Predictive and Causal Implications of using Shapley Value for Model Interpretation

Sisi Ma, Roshan Tourani

University of Minnesota

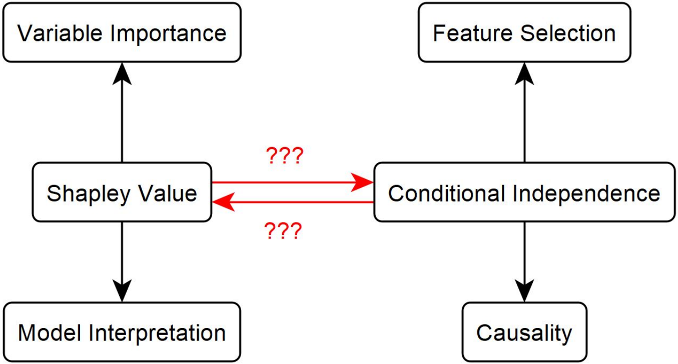
sisima@umn.edu

July 25, 2020

Background

- Complex predictive models can be highly accurate, but difficult to interpret.
- Shapley value based model interpretation tools has gained popularity.

Overview


Variable Importance
Feature Selection
???
Shapley Value
Conditional Independence
???
Model Interpretation
Causality

Shapley Value: Definition

The Shapley value is a solution concept in game theory [Shapley, 1953]. The Shapley value defines the division of total payoff generated by all players to individual players according to their contribution.

Definition

Coalitional game. A coalitional game is a tuple $\langle \mathbf{N}, m \rangle$ where $\mathbf{N} = \{1, 2, \ldots, n\}$ is a finite set of n players, and $m : 2^N \rightarrow \mathbb{R}$ is a characteristic function such that $m(\emptyset) = 0$.

m is a function defined over subsets of $\mathbf{N}$ that describes the value of each subset of $\mathbf{N}$.

Shapley Value: Definition

Definition

Shapley value. For a game $\langle \mathbf{N}, m \rangle$, Shapley value for a player i is defined as:

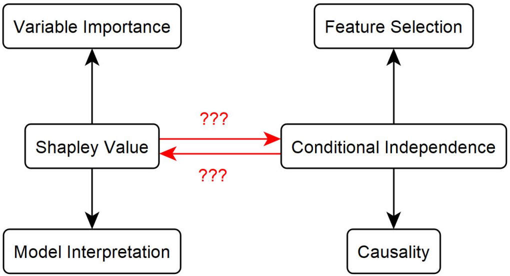
$$\phi_i(m) = \sum_{\mathbf{S} \subseteq \mathbf{N} - \{i\}} \frac{(|\mathbf{N}| - |\mathbf{S}| - 1)!|\mathbf{S}|!}{|\mathbf{N}|!} \left[m(\mathbf{S} \cup \{i\}) - m(\mathbf{S}) \right].$$

Briefly, the Shapley value of a variable i is the weighted sum of the contribution of i in each subset of $\mathbf{N}$. The contribution of i with respect to a subset $\mathbf{S}$ is computed by $m(\mathbf{S} \cup \{i\}) - m(\mathbf{S})$.

Common Ways to use Shapley Value for Predictive Modeling

- For individual observations.
- For collection of observations.

Shapley Value and Conditional Independence



- Shapley value summand for variable i given a subset $\mathbf{S}$ with respect to predicting T is $m(\mathbf{S} \cup \{i\}) - m(\mathbf{S})$;
- The conditional independence between i and T given a subset $\mathbf{S}$ is determined by comparing $p(T|i, \mathbf{S})$ and $p(T|\mathbf{S})$

Shapley Value from the Bayesian Network Perspective

If we constrain function m to be maximized for each $\mathbf{S} \subseteq \mathbf{V}$ only when $p(T|\mathbf{S})$ is estimated accurately, then, we have:

- the conditional independence relationship $p(T|i, \mathbf{S}) = p(T|\mathbf{S})$ corresponds to $m(\mathbf{S} \cup \{i\}) - m(\mathbf{S}) = 0$
- the conditional dependence relationship $p(T|i, \mathbf{S}) \neq p(T|\mathbf{S})$ corresponds to $m(\mathbf{S} \cup \{i\}) - m(\mathbf{S}) > 0$.

