Interpretable Machine Learning

Shapley Values





Learning goals

- Learn what game theory is
- Understand the concept behind cooperative games
 Understand the concept behind cooperative games
- Understand the Shapley value in game theory me theory

Game theory is the study of strategic games between players/egame arefers for s to any series of any series of interactions between actors/agents with gains and losses of lable utility value quantifiable utility value

- Game theory is the study of strategic games between players egame arefers to rs to any series of any series of interactions between actors/agents with gains and losses of lable utility value
 quantifiable utility value For all possible players P = {1,...,p}, each subset of players
- Cooperative games: For all possible players: Pess { tertairp} aeach subset of players S ⊆ P forms a coalition each coalition S achieves a certain payout

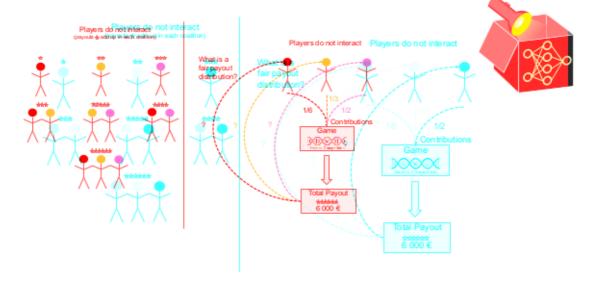
- Game theory is the study of strategic games between players/game arefers to any series of any series of interactions between actors/agents with gains and losses of lable utility value
 quantifiable utility Yalue For all possible players P = {1,...,p}, each subset of players S
- A value function v : 2^P → ℝ maps all 2^{|P|} possible coalitions to their payout (or gain)

- Game theory is the study of strategic games between players/game arefers to any series of any series of interactions between actors/agents with gains and losses of lable utility value
 quantifiable utility Yalue For all possible players P = {1,...,p}, each subset of players S
- Cooperative glames: For all possible players: Pes { tertairp} agach subset of eplayers S = Peterms a coalition reach coalition Sachieves a cortain payout ayout (or gain)
- A value function $v \in 2^{n} \to \mathbb{R}$ maps all $2^{\lfloor p \rfloor}$ possible coalitions to their payout forero: $v(\emptyset) = 0$) gain)
- v(S) is the payout of coalition $S \subseteq P$ (payout of empty coalition must be zero: $v(\emptyset) = 0$)

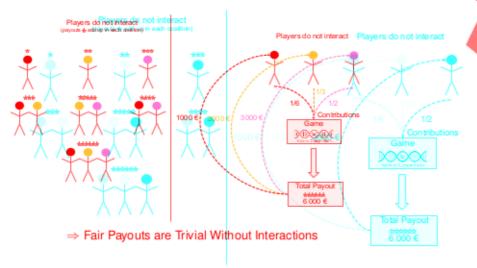
- Game theory is the study of strategic games between players "game" arefers to any series of any series of interactions between actors/agents with gains and losses of lable utility value
 quantifiable utility value For all possible players P = {1,...,p}, each subset of players \$2.
- Cooperative games: For all possible players: Pes { tertairp} agach subset of
 players S = Reforms a coalition reach coalition Sachieves a certain payout ayout (or gain)
- A value function $v_0 : 2^0 \mapsto \mathbb{R}$ maps all $2^{|P|}$ possible coalitions to their payout forero: $v(\emptyset) = 0$
- Parameter v(S) is the payout of coalition $S \subseteq R$ cleavout of empty coalition must be zero uton $v(\emptyset) = 0$
- As some players contribute more than others, we want to fairly divide the total achievable payout ν(P) among the players according to a player's individual contribution

- Game theory is the study of strategic games between players regame are fers to any series of any series of interactions between actors/agents with gains and losses of able utility value **equantifiable willity yalue** For all possible players $P = \{1, \dots, p\}$, each subset of players S
- Cooperative games: Forcall possible players: Pess { tertairp} aeach subset of players S - P forms a coalition - each coalition Sachieves a certain payout ayout (or gain)
- A value function $v_{i,j} 2^{\rho}_{i,j} \to \mathbb{R}$ maps all $2^{j,\rho}$ possible coalitions to their payout (or $v_{i,j} = 0$) gain) As some players contribute more than others, we want to fairly divide the total achievable
- v(S) is the payout of coalition $S \subseteq P$ (payout of empty coalition must be zero: tion
- We call the individual payout per player ϕ_j , $j \in P$ (later: Shapley value) As some players contribute more than others, we want to fairly divide the total achievable payout v(P) among the players according to a player's individual contribution
- We call the individual payout per player φ_i, j ∈ P (later: Shapley value)

