Feature Attribution

Lecture 20

Given an AI prediction model $f : \mathbb{R}^d \to \mathbb{R}$ that takes points in a high-dimensional space, $\mathbb{R}^d$, and returns a 'prediction' in $\mathbb{R}$, we want to quantify the relevance of each of the d dimensions for the prediction of f. We will explore one solution to this problem based on cooperative game theory.

1 Cooperative Games

Definition 1 (Cooperative Games)**.** A *cooperative game* is a pair (N, v), where

- $N = \{1, \ldots, n\}$ is the set of players, and
- $v : 2^N \to \mathbb{R}_{\geq 0}$ is the *value function* that assigns a non-negative, real value $v(S)$ to every coalition $S \subseteq N$. We assume that $v(\emptyset) = 0$.

The value function $v(S)$ specifies the total value that a set S of players obtains if they 'work together' as a group and ignore all other players. The members of the group can divide the value between them as they wish. Our goal is to find the 'best'[1] partition of the n players into disjoint coalitions $S_1, ..., S_k$, together with a distribution of $v(S_i)$ over the members of S_i for every coalition.

Cooperative games model situations where players can form binding agreements to cooperate to generate value that they then distribute among them. Unlike the games we've studied so far, the focus here is on the formation of coalitions and fair allocation of the resulting value, rather than strategic behavior. We'll now see some examples of cooperative games.

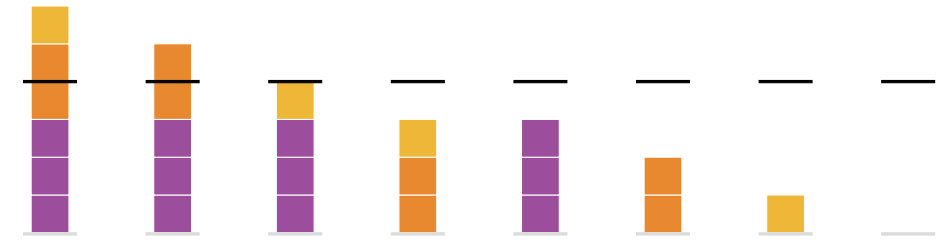
Definition 2 (Weighted Voting Games)**.** In a *weighted voting game*, given is a weight $w_i \in N$ for each player $i \in N$ and a *threshold* $q > 1/2 \cdot \sum_{i \in N} w_i$. The value function for any coalition $S \subseteq N$ is

$$v(S) = \begin{cases} 1 & \text{if } \sum_{i \in S} w_i \geq q, \\ 0 & \text{else,} \end{cases}$$

so 1 if and only if the total weight of the members of S meets the threshold q.

Weighted voting games can be used to model a bill in a multi-party parliament. Suppose there are n parties in a parliament that have $w_1, ..., w_n$ seats, respectively. To pass a bill, a coalition of parties must collectively hold at least a certain number q of seats in the parliament.

Example 1 (Weighted Voting Game)**.** Consider a parliament with three parties and six seats, where three seats are held by party 1 (purple), two are held by party 2 (orange), and one is held by party 3 (yellow). Let's assume a 2/3-majority is needed to pass a bill, so at least $q = 4$ of the members need to vote in favor.



$S =$	$\{1,2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\emptyset$
$v(S) =$	1	1	1	0	0	0	0	0

[1] As usual, 'best' here may have many meanings like stable, fair, or welfare-maximizing.

Assuming parties always vote as a block, the outcomes of the election shown above are possible (where the total weight of the coalition of parties voting in favor is shown). There are three coalitions, $\{1,2,3\}$, $\{1,2\}$, and $\{1,3\}$, that reach the threshold and thus can achieve value 1. All other coalitions don't meet the threshold, so have value 0.

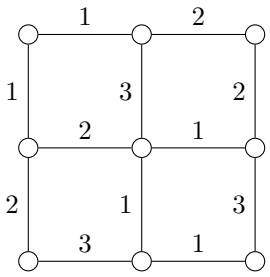
Real-world examples of weighted voting games include the electoral college, multi-party parliaments, the International Monetary Fund, the European Union, and the United Nations Security Council.

Definition 3 (Induced Subgraph Games)**.** In an *induced subgraph game*, given is a weighted, undirected graph $G = (V, E, w)$. Each vertex corresponds to one player, $N = V$, and the value of a coalition S is

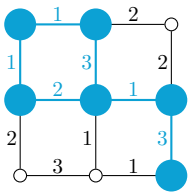
$$v(S) = \sum_{(i,j)\in E: i,j\in S} w(i,j),$$

the total weight of the edges between players in S (i.e., in the *induced subgraph* on S).

