GAME THEORY IN MACHINE LEARNING

DOMINIK KOZŁOWSKI

ABSTRACT. In recent years machine learning models and deep neural networks became more common within business enterprises due to their ability to make high-accuracy predictions from large datasets. However, it is often achieved by models that even experts struggle to interpret why certain predictions are being made. This imposes limitations on where this technology can be applied. It resulted in a need to find a method that will help users interpret these predictions. In this paper I will introduce the Shapley values – a method from coalitional game theory which tells us how to fairly distribute the "payout" among the players and how it is used in Machine Learning.

INTRODUCTION

It is crucial to be able to appropriately interpret the results of a prediction model. It promotes proper user trust, offers suggestions on how a model could be improved and aids comprehension of the process replication. In some cases, simple models, like linear or tree models are frequently favoured in particular applications because they are easier to interpret, even if they may be less accurate than complex ones. However, the advantages of utilizing complicated models have grown as a result of the increased accessibility of big data, highlighting the tension between the model's accuracy and interpretability. So far, the most popular frameworks to this problem comes from cooperative game theory: The Shapley value, introduced by Lloyd Shapley in 1953. He was later awarded the Nobel Prize in Economics for this work.

1. PRELIMINARIES

Let Γ be an infinite set, the universe of all players, and denote by $\aleph$ the set of all finite subsets of Γ. A cooperative game with transferable utility (TU - game) is a pair (N, v) consisting of a set of players $N \in \aleph$ and coalition function $v \in \{f : 2^N \to \mathbf{R} | f(\emptyset) = 0\}$ where 2^N is the power set of N. In fact, a TU-game can be uniquely identified by its coalition function v. The subsets $S \subseteq N$ are called coalitions and $v(S)$ is called the worth of coalition S. The set of all TU- games with player set N is denoted by $\mathcal{G}^N$ and the set of all TU-games on N where value of all singletons are all positive by $\mathcal{G}_0^N = \{v \in \mathcal{G}^N : v(\{i\}) > 0 \text{ for all } i \in N\}$.If the value of the singletons can only be all positive rational we mark this set by $\mathcal{G}_0^N = \{v \in \mathcal{G}^N : v(\{i\}) \in \mathbb{Q}_+ \text{ for all } i \in N\}$. A TU-value on $\mathcal{G}^N$ is an operator ϕ, which assigns any $v \in \mathcal{G}^N$ a payoff vector $\phi(N, v) \in \mathbb{R}^N$ or $\phi(v)$ in short for all $N \in \aleph$, with the meaning that $\phi_i(v)$ is the payoff to player i in the TU-game v.

.

2. Shapley's Axioms

Given a game in coalitional form (N, v), the Shapley value is denoted by $\phi(v) = (\phi_1(v), ..., \phi_n(v))$ where $\phi_i(v)$ is the expected payoff to player i. The Shapley value tries to capture how coalitional competitive forces influence the possible outcomes of a game. It describes a "fair way" of dividing the gains from cooperation given the strategic realities captured by the characteristic function. Let π be a permutation on the set N. Let $(N, \pi v)$ be the coalitional game such that

$$\pi v(\{\pi(i) : i \in C\}) = v(C) \quad \forall C \subseteq N.$$

This means that the role of any player $i \in N$, in the game (N, v) is essentially the same as the role of player $\pi(i)$ in $(N, \pi v)$.
Example
For example, suppose $N = \{a, b, c\}$. Consider the permutation π on N defined by $\pi(a) = c; \pi(b) = a; \pi(c) = b$. Then the game $(N, \pi v)$ will be described by the following: $\pi v(a) = v(b); \pi v(b) = v(c); \pi v(c) = v(a); \pi v(ab) = v(bc); \pi v(bc) = v(ac); \pi v(ac) = v(ab); \pi v(abc) = v(abc)$, where $v(abc)$ denotes $v(\{a, b, c\})$, etc.