COOPERATIVE GAMES WITHOUT INTERACTIONS

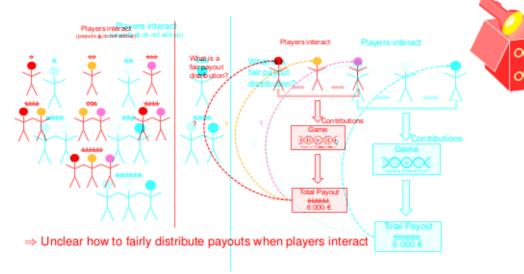


COOPERATIVE GAMES WITHOUT INTERACTIONS



⇒ Fair Payouts are Trivial Without Interactions

COOPERATIVE GAMES WITH INTERACTIONS



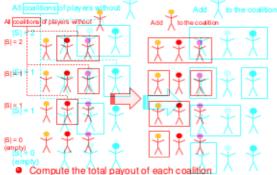
⇒ Unclear how to fairly distribute payouts when players interact

COOPERATIVE GAMES WITH INTERACTIONS

Question: What is a fair payout for player, yellow ??

Idea: Compute marginal contribution of the player of interest across differenterent coalitions

coalitions



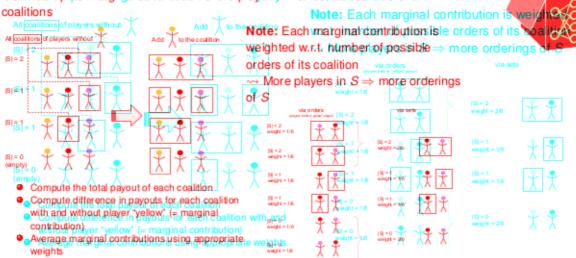
- Compute difference in payouts for each coalition with and without player, yellow, to marginal altion with and
 - contribution ver "yellow" (= marginal contribution)
- Average marginal contributions using appropriate e weights weights



COOPERATIVE GAMES WITH INTERACTIONS

Question: What is a fair payout for player, yellow ??

Idea: Compute marginal contribution of the player of interest across different coalitions



SHAPLEY VALUE - SET DEFINITION

This idea refers to the Shapley value which assigns a payout value to each player according to its according to its marginal contribution in all possible coalitions.

- ••Let $v(SS|\{j\}) \rightarrow v(S)$ be the marginal contribution of player j to coalition S ion S
 - --- measures how much a player/j increases the value of a coalition Son S

SHAPLEY VALUE - SET DEFINITION

This idea refers to the Shapley value which assigns grayout value to each player according to its according to its marginal contribution in all possible coalitions.

- Létev(SSU(j)))→ v(S) bè the imarginal contribution of player/j to coalition/Sion S
 measures how much a player/jeinéreases the value of a coalition/Son S
- Average marginal contributions for all possible coalitions is S ⊆ P ⊈ {f} \ {j}
 → order of thow players join the localition matters test different weights depending on size of S on size of S

ISHAPLEY VALUE - SET DEFINITION

This idea refers to the Shapley value which assigns a payout value to each player layer according to its according to its marginal contribution in all possible coalitions.

- Létev(SSU(j)))→ v(S) bè the imarginal contribution of player ji to coalition Sion S
 measures how much a player jeinéreases the value of a coalition iSon S
- Average marginal contributions for all possible coalitions is S ⊆ P ⊈ {f} \ {j} \
 ordersofdidw players join the localition imatters test different weights depending on size of S
 ordersofdidw players join the localition imatters test different weights depending on size of S
 ordersofdidw players join the localition imatters test different weights depending on size of S
- Shapley value via set definition (weighting via multinomial coefficient):

$$\phi_{j} = \sum_{S \subseteq P \setminus \{j\}}^{\phi_{j}} \frac{|S|!(|P| - |S| - 1)!}{|S|!(|P| - |S| - 1)!} (v(S \cup \{j\}) - v(S))$$

SHAPLEY VALUE - ORDER DEFINITION

The Shapley-value was introduced as summation lover sets $S \subseteq P \subseteq \{j\}$, but it can be equivalently defined as a summation of all orders of players:

$$\phi_{j} = \frac{1}{|P|!} \sum_{\tau \in \Pi} \left(v(\underbrace{S_{j}^{\tau}}_{P|!} \underbrace{\bigcup_{\tau \in \Pi}}_{P|!}) \left(S_{j}^{v}(S_{j}^{\tau}) \right) \right) - v(S_{j}^{\tau}) \right)$$