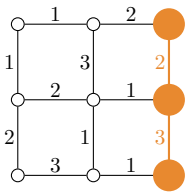
Example 2 (Induced Subgraph Games)**.** Consider the following weighted graph.



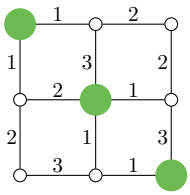
Four possible coalitions are shown below as colored vertices. For example, the value of the blue coalition is the total weight of edges between two blue vertices, so $1+1+2+3+1+3 = 11$. We can similarly calculate the value of the other three coalitions.



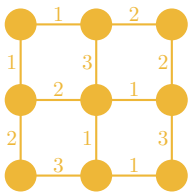
$v(S) = 11$



$v(S) = 5$



$v(S) = 0$



$v(S) = 22$

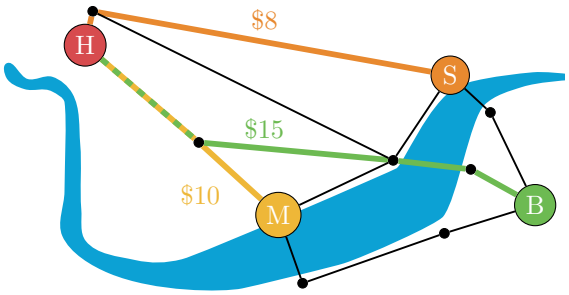
Definition 4 (Taxi Fare Games)**.** A group N of players needs to travel from a common source x to their respective destinations, y_i for $i \in N$. The value of any coalition S is

$$v(S) = \max\{0, \sum_{i\in S} c(x, y_i) - c(x, S)\},$$

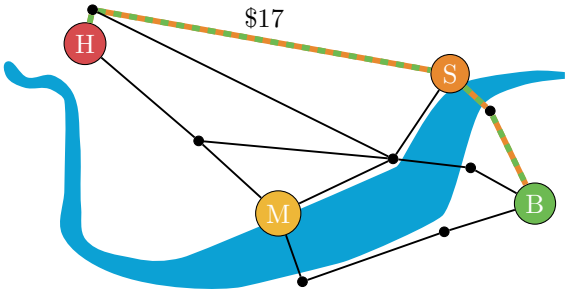
where for any set of players $S' \subseteq N$, $c(x, S')$ is the cheapest taxi ride from x to all destinations y_i of players $i \in S'$.

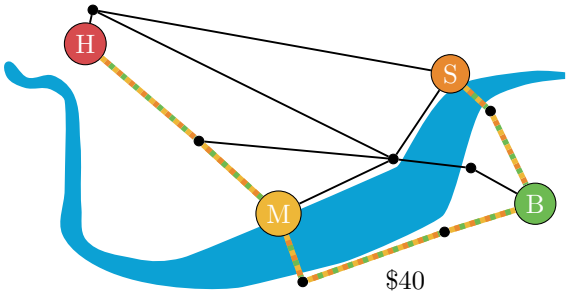
Intuitively, $\sum_{i\in S} c(x, y_i)$ is to the total cost if all players took separate taxis, while $c(x, S)$ is the cost if all players in S took a taxi together. Thus, if it is cheaper for all players in S to take a taxi together, $v(S)$ are the savings, while otherwise $v(S) = 0$.

Example 3 (Taxi Fare Games)**.** Assume there are three students leaving from Harvard Square ($x = H$) that want to go to MIT ($y_1 = M$), Boston Common ($y_2 = B$), and the Museum of Science ($y_3 = S$). They would ideally like to share a taxi, but it is unclear how they should optimally group and split the payments. If they each take a taxi by themselves, the prices are \$10, \$15, and \$8, respectively.



If the student going to the Museum of Science and the student going to Boston Common ride together, their ride costs, say, \$17. The value of this coalition thus is $v(\{2,3\}) = \max\{0, 15+8-17\} = 6$, since these two students could save \$6 by riding together. If all three students ride together, their ride costs, say, \$40, because of the large detour needed. The value of this coalition thus is $v(\{1,2,3\}) = \max\{0, 10+15+8-40\} = 0$, since the students don't have any savings over riding by themselves.





