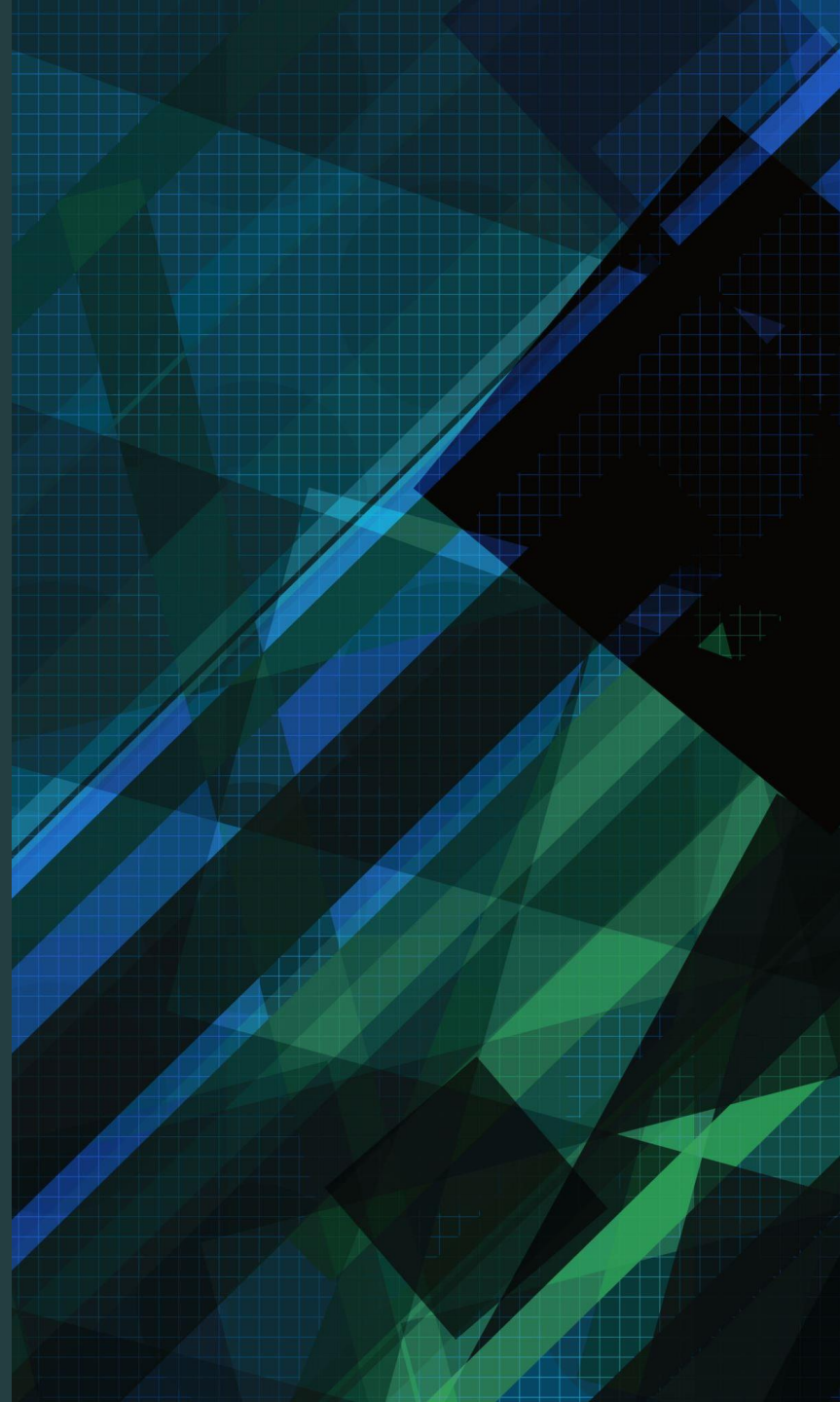
SHAPLEY VALUE

XIUYI FAN

CSCM23



SHAPLEY VALUE

- Solution concept in cooperative game theory
- Named after Lloyd Shapley – introduced in 1951
- To each cooperative game Shapley value assigns a unique distribution (among the players) of a total surplus generated by the coalition of all players

SHAPLEY VALUE

- Coalition of players cooperates and obtains a certain overall gain from that cooperation. Since some players may contribute more to the coalition than others or may possess different bargaining power, what final distribution of generated surplus among the players should arise in any particular game?

