Journal of Artificial Intelligence Research (2024) Submitted 03/2024; published 03/2024

Errata for: "On the Tractability of SHAP Explanations"

Guy Van den Broeck GUYVDB@CS.UCLA.EDU
University of California
404 Westwood Plaza, Los Angeles, CA 90095, USA

Anton Lykov ANTONY.LYKOV@GMAIL.COM
Maximilian Schleich MAXIMILIAN.SCHLEICH@RELATIONAL.AI
Dan Suciu SUCIU@CS.WASHINGTON.EDU
University of Washington
185 E Stevens Way NE, Seattle, WA 98195, USA

The following is a correction to the proof of Proposition 4 of den Broeck et al. (2022) The notations preceding Claim 3 define incorrectly the probability distribution p'_{ij}, because they do not sum to 1 for each i, i.e. they do not form a probability space on $\text{dom}(X_i)$. The corrected (and much simplified) definitions are the following, for all $i = 1, n$ and $j = 2, m_i$:

$$p'_{i1} \stackrel{\text{def}}{=} q_i \qquad\qquad p'_{ij} \stackrel{\text{def}}{=} \frac{1-q_i}{1-p_{i1}} p_{ij}$$

We check that for each $i = 1, n$, the numbers p'_{ij} sum up to 1:

$$\sum_{j=1,m_i} p'_{ij} = p'_{i1} + \sum_{j=2,m_i} p'_{ij} = q_i + \frac{1-q_i}{1-p_{i1}} \sum_{j=2,m_i} p_{ij} = q_i + \frac{1-q_i}{1-p_{i1}}(1-p_{i1}) = 1$$

In Claim 3 the factor $Z \cdot W$ is removed, and Claim 3 becomes:

Claim 3. $\mathbf{E}[F_\pi] = \mathbf{E}'[F]$

The proof of the claim is updated as follows. The derivation of $F_\pi[\mathbf{x}]$ remains unchanged from den Broeck et al. (2022). The derivation of $\mathbf{E}[F_\pi]$ is modified as follows:

$$\begin{aligned}
\mathbf{E}[F_\pi] &= \sum_{\mathbf{x}\in\{0,1\}^n} F_\pi(\mathbf{x}) \prod_{i=1,n:\mathbf{x}(i)=1} q_i \prod_{i=1,n:\mathbf{x}(i)=0} (1-q_i) \\
&= \sum_{\mathbf{x}\in\{0,1\}^n} \left(\sum_{\tau\in\mathcal{X}:\mathbf{x}^{-1}(1)=\tau^{-1}(1)} F(\tau) \cdot \prod_{i=1,n:\tau_i\neq 1} \frac{p_{i\tau_i}}{1-p_{i1}} \right) \prod_{i=1,n:\mathbf{x}(i)=1} q_i \prod_{i=1,n:\mathbf{x}(i)=0} (1-q_i) \\
&= \sum_{\mathbf{x}\in\{0,1\}^n} \left(\sum_{\tau\in\mathcal{X}:\mathbf{x}^{-1}(1)=\tau^{-1}(1)} F(\tau) \cdot \prod_{i=1,n:\tau_i\neq 1} \frac{p_{i\tau_i}}{1-p_{i1}} \left(\prod_{i=1,n:\tau_i=1} q_i \prod_{i=1,n:\tau_i\neq 1} (1-q_i) \right) \right) \\
&= \sum_{\tau\in\mathcal{X}:\mathbf{x}^{-1}(1)=\tau^{-1}(1)} F(\tau) \cdot \prod_{i=1,n:\tau_i\neq 1} \frac{p_{i\tau_i}(1-q_i)}{1-p_{i1}} \prod_{i=1,n:\tau_i=1} q_i \\
&= \sum_{\tau\in\mathcal{X}:\mathbf{x}^{-1}(1)=\tau^{-1}(1)} F(\tau) \cdot \prod_{i=1,n:\tau_i\neq 1} p'_{i\tau_i} \prod_{i=1,n:\tau_i=1} p'_{i1} = \sum_{\tau\in\mathcal{X}:\mathbf{x}^{-1}(1)=\tau^{-1}(1)} F(\tau) \cdot \prod_{i=1,n} p'_{i\tau_i} = \mathbf{E}'[F]
\end{aligned}$$

For a simple illustration, consider the case when the variable X_i has a domain of size 2, $\text{dom}(X_i) = \{1, 2\}$, in other words $m_i = 2$. Then a binary variable with outcomes $\{0, 1\}$ and probabilities $1 - q_i$ and q_i respectively is converted into a variable with outcomes $\{1, 2\}$ and probabilities p'_{i1}, p'_{i2} :

$$p'_{i1} = q_i \qquad\qquad p'_{i2} = \frac{1 - q_i}{1 - p_{i1}} p_{i2} = 1 - q_i$$

since $p_{i1} + p_{i2} = 1$. In other words, the new distribution on $\{1, 2\}$ is the same as the distribution on $\{0, 1\}$, up to the renaming of 0 to 2. More generally, we convert a binary variable with outcomes $\{0, 1\}$ into one with m_i outcomes, $\text{dom}(X_i) = \{1, \ldots, m_i\}$, by setting $p'_{i1} = q_i$, and distributing the remaining probability mass $1 - q_i$ across the outcomes $\{2, 3, \ldots, m_i\}$, proportionally to the initial probabilities $p_{i2}, p_{i3}, \ldots, p_{im_i}$.

References

den Broeck, G. V.; Lykov, A.; Schleich, M.; and Suciu, D. 2022. On the Tractability of SHAP Explanations. *J. Artif. Intell. Res.* 74: 851–886. doi:10.1613/JAIR.1.13283. URL `https://doi.org/10.1613/jair.1.13283`.