• □ Π: All possible orders of players (we have P! in total)otal)

SHAPLEY VALUE - ORDER DEFINITION

The Shapley evalue was introduced as summation over sets $S \subseteq P \subseteq \{f\}$, but it can be equivalently defined as a summation of all orders of players:

$$\phi_{j} = \frac{1}{|P|!} \sum_{\tau \in \Pi} \left(v(\underbrace{S_{j}^{\tau}}_{P|!} \underbrace{S_{j}^{\tau}}_{\tau \in \Pi})) (S_{j}^{\tau}(S_{j}^{\tau})) - v(S_{j}^{\tau}) \right)$$

- □ Π: All possible orders of players (we have P! in total) or all
- Sp. Set of players before player in order $\sigma_{\mathbb{C}} = (\tau_{\mathbb{C}}^{(1)}, \tau_{\mathbb{C}}^{(p)})$, where $\sigma_{\mathbb{C}}^{(1)}$ is *i*-th element element mple: Players 1, 2, 3 $\Rightarrow \Pi = \{(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)\}$
 - \Rightarrow Example of Players 1, 2,31 \Rightarrow 0) and player of interest $j=3\Rightarrow S_i^{\tau}=\{2,1\}$
 - $\Pi = \{(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)\}_{1} \Rightarrow S_{1}^{T} = \{3\}$
 - \rightarrow For order $\sigma = (2, 1, 3)$ and player of interest $j = 3 \Rightarrow S_j^T = \{2, 1\}$
 - \rightarrow For order $\tau = (3, 1, 2)$ and player of interest $j = 1 \Rightarrow S_j^{\tau} = \{3\}$
 - \rightsquigarrow For order $\tau = (3, 1, 2)$ and player of interest $j = 3 \Rightarrow S_j^{\tau} = \emptyset$

SHAPLEY VALUE - ORDER DEFINITION

The Shapleyeyalue was introduced as summation over sets $S \subseteq P \subseteq \{f\}$, but it can be equivalently defined as a summation of all orders of players:

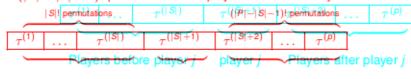
$$\phi_{j} = \frac{1}{|P|!} \sum_{\tau \in \Pi} \left(v(S_{j}^{\tau} \bigcup_{\tau \in \Pi} \{) (S_{j}^{\tau}(S_{j}^{\tau}))\} - v(S_{j}^{\tau}) \right)$$

- □ Π: All possible orders of players (we have P! in total) or all
- Sp. Set of players before player in order $\sigma = (\tau^{(1)}, \tau^{(p)})$, where $\sigma^{(i)}$ is i-th element element mple: Players 1, 2, 3 $\Rightarrow \Pi = \{(-3), (1,3,2), (-3), (2,3,1), (3,1,2), (3,2,1)\}$
 - \Rightarrow Example: Players 1, 23 \Rightarrow and player of interest $i = 3 \Rightarrow S_i^T = \{2, 1\}$
 - $\Pi = \{(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)\} \Rightarrow S^{T} = \{3\}$
 - \rightarrow For order $\sigma = (2,1,3)$ and player of interest $j = 3 \Rightarrow S_i^T = \{2,1\}$
- For order $\tau = (3, 1, 2)$ and player of interest $j = 1 \Rightarrow S^T = \{3\}$ For order $\tau = (3, 1, 2)$ and player of interest $j = 3 \Rightarrow S^T = \emptyset$

- is summed twice
- Order definition: Marginal contribution of orders that yield set S = {1,2} is summed twice
 - \leadsto In set definition, it has the weight $\frac{2!(3-2-1)!}{3!}=\frac{2\cdot 0!}{6}=\frac{2}{6}$

SHAPLEY VALUE - COMMENTS ON ORDER DEFINITION DEFINITION

- Order and set definition are equivalent
- Order and set definition are equivalent The number of orders which yield the same coalition S is |S|!(|P|-|S|-P). Reason: The number of orders which yield the same coalition S is
- - $|S|!(|P||_{ETe}|S|_{ETe})|S| = (1)! |S| 1)!$ possible orders of players without S and j \Rightarrow There are |S|! possible orders of players within coalition S
 - \Rightarrow There are (|P| |S| 1)! possible orders of players without S and