2 Properties of Cooperative Games

We'll introduce two properties of cooperative games that will prove useful later.

Definition 5 (Superadditivity)**.** A cooperative game is *superadditive* if for every pair of disjoint coalitions $S, T \subseteq N$, it holds that $v(S \cup T) \geq v(S) + v(T)$.

Let's consider which of the cooperative games we have seen so far are superadditive. First, weighted voting games are superadditive. If the union of two coalitions meets the threshold, so has value 1, then at most one of the individual coalitions can meet the threshold since $q > 1/2 \cdot \sum_{i \in N} w_i$. If the union of two coalitions does not meet the threshold, so has value 0, then neither of the individual coalitions can meet the threshold.

Induced subgraph games are also superadditive. If we combine two coalitions, all of the edges that were included in either of the coalitions before now count to the value and each of those edges counted to at most one value of the two individual coalitions, so the total value cannot decrease from combining two disjoint coalitions.

Finally, taxi fare games are not superadditive. Assume, for example, that we have two people going east and one person going west. If the two people going east form a coalition, they have a positive value since they are saving by commuting together. However, if the person going west joins the coalition, the trip becomes a lot more expensive because the taxi needs to go all the way to the east before heading all the way west, giving this coalition value 0. Thus, combining two coalitions can decrease their total value.

Definition 6 (Supermodularity)**.** A cooperative game is *supermodular* (or *convex*) if for all coalitions $S, T \subseteq N$ with $S \subseteq T$ and player $i \in N \setminus T$, it holds that $v(S \cup \{i\}) - v(S) \leq v(T \cup \{i\}) - v(T)$.

This definition is the same as *sub*modularity from the lecture on influence maximization on cascade networks, with the direction of the inequality flipped. A cooperative game is supermodular if the marginal benefit from a player joining the coalition can only *increase* if the coalition gets larger.

One can verify that if a cooperative game is supermodular, it is also superadditive. Thus, we already know that taxi fare games are not supermodular. Let's examine whether the other two types of games are.

First, weighted voting games are not supermodular. Consider S being a coalition with weight just below the threshold such that i adds enough weight to meet the threshold, while T is a larger coalition with enough weight to meet the threshold even without i. Thus, the gain in value from adding i to S is 1, strictly greater than the gain of 0 from adding i to T.

In contrast, induced subgraph games are supermodular. All edge weights that get added to the value from combining i with S also get added when we combine i with T, so the increase in value from adding i to T is at least as big as the increase in value from adding i to S.

3 The Shapley Value

If a cooperative game is superadditive or supermodular, two coalitions of players merging into a new coalition can only increase the total value. Thus, it is rational for the *grand* coalition $S = N$ to form, meaning that all players will collaborate. Therefore, in superadditive games it suffices to focus on how to divide the payoffs of the grand coalition, and we no longer need to worry about which coalition structures will form.

Definition 7 (Payoff Division)**.** Given a cooperative game (N, v), a *payoff division* is a vector $\boldsymbol{p} \in (\mathbb{R}_{\geq 0})^n$ such that $\sum_{i \in N} p_i = v(N)$. We say that p_i is the payoff of player i.

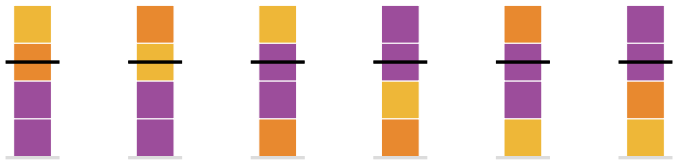
The definition of a payoff division assumes that the grand coalition has formed. We'll now see one approach to constructing a reasonable payoff division that satisfies a strong notion of stability.

Definition 8 (Shapley Value)**.** Given a permutation π over N, let S^i_π be the coalition that consists of the players before i in the permutation π. The *Shapley value* of player i in a cooperative game (N, v) is

$$\sigma_i(N, v) = \frac{1}{n!} \sum_{\pi} [v(S^i_\pi \cup \{i\}) - v(S^i_\pi)].$$

In other words, the Shapley value is the average (over all possible permutations π of N) value that player i adds to the coalition of players that preceded i in π.[2]

Example 4 (Shapley Value)**.** Consider a weighted voting game with $n = 3$ parties with weights $w_1 = 2$ (purple), $w_2 = 1$ (orange), $w_3 = 1$ (yellow) and quota $q = 2.5$. There are six possible permutations over the voters. Let's check for each of them what the change in value due to player 1 joining the coalition of players preceding them is. Since the value function in weighted voting games is 1 if and only if the coalition meets the threshold, this is equivalent to checking for which permutations the coalition of players preceding player 1 don't meet the threshold but do once player 1 joins them.