- *How important is each player to the overall cooperation, and what payoff can he or she reasonably expect?*

DEFINITION (WIKIPEDIA)

- A coalitional game is defined as: There is a set N (of n players) and a function v that maps subsets of players to the real numbers: $v : 2^N \rightarrow R$, with $v(\{\}) = 0$. The function v is called a characteristic function.

- The function v has the following meaning: if S is a coalition of players, then $v(S)$, called the worth of coalition S, describes the total expected sum of payoffs the members of S can obtain by cooperation.

DEFINITION

According to the Shapley value, the amount that player *i* gets given in a coalitional game (*v, N*) is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where *n* is the total number of players and the sum extends over all subsets *S* of *N* not containing player *i*.

The formula can be interpreted as follows: imagine the coalition being formed one actor at a time, with each actor *i* demanding their contribution v(S∪{i}) − v(S) as a fair compensation, and then for each actor *i* take the average of this contribution over the possible different permutations in which the coalition can be formed.

DEFINITION

According to the Shapley value, the amount that player *i* gets given in a coalitional game (*v, N*) is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where *n* is the total number of players and the sum extends over all subsets *S* of *N* not containing player *i*.

marginal contribution of *i* to coalition S

DEFINITION

According to the Shapley value, the amount that player *i* gets given in a coalitional game (*v, N*) is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where *n* is the total number of players and the sum extends over all subsets *S* of *N* not containing player *i*.

ordering of coalitions formed before including *i*

DEFINITION

According to the Shapley value, the amount that player *i* gets given in a coalitional game (*v, N*) is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where *n* is the total number of players and the sum extends over all subsets *S* of *N* not containing player *i*.

ordering of coalitions formed after including *i*

DEFINITION

According to the Shapley value, the amount that player *i* gets given in a coalitional game (*v, N*) is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where *n* is the total number of players and the sum extends over all subsets *S* of *N* not containing player *i*.

total number of orderings

DEFINITION

According to the Shapley value, the amount that player *i* gets given in a coalitional game (*v, N*) is:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

where *n* is the total number of players and the sum extends over all subsets *S* of *N* not containing player *i*.

Sum of all computed values

DEFINITION (ALTERNATIVE)

Alternatively, Shapely Value can be equivalently defined as:

$$\varphi_i(v) = \frac{1}{n!} \sum_R \left[v(P_i^R \cup \{i\}) - v(P_i^R) \right]$$

where orders R is the order of the players and P_i^R is the set of players in N which precede i in the order R.

EXAMPLE

$$\varphi_i(v) = \frac{1}{n!} \sum_R \left[v(P_i^R \cup \{i\}) - v(P_i^R) \right]$$

Orders *R* is the order of the players and P_i^R is the set of players in *N* which precede *i* in the order *R*.

Two players: 1, 2

Characteristic function: v({1}) = 1, v({2}) = 2, v({1, 2}) = 4

For player 1:

Order R	Marginal Contribution
1, 2	v({1}) – v({}) = 1 – 0 = 1
2, 1	v({1,2}) – v({2}) = 4 – 2 = 2

$\Phi_1(v) = (1 + 2) / 2! = 3 / 2 = 1.5$

EXAMPLE

$$\varphi_i(v) = \frac{1}{n!} \sum_R \left[v(P_i^R \cup \{i\}) - v(P_i^R) \right]$$

Orders *R* is the order of the players and P_i^R is the set of players in *N* which precede *i* in the order *R*.