Players before player j

player j

Players after player j



SHAPLEY/VALUE - COMMENTS: ON ORDER DEFINITION DEFINITION

- Order and set definition are equivalent
- Order and set definition are equivalent heason set definition are equivalent heason set definition are equivalent.
- Reason: The number of orders which yield the same coalition S is |S||(|P||_v |S| 1)|| ||S|| ||1)|| ||s|| ||1)|| ||s|| ||s|
 - $|S| \cdot (|F| |S| |S| |S| 1)!$ possible orders of players without S and j \Rightarrow There are |S|! possible orders of players within coalition S (|P| |S| 1)! permutation
 - \Rightarrow There are (|P| |S| 1)! possible orders of players without S and j



- Relevance of the order definition: Approximate Shapley values by sampling permutations player;
- Relevance of the order definition: Approximate Shapley values by sampling empermutations
 - --- randomly sample a fixed number of M permutations and average them:

$$\phi_{j} = \frac{1}{M} \sum_{\substack{T \in \Pi_{M} \\ \text{where } \Pi_{M} \subset \Pi \text{ is a rand } M}} \underbrace{(v(S_{j}^{T} \cup \{j\}) - v(S_{j}^{T}))}_{\text{where } \Pi_{M} \subset \Pi \text{ is a rand } M} \underbrace{v(S_{j}^{T} \cup \{j\}) - v(S_{j}^{T}))}_{\text{where } \Pi_{M} \subset \Pi \text{ is a rand } M}$$

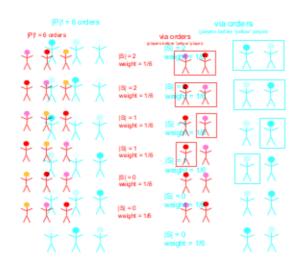
where $\Pi_M \subset \Pi$ is a random subset of Π containing only M orders of players

WEIGHTS FOR MARGINAL CONTRIBUTION - ILLUSTRATION ILLUSTRATION

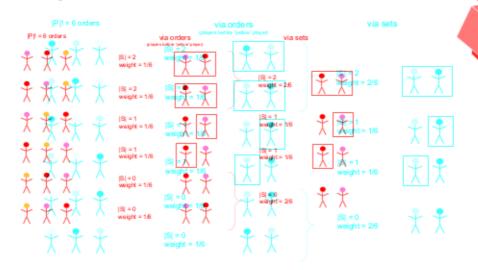




WEIGHTS FOR MARGINAL CONTRIBUTION - ILLUSTRATION ILLUSTRATION



WEIGHTS FOR MARGINAL CONTRIBUTION - ILLUSTRATION ILLUSTRATION

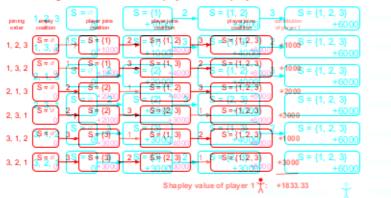


Shaplel value of player fis the marginal contribution to the value when it enters nters any coalition
 any coalition

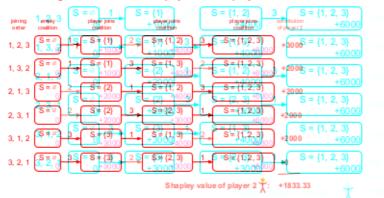
Produce all possible joining orders of player coalitions



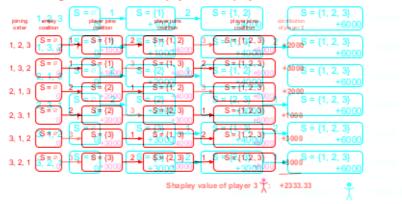
- Shaples value of player fis the marginal contribution to the value when it enters nters any sealing in
- •any coalition possible joining orders of player coalitions
- Measure and average the difference in payout after player 1 enters the coalition
- Measure and average the difference in payout after player 1 enters the coalition



- Shaples value of player f is the marginal contribution to the value when it enters need any prairie.
- •any coalition possible joining orders of player coalitions
- Measure and average the difference in payout after player 2 enters the coalition
- Measure and average the difference in payout after player 2 enters the coalition

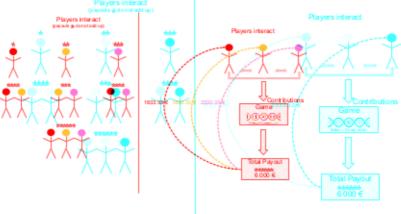


- Shaples value of player fis the marginal contribution to the value when it enters need any position
- •any coalition possible joining orders of player coalitions
- Measure and average the difference in payout after player 3 enters the coalition
- Measure and average the difference in payout after player 3 enters the coalition



Shapley-value of player is the marginal contribution to the value when then ters nters any coalition any coalition possible joining orders of player coalitions

Produce all possible joining orders of player coalitions



Why is this a fair payout solution on?

