated decision-making one needs to be able provide "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" in Art. 13 section 2 under f [16], Art. 14 section 2 under g [17] and Art. 15 section 1 under g [18]. These requirements coupled with our shallow understanding of model behaviour cause a need to find ways to better interpret our models.

Shapley values aim at providing us such an interpretation and are presumably the most popular method for interpretable machine learning nowadays. However, clarity lacks on how to take care of available knowledge on causal relationships between input variables when calculating Shapley values. When this is handled inadequately, it can leave us with counter-intuitive explanations for our models' behaviour. Since methods for interpretable machine learning are in demand and will have a significant impact once they find their way into practice, we need them to be able to deal with causal knowledge in a proper way.

Aas et al. [1] came up with a method to take into account the correlation between features when calculating Shapley values, although without considering the causal structure. Janzing et al. [5] argue that do-calculus should be used in order to derive the Shapley values in a way that respects the

causal structure. Remarkably, they conclude based on this that the ‘standard’ Shapley values are the ones preferred after all. Frye et al. [2] propose asymmetric Shapley values, which we will see later on is a different type of solution than the one of Aas et al. [1], Janzing et al. [5] and our selves. In this work, we follow the line of thinking of Janzing et al. [5] in the sense of using do-calculus to derive the Shapley values. However, we conclude that this does not lead to the ‘standard’ Shapley values, but to ‘causal’ Shapley values. We outline how these so-called causal Shapley values can be derived instead.

Chapter 2

Background

Section 2.1 describes the idea behind Shapley values, why Shapley values are considered to be the fair distribution value and the build-up of the Shapley equation. Section 2.2 provides background knowledge about probability distributions and causal structures in order to understand the idea behind conditional and causal Shapley values.

2.1 Shapley values in game theory

Shapley values are a game theory concept which expresses a fair way to divide the pay-out of a game amongst its players [11]. It calculates the difference in outcome when one specific player does not perform in the game versus when the player does. An example is given in figure 2.1.

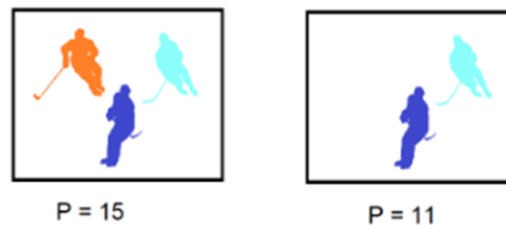


Figure 2.1: Three ice hockey players play a game and get a pay-out of 15. When orange does not play, the pay-out is 11. The marginal contribution of the orange player in this example is 4.

The difference in pay-out can be calculated for all possible coalitions of players, as denoted in figure 2.2.

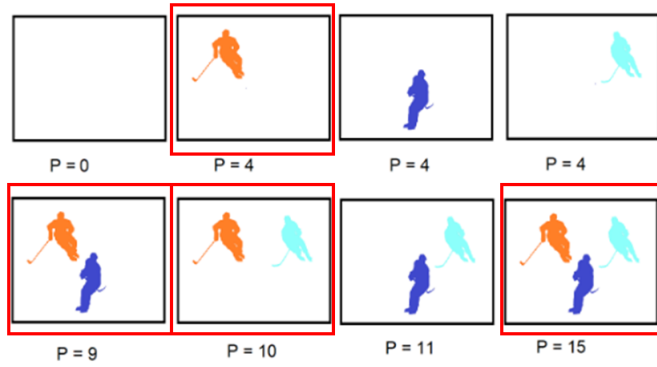


Figure 2.2: All possible coalitions of three players. The coalitions in which we can leave out orange to calculate the marginal contribution are outlined in red.

The Shapley equation (specifics are elaborated in section 2.1.1) sums over all these marginal contributions and weighs them in such a way that the resulting value satisfies the following properties [11]:

- Efficiency: the sum of the contribution values should add up to the pay-out of the game with all players.
- Symmetry: if two players contribute equally to all coalitions, they should get the same contribution value.
- Dummy: if a player does not affect the outcome in all possible coalitions, this player should have a contribution value of zero.
- Additivity: if we run two separate games, the contribution values of the players are just the sum of the contribution value of the individual games.

These properties make them to be considered the fair way to divide the pay-out.











2.1.1 Shapley equation

The Shapley value of player i of a game with value function v is,

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \underbrace{\frac{|S|!(n - |S| - 1)!}{n!}}_{\text{weight factor}} \underbrace{(v(S \cup \{i\}) - v(S))}_{\text{marginal contribution}}. \quad (2.1)$$

Here N is the set of all players, n is the total number of players and $|S|$ is the number of players in subset $S \subseteq N$. We sum over all subsets of N that

do not contain i , in which we add the weighted marginal contribution of i . Figure 2.3 shows an example in which we calculate the fair pay-out ($P = 15$) of three ice-hockey players based on Shapley values.