$\pi =$	$(1,2,3)$	$(1,3,2)$	$(2,1,3)$	$(2,3,1)$	$(3,1,2)$	$(3,2,1)$
$v(S^1_\pi \cup \{i\}) - v(S^1_\pi) =$	0	0	1	1	1	1

Thus, the Shapley value of player 1 is $\sigma_1(N, v) = 1/6(0 + 0 + 1 + 1 + 1 + 1) = 2/3$. Similarly, we can find that the Shapley value of players 2 and 3 is $\sigma_2(N, v) = \sigma_3(N, v) = 1/6$. Thus, $\sigma_1(N, v) + \sigma_2(N, v) + \sigma_3(N, v) = 1 = v(N)$, so $\boldsymbol{p} = \boldsymbol{\sigma}(N, v)$ is a valid payoff division.

We'll now prove that the fact that $\boldsymbol{\sigma}(N, v)$ is a payoff division in the example above is not a coincidence.

[2] It may seem arbitrary to average the value a player adds over coalitions induced by a uniformly random *permutation* instead of just averaging over all possible coalitions $S \subseteq N$. Intuitively, the reason for this is that averaging over permutations leads to the size of the coalition (with respect to which we measure the impact of adding player i) being uniformly distributed on $\{0, ..., n-1\}$, while averaging over all possible coalitions puts almost all 'averaging weight' on coalitions of size $n/2 \pm \Theta(\sqrt{n})$ while essentially ignoring very small or very large coalitions.

Theorem 1. *The vector $\boldsymbol{\sigma}(N,v)$ is a payoff division for the cooperative game (N,v).*

Proof. We need to show that $\sum_{i\in N}\sigma_i(N,v) = v(N)$. Plugging in the definition of $\sigma_i(N,v)$ and switching the order of summation, to obtain a telescoping sum that we simplify, to get

$$\begin{aligned}
\sum_{i\in N}\sigma_i(N,v) &= \frac{1}{n!}\sum_{\pi}\sum_{i\in N}[v(S^i_\pi\cup\{i\}) - v(S^i_\pi)] \\
&= \frac{1}{n!}\sum_{\pi}\left([v(S^n_\pi) - v(S^{n-1}_\pi)] + \ldots + [v(S^1_\pi) - v(S^0_\pi)]\right) \\
&= \frac{1}{n!}\sum_{\pi}\left(v(S^n_\pi) - v(S^0_\pi)\right) \\
&= \frac{1}{n!}\sum_{\pi}(v(N) - v(\emptyset)) \\
&= v(N)
\end{aligned}$$

□

One avenue to argue that the Shapley value is a reasonable payoff division is by showing that it is the unique payoff division that satisfies a handful of, arguably, reasonable axioms.

Definition 9 (Properties of Payoff Divisions)**.** A *payoff division rule* $\boldsymbol{\phi}$ is a function that assigns to each cooperative game (N,v) a payoff division $\boldsymbol{\phi}(N,v)$.

- $\boldsymbol{\phi}$ is *symmetric* if for any $i,j\in N$ such that for all $S\subseteq N\setminus\{i,j\}$, $v(S\cup\{i\}) = v(S\cup\{j\})$, it holds that $\phi_i(N,v) = \phi_j(N,v)$. In other words, if two players can be exchanged in any coalition without changing the value (i.e., they are essentially identical), then they should get the same payoff.
- $\boldsymbol{\phi}$ satisfies the *null player property* if for any $i\in N$ such that for all $S\subseteq N\setminus\{i\}$, $v(S\cup\{i\}) = v(S)$ it holds that $\phi_i(N,v) = 0$. In other words, if a player never changes the value of a coalition, their payoff should be 0.
- $\boldsymbol{\phi}$ satisfies *additivity* if for any two value functions $v_1, v_2 : 2^N\to\mathbb{R}_{\geq 0}$, it holds that $\boldsymbol{\phi}(N,v_1+v_2) = \boldsymbol{\phi}(N,v_1)+\boldsymbol{\phi}(N,v_2)$, where the value function v_1+v_2 is defined as $(v_1+v_2)(S) = v_1(S)+v_2(S)$. In other words, if we take two cooperative games over the same set of players and combine them by summing up the value functions, then the Shapley values should be the sum of the Shapley values in the original games.