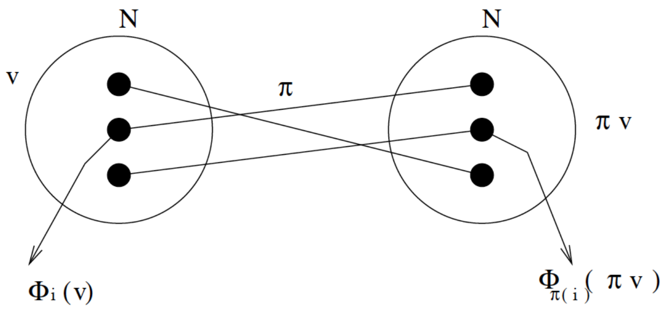
Shapley put out three axioms to outline the desired characteristics that a "fair solution" should possess

Axiom 1: Symmetry.
Axiom 2: Linearity.
Axiom 3: Carrier.

2.1. **Symmetry.** For all $v \in \mathbf{R}^{2^n-1}$, any permutation π on N, and all player $i \in N$,

$$\phi_{\pi(i)}(\pi v) = \phi_i(v)$$

In other words, Shapley value of a player relabeled by a permutation, under the permuted value function is the same as the Shapley value of the original player in the initial game. According to this principle, a player's specific role in the game should be what counts. It does not matter whether labels or exact names are used in N.



2.2. **Linearity.** Let (N, v) and (N, w) be any two coalitional games. Suppose $p \in [0, 1]$. Define a new coalitional game $(N, pv + (1 - p)w)$ as follows.

$$(pv + (1 - p)w)(C) = pv(C) + (1 - p)w(C) \quad \forall C \subseteq N$$

Linearity states that, given any two coalitional games (N, v) and (N, w), any number $p \in [0, 1]$, and any player $i \in N$,

$$\phi_i(pv + (1-p)w) = p\phi_i(v) + (1-p)\phi_i(w)$$

In other words, the Shapley value of a player for a convex combination of coalitional games is the convex combination of Shapley values of the player in the individual games. This axiom asserts that the expected payoff to each player is the same before resolution of uncertainty and after resolution of uncertainty.

Example Imagine beginning of football match. The toss decide if team will start the match. The mapping v corresponds to game when the team has won the toss while w when the team has lost.

- $\phi_i(pv + (1-p)w)$ is the expected payoff to player i before the toss takes place.
- $p\phi_i(v) + (1-p)\phi_i(w)$ is the expected payoff to player i after the toss takes place.

Linearity in this case means that these two expected payoffs are the same.

2.3. **Carrier.** A coalition D is said to be a *carrier* of a coalitional game (N, v) if

$$v(C \cap D) = v(C) \quad \forall C \subseteq N$$

If D is a carrier and $i \notin D$, then

$$v(\{i\}) = v((\{i\}) \cap D) = v(\phi) = 0$$

If D is a carrier of (N, v), all players $j \in (N - D)$ are called dummies in (N, v) because their entry into any coalition cannot change the worth of the coalition. Also, for any $C \subseteq N$ and $i \notin D$,

$$v(C \cup \{i\}) = v((C \cup \{i\}) \cap D) = v(C \cap D) = v(C).$$

Set D includes all non dummy players. If D is a carrier and $i \in N$, then for any $C \subseteq N$,

$$v(C) = v(C \cap D) = v(C \cap (D \cup \{i\}))$$

Hence, $D \cup \{i\}$ for any $i \in N$ is also a carrier. In fact, the set N is always a carrier. This means that $v(D) = v(N)$. This can also be seen from

$$v(D) = v(DN) = v(N)$$

It is however possible that no proper subset of N is a carrier. The Carrier Axiom states that, for any $v \in \mathbf{R}^{2^n - 1}$ and any coalition D that is a carrier of (N, v),

$$\sum_{i \in D} \phi_i(v) = v(D) = v(N)$$

The carrier axiom immediately implies that

$$\phi_i(v) = 0 \quad \forall i \notin D$$

$$\sum_{i \in N} \phi_i(v) = v(N) \quad i \in N$$

According to this axiom, the participants in a carrier set should distribute among themselves their total worth, which is equivalent to the value of the great coalition. Shapley value always distributes the grand coalition's value across the game's participants. However, if player is a dummy he does not receive anything.