Subgroup	Subgroup output	Marginal contribution of 	Marginal contribution of 	Marginal contribution of 
	0	-	-	-
	4	4 - 0 = 4	-	-
	4	-	4 - 0 = 4	-
	4	-	-	4 - 0 = 4
	9	9 - 4 = 5	9 - 4 = 5	-
	10	10 - 4 = 6	-	10 - 4 = 6
	11	-	11 - 4 = 7	11 - 4 = 7
	15	15 - 11 = 4	15 - 10 = 5	15 - 9 = 6

Shapley value

$$\text{orange} = \binom{4}{3} + \frac{5}{6} + \frac{6}{6} + \frac{4}{3} = 4.5 \quad \text{purple} = \binom{4}{3} + \frac{5}{6} + \frac{7}{6} + \frac{5}{3} = 5 \quad \text{blue} = \binom{4}{3} + \frac{6}{6} + \frac{7}{6} + \frac{6}{3} = 5.5$$


























































Figure 2.3: Calculation of fair pay-out per player based on Shapley values. The first column denotes which players participate, the second column denotes the pay-out for that set of players, and the third, fourth and last column denote the marginal contribution of the orange player, purple player and blue player respectively (if relevant). Beneath the table we see the Shapley value which is an addition of all the separate marginal contributions in which we weigh them by a factor.

2.1.2 Permutation-based Shapley equation

Equivalently, the Shapley equation can be written in another form using permutations,

$$\phi_i(v) = \frac{1}{n!} \sum_{r \in R} [v(P_i^r \cup \{i\}) - v(P_i^r)]. \quad (2.2)$$

Here P_i^r is the set of players that precedes i in permutation r , and $R = \mathfrak{S}(N)$ is the set of all permutations of players N . When calculating the Shapley value of i with this variant, we sum over all possible permutations $r \in R$ in which we add the marginal contribution of i . The permutation-based equation is more intuitive in the sense of how we weigh the marginal contributions, as we just divide by the number of permutations. In figure 2.4 the same example as in figure 2.3 is shown, but this time the permutation-based variant of the Shapley equation is used.

Order	Marginal contribution of 	Marginal contribution of 	Marginal contribution of 
  	- = 4 - 0 = 4	   = 15 - 10 = 5	   = 10 - 4 = 6
  	- = 4 - 0 = 4	   = 9 - 4 = 5	   = 15 - 9 = 6
  	   = 10 - 4 = 6	   = 15 - 10 = 5	- = 4 - 0 = 4
  	   = 15 - 11 = 4	   = 11 - 4 = 7	- = 4 - 0 = 4
  	   = 9 - 4 = 5	- = 4 - 0 = 4	   = 15 - 9 = 6
  	   = 15 - 11 = 4	- = 4 - 0 = 4	   = 11 - 4 = 7

Shapley value




 = $\frac{4+4+6+4+5+4}{6} = \frac{27}{6} = 4.5$  = $\frac{5+5+5+7+4+4}{6} = \frac{30}{6} = 5$  = $\frac{6+6+4+4+6+7}{6} = \frac{33}{6} = 5.5$

Figure 2.4: Calculation of fair pay-out per player based on Shapley values (permutation based variant). The first column denotes all different orders of players orange, purple and blue. The second, third and last column denote the marginal contribution of the orange, purple and blue player respectively for that permutation. Beneath the table we see the Shapley value which is an addition of all the separate marginal contributions in which we divide by the total number of permutations.

2.2 Probability distributions and causal structures

This section provides background materials in order to understand the conditional en causal Shapley values.

2.2.1 Conditional probability

When we want to know the probability of an event, given that another event has occurred, we use the conditional probability. This is defined as the probability that both events occur divided by the probability that the conditional event occurs,

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}.$$

These probabilities can also be called the observational conditional probabilities, since we calculate the probability of X when we observe Y .

2.2.2 Conditionally independent

Two variables X and Y are called independent, denoted $X \perp\!\!\!\perp Y$, when X does not provide additional information about Y , and vice versa, and therefore, $P(X, Y) = P(X)P(Y)$. Furthermore, two variables can also be independent given a certain set of variables, $X \perp\!\!\!\perp Y \mid Z$, subsequently, $P(X, Y|Z) = P(X|Z)P(Y|Z)$.

When $X \perp\!\!\!\perp Y \mid Z$ we say that X and Y are conditionally independent given Z .

2.2.3 Causal structure

Pearl defines a causal structure as a directed acyclic graph (DAG) in which the nodes represent variables and the edges represent a direct functional relationship between the two variables it connects [9]. This causal structure defines the architecture of the model, but does not specify the precise function by which these variables are linked.

