# The monotonicity axiom implies the symmetry axiom for Shapley values

October 31, 2017

## 1 The monotonicity axiom implies the symmetry axiom for Shapley values

The Shapley values are a well known way in coalitional game theory of assigning credit for an outcome among a set of players. Since any arbitrary mapping between player sets and outcomes is possible, how to assign individual credit can be unclear. The Shapley values are one such way of assigning credit and, importantly, they are the only way that satifies some very basic and desirable properties. Typically these properties are given as four axioms, where the Shapley values are the only credit assignment method that satisfies all four axioms:

1. **Efficiency**

$$f(x) = \sum_{i=0}^{M} \phi_i \tag{1}$$

   This assumption forces the model to correctly capture the original predicted value.
2. **Symmetry**. Let $1_S \in \{0,1\}^M$ be an indicator vector equal to 1 for indexes $i \in S$, and 0 elsewhere, and let $f_x(S) = f(h_x^{-1}(1_S))$. If for all subsets $S$ that do not contain $i$ or $j$

$$f_x(S \cup \{i\}) = f_x(S \cup \{j\}) \tag{2}$$

   then $\phi_i(f,x) = \phi_j(f,x)$. This states that if two features contribute equally to the model then their effects must be the same.
3. **Null effects**. If for all subsets $S$ that do not contain $i$

$$f_x(S \cup \{i\}) = f_x(S) \tag{3}$$

   then $\phi_i(f,x) = 0$. A feature ignored by the model must have an effect of 0.
4. **Linearity**. For any two models $f$ and $f'$

$$\phi_i(f + f', x) = \phi_i(f,x) + \phi_i(f',x). \tag{4}$$

   This states that the effect a feature has on the sum of two functions is the effect is has on one function plus the effect it has on the other.

### 1.1 Young showed in 1985 that linearity and null effects can be eliminated using a monotonicity axiom

**Monotonicity** For any two model functions $f$ and $f'$ if for all subsets $S$ of the simplified input features $Z$ that do not contain $i$

$$f_x(S \cup \{i\}) - f_x(S) \geq f'_x(S \cup \{i\}) - f'_x(S) \tag{5}$$

then

$$\phi_i(f, x) \geq \phi_i(f', x) \tag{6}$$

### 1.2 Here we show how the symmetry axiom is also implied by the monotonicity axiom for models

Assume that $f'$ is the same as $f$ except the inputs $i$ and $j$ are swapped. The means for all subsets $S$ that do not contain $i$ or $j$ that $f'_x(S \cup \{i\}) = f_x(S \cup \{j\})$ and $f'_x(S) = f_x(S)$. If $S$ does contain $j$ then $f'_x(S \setminus \{j\} \cup \{i, j\}) = f_x(S \setminus \{j\} \cup \{j, i\})$, which is garunteed to hold so we can ignore it in the implication below. Starting with the monotonicity axiom

$$\forall_{S \subset Z \setminus \{i\}} \quad f_x(S \cup \{i\}) - f_x(S) \geq f'_x(S \cup \{i\}) - f'_x(S) \implies \phi_i(f, x) \geq \phi_i(f', x) \tag{7}$$

can be transformed into

$$\forall_{S \subset Z \setminus \{i,j\}} \quad f_x(S \cup \{i\}) \geq f_x(S \cup \{j\}) \implies \phi_i(f, x) \geq \phi_i(f', x) \tag{8}$$

by using $f'_x(S \cup \{i\}) = f_x(S \cup \{j\})$, $f'_x(S) = f_x(S)$, and ignoring the terms that include $j$. Swapping $i$ and $j$ and then repreating the process shows that

$$\forall_{S \subset Z \setminus \{i,j\}} \quad f_x(S \cup \{i\}) = f_x(S \cup \{j\}) \implies \phi_i(f, x) = \phi_j(f, x) \tag{9}$$

which is the symmetry axiom. This shows that we only need efficiency and monotonicity to uniquely constrain ourselves to using the Shapley values.