3. Shapley's Theorem

With all three axioms we can state Shapley's theorem.

Theorem 3.1 (Shapley)**.** *There is exactly one mapping* $\phi : \mathbf{R}^{2^n-1} \to \mathbf{R}^n$ *that satisfies all three axioms. This mapping satisfies:* $\forall i \in N, \forall v \in \mathbf{R}^{2^n-1}$*,*

$$\phi_i(v) = \sum_{C \subseteq N-i} \frac{|C|!(n-|C|-1)!}{n!}\{v(C \cup \{i\}) - v(C)\}$$

The notation $N - i$ above denotes the set $N \ \setminus \{i\}$. The term $\frac{|C|!(n-|C|-1)!}{n!}$ can be interpreted as the probability that in any permutation, the members of C are ahead of a distinguished player i. The term $v(C\{i\}) - v(C)$ gives the marginal contribution of player i to the worth of the coalition C. Thus the above formula for $\phi_i(v)$ gives the expected contribution of player i to the worth of any coalition. Suppose there is a collection of n resources and each resource is useful in its own way towards executing a certain service. Suppose $v(N)$ is the total value that this collection of resources would create if all the resources are deployed for the service accomplishment. Let us focus on a certain resource, say resource i. Now, this resource will make a marginal contribution to every subset C of $N - i$ when it is included to the set C. We can choose the set C in $(|C|!|n-|C|-1|!)$ ways and when this is divided by $n!$, we obtain the probability of choosing a particular subset C. Thus the Shapley value of player i is the average marginal contribution that player i will make to any arbitrary coalition that is a subset of $N - i$.

Proof. In this paper I will only show that ϕ satisfies all three axioms. Proof of uniqueness of the mapping can be found in an article: *A value for n-person games* written by Lloyd S. Shapley. on pages 309–317.

Symmetry

We can observe that in the formula for $\phi_i(v)$, what only matters about a coalition is the number of players it contains and does it contains i. Thus relabeling does not affect the value in any way. This observation clearly shows that symmetry is satisfied.

Linearity

We have, by definition, for any $p \in [0,1]$,

$$(pv + (1-p)w)(C) = pv(C) + (1-p)w(C) \quad \forall C \subseteq N$$

Note that

$$\phi_i(v)(pv+(1-p)w) = \sum_{C \subseteq N-i} \frac{|C|!(n-|C|-1)!}{n!}((pv+(1-p)w))(C \cup \{i\}) - (pv+(1-p)w)(C))$$

$$= pv(C) + (1-p)w(C) \quad \forall C \subseteq N$$

Carrier

Suppose D is a carrier of (N, v). We can see that $v(D) = v(N)$. Then, we know that

$$v(C) = v(C \cap D) \quad \forall C \subseteq N$$

We have also seen that

$$v(\{i\}) = 0 \quad \forall i \in N \setminus D$$

If we take a look at the formula for the Shapley value, it is very clear that

$$\phi_i(v) = 0 \quad \forall i \in N \setminus D$$

since

$$v(C \cup \{i\}) = v((C \cup \{i\}) \cap D) = v(C \cap D) = v(C).$$

We have to show for all carriers D that

$$\sum_{i \in D} \phi_i(v) = v(D)$$

Substituting the formula for $\phi_i(v)$ from the Shapley theorem and simplifying, we can show that

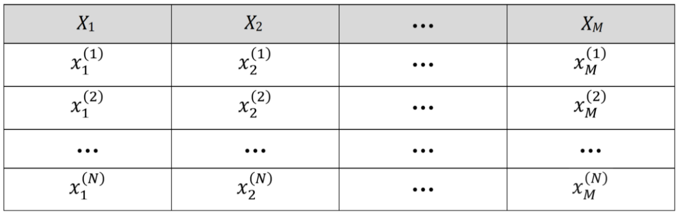
$$\sum_{i \in N} \phi_i(v) = v(N)$$

Since D is a carrier, we have $v(D) = v(N)$, hence the carrier axiom follows.

