I Ranking and Sign Estimation

We next propose a simple and (as we shall experimentally show) effective heuristic solution for estimating the order of Shapley values, without computing the actual values. The solution uses $A_{d,s}$ as a proxy for the Shapley value of d w.r.t. s, namely instead of ranking based on the (unknown) Shapley values, it ranks based on $A_{d,s}$. Similarly, to estimate the sign of the Shapley value of a data point $d \in D$ with respect to a label $s \in S$, we output the sign of $A_{d,s} - AVG_{i \in S}(A_{d,i})$ (if the value is greater than the average, we output "positive", otherwise we output "negative").

This solution is based on the following observation: for $S = \{s_1, s_2\}$ and $A_{d,s_1} > A_{d,s_2}$ for some sequence $d \in D$, then the Shapley value of d with respect to s_1 is positive, and with respect to s_2 is negative. While a generalization to cases where $\mid S \mid > 2$ does not necessarily hold, there is an intuitive correlation between the relative magnitudes of $A_{d,s}$ and the relative magnitudes of Shapley values. Indeed, we will experimentally show that the simple algorithm ranking data items based on $A_{d,s}$ values generally achieves good ranking accuracy, and the simple algorithm for the sign estimation attains good sign accuracy.

Data: Dataset D
Result: Estimations to Shapley values for each $d \in D$ (serves as attribution explanation for d).
Run ComputeDataProb() step of Algorithm 1 on D to get $A_{d,s}$ values for each $d \in D, s \in S$;
return $(A_{d,s} - AVG_{i \in S}[A_{d,i}])$ as explanation value for data point d on label s.

Algorithm 1: Attribution explanations (estimating Shapley values order and sign using $A_{d,s}$ values)

Proposition 1. *If* $\mid S \mid = 2$, *and* $A_{d,s_1} > A_{d,s_2}$ *for a data point* d *and the two labels* s_1, s_2, *the Shapley value of* d *with respect to the label* s_1 *is greater than its Shapley value with respect to* s_2, *and also* A_{d,s_1} *is positive and* A_{d,s_2} *is negative.*

Proof. We will show that when $|S| = 2$, if $A_{i,s_1} \geq A_{i,s_2}$, then $Shapley(i)[s_1] \geq Shapley(i)[s_2]$. Moreover, we will show that the Shapley value of i with respect to s_1 is positive, and the Shapley value of i with respect to s_2 is negative. The value function is:

$$v(D) = argmax(\prod_{d \in D} \sum_{s \in S} \alpha_s \cdot A_{d,s}) \tag{1.1}$$

And in the case of 2 labels:

$$v(D) = argmax(\prod_{d \in D} [\alpha_{s_1} * A_{d,s_1} + \alpha_{s_2} * A_{d,s_2}]) \tag{1.2}$$

And since $\alpha_{s_1} + \alpha_{s_2} = 1$:

$$v(D) = argmax(\prod_{d \in D} [\alpha_{s_1} * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}]) \tag{1.3}$$

The Shapley value of a data point $i \in D$ is:

$$\frac{1}{N!} \sum_{G \subseteq [N] \setminus \{i\}} [v(G \cup i) - v(G)] * |G|! * (N - |G| - 1)! \tag{1.4}$$

We will show that vector $v(G \cup i) - v(G)$ is always positive in the index s_1 and always negative in the index s_2. The value $v(G \cup i) - v(G)$ is:

$$\begin{aligned} argmax(\prod_{d \in G \cup i} [\alpha_{s_1} * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}]) - \\ argmax(\prod_{d \in G} [\alpha_{s_1} * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}]) \end{aligned} \tag{1.5}$$

Equals:

$$\begin{aligned} argmax(\prod_{d \in G} [\alpha_{s_1} * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}] * \\ (\alpha_{s_1} * (A_{i,s_1} - A_{i,s_2}) + A_{i,s_2})) - \\ argmax(\prod_{d \in G} [\alpha_{s_1} * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}]) \end{aligned} \tag{1.6}$$

Assume that the vector that is maximizing vector $v(G \cup i)$ is (m_1, m_2) and that that vector that is maximizing $v(G)$ is (n_1, n_2). Note that what we need to prove is that $m_1 \geq n_1$ and that $m_2 \leq n_2$ (to show that the maximizing vector after adding data point i is higher in the index of s_1 and smaller in the index of s_2).

Assume for the way of contradiction that $m_1 < n_1$. We know that (m_1, m_2) is maximizing $v(G \cup i)$ so we have that the likelihood function on $G \cup i$ is higher in the point (m_1, m_2) than in any other point, specifically (n_1, n_2), so:

$$\prod_{d\in G}[m_1 * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}] * (m_1 * (A_{i,s_1} - A_{i,s_2}) + A_{i,s_2}) \geq \\ \prod_{d\in G}[n_1 * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}] * (n_1 * (A_{i,s_1} - A_{i,s_2}) + A_{i,s_2}) \tag{1.7}$$

But we know that (n_1, n_2) is maximizing the likelihood function on G, so:

$$\prod_{d\in G}[m_1 * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}] \leq \\ \prod_{d\in G}[n_1 * (A_{d,s_1} - A_{d,s_2}) + A_{d,s_2}] \tag{1.8}$$

Which means that

$$(m_1 * (A_{i,s_1} - A_{i,s_2}) + A_{i,s_2}) \geq (n_1 * (A_{i,s_1} - A_{i,s_2}) + A_{i,s_2}) \tag{1.9}$$

Which is equivalent to:

$$m_1 * (A_{i,s_1} - A_{i,s_2}) \geq n_1 * (A_{i,s_1} - A_{i,s_2}) \tag{1.10}$$

But since we assumed that $A_{i,s_1} > A_{i,s_2}$, the value $A_{i,s_1} - A_{i,s_2}$ is positive, hence

$$m_1 \geq n_1 \tag{1.11}$$

And this is a contradiction.

This means that for each subset G, $v(G \cup i)[s_1] \geq v(G)[s_1]$, so $v(G \cup i)[s_1] - v(G)[s_1]$ is always positive, and since Shapley value is just summing over these values and multiplying with positive values, the Shapley value $Shapley(i, s_1)$ will be positive.

The proof of the second direction is symmetric (that $Shapley(i, s_2)$ will be negative).

□

On the other hand:

Proposition 2. *A generalization of Proposition 1 to* $\mid S \mid \geq 3$ *does not hold.*

Proof. It might be the case that $A_{i,s_1} > A_{i,s_2}$, but still $Shapley(i)[s_1] < Shapley(i)[s_2]$, for example, in a case where the algorithm is debating between s_2 and s_3, and data point i is making it decide on s_2, because $A_{i,s_2} > A_{i,s_3}$. So the Shapley of i on s_2

would be very high, and the Shapley of i on s_1 would be 0, because anyway the algorithm was not even considering label s_1. A formal counter example would be when the likelihood function is:

$$L(D, \alpha) = (10^8\alpha_1 + 10^8\alpha_2 + 0\alpha_3)^{1000} \cdot (100\alpha_1 + 0\alpha_2 + 101\alpha_3) \tag{1.12}$$

For the last data point, even though $A_{i,s_1} < A_{i,s_3}$. this is the data point that is making the algorithm decide between s_1 and s_2, hence it will have a high value in $Shapley(i)[s_1]$, and $Shapley(i)[s_3]$ will be 0. □