Shapley Value from the Bayesian Network Perspective

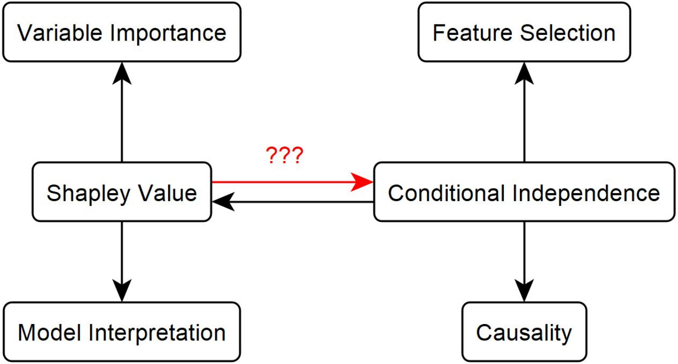
Theorem

If $V_i \perp T | V_j$ and $V_j \not\perp T | V_i$ in a faithful (causal) Bayesian network $\langle \mathbf{V} \cup T, G, p \rangle$, then the Shapley value of V_j is larger than the Shapley value of V_i, i.e. $\phi_j > \phi_i$.

key step of the proof:

$$
\begin{aligned}
&\phi_i - \phi_j \\
= &\sum_{\mathbf{S} \subseteq \mathbf{V} - \{V_i, V_j\}} \left(\frac{(|\mathbf{N}| - |\mathbf{S}| - 1)!|\mathbf{S}|!}{|\mathbf{N}|!} + \frac{(|\mathbf{N}| - |\mathbf{S}| - 2)!(|\mathbf{S}| + 1)!}{|\mathbf{N}|!} \right) \\
&\times \left[m(\mathbf{S} \cup \{V_i\}) - m(\mathbf{S} \cup \{V_j\}) \right],
\end{aligned}
$$

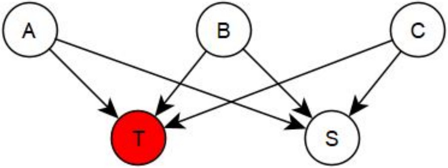
Shapley Value from the Bayesian Network Perspective



Implications of using Shapley Value for Model Interpretation

- Markov Boundary is the optimal feature set for prediction, they are also causal if data is generated from faithful Causal Bayesian Network.
- Questions: How does the Shapley value of Markov Boundary members compared to non-Markov Boundary members of the target.

Implications of using Shapley Value for Model Interpretation



$A \sim \mathcal{N}(0, 2^2)$, $B \sim \mathcal{N}(0, 2^2)$, $C \sim \mathcal{N}(0, 2^2)$
$T = A + B + C + \mathcal{N}(0, 2^2)$
$S = A + B + C + \mathcal{N}(0, 2^2)$
We use a linear regression model and the ordinary R^2 as the $m(\cdot)$ function for Shapley value computation.

Implications of using Shapley Value for Model Interpretation

$$m(\emptyset) = 0$$

$$m(\{A\}) = m(\{B\}) = m(\{C\}) = \frac{1}{4}$$

$$m(\{S\}) = \frac{9}{16}$$

$$m(\{A, B\}) = m(\{B, C\}) = m(\{A, C\}) = \frac{1}{2}$$

$$m(\{A, S\}) = m(\{B, S\}) = m(\{C, S\}) = \frac{7}{12}$$

$$m(\{A, B, S\}) = m(\{A, C, S\}) = m(\{B, C, S\}) = \frac{5}{8}$$

$$m(\{A, B, C\}) = m(\{A, B, C, S\}) = \frac{3}{4}$$

Implications of using Shapley Value for Model Interpretation

Applying Shapley value formular,

$$\phi_i(v) = \sum_{\mathbf{S} \subseteq \mathbf{N} - \{i\}} \frac{(|\mathbf{N}| - |\mathbf{S}| - 1)!|\mathbf{S}|!}{|\mathbf{N}|!} \left[v(\mathbf{S} \cup \{i\}) - v(\mathbf{S}) \right].$$

we have:

$$\phi_A = \phi_B = \phi_C = 95/576 = 0.1649\ldots$$

$$\phi_S = 49/192 = 0.2552\ldots$$

The Markov Boundary members A, B, and C all have smaller Shapley value compared to non-Markov boundary member S.

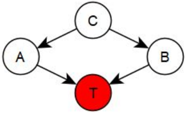
Theorem

Markov boundary members of T can have smaller Shapley value compared to non-Markov boundary members.