Two players: 1, 2

Characteristic function: v({1}) = 1, v({2}) = 2, v({1, 2}) = 4

For player 2:

Order R	Marginal Contribution
1, 2	v({1,2}) – v({1}) = 4 – 1 = 3
2, 1	v({2}) – v({}) = 2 – 0 = 2

$\Phi_2(v) = (3 + 2) / 2! = 5 / 2 = 2.5$

SHAPLEY VALUE PROPERTIES

Efficiency

- The sum of the Shapley values of all agents equals the value of the grand coalition, so that all the gain is distributed among the agents:

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

In our example:

- $v(\{1, 2\}) = 4$
- $\Phi_1(v) + \Phi_2(v) = 1.5 + 2.5 = 4$

SHAPLEY VALUE PROPERTIES

Symmetry

- If *i* and *j* are two actors who are equivalent in the sense that

$$v(S \cup \{i\}) = v(S \cup \{j\})$$

for every subset *S* of *N* which contains neither *i* nor *j*, then $\phi_i(v) = \phi_j(v)$.

SHAPLEY VALUE PROPERTIES

Linearity

If two coalition games described by characteristic functions *v* and *w* are combined, then the distributed gains should correspond to the gains derived from *v* and the gains derived from *w*:

$$\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$$

for every *i* in *N*.

SHAPLEY VALUE PROPERTIES

Null player

The Shapley value $\phi_i(v)$ of a null player i in a game v is zero. A player i is null in v if $v(S \cup \{i\}) = v(S)$ for all coalitions S that do not contain i.

Given a player set N, the Shapley value is the only map from the set of all games to payoff vectors that satisfies all four properties: ***Efficiency, Symmetry, Linearity, Null player***.

SHAPLEY VALUE FOR MACHINE LEARNING

A prediction can be explained by assuming that each feature value of the instance is a “player” in a game where the prediction is the pay-out.

The “game” is the prediction task for a single instance of the dataset. The “gain” is the actual prediction for this instance minus the average prediction for all instances. The “players” are the feature values of the instance that collaborate to receive the gain (= predict a certain value).

SHAPLEY VALUE FOR MACHINE LEARNING

An intuitive way to understand the Shapley value is the following illustration: The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them.

SHAPLEY VALUE FOR MACHINE LEARNING

- For a certain apartment it predicts €300,000 and you need to explain this prediction. The apartment has a size of 50 m^2, is located on the 2nd floor, has a park nearby and cats are banned:

- Features: *park-nearby, cat-banned, area-50, floor-2nd*

SHAPLEY VALUE FOR MACHINE LEARNING

- Features: *park-nearby, cat-banned, area-50, floor-2nd*
- To compute the "*contribution*" of each feature, we need the characteristic function v, e.g.,

  v({}),

  v({park-nearby}), v({cat-banned}), v({area-50}), v({floor-2nd}),

  v({park-nearby, cat-banned}), v({park-nearby, floor-2nd}), …,

  v({park-nearby, cat-banned, area-50}), v({park-nearby, cat-banned, floor-2nd}), …,

  v({park-nearby, cat-banned, area-50, floor-2nd})

SHAPLEY VALUE FOR MACHINE LEARNING

- Features: *park-nearby, cat-banned, area-50, floor-2nd*
- To compute the "*contribution*" of each feature, we need the characteristic function v, e.g.,

  v({}) = 0,

  v({park-nearby}), v({cat-banned}), v({area-50}), v({floor-2nd}) = ??

  v({park-nearby, cat-banned}), v({park-nearby, floor-2nd}), … = ??

  v({park-nearby, cat-banned, area-50}), v({park-nearby, cat-banned, floor-2nd}), … == ??

  v({park-nearby, cat-banned, area-50, floor-2nd}) = 300,000

SHAPLEY VALUE FOR MACHINE LEARNING

To estimate v({x}), use v({x}) = v({x,rand$_1$, rand$_2$, …, rand$_n$})

- Randomly select feature values for features that are NOT presented in the current coalition
- Do this multiple times and compute the average
- For example:

v({park-nearby, cat-banned, area-50}) = ½ * (v({park-nearby, cat-banned, area-50, floor-2nd}) +
v({park-nearby, cat-banned, area-50, floor-1st}))

Do this for all coalitions to estimate v

Then carry out the standard Shapley Value calculation for every feature

REFERENCE

- https://christophm.github.io/interpretable-ml-book/shapley.html
- https://youtu.be/qcLZMYPdpH4

SUMMARY

- Derived from Game Theory
- Four axioms: efficiency, symmetry, Linearity, Null player
- Explain a prediction as a game