Fairness Through Uncertainty

Theo Stephens Kehoe, Master of Science in Computer Science

University of Dublin, Trinity College, 2024

Supervisor: Andrea Patane

Abstract As machine learning models become more prevalent in our lives and take on critical decision-making roles across various sectors of society, ensuring their decisions are not only accurate but also fair has become ever more crucial.

This paper focuses on individual fairness in Bayesian Neural Networks (BNNs). Individual fairness ensures that similar individuals receive similar outcomes by a model. Specifically, we undertake the definition of $\epsilon - \delta$ individual fairness. BNNs offer advantages over deterministic neural networks due to their ability to quantify uncertainty and effectively handle smaller datasets; a capability that has led to their adoption in critical fields such as medicine.

In this paper, we construct a fairness regularisation method for BNNs by transferring techniques from adversarial robustness training and employing the Fair-FGSM algorithm [17] for generating similar inputs. This approach draws on existing research that highlights the similarity between adversarial robustness and individual fairness definitions. The regulariser is designed to be simple, facilitating its integration into existing training procedures without extensive modifications. We also introduce a simple metric for measuring and comparing $\epsilon - \delta$ individual fairness models in an intuitive manner, the *Threshold-Fairness* metric.

Through our experimentation on various model architecture sizes and similarity metric parameters, encompassing a total of three-thousand models, we can attest that our devised regulariser is effective at improving the individual fairness of a BNN. However, due the to fairness-accuracy trade-off, there is a small degradation of model accuracy imposed by the regulariser. Additionally, we empirically evaluate that there is a relationship between the set of non-protected ϵ parameter values and the effectiveness of the regulariser at improving fairness. We hope that our devised regulariser acts a starting point for more sophisticated and adaptable individual fairness mechanisms, and that these findings will serve as a foundational piece for future research into individual fairness in BNNs.