**Abstract**

With growing awareness of societal impact of artificial intelligence, fairness has become an important aspect of machine learning algorithms. The issue is that human biases towards certain groups of population, defined by sensitive features like race and gender, are introduced to the training data through data collection and labeling. Two important directions of fairness ensuring research have focused on (i) instance weighting in order to decrease the impact of more biased instances and (ii) adversarial training in order to construct data representations informative of the target variable, but uninformative of the sensitive attributes. In this paper we propose a Fair Adversarial Instance Re-weighting (FAIR) method, which uses adversarial training to learn instance weighting function that ensures fair predictions. Merging the two paradigms, it inherits desirable properties from both interpretability of reweighting and end-to-end trainability of adversarial training. We propose four different variants of the method and, among other things, demonstrate how the method can be cast in a fully probabilistic framework. Additionally, theoretical analysis of FAIR models properties have been studied extensively. We compare FAIR models to six other related and state-of-the-art models and demonstrate that FAIR is able to achieve a better trade-off between accuracy and unfairness. To the best of our knowledge, this is the first model that merges reweighting and adversarial approaches by means of a weighting function that can provide interpretable information about fairness of individual instances.