How to compute Shapley Values

1 Introduction

During the course, we introduced the concept of Shapley values and said that due to the great amount of calculation needed for big data sets with several features, an efficient way of computing them is needed, and for this, the SHAP library comes to help.
To summarize, focusing on a specific instance j, we can compute the Shapley value for each feature $k = 1, ..., K$. Let's choose feature A (a player in game theory) and compute the Shapley values, ϕ_A, as the weighted sum of the marginal contributions on different sets s:

$$\phi(x_A) = \sum_{s \in S} w_s \phi_s(x_A) \tag{1}$$

where the marginal contributions are the difference between the prediction on a subset containing the feature of interest, $f(s_{+x_A})$, and the subset without that feature $f(s_{-x_A})$.

$$\phi_s(x_A) = f(s_{+x_A}) - f(s_{-x_A}) \tag{2}$$

For the weights we need to consider that the larger and smaller set have the highest weight [1]. Once we compute the Shapley value for each feature k, we can use them to understand the impact they have on the prediction for the specific instance j we are focusing on. Moreover, a very important property of the Shapley values is that they add to the difference between the model prediction (with all the features, and the baseline), giving a way to estimate how each feature contributes to deviating the prediction compared to the baseline.

$$f(j) - \mu(f(J)) = \sum_{k \in K} \phi(x_A) + ... + \phi(x_K) \tag{3}$$

where $f(j)$ is the prediction for the instance (the one considering all the available features) and $\mu(f(J))$ is called **baseline**, i.e. the prediction computed when all features are excluded. In a tabular data set is often computed as the average prediction among all the instances in the data set.

But to check if we completely understand the Shapley values themselves, let's compute them in the context of a really small data set. Try to solve the following exercise [2] with pen and paper, and check if your solution is correct.

[1] For more detail on how the weights are computed, watch the following video `https://www.youtube.com/watch?v=UJeu29wq7d0`

[2] This exercise is taken from the blog post `https://www.aidancooper.co.uk/`

2 Exercise Description

You are working with a simplified version of the Boston housing data set[3] that collects information about the percentage of the population that is working class, the number of rooms and the nitric oxides concentration (parts per million) of a house. For simplicity, we will call the features A, B, and C. Table 1 shows a small example of the dataset.

Instance	**Feature A**	**Feature B**	**Feature C**
$house_1$	10.0	6.50	0.5
$house_2$	21.0	6.22	0.6
$house_3$	39.5	5.34	0.5
$house_4$	24.7	6.23	0.69

Table 1: Feature description

Let's imagine you trained a machine learning model to predict the car price (regression problem) and you want to explain the results. In the specific, you want to understand which features are impacting your new prediction. Considering that you trained the model and you are able to run inference, you can have the value of the prediction for all the combinations of features (and you can exclude some by shuffling them); the values of the prediction on all the possible subsets for features are summarized in Figure 1.

Task: Compute the Shapley value for the instance $house_1$, $\phi(house_1)$

how-shapley-values-work/ and developed to contain all the steps and more detailed explanations

[3] https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html?xgtab=&ref=aidancooper.co.uk

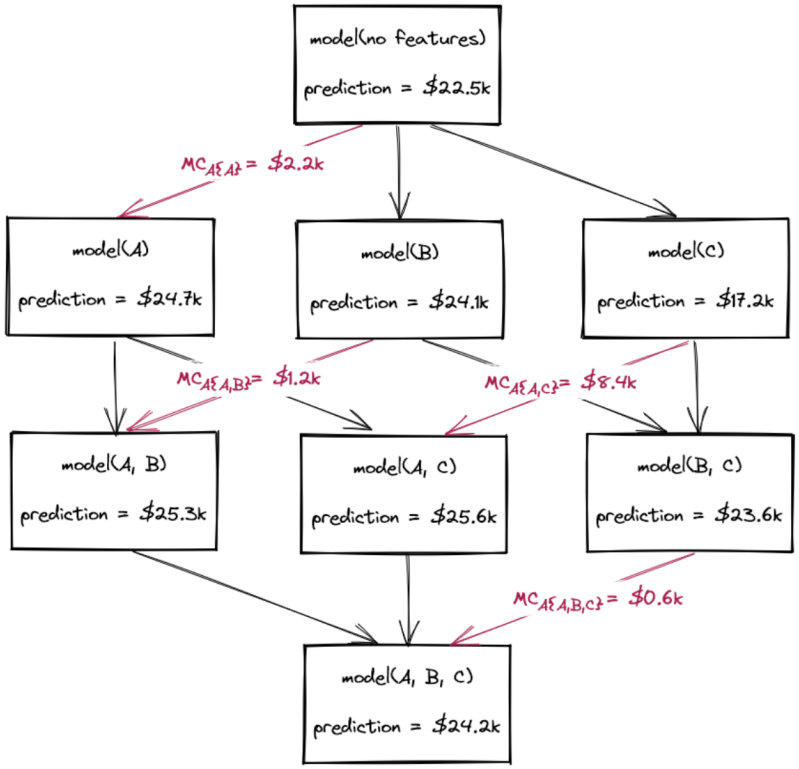

Figure 1: Values of inference for different feature combinations

3 Solution

Let's compute the Shapley value for each feature and the first instance $house_1$, knowing that the model prediction is $f(house_1) = 24.2\ K\$$, and this corresponds to the prediction made considering all the available features ($model(A, B, C)$).