2.2.4 Causal model

Subsequently, Pearl defines a causal model as a combination of a causal structure D and a set of parameters Θ_D , where these parameters Θ_D should obviously be compatible with the causal structure D [9]. These parameters assign a function $v_i = f_i(pa_i, u_i)$ to each vertex $V_i \in V$, where pa_i are the parents of V_i and u_i is a random disturbance factor. This definition is in line with the causal Markov condition or Markov assumption, which says that every node $V_i \in V$ is independent of all non-descendants of V_i given the parents of V_i [13].

2.2.5 D-separation

If Z d-separates X and Y in causal graph G , then $X \perp\!\!\!\perp Y \mid Z$ is guaranteed to hold. Z d-separates X and Y when every path $\pi = (X, V_1, \dots, V_n, Y)$ with $n \geq 0$, satisfies one of the following conditions.

- π contains a collider $X \rightarrow V_i \leftarrow Y$, with $V_i \notin Z$ and V_i is not an ancestor of any node in Z ;
- π contains a non-collider V_i where $V_i \in Z$.

Two examples on graphs and d-separation are given in figure 2.5.



Figure 2.5: In model A, X and Y are d-separated by the empty set, i.e., $Z = \{\}$. The only path between X and Y is via V_1 which is a collider, which makes that the empty set already satisfies the conditions for d-separation. Notice that by adding V_1 to Z one would open the path and X and Y would then not be d-separated anymore. In model B, X and Y are d-separated by $Z = \{V_1\}$. The only path between X and Y is via V_1 which is a confounder, this makes that we need to add it to Z in order to satisfy the conditions for d-separation.

2.2.6 Do-operator: interventional conditional probability

Pearl's do-operator, $P(Y|do(X = x))$, is used to compute the effect of a certain intervention $X = x$ on another variable Y [8]. This operator informs us how we should calculate the probability distribution when we are dealing with an intervention rather than with a passive observation. It enforces that the effect we measure is due to the effect of the variable we intervene on and is not confounded by other factors.

For some cases we can rewrite the do-operator and thereby calculate the probabilities by intervention from the probabilities by observation. Do-calculus specifies the three rules stated below to do so [10]. It was proven to be complete to the identifiability of causal effects [4][12]. This means that if the do-operator can not be removed by applying these rules, then the causal effect is not identifiable. We can then not calculate the probabilities by intervention based on the probabilities by observation.

Do-calculus

Let X, Y, Z and W be sets of nodes in a causal DAG G . We specify $G_{\overline{W}}$ as a copy of graph G in which we delete all arrows in graph G that point into W . Furthermore, we define $G_{\underline{W}}$ as a copy of graph G in which we delete all arrows in graph G that emerge from W .

Rule 1: insertion/deletion of observation. If $X \perp\!\!\!\perp Y|Z, W$ holds in graph $G_{\overline{W}}$ then we can apply rule 1,

$$P(Y|X, Z, do(W)) = P(Y|Z, do(W)).$$



Figure 2.6: Example in which rule 1 applies. We can see that X and Y are d-separated by W and Z in graph $G_{\bar{W}}$.

Rule 2: action/observation exchange. If $X \perp\!\!\!\perp Y|Z, W$ holds in graph $G_{\bar{W}\underline{X}}$ then we can apply rule 2,

$$P(Y|do(X), Z, do(W)) = P(Y|X, Z, do(W)).$$



Figure 2.7: Example in which rule 2 applies. We can see that X and Y are d-separated by W and Z in graph $G_{\bar{W}\underline{X}}$.

Rule 3: insertion/deletion of action. If $X \perp\!\!\!\perp Y|Z, W$ holds in graph $G_{\bar{W}, \overline{X(Z)}}$ in which $X(Z)$ is the set of nodes of X which are not ancestors of any Z -node in $G_{\bar{W}}$ then we can apply rule 3,

$$P(Y|do(X), Z, do(W)) = P(Y|Z, do(W)).$$

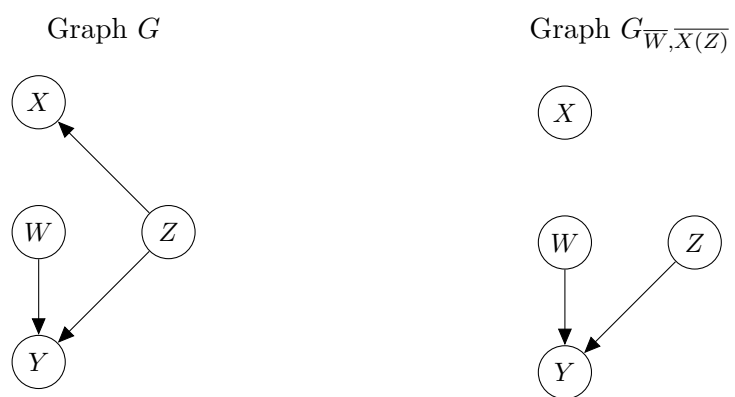


Figure 2.8: Example in which rule 3 applies. We can see that X and Y are d-separated by W and Z in graph $G_{\overline{W}, \overline{X(Z)}}$.

