Shapley Value

Summary for *A value for n-person games*

Lloyd S. Shapley

Presenter: Zijie Pan

https://qdata.github.io/deep2Read/

2/28/2020

Motivation

- Concept in Game Theory

- To each cooperative game it assigns a unique distribution (among the players) of a total surplus generated by the coalition of all players

Example:

A: Boss B: Engineer C and D are workers

only A： 0

A+B working together: Total monthly profit : $30k

A+B+C working together: Total monthly profit : $60k

A+B+C+D working together: Total monthly profit : $90k

A+C+D (no engineer) : 0

Question : How to distribute $90k fairly.

It seems like the marginal profit for each worker is $30k, they deserve more money

By Shapley (Based on contribution): A:3.5 B:3.5 C:1 D:1

Shapley Value Formula:

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

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

$\phi_i(v)$ is the shapley value for i, which quantify the contribution

v is game function, like v(A+B+C+D) = $90k (total value for the coalition)

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

In previous example: N ={A,B,C,D}

Let's focus on D, so i =D

S are the potential subsets that exclude D, and we need to sum and get average

$$A \quad AB$$

$$\varnothing \quad B \quad BC \quad ABC$$

$$C \quad CA$$

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

The change of value if we add i in:

$$\begin{array}{cccc} & \Delta v_{A,D} & \Delta v_{AB,D} & \\ \Delta v_{\varnothing,D} & \Delta v_{B,D} & \Delta v_{BC,D} & \Delta v_{ABC,D} \\ & \Delta v_{C,D} & \Delta v_{CA,D} & \end{array}$$

$$\begin{array}{cccc} & \frac{1}{3}\Delta v_{A,D} & \frac{1}{3}\Delta v_{AB,D} & \\ 1\Delta v_{\varnothing,D} & \frac{1}{3}\Delta v_{B,D} & \frac{1}{3}\Delta v_{BC,D} & 1\Delta v_{ABC,D} \\ & \frac{1}{3}\Delta v_{C,D} & \frac{1}{3}\Delta v_{CA,D} & \end{array}$$

1 ⅓ ⅓ 1

Axiom of Shapley:

- Symmetry :

For any v, if i and j are interchangeable, then they should receive the same payments or share the same contribution.

$$\phi_i(v) = \phi_j(v)$$

- Linearility

If the game can be separated into two parts, the payments can also be divided

$$\phi[v + w] = \phi[v] + \phi[w]$$

- Efficiency or Completeness

$$\sum_N \phi_i[v] = v(N)$$

- **Dummy player should receive 0 payments**

Connect Shapley with Attribution

Shapley: Assign results based on individual contribution

Attribution: Quantify the feature importance to the result

Shapley -> additive feature attribution methods:

Satisfy: completeness, dummy, implementation invariance

Shapley is used in machine learning interpretability

Using sampling to approximate Shapley estimation for single feature value

Algorithm 1 Approximating the contribution of the i-th feature's value φ_i for instance $y \in \mathcal{X}$ and model f

Require: model f, instance y
Require: number of samples N

1: $\varphi_i \leftarrow 0$
2: **for** $j = 1$ to N **do**
3: choose a random permutation of features $\mathcal{O} \in S_p$
4: choose a random instance $z \in \mathcal{X}$
5: $x' \leftarrow$ **if** $k \in Pre^i(\mathcal{O}) \cup \{i\}$ **then** $x'_k = y_k$ **else** $x'_k = z_k$
6: $x'' \leftarrow$ **if** $k \in Pre^i(\mathcal{O})$ **then** $x''_k = y_k$ **else** $x''_k = z_k$
7: $\varphi_i \leftarrow \varphi_i + f(x') - f(x'')$
8: $\varphi_i \leftarrow \frac{\varphi_i}{N}$

With feature j: $x_{+j} = (x_{(1)}, \ldots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$

Without feature j: $x_{-j} = (x_{(1)}, \ldots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)})$

The Shapley value of a feature value is not the difference of the predicted value after removing the feature from the model training. The interpretation of the Shapley value is: Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction

Other Work

SHAP (SHapley Additive exPlanations), DeepSHAP(deeplift+shapley),KernelSHAP(LIME + shapley), TreeSHAP in *A Unified Approach to Interpreting Model Predictions*