Errata for A Unified Approach to Interpreting Model Predictions

This is a working list of corrections for issues in the paper that we have found after final publication [1]. We are grateful to those have carefully read the paper and pointed out these issues.

1. Equation 8 in the original paper has an off-by-one error in the weighting term and should instead be written as:

   $$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{(|z'| - 1)!(M - |z'|)!}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{8}$$

   The reason Equation 8's weighting term should be slightly different than Equation 4 is that in Equation 4 we add the element i to the set while in Equation 8 we are removing it from the set.
2. In Corollary 1 (Linear SHAP) There are two typos in the equation: First, $\phi_i(f, x)$ should have a j instead of an i and so should instead be written as $\phi_j(f, x)$. Second, $\phi_0(f, x) = b$ is incorrect when the mean of the features are non-zero. It should instead be written as $\phi_0(f, x) = b + \sum_{j=1}^{M} w_j E[x_j]$.
3. The supplementary proof "The monotonicity axiom implies the symmetry axiom for Shapley values" is incorrect because it implicitly assumes feature *anonymity* (which means the names of the features don't impact the credit we assign to them). While this is a reasonable assumption in many cases, it is none-the-less an assumption that should be stated. Thanks to Aaron Fisher for pointing this out (and Lizao Li).

References

[1] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4765–4774.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.