Chapter 3

Shapley values for machine learning

A translation of the game-theory concept Shapley values can be made for machine learning models. In this translation, we see the outcome of a model as the pay-out, and the features as players. The Shapley values are used to distribute contribution to the models outcome fairly among the features. However, to do so we need to calculate the expected outcome of the model when we only know the values of some of the features. The next section shows how we can do this.

3.1 Expected value of a model

If we want to use the Shapley equation for machine learning models we somehow need to calculate the outcome of the model without knowing all the feature values. The most natural way to do this, will probably be to use the expected outcome of the model as value function;

$$v(S) = E[f(X_{\bar{S}}, x_S)].$$

Here $X_{\bar{S}}$ is the set of (unknown) stochastic variables and x_S is the set of (known) realizations of variables. In order to calculate the expected outcome of a model f we need to integrate over all possible values of the unknown features $X_{\bar{S}}$, weigh them by the probability that they appear $P(X_{\bar{S}})$ and multiply them with the outcome for that specific realization $f(X_{\bar{S}}, x_S)$. Here, the known feature values are on indices $S \subseteq N$ with $N = \{1, \dots, n\}$ the set of all feature indices and unknown feature values are on $\bar{S} = N \setminus S$. Using the marginal probability distribution results in,

$$E[f(X_{\bar{S}}, x_S)] = \int dX_{\bar{S}} P(X_{\bar{S}}) f(X_{\bar{S}}, x_S).$$

We will refer to the case of using this as the value function as the *marginal* Shapley values.

3.2 Properties of Shapley values for machine learning

Notice that specifying,

$$v(S) = E[f(X_{\bar{S}}, x_S)],$$

does not change the structure of the Shapley equation and therefore the properties as stated in section 2.1 are preserved. When we translate those properties to ‘machine learning language’ this results in,

- Efficiency: the sum of the contribution values of all features add up to the difference between the outcome of the model for this specific realisation and the overall expected value of the model. Likewise,

$$\sum_{i=1}^n \phi_i = f(x) - E[f(X)].$$

- Symmetry: if two features contribute equally to all coalitions, they get the same contribution value. So if,

$$E[f(X_{\overline{S \cup i}}, x_{S \cup i})] - E[f(X_{\bar{S}}, x_S)] = E[f(X_{\overline{S \cup j}}, x_{S \cup j})] - E[f(X_{\bar{S}}, x_S)],$$

for all $S \subseteq N \setminus \{i, j\}$ then,

$$\phi_i = \phi_j.$$

- Dummy: if a feature does not affect the outcome in all possible coalitions, this feature has a contribution value of zero. So if,

$$E[f(X_{\overline{S \cup i}}, x_{S \cup i})] = E[f(X_{\bar{S}}, x_S)]$$

for all $S \subseteq N \setminus \{i\}$ then,

$$\phi_i = 0.$$

- Additivity: if we run two separate models, the contribution values of the features are just the sum of the contribution value of the individual models. This enables us to easily calculate the Shapley values over ensemble models.

3.3 Approximating Shapley values

Štrumbelj et al. [14][15] proposed an algorithm to approximate the marginal Shapley value, as displayed in algorithm 1. The basic idea is to sample random instances of the data, mix the random instance with the instance of the data at hand, and calculate the marginal contribution of i . This is repeated for random permutations of the features and random data instances. The sampling of random instances is a replacement for integrating over all possible values of the unknown features. Since the data instances are randomly sampled they respect the probability distribution of the data.

Algorithm 1 Approximating marginal Shapley value ϕ_i of the feature i for model f for instance of data $x \in X$, given features N , data X , and parameter m .

```

1: procedure APPROXIMATESHAPLEY( $i, N, X, x, m$ )
2:    $R \leftarrow \mathfrak{S}(N)$  ▷ Set of all permutations of features  $N$ 
3:    $\phi_i \leftarrow 0$ 
4:   for 1 to  $m$  do
5:      $r \leftarrow \text{getElement}(R)$  ▷ Select random permutation of features.
6:      $S \leftarrow P_i^r$  ▷ Features that precede  $i$  in  $r$ 
7:      $z \leftarrow \text{getElement}(X)$  ▷ Select random instance from data.
8:      $c \leftarrow f(z_{\overline{S \cup i}}, x_{S \cup i}) - f(z_{\overline{S}}, x_S)$  ▷ Contribution of  $i$ 
9:      $\phi_i \leftarrow \phi_i + c$ 
10:  end for
11:   $\phi_i \leftarrow \frac{\phi_i}{m}$ 
12:  return  $\phi_i$ 
13: end procedure

```

3.4 Taking dependencies into account

This section discusses three approaches to calculate Shapley values, all of which take dependencies between variables into account. The first approach uses conditional probability distributions, the second the interventional conditional probability distributions, and the third abandons the symmetry axiom of Shapley values.