□

4. Shapley values in Machine learning

How a theorem from Game Theory changed Machine Learning industry? Suppose that we have a data set with N rows and M features.



X_1	X_2	...	X_M
$x_1^{(1)}$	$x_2^{(1)}$	...	$x_M^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	...	$x_M^{(2)}$
...	...	...	...
$x_1^{(N)}$	$x_2^{(N)}$	...	$x_M^{(N)}$

X_i is the i-th feature of the data set,$x_i^{(j)}$ is the value of the i-th feature in the j-th sample, and $y^{(j)}$ is the target of the j-th row. The values of the features form a feature vector:

$$\boldsymbol{x} = [x_1, x_2, ..., x_M]$$

Here we have $X_1 = x_1, X_2 = x_2, \dots X_M = x_M$. A feature vector can be also the j-th row of the data set. In that case, we can write:

$$\boldsymbol{x}^{(j)} = [x_1^{(j)}, x_2^{(j)}, ..., x_M^{(j)}]$$

It can be a test data point that is not present in the data set. A pair $(\boldsymbol{x}^{(j)}, y^{(j)})$ is called a training example of this data set. Now we can use a model to learn this data set:

$$f(\boldsymbol{x}) = f(x_1, x_2, ..., x_M)$$

The function takes the feature vector $\boldsymbol{x}$ which means that it is applied to all the elements of $\boldsymbol{x}$.

Example

Let $\boldsymbol{x}$ be a feature vector from training or test data set. For linear regression model:

$$f(\boldsymbol{x}) = c_0 + c_1 x_1 + c_2 x_2 + ... + c_M x_M$$

For each value of vector $\boldsymbol{x}$, model predicts a target value $f(\boldsymbol{x})$. We we can predict the target of one of the training examples:

$$f(\boldsymbol{x}^{(j)}) = c_0 + c_1 x_1^{(j)} + c_2 x_2^{(j)} + ... + c_M x_M^{(j)}$$

So, $f(x^{(j)})$ is the model prediction for the j-th row of the data set, and the difference between $f(\boldsymbol{x}^{(j)})$ and $y^{(j)}$ is the prediction error of the model for the j-th training example.

We assume that a coalition game if the prediciton task for a single instance. The "players" are the feature values of the instance that collaborate to receive the gain. But what should be the characteristic function of this game? $f(\boldsymbol{x})$ itself is a good starting point. We have to remember that a characteristic function has to fulfill Carrier axiom, which states that the sum gain is zero in the absence of players. How can we evaluate $f(\boldsymbol{x})$ when we have no features?

The absence of a feature from the game indicates that its current value is unknown, and we want to estimate the model's target without having the knowledge of that feature value. When there are no features in the game, this indicates that none of the features' current values are known. In that case, we can only use the training set for prediction. In this case, we can use the average of $f(\boldsymbol{x}^{(j)})$ for a sample of the training examples (or all of them) as our best estimate.Therefore, when there are no features, our forecast is:

$$f(NA, NA, ..., NA) = E[f(\mathbf{x})] = \frac{1}{k}\sum_{j=1}^{k} f(\mathbf{x})^{(j)}$$

NA stands for a feature that is not available. From the training data set ($k \leq N$), we also sampled k data feature vectors. We now specify characteristic function for the great coalition as:

$$v(\mathbf{N}) = v(\{X_1, X_2, ..., X_m\}) = f(\mathbf{x}) - E[f(\mathbf{x})] = f(\mathbf{x}) - \sum_{j=1}^{k} f(\mathbf{x})^{(j)}$$

