Shapley Kernel Proof

November 3, 2017

The Shapley kernel is the sample weight given to each binary vector $z' \in \{0,1\}^M$:

$$k(z') = k(M,s) = \frac{M-1}{(M\ choose\ s)s(M-s)}$$

where $s = |z'|$, the number of ones in z'.

Let X be the matrix of all possible binary vectors of length M with 2^M rows and M columns. We use the Shapley kernel to compute the Shapley values using weighted linear regression:

$$\phi = (X^TWX)^{-1}X^TWy$$

where W is a diagonal matrix with the Shapley kernel weights for each row of X, and the $y_i = f_x(S_i)$ values are the function outputs for each row of X (where S_i is the set of ones in $X_{i,*}$). Note that $k(M,0) = k(M,M) = \infty$, so W is infinity for the all zero row of X and the row of all ones. However, if we set these infinite weights to a large constant, then $X^TWX = \frac{1}{M-1}I + cJ$ for some positive constant c (where I is the identity matrix and J is the matrix of all ones). As $c \to \infty$ the inverted form becomes $(X^TWX)^{-1} = I + \frac{1}{M-1}(I-J)$

The term X^TW is a matrix where all the ones in X^T have been replaced with $k(M,s)$, where s is the number of ones in that column of X^T. Multiplying X^TW by $(X^TWX)^{-1}$ creates a matrix of weights to apply to the function outputs in y. If we only consider the Shapley value of a single feature ϕ_j, then we only need to consider a single row of this $2^M \times M$ matrix, which is equivelent to only using the j'th row of $(X^TWX)^{-1}$. When we do this we see that the value of the weight for row i is

$$\begin{aligned}
k(M,s_i)[\mathbf{1}_{X_{i,j}=1} - \frac{(s_i - \mathbf{1}_{X_{i,j}=1})}{M-1}] &= \frac{M-1}{(M\ choose\ s_i)s_i(M-s_i)}\mathbf{1}_{X_{i,j}=1} - \frac{(s_i - \mathbf{1}_{X_{i,j}=1})}{(M\ choose\ s_i)s_i(M-s_i)} && (1)\\
&= \frac{(M-1)(M-s_i)!s_i!}{M!s_i(M-s_i)}\mathbf{1}_{X_{i,j}=1} - \frac{(s_i - \mathbf{1}_{X_{i,j}=1})(M-s_i)!s_i!}{M!s_i(M-s_i)} && (2)\\
&= \frac{(M-1)(M-s_i-1)!(s_i-1)!}{M!}\mathbf{1}_{X_{i,j}=1} - \frac{(s_i - \mathbf{1}_{X_{i,j}=1})(M-s_i-1)!(s_i-1)!}{M!} && (3)\\
&= \frac{(M-s_i-1)!(s_i-1)!}{M!}[(M-1)\mathbf{1}_{X_{i,j}=1} - (s_i - \mathbf{1}_{X_{i,j}=1})] && (4)
\end{aligned}$$

where s_i is the number of ones in the i'th row of X, and $\mathbf{1}_{X_{i,j}=1}$ is one if $X_{i,j} = 1$ and zero otherwise. When $\mathbf{1}_{X_{i,j}=1} = 0$ we get

$$-\frac{(M-s_i-1)!s_i!}{M!}$$

When $\mathbf{1}_{X_{i,j}=1} = 1$ we get

$$\frac{(M-s_i-1)!(s_i-1)!}{M!}[(M-1)-(s_i-1)] = \frac{(M-s_i-2)!(s_i-1)!}{M!}$$

Taking the dot product of these values with y leads to the following equation

$$\phi_j = \sum_{S \subseteq N \setminus j} \frac{(M - s_i - 1)! s_i!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$

which is a classic form of estimating the Shapley value ϕ_j (N is the set of all features).