3.4.1 Conditional probability distribution

Aas et al. [1] argue that dependencies between features should be taken into account by using the conditional probability distribution, to which we will refer to as the *conditional* Shapley values,

$$v(S) = E[f(X_{\overline{S}}, x_S) | X_S = x_S] = \int dX_{\overline{S}} P(X_{\overline{S}} | x_S) f(X_{\overline{S}}, x_S).$$

They compare different methods for approximating the conditional probability distribution in their work. The next section will show how using the conditional probability distributions can result in unwanted effects.

3.4.2 Interventional conditional probability distribution

Janzing et al. [5] argue that the interventional conditional probability distributions should be used instead of the observational (conditional) distributions. They base this argumentation on the do-calculus as introduced in section 2.2.6. They illustrate how the observational conditional distributions gives problems with the following example. Consider the function $f(x_{1,2}) = x_1$, in which x_1, x_2 are binary variables. In addition,

$$P(x_1, x_2) = \begin{cases} \frac{1}{2} & \text{if } x_1 = x_2 \\ 0 & \text{otherwise} \end{cases}$$

which creates a confounding effect between x_1 and x_2 . The overall causal structure is therefore as in figure 3.1. We see that x_2 is irrelevant to the output.

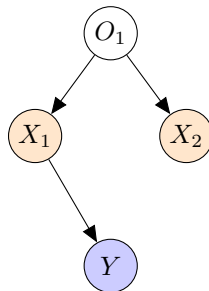


Figure 3.1: Causal structure with X_1, X_2 the input variables, O_1 an unobserved variable and $Y = f(x)$ the output of the model.

This gives the following conditional expectations for f

$$\begin{aligned} E[f(X_{1,2})] &= \frac{1}{2} \\ E[f(x_1, X_2)|x_1] &= x_1 \\ E[f(X_1, x_2)|x_2] &= x_2 \\ E[f(x_{1,2})] &= x_1. \end{aligned}$$

Hence,

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

When using marginal expectations we get,

$$\begin{aligned} E[f(X_{1,2})] &= \frac{1}{2} \\ E[f(x_1, X_2)] &= x_1 \\ E[f(X_1, x_2)] &= \frac{1}{2} \\ E[f(x_{1,2})] &= x_1. \end{aligned}$$

Hence,

$$\phi_2 = \frac{1}{2} \left(\frac{1}{2} - \frac{1}{2} \right) + \frac{1}{2} (x_1 - x_1) = 0.$$

When we use the conditional expectations, it would seem as if x_2 has an influence on the output. This is a result of confounding between x_1 and x_2 . This is in contrast with the marginal expectations in which confounding does not affect the Shapley values. With this example they show that the observational conditional expectations might not always create the desired effect.

Surprisingly, after Janzing et al. [5] make this observation, they formally separate the real-world object features from the input features to the model. This leads them to conclude that the marginal expectations coincide with the interventional ones. Based on this research, Lundberg et al. [6] justify an interventional interpretation of their TreeSHAP algorithm that uses marginal expectations.

In this work we follow the line of thinking from Janzing et al. [5] that the interventional conditional distributions should be used instead of the observational ones. However, in contrast with their work, we want to know the effect of the real-world features on the models' output. This entails that we do take into account the effect that a variable has on the output via other input variables, but we do not take into account confounding effects of other variables.

3.4.3 Asymmetric Shapley values

Frye et al. [2] propose asymmetric Shapley values in which they implement their philosophy that “if X_i is known to be the deterministic causal ancestor of X_j , one might want to attribute all the importance to X_i and none to X_j ”. They do this by only considering the permutations that are consistent with the causal ordering. In addition, they propose on-manifold data sampling which is equivalent to using the conditional probability distribution. This is not appropriate for all cases from a causal perspective as we have seen in the section 3.4.2. Nonetheless, the concept of asymmetric Shapley values can be separately applied from choosing the value function $v(S)$, therefore

we could apply this asymmetry concept to causal Shapley values as well. Although the idea is interesting, it is not necessary to use the asymmetry in order to calculate the Shapley values in a causal way as we will see in the next chapter.