If we have no features, then we get:

$$v(\emptyset) = E[f(\mathbf{x})] - E[f(\mathbf{x})] = 0$$

This characteristic function now satisfies Carrier axiom and is able to present the great coalition's value. $N = X, X, \ldots, X_M$. Now we will show how can we apply the function f for any coalition of $N - i$. We begin with retraining the model only on a subset of the original features. Let say that coalition $\mathbf{C}$ contains the features:

$$\mathbf{C} = \{X_{c1}, X_{c2}, ..., X_{cp}\}$$

The marginal value $f_{\mathbf{C}}(\mathbf{x}_C)$ for these attributes is then required:

$$f_{\mathbf{C}}(\mathbf{x}_C) = f_{\mathbf{C}}(x_{c1}, x_{c2}, ..., x_{cp})$$

Here $\mathbf{x_C}$ is a vector that only contains the values of the features present in $\mathbf{C}$. The original function f can be used to calculate $f_{\mathbf{C}}$. Alternatively we can retrain similar model on the features present in the coalition C, to get $f_{\mathbf{C}}(\mathbf{x_C})$. When a feature is not present in C, it is replaced with NA.

$$f_{\mathbf{C}}(\mathbf{x_C}) = f(x_1, x_2, ..., x_M) \text{ where } x_i = NA \text{ if } x_i \notin \mathbf{C}$$

Now we can assume that the model represented by f can handle NA values. Hence the worth of this coalition is:

$$v(\mathbf{C}) = v(\{x_{\mathbf{C}_1}, x_{\mathbf{C}_2}, ..., x_{\mathbf{C}_p}\}) = f_{\mathbf{C}}(\mathbf{x}_{\mathbf{C}}) - E[f(\mathbf{x})]$$

Now we can calculate the Shapley value of the feature X_i:

$$\begin{aligned}\phi_i = &\sum_{\mathbf{C} \subseteq \mathbf{N} - \{i\}} \frac{|\mathbf{C}|!(|N| - |C| - 1)!}{|N|!}(f_{\mathbf{C}\cup\{i\}}(\mathbf{x}_{\mathbf{C}\cup\{i\}})) - E[f(\mathbf{x})] - (f_{\mathbf{C}}(\mathbf{x}_{\mathbf{C}}) - E[f(\mathbf{x})])) \\ &= \sum_{\mathbf{C} \subseteq \mathbf{N} - \{i\}} \frac{|\mathbf{C}|!(|N| - |C| - 1)!}{|N|!}(f_{\mathbf{C}\cup\{i\}}(\mathbf{x}_{\mathbf{C}\cup\{i\}})) - f_{\mathbf{C}}(\mathbf{x}_{\mathbf{C}}).\end{aligned}$$

where $f_{\boldsymbol{C}}(\mathbf{x}_{\boldsymbol{C}})$ is the marginal value of f for the features present in the coalition C, $f_C\{i\}(x_C\{i\})$ is the marginal value of f for the features present in the coalition C plus feature i.Now we can write:

$$\sum_{i=1}^{|N|} \phi_i = v(\mathbf{N}) = f(\mathbf{x}) - E[f(\mathbf{x})]$$

which means that the sum of the Shapely values of all the features gives the difference between the prediction of the model with the current value of features and the average prediction of the model for all the training examples.

5. Summary

In this paper I have introduce Shapley values and how they are used in Machine Learning. I encourage the insightful reader to read *A Unified Approach to Interpreting Model Predictions* by Scott M. Lundberg, where he introduced *SHAP* - state of the art framework based on Shapley values.

References

1. Lloyd S. Shapley. *A value for n-person games. In Robert Kuhn and Albert Tucker, editors, Contributions to the Theory of Games*, pages 307–317. Princeton University Press, 1953.
2. Scott M. Lundberg, Su-In Lee *A Unified Approach to Interpreting Model Predictions* 2017.
3. Y. Narahari, *Game Theory, Lecture Notes By Y. Narahari, Chapter 32,* 2012.
4. Reza Bagheri, *Introduction to SHAP Values and their Application in Machine Learning*, medium.com 2022.

Department of Mathematics, Lodz University of Technology, Lodz, Poland
Email address: 244075@edu.p.lodz.pl