Implications of using Shapley Value for Model Interpretation

What about the summation of the Shapley values of all the Markov Boundary members compared to any non-Markov boundary members?

Implications of using Shapley Value for Model Interpretation



$$P(C=1)=P(C=2)=P(C=3)=P(C=4)$$

$$P(A=1|C=1)=0.05; P(B=1|C=1)=0.05$$

$$P(A=1|C=2)=0.05; P(B=1|C=2)=0.95$$

$$P(A=1|C=3)=0.95; P(B=1|C=3)=0.05$$

$$P(A=1|C=4)=0.95; P(B=1|C=4)=0.95$$

$$P(T=1|A=0,B=0)=0.9$$

$$P(T=1|A=0,B=1)=0.05$$

$$P(T=1|A=1,B=0)=0.15$$

$$P(T=1|A=1,B=1)=0.9$$

Implications of using Shapley Value for Model Interpretation

Use the m function:
$\sum_{\mathbf{s}} P(\mathbf{S}=\mathbf{s}) \max(P(T=1|\mathbf{S}=\mathbf{s}), P(T=0|\mathbf{S}=\mathbf{s}))$.

$$m(\emptyset)=0.5; m(\{A\})=m(\{B\})=0.0525; m(\{C\})=0.8235$$

$$m(\{A,B\})=0.9; m(\{A,C\})=m(\{B,C\})=0.8597;$$
$$m(\{A,B,C\})=0.9;$$

and the Shapley values are:

$$\phi_A=\phi_B=0.0903\ldots; \phi_C=0.2194\ldots$$

$(\phi_A+\phi_B)<\phi_C$, i.e. the sum of the Shapley values of all Markov boundary members of T can be smaller than that of a non-Markov boundary member.

Implications of using Shapley Value for Model Interpretation

Use the m function:
$\sum_{\mathbf{s}} P(\mathbf{S}=\mathbf{s}) \max(P(T=1|\mathbf{S}=\mathbf{s}), P(T=0|\mathbf{S}=\mathbf{s}))$.

$$m(\emptyset)=0.5; m(\{A\})=m(\{B\})=0.0525; m(\{C\})=0.8235$$

$$m(\{A,B\})=0.9; m(\{A,C\})=m(\{B,C\})=0.8597;$$
$$m(\{A,B,C\})=0.9;$$

and the Shapley values are:

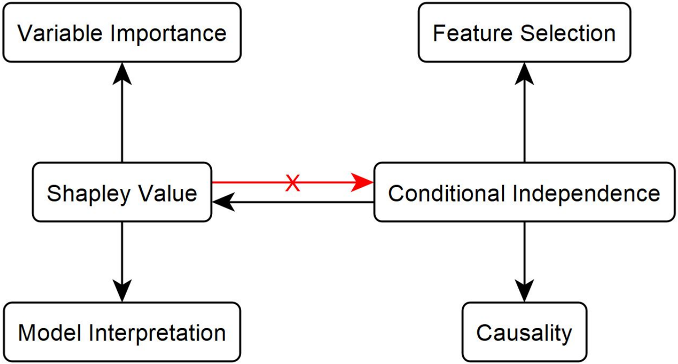
$$\phi_A=\phi_B=0.0903\ldots; \phi_C=0.2194\ldots$$

$(\phi_A+\phi_B)<\phi_C$, i.e. the sum of the Shapley values of all Markov boundary members of T can be smaller than that of a non-Markov boundary member.

Theorem

The sum of the Shapley values of all Markov boundary members of T can be smaller than the Shapley value of a non-Markov boundary member.

Implications of using Shapley Value for Model Interpretation



Why didn't this work out???

Implications of using Shapley Value for Model Interpretation

- Using Shapley value for feature selection do not guarantee obtaining the minimal optimal feature set.
- Magnitude of Shapley value of variables do not necessarily correspond to causality.
- Variables that are in the local causal neighborhood of T do not necessarily have larger Shapley value compared to other variables.

Future Work

- More general theoretical results.
- Empirical analytical experiments.

References I

Owen, A. B. and Prieur, C. (2017).
On shapley value for measuring importance of dependent inputs.
SIAM/ASA Journal on Uncertainty Quantification, 5(1):986–1002.

Shapley, L. S. (1953).
A value for n-person games.
Contributions to the Theory of Games, 2(28):307–317.

Shorrocks, A. F. (1999).
Decomposition procedures for distributional analysis: a unified framework based on the shapley value.
Technical report, mimeo, University of Essex.