Chapter 4

Causal Shapley values

To deal with relationships between input features we propose *causal* Shapley values,

$$v(S) = E[f(X_{\bar{S}}, x_s) | do(X_s = x_s)] = \int dX_{\bar{S}} P(X_{\bar{S}} | do(X_s = x_s)) f(X_{\bar{S}}, x_s).$$

In other words, we are going to use the interventional conditional probability distributions to calculate the expected outcome of the model. Hereby, we take into account that we intervene on the variables X_S when we calculate the probability of the unknown features having a specific value. By doing so, we do take into account the effect that a variable has on the output via other input variables, but we do not take into account confounding effects of other variables. In case that there are no causal paths between the input variables, we can rewrite the do-operator in the causal Shapley values calculation using rule 3 of the do-calculus (in 2.2.6). This then results in the marginal Shapley values. When we do have causal paths however, this leads to something different as we will see in section 4.1, which gives an algorithm for approximating the causal Shapley values. Finally section 4.2 will illustrate the results of the marginal, conditional and causal Shapley values for some examples.

4.1 Approximating causal Shapley values

Algorithm 2 shows a way to approximate the causal Shapley values. This algorithm uses the rules of do-calculus to convert conditioning by intervention to conditioning by observation (if possible). Hereafter, we can use the approach presented in [1], in which a multivariate Gaussian distribution is assumed, to sample from the conditional distribution. As the resulting conditioning set may differ per intervention, we sample all feature values separately from each other. And as the value of the other feature values may differ depending on whether i needs to be sampled, we should sample

all features other than i two times. It is good to note that this algorithm only works when it is possible to convert conditioning by intervention to conditioning by observation. This is typically not the case when there is a causal path between two variables and also a confounder between those two variables. For the causal structures of the examples in the section 4.2 this algorithm does work.

Algorithm 2 Approximating causal Shapley value ϕ_i of the i^{th} feature for model f for instance of data $x \in X$, given features N , data X , causal structure G and parameter m .

```

1: procedure APPROXIMATECAUSALSHAPLEY( $i, N, X, x, G, m$ )
2:    $R \leftarrow \mathfrak{S}(N)$  ▷ Set of all permutations of features  $N$ 
3:    $\phi_i \leftarrow 0$ 
4:   for 1 to  $m$  do
5:      $r \leftarrow \text{getElement}(R)$  ▷ Select random permutation of features.
6:      $S \leftarrow P_i^r$  ▷ Features that precede  $i$  in  $r$ 
7:      $O \leftarrow \text{orderOnCausality}(\overline{S \cup i}, G)$  ▷ Ancestors before children.
8:      $\alpha_{\overline{S \cup i}} \leftarrow \text{SAMPLEFEATUREVALUES}(X, x, O, S, G)$ 
9:      $\bar{O} \leftarrow \text{orderOnCausality}(\bar{S}, G)$  ▷ Ancestors before children.
10:     $\beta_{\bar{S}} \leftarrow \text{SAMPLEFEATUREVALUES}(X, x, O, S, G)$ 
11:     $c \leftarrow f(\alpha_{\overline{S \cup i}}, x_{S \cup i}) - f(\beta_{\bar{S}}, x_S)$  ▷ Contribution of  $i$ 
12:     $\phi_i \leftarrow \phi_i + c$ 
13:  end for
14:   $\phi_i \leftarrow \frac{\phi_i}{m}$ 
15:  return  $\phi_i$ 
16: end procedure

17: procedure SAMPLEFEATUREVALUES( $X, x, O, S, G$ )
18:  for  $j$  in  $O$  do
19:     $A \leftarrow P_j^O$  ▷ Features that precede  $j$  in  $O$ 
20:     $Z \leftarrow \text{applyDoCalculus}(j, A \cup S, G)$  ▷ Rewrite do-operator
21:     $z_j \leftarrow \text{Sample}(x, X, j, Z)$  ▷ Sample from conditional distribution.
22:  end for
23:  return  $\mathbf{z}$  ▷ Return all sampled features
24: end procedure

```

4.2 Results of different approaches

A few examples of combinations of causal structures, functions and their marginal, conditional and causal Shapley values are shown in this section.

4.2.1 Example 1a: direct relationship

Table 4.1 gives the different Shapley values for the causal structure as described in figure 4.1. In this structure there is a direct relationship between the input variables. The results in table 4.1 show that the marginal Shapley values do not attribute any contribution value to X_1 , whereas the conditional Shapley values and the causal Shapley values do. As X_1 and X_2 both cause the same effect in the output, the attribution of the conditional and causal Shapley values in which each of the two input variables gets half seems intuitive as they provide the same information about the output.

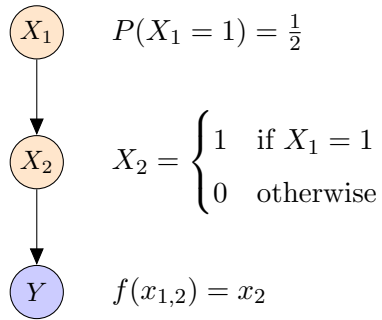


Figure 4.1: Causal structure with X_1, X_2 binary input variables and $Y = f(x)$ the output of the model.

