Shapley values

Aaditya Ramdas

Carnegie Mellon University

Thanks, Wikipedia

Incentive!

Now all of you know enough basic ML to start contributing material to the class.

At the start of each class, any student may choose to present anything for 4-5mins to the class.

For eg: prepare 2-3 slides on something interesting (1hr prep time?):
a) a "better" variant of an algorithm taught in class
b) a simulation that is "revealing"
c) behavior on real data that is "funky"
d) a theorem that is "insightful"

Before each class, a student may email me to "volunteer".
Each class will have at most one student (first come first serve).
Each student can volunteer at most once.

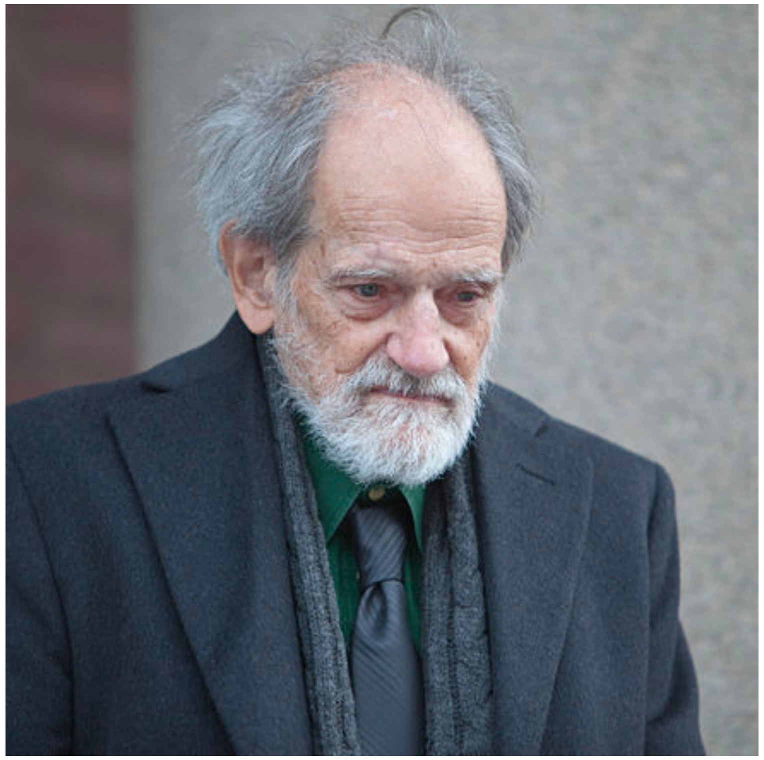
Up to +2 points on any "quiz" (judged on conciseness, timing, relevance, quality of presentation, etc).
Equivalently, "grade bump" (just below cutoff to just above cutoff).

Outline

1. *Shapley values (2/3 class)*

2. *Variable and datapoint importance (1/3 class)*

Cooperative game theory



Theory introduced: 1951, Nobel prize: 2012

Problem setup

N players cooperate to produce something of "value".
Value function $v : 2^N \to \mathbb{R}, v(\varnothing) = 0$.
v(N) is the (dollar, say) value actually obtained.
v(S) is the (hypothetical, assumed known) value obtained when subset S work together

How should v(N) be split up amongst the N players?
i.e. how much money $\phi_i(v)$ should player i get?

How about everybody gets 0 dollars?

How about $\phi_1(v) = v(N), \phi_j(v) = 0$ for $j > 1$?

How about $\phi_i(v) = v(N)/N$?

Desiderata (“axioms”)

“Efficiency”

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

“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 $\varphi_i(v) = \varphi_j(v)$.

This property is also called *equal treatment of equals*.

Desiderata ("axioms")

"Efficiency"

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

"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 $\varphi_i(v) = \varphi_j(v)$.

This property is also called *equal treatment of equals*.

"Linearity"

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

for every i in N. Also, for any real number a,

$$\varphi_i(av) = a\varphi_i(v)$$

for every i in N.

"Null player" (freeloader)

Player i is "null" if $\forall S$ with $i \notin S, v(S \cup i) = v(S)$

$\varphi_i(v)$ of a null player i in a game v is zero.

The Shapley value

There is a unique function satisfying all four axioms.

$$\boxed{\varphi_i(v) = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} \left(v(S \cup \{i\}) - v(S)\right)}$$

$$\varphi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

The Shapley value

There is a unique function satisfying all four axioms.

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

$$\varphi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions excluding } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

An alternative equivalent formula for the Shapley value is:

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

where the sum ranges over all $n!$ orders R of the players

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

Eg: the "business game"

Owner provides initial capital, workspace, vision, etc.

Each worker provides additional profit of p.

The value function for this coalitional game is

$$v(S) = \begin{cases} mp, & \text{if } o \in S \\ 0, & \text{otherwise} \end{cases}$$

where m is the cardinality of $S \setminus \{o\}$.

Total profit: $(N-1)p$

$\phi_o(v) = (N-1)p/2, \quad \phi_w(v) = p/2$

Eg: the “glove game”

Player 1 and 2 have left-hand gloves, player 3 has right-hand glove

A coalition has value one if they have a complete pair, else zero

The value function for this coalitional game is

$$v(S) = \begin{cases} 1 & \text{if } S \in \{\{1,3\},\{2,3\},\{1,2,3\}\}\,; \\ 0 & \text{otherwise.} \end{cases}$$

$$\boxed{\phi_1(v) = \phi_2(v) = 1/6, \quad \phi_3(v) = 2/3}$$

More properties

"Negative externality"

If *v* is a subadditive set function, i.e., $v(S \sqcup T) \leq v(S) + v(T)$,

then for each agent *i*: $\varphi_i(v) \leq v(\{i\})$.

"Positive externality"

if *v* is a superadditive set function, i.e., $v(S \sqcup T) \geq v(S) + v(T)$,

then for each agent *i*: $\varphi_i(v) \geq v(\{i\})$.

More properties

“Negative externality”

If v is a subadditive set function, i.e., $v(S \sqcup T) \leq v(S) + v(T)$,

then for each agent i: $\varphi_i(v) \leq v(\{i\})$.

“Positive externality”

if v is a superadditive set function, i.e., $v(S \sqcup T) \geq v(S) + v(T)$,

then for each agent i: $\varphi_i(v) \geq v(\{i\})$.

“Anonymity”

Relabeling the indices of the players leaves their Shapley value unchanged.

“Coalitions”

$$\varphi_C(v) = \sum_{T \subseteq N \setminus C} \frac{(n - |T| - |C|)!\ |T|!}{(n - |C| + 1)!} \sum_{S \subseteq C} (-1)^{|C|-|S|} v(S \cup T) .$$

Variable importance

How would you define v(S)?

Datapoint importance

How would you define v(S)?
(regression vs classification)

Breakout rooms

Incentive!

Now all of you know enough basic ML to start contributing material to the class.

At the start of each class, any student may choose to present anything for 4-5mins to the class.

For eg: prepare 2-3 slides on something interesting (1hr prep time?):
a) a "better" variant of an algorithm taught in class
b) a simulation that is "revealing"
c) behavior on real data that is "funky"
d) a theorem that is "insightful"

Before each class, a student may email me to "volunteer".
Each class will have at most one student (first come first serve).
Each student can volunteer at most once.

Up to +2 points on any "quiz" (judged on conciseness, timing, relevance, quality of presentation, etc).
Equivalently, "grade bump" (just below cutoff to just above cutoff).