Feature A

We compute all the marginal contributions first:

$$\phi_{s_1}(A) = f(m(A,B,C)) - f(m(B,C)) = (24.2 - 23.6)\ K\$ = 0.6\ K\$$$

$$\phi_{s_2}(A) = f(m(A,B)) - f(m(B)) = (25.3 - 24.1)\ K\$ = 1.2\ K\$$$

$$\phi_{s_3}(A) = f(m(A,C)) - f(m(C)) = (25.6 - 17.2)\ K\$ = 8.4k\ K\$$$

$$\phi_{s_4}(A) = f(m(A)) - f(m(\{\})) = (24.7 - 22.5)\ K\$ = 2.2\ K\$$$

Therefore using Eq. 1

$$\begin{aligned}\Rightarrow \phi(A) &= w_1\phi_{s_1}(A) + w_3\phi_{s_2}(A) + w_3\phi_{s_3}(A) + w_4\phi_{s_4}(A) = \\ &= \left(\frac{1}{3} \times 0.6 + \frac{1}{6} \times 1.2 + \frac{1}{6} \times 8.4 + \frac{1}{3} \times 2.2\right)\ K\$ = \\ &= 2.5\ K\$\end{aligned}$$

where the sets s_1 and s_4 have higher weights (one possible combination over the three possible, therefore weights equals 1/3).

Feature B

We compute all the marginal contributions first:

$$\phi_{s_1}(B) = f(m(A,B,C)) - f(m(A,C)) = (24.2 - 25.6)\ K\$ = -1.4\ K\$$$

$$\phi_{s_2}(A) = f(m(A,B)) - f(m(A)) = (25.3 - 24.7)\ K\$ = 0.6\ K\$$$

$$\phi_{s_3}(A) = f(m(B,C)) - f(m(C)) = (23.6 - 17.2)\ K\$ = 6.4k\ K\$$$

$$\phi_{s_4}(A) = f(m(B)) - f(m(\{\})) = (24.1 - 22.5)\ K\$ = 1.6\ K\$$$

Therefore, using Eq. 1

$$\begin{aligned}\Rightarrow \phi(B) &= w_1\phi_{s_1}(B) + w_3\phi_{s_2}(B) + w_3\phi_{s_3}(B) + w_4\phi_{s_4}(B) = \\ &= \left(\frac{1}{3} \times (-1.4) + \frac{1}{6} \times 0.6 + \frac{1}{6} \times 6.4 + \frac{1}{3} \times 1.6\right)\ K\$ = \\ &= 1.25\ K\$\end{aligned}$$

Feature C

We compute all the marginal contributions first:

$$\phi_{s_1}(C) = f(m(A,B,C)) - f(m(A,B)) = (24.2 - 25.3)\ K\$ = -1.1\ K\$$$
$$\phi_{s_2}(C) = f(m(A,C)) - f(m(A)) = (25.6 - 24.7)\ K\$ = 0.9\ K\$$$
$$\phi_{s_3}(A) = f(m(B,C)) - f(m(B)) = (23.6 - 24.1)\ K\$ = -0.5k\ K\$$$
$$\phi_{s_4}(A) = f(m(C)) - f(m(\{\})) = (17.2 - 22.5)\ K\$ = -5.3\ K\$$$

Therefore, using Eq. 1

$$\Rightarrow \phi(B) = w_1\phi_{s_1}(C) + w_3\phi_{s_2}(C) + w_3\phi_{s_3}(C) + w_4\phi_{s_4}(C) =$$
$$= \left(\frac{1}{3} \times (-1.1) + \frac{1}{6} \times 0.9 + \frac{1}{6} \times (-0.5) + \frac{1}{3} \times (-5.3)\right)\ K\$ =$$
$$= -2.10\ K\$$$

Let's put all of them together to understand the effect of each feature on the prediction compared to the baseline. Using Eq. 3, we use $f(house_1) = f(m(A,B,C))$ as a prediction for the instance $house_1$ and $f(m(\{\})$ as baseline:

$$f(house_1) = f(m(\{\}) + \phi(A) + \phi(B) + \phi(C) =$$
$$= (22.5 + 2.5 + 1.25 - 2.1)\ K\$ =$$

We can conclude that the feature A and B contribute positively, increasing the predicted value for $house_1$, while feature C tends to reduce the predicted value. In the Figure 2

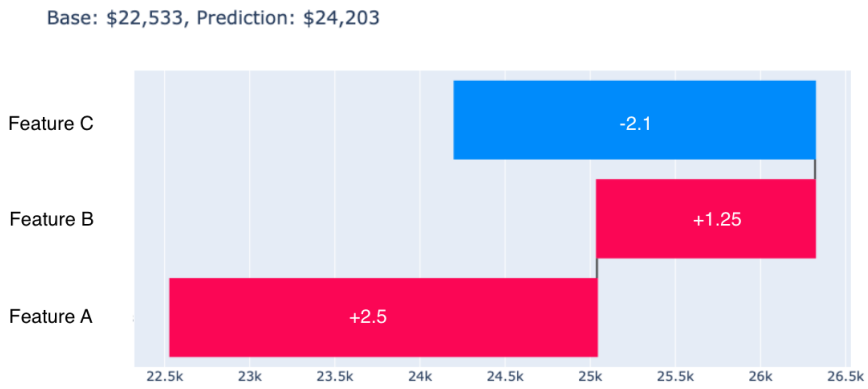

Figure 2: SHAP values plotted in red (positive contributions) and blue (negative contributions)