One possibility to define fair payouts are the following axioms for a given value rate function v:

- Efficiency: Player contributions add up to the total payout of the game: $\sum_{j=1}^{\rho} \phi_j = v$
- **Efficiency**: Player contributions add up to the total payout of the game: $\sum_{i=1}^{p} \phi_i = v(P)$





Why is this a fair payout solution?n?

One possibility to define fair payouts are the following axioms for a given value alue function value. function v:

- Efficiency: Player contributions add up to the total payout of the game: $\sum_{i=1}^{p} \phi_i =$
- Efficiency: Player contributions add up to the total payout of the game:
- Symmetry p ayers $j, k \in P$ who contribute the same to any coalition get the same payout: $(S \cup \{j\}) = v(S \cup \{k\})$ for all $S \subseteq P \setminus \{j, k\}$, then $\phi_i = \phi_k$
- Symmetry: Players $j, k \in P$ who contribute the same to any coalition get the same payout:

If
$$v(S \cup \{j\}) = v(S \cup \{k\})$$
 for all $S \subseteq P \setminus \{j, k\}$, then $\phi_j = \phi_k$

Why is this a fair payout solution?n?

One possibility to define fair payouts are the following axioms for a given value alue function val function v:

- Efficiency: Player contributions add up to the total payout of the game: $\sum_{i=1}^{p} \phi_i = 0$
- Efficiency: Player contributions add up to the total payout of the game:
- Symmetry p ayers $j, k \in P$ who contribute the same to any coalition get the same payout: If $v(S \cup \{i\}) = v(S \cup \{k\})$ for all $S \subseteq P \setminus \{i, k\}$, then $\phi_i = \phi_k$
- Symmetry: Players $j, k \in P$ who contribute the same to any coalition get the
- •same payout II Player: Payout is 0 for players who don't contribute to the value of any coalition:

If
$$V(S \cup \{j\}) \Rightarrow \overline{V}(S \cup \{k\})$$
 for all $S \subseteq P \setminus \{j, k\}$, then $\phi_j = \phi_k$

Dummy/Null Player: Payout is 0 for players who don't contribute to the value of any coalition:

If
$$v(S \cup \{j\}) = v(S) \quad \forall \quad S \subseteq P \setminus \{j\}$$
, then $\phi_j = 0$

Why is this a fair payout so lution?n?

One possibility to define fair payouts are the following axioms for a giveryvalue/alue function ν function ν :

- Efficiency: Player contributions add up to the total payout of the game: $\sum_{j=1}^{p} \phi_j = 0$
- Efficiency: Player contributions add up to the total payout of the game:
 - **Symmetry**(\not b) layers $j, k \in P$ who contribute the same to any coalition get the same payout: If $v(S \cup \{j\}) = v(S \cup \{k\})$ for all $S \subseteq P \setminus \{j, k\}$, then $\phi_i = \phi_k$
- Symmetry: Players $j, k \in P$ who contribute the same to any coalition get the
- •same payout II Player: Payout is 0 for players who don't contribute to the value of any coalition: If $|V(S \cup \{j\})| \Rightarrow \overline{V(S \cup \{j\})}$ for all $S \subseteq R \setminus \{j \mid k\}$, then $\phi_j = \phi_k$
- **Dummy/Null Player:** Payout is to complayers who don't contribute to the value of a payout is the sum any coalition: $\phi_{j,v} = \phi_{j,v_1} + \phi_{j,v_2}$ If $v(S \cup \{j\}) = v(S) \quad \forall \quad S \subseteq P \setminus \{j\}$, then $\phi_j = 0$
- **Additivity**: For a game v with combined payouts $v(S) = v_1(S) + v_2(S)$, the payout is the sum of payouts: $\phi_{i,v} = \phi_{i,v} + \phi_{i,v_2}$