Table 4.1: The Shapley values for the possible data points regarding the structure in figure 4.1. For the first data point we need to explain the difference between $f(x_{1,2}) = 0$ and the average $E(f(X)) = \frac{1}{2}$, which is $-\frac{1}{2}$. When using the marginal expectations, all of this is attributed to X_2 and in the cases of the conditional expectations and the causal expectations this is divided over both X_1 and X_2 . For the second data point the same applies but then for a difference of $\frac{1}{2}$.

	Marginal		Conditional		Causal	
	ϕ_1	ϕ_2	ϕ_1	ϕ_2	ϕ_1	ϕ_2
$X_1 = 0, X_2 = 0, f(x_{1,2}) = 0$	0	$-\frac{1}{2}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
$X_1 = 1, X_2 = 1, f(x_{1,2}) = 1$	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

4.2.2 Example 1b: direct relationship

Table 4.2 gives the different Shapley values for the causal structure as described in figure 4.2. In this structure there is a direct relationship between the input variables. Different from example 1a, however, this time X_1 as well as X_2 influence the output. The results in table 4.2 show that the marginal Shapley values and the conditional Shapley values attribute the

same contribution value to X_1 and X_2 . The causal Shapley values attribute more to X_1 than to X_2 . It seems fair to attribute more to X_1 than to X_2 , since X_1 influences the model output both directly and via X_2 .

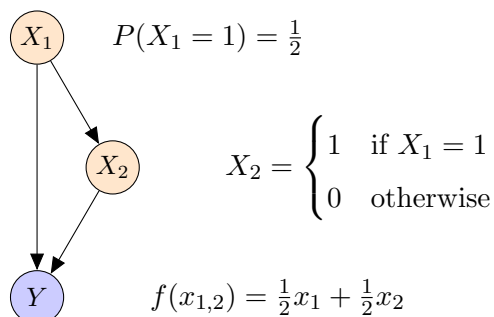


Figure 4.2: Causal structure with X_1, X_2 binary input variables and $Y = f(x)$ the output of the model.

Table 4.2: The Shapley values for the possible data points regarding the structure in figure 4.2.

	marginal		conditional		causal	
	ϕ_1	ϕ_2	ϕ_1	ϕ_2	ϕ_1	ϕ_2
$X_1 = 0, X_2 = 0, f(x_{1,2}) = 0$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{3}{8}$	$-\frac{1}{8}$
$X_1 = 1, X_2 = 1, f(x_{1,2}) = 1$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{8}$

4.2.3 Example 2: multiple direct relationships

Table 4.3 gives the different Shapley values for the causal structure as described in figure 4.3. In this structure there are multiple direct relationships between the input variables. The results in table 4.3 show that the marginal Shapley values do not attribute any contribution value to X_1 or X_2 , whereas the conditional Shapley values and the causal Shapley values do. This seems more intuitive for the same reason as in example 1.

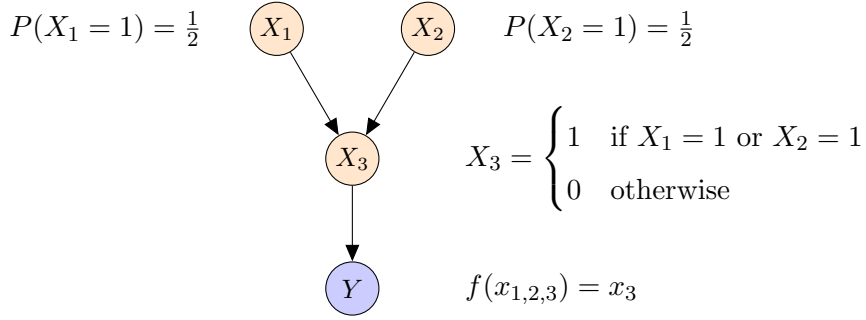


Figure 4.3: Causal structure with X_1, X_2, X_3 binary input variables and $Y = f(x)$ the output of the model.

Table 4.3: The Shapley values for the possible data points regarding the structure in figure 4.3.

	marginal			conditional			causal		
	ϕ_1	ϕ_2	ϕ_3	ϕ_1	ϕ_2	ϕ_3	ϕ_1	ϕ_2	ϕ_3
$X_1 = 0, X_2 = 0, X_3 = 0, f(x_{1,2,3}) = 0$	0	0	$-\frac{3}{4}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$
$X_1 = 0, X_2 = 1, X_3 = 1, f(x_{1,2,3}) = 1$	0	0	$\frac{1}{4}$	$-\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$	$-\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$
$X_1 = 1, X_2 = 0, X_3 = 1, f(x_{1,2,3}) = 1$	0	0	$\frac{1}{4}$	$\frac{1}{6}$	$-\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$	$-\frac{1}{12}$	$\frac{1}{6}$
$X_1 = 1, X_2 = 1, X_3 = 1, f(x_{1,2,3}) = 1$	0	0	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