Theorem 2. *The Shapley value is the unique payoff division rule that satisfies symmetry, the null player property, and additivity.*

It is not hard to argue that symmetry and the null player property are reasonable properties to impose. On the other hand, additivity is a bit harder to argue for. While it certainly is a nice property, one could conceive a reasonable payoff division rule that does not satisfy additivity. However, we'll now see a second avenue to argue in favor of the Shapley value: It satisfies a strong notion of stability in a large class of cooperative games.

4 The Core

So far, we have not defined what form of stability or equilibrium we are hoping for. We now do so.

Definition 10 (Core)**.** The *core* of a cooperative game (N,v) is the set of all payoff divisions $\boldsymbol{p}$ such that for all coalitions $S\subseteq N$, it holds that

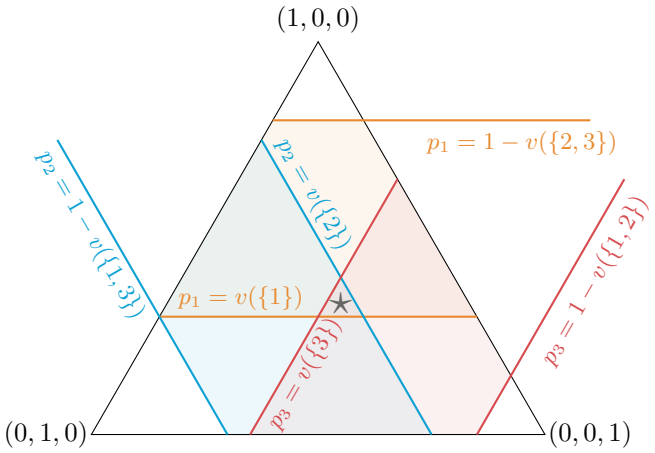
$$\sum_{i\in S} p_i \geq v(S).$$

In other words, if a payoff division $\boldsymbol{p}$ is in the core, no coalition S can deviate to work together (and ignore the other players) to get a value $v(S)$ larger than their total payment under $\boldsymbol{p}$. In particular, this is the case if and only if the value $v(S)$ is large enough so that S can distribute it among its members in a way that

every player i in S gets more than p_i; thus, $\boldsymbol{p}$ is in the core unless a coalition S can deviate to make *all* players strictly better off. A payoff division in the core ensures that every coalition receives at least as much as it could obtain on its own.

Example 5 (Core)**.** The ternary plot below shows the possible payoff divisions for a given v for $n = 3$ players with $v(N) = 1$. For each $S \subseteq N$, $S \neq \emptyset, N$, we get a restriction on the parts of the plot in which the payoff division has to lie to be in the core. For example, consider the coalition $S = \{1\}$. This coalition can by themselves achieve a value of $v(\{1\})$, so any payoff division in the core needs to have $p_1 \geq v(\{1\})$, so lie on or above the orange line labeled with $p_1 \geq v(\{1\})$ in the ternary plot. As another example, consider the coalition $S = \{2, 3\}$. This coalition can by themselves achieve a value of $v(\{2, 3\})$, so any payoff division in the core needs to have $p_2 + p_3 \geq v(\{2, 3\})$, or equivalently, $p_1 \leq 1 - v(\{2, 3\})$, so lie on or below the orange line labeled with $p_1 \leq 1 - v(\{2, 3\})$ in the ternary plot.

A payoff division is in the core if and only if it is in the small area at the center marked with $\star$, that lies on the 'right' side of all six lines (i.e., satisfies the core condition for all S).



Unfortunately, the core may be *empty*, meaning that there exists no payoff division in the core. This corresponds to the area marked with $\star$ above not existing. For example, in any non-superadditive cooperative game in which there exists a partition of the players N into sets $S_1, ..., S_k$ so that $v(S_1) + ... + v(S_k) > v(N)$, the core is empty since the grand coalition N is suboptimal. Furthermore, the core can also be empty in superadditive games.

