Abstract

Interpretability and Fairness in Machine Learning: A Formal Methods Approach

by

Bishwamittra Ghosh

Doctor of Philosophy in Computer Science

National University of Singapore

The significant success of machine learning in past decades has led to a host of applications of algorithmic decision-making in different safety-critical domains. The high-stake predictions of machine learning in medical, law, education, transportation and so on have far-reaching consequences on the end-users. Consequently, there has been a call for the regulation of machine learning by defining and improving the interpretability, fairness, robustness, and privacy of predictions. In this thesis, we focus on the interpretability and fairness aspects of machine learning, particularly on *learning interpretable rule-based classifiers*, *verifying fairness*, and *interpreting sources of unfairness*. Prior studies aimed for these problems are limited by either scalability or accuracy or both. To alleviate these limitations, we integrate formal methods and automated reasoning with interpretability and fairness in machine learning and provide scalable and accurate solutions to the underlying problems.

In interpretable machine learning, rule-based classifiers are particularly effective in representing the decision boundary using a set of rules. The interpretability of rule-based classifiers is generally related to the size of the rules, where smaller rules with higher accuracy are preferable in practice. As such, interpretable classification learning becomes a combinatorial optimization problem suffering from poor scalability in large datasets. To this end, we discuss an incremental learning framework, called IMLI, which applies an iterative solving of maximum satisfiability (