4.2.4 Example 3: direct relationship and confounder

Table 4.4 gives the different Shapley values for the causal structure as described in figure 4.4. In this structure there is both a direct relationship and a confounding relationship between the input variables. The results in table 4.4 show that the marginal Shapley values do not attribute any contribution value to X_1 , the conditional Shapley values attribute value to all three variables and the causal Shapley values attribute the contribution to both X_1 and X_3 . The causal Shapley values seem the most logical for this example. As X_1 causes change in the output, so it should get some attribution value. X_2 however does not cause change in the output, it only correlates with X_1 because of the confounder O_1 and should therefore not get any contribution value.

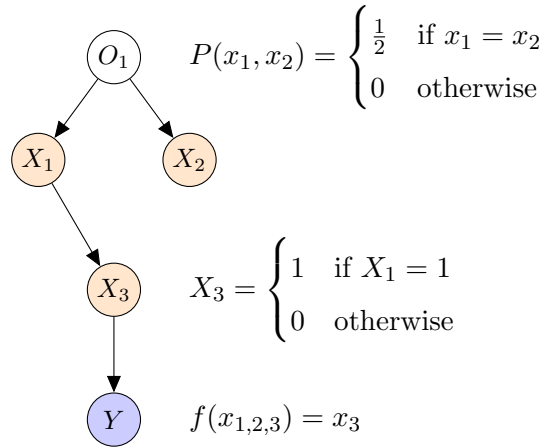


Figure 4.4: Causal structure with X_1, X_2, X_3 binary input variables, O_1 an unobserved variable and $Y = f(x)$ the output of the model.

Table 4.4: The Shapley values for the possible data points regarding the structure in figure 4.4.

	Marginal			Conditional			Causal		
	ϕ_1	ϕ_2	ϕ_3	ϕ_1	ϕ_2	ϕ_3	ϕ_1	ϕ_2	ϕ_3
$X_1 = 0, X_2 = 0, X_3 = 0, f(x_{1,2,3}) = 0$	0	0	$-\frac{1}{2}$	$-\frac{1}{6}$	$-\frac{1}{6}$	$-\frac{1}{6}$	$-\frac{1}{4}$	0	$-\frac{1}{4}$
$X_1 = 1, X_2 = 1, X_3 = 1, f(x_{1,2,3}) = 1$	0	0	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{4}$	0	$\frac{1}{4}$

4.2.5 Conclusion

The previous examples have shown that the causal Shapley values provide intuitive explanations as they attribute value to the features that cause a difference in output and do not attribute to features that accidentally correlate with other features that influence the output.

Chapter 5

Conclusions

This work has introduced causal Shapley values. They show how to take into account the causal structure of the data when calculating the Shapley value. In case of no causal paths, the causal Shapley values boil down to the marginal Shapley values. But in case that we do have causal paths, the causal Shapley values do not boil down to either the marginal Shapley values or the conditional Shapley values. In order to understand what this does boil down to, we need to have some idea of the underlying causal structure, because then we can try to figure out how to rewrite the do-operator in the calculation of the causal Shapley values. Rewriting the do-operator, however, might not always be possible. Furthermore, the user may not always be able to fully specify the causal structure of the input variables. Therefore future work involves finding methods to approximate the causal Shapley values in such a way that it is easy for the user to provide the necessary knowledge on the causal structure and such that the effect is always identifiable.

Bibliography

- [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] Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.
- [3] AI Hleg. Ethics guidelines for trustworthy AI. *B-1049 Brussels*, 2019.
- [4] Yimin Huang and Marco Valtorta. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the national conference on artificial intelligence*, volume 21, page 1149. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
- [5] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causality problem. *arXiv preprint arXiv:1910.13413*, 2019.
- [6] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.
- [7] Cathy O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.
- [8] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
- [9] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [10] Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.
- [11] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

- [12] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
- [13] Peter Spirtes, Clark Glymour, and Richard Scheines. Causation and prediction: axioms and explications. In *Causation, prediction, and search*, pages 41–86. Springer, 1993.
- [14] Erik Štrumbelj and Igor Kononenko. A general method for visualizing and explaining black-box regression models. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 21–30. Springer, 2011.
- [15] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.
- [16] Nicholas Vollmer. Information to be provided where personal data are collected from the data subject. -<https://www.privacy-regulation.eu/en/13.htm>. Accessed: May 2020.
- [17] Nicholas Vollmer. Information to be provided where personal data have not been obtained from the data subject. -<https://www.privacy-regulation.eu/en/14.htm>. Accessed: May 2020.
- [18] Nicholas Vollmer. Right of access by the data subject. -<https://www.privacy-regulation.eu/en/15.htm>. Accessed: May 2020.