Example 6 (Empty Core in Weighted Voting Games)**.** Consider a weighted voting game with $n = 3$ players where $w_1 = w_2 = w_3 = 1$ and the threshold is $q = 1.5$. Thus, any coalition of two players has enough weight to meet the threshold, so $v(S) = 1$ for any S with $|S| \geq 2$. Consider any payoff division p and suppose w.l.o.g. that $p_1 > 0$. Then, the total payoff to players 2 and 3 is $p_2 + p_3 = 1 - p_1 < 1 = v(\{2, 3\})$; so p is not in the core. Players 2 and 3 have an incentive to form their own coalition. Therefore, the core is empty, so there are no payoff divisions that satisfy the core constraints for this weighted voting game.

However, for supermodular games, the core exists, and we already developed the tools to find a payoff division in the core:

Theorem 3. *In any supermodular cooperative game, the core is nonempty and contains the Shapley value.*

There is one additional obstacle to finding outcomes in the core via the Shapley value in practice: Computing the Shapley is computationally hard (as we may expect, since it is defined as a sum of $n!$ elements). You will alleviate this concern on the homework assignment by showing that we can estimate the Shapley value very well by sampling random permutations π.

In games that are not supermodular and thus may have an empty core, there exists a relaxed notion of stability that is always satisfiable:

Definition 11 (Least Core)**.** The *least core* is the set of payoff divisions $\boldsymbol{p}$ that achieve the optimal (minimal) ε in the (exponential-size) linear program

$$\begin{aligned} \min \quad & \epsilon \\ \text{s.t.} \quad & \forall S \subseteq N : \sum_{i \in S} p_i \geq v(S) - \epsilon \\ & \sum_{i \in N} p_i = v(N) \\ & \forall i \in N : p_i \geq 0 \\ & \epsilon \geq 0 \end{aligned}$$

In other words, we find the smallest $\varepsilon \geq 0$ so that there exists some payoff division $\boldsymbol{p}$ that satisfies the relaxed core constraints

$$\sum_{i \in S} p_i \geq v(S) - \varepsilon$$

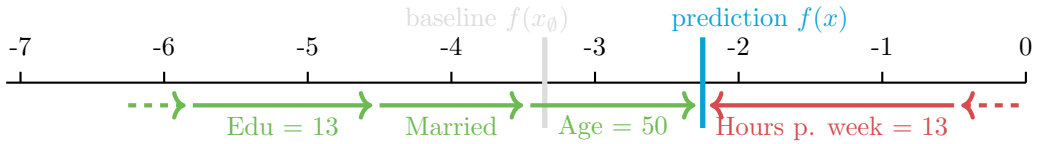
for all $S \subseteq N$. The least core is all payoff divisions that satisfy all relaxed core constraints for this smallest possible ε.

5 Feature Attribution

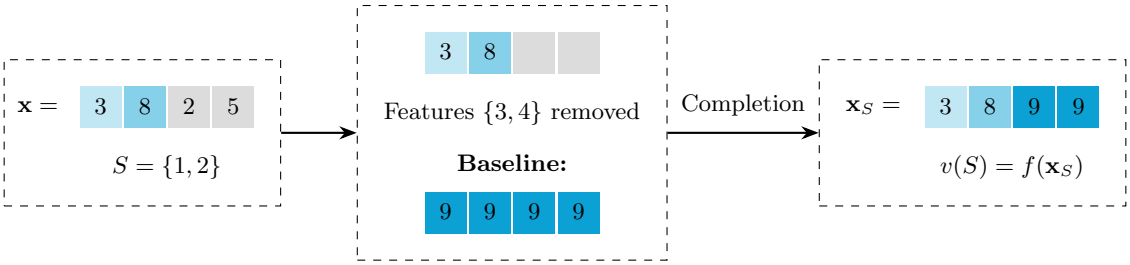
An important subfield of research on AI is *interpretability* that seeks to understand how the 'black-box' AI models arrive at their predictions. One relevant question in the field is determining the importance of individual features in the input on the output. In particular, given an AI model $f : \mathbb{R}^d \to \mathbb{R}$ and a point $\boldsymbol{x} \in \mathbb{R}^d$, what is the influence of each of the d features in $\boldsymbol{x}$ on the model's prediction $f(\boldsymbol{x})$?

The Shapley value is a popular tool to give answers to this question. In particular, for the model f and a data point $\boldsymbol{x}$, we can define the following cooperative game: Let the players N be the d features. Let $v(S) = f(\boldsymbol{x}_S)$, where $\boldsymbol{x}_S$ is $\boldsymbol{x}$ with the features in $N \setminus S$ 'removed' (we'll return to what exactly 'removed' means later). By definition, $v(S)$ is the value of the game with only players in S; correspondingly, $f(\boldsymbol{x}_S)$ is the prediction with only features in S. The Shapley value captures how much a player $i \in N$ influences the value of the game; correspondingly, the Shapley value here captures how much a feature $i \in N$ influences the prediction. For the prediction $f(\boldsymbol{x})$ and the baseline $f(\boldsymbol{x}_\emptyset)$ (i.e., on $\boldsymbol{x}$ with all information removed), the Shapley values, as we have seen, sum up to $\sum_{i \in N} \sigma_i(N, v) = v(N) - v(\emptyset) = f(\boldsymbol{x}) - f(\boldsymbol{x}_\emptyset)$. Thus, we can explain the difference from the prediction to the baseline exactly with the Shapley values of each feature, as shown in the example below.

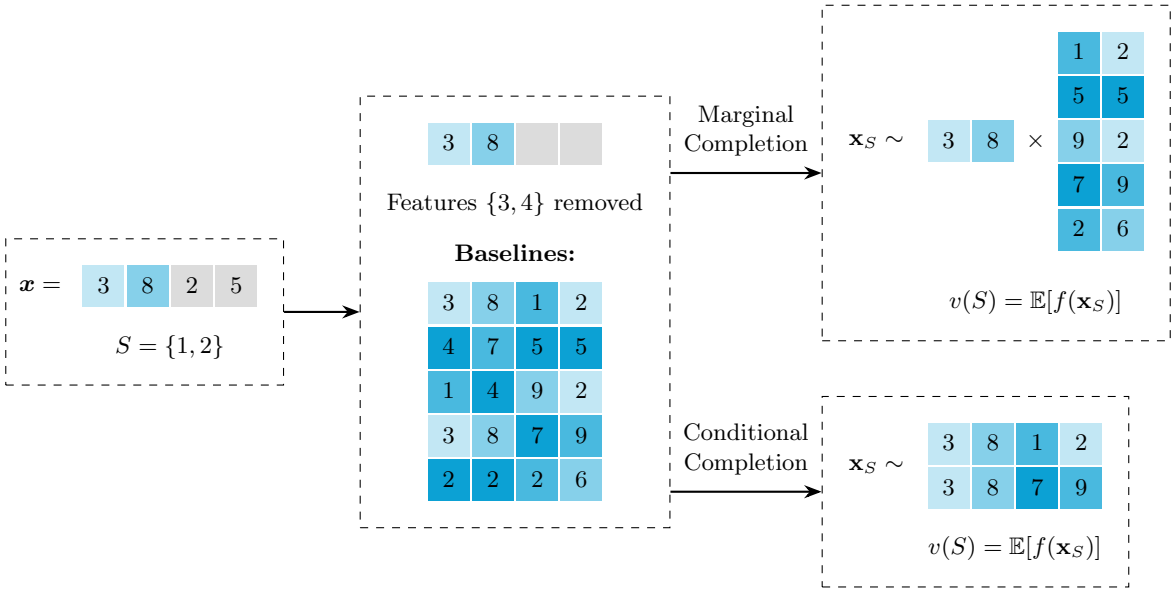
In their 2022 paper, Hugh Chen, Scott M. Lundberg and Su-In Lee apply this to the following setting: Assume we have a model f trained to predict whether individuals have income greater than \$50k solely based on their U.S. census data. Thus, the input features might include age, education level, marital status, and hours worked per week. The model outputs $f(\boldsymbol{x})$ to predict the probability that the income is greater than \$50k to be $e^{f(\boldsymbol{x})}$. For each feature, the Shapley value gives us its contribution to $f(\boldsymbol{x})$, as shown in the following (made up) example:



We now revisit what it actually means to 'remove' features to get $f(x_S)$. There are several reasonable ways to do this. First, we can set a *baseline* value for each feature. If we 'remove' a feature, we simply replace it with its baseline value:



Alternatively, we can use entries from the dataset to replace the 'removed' features with. We can do this *marginally* by sampling a random entry from the dataset to replace the missing entries with, agnostic of the not removed entries of x. We can also do this *conditionally*, by sampling an entry from the rows in the dataset where the not removed features match the ones in x. This respects correlations in the dataset more, but may have no matches and be computationally expensive.



Each of these methods defines a slightly different cooperative game and Shapley values. There is ongoing discussion about which approach is the most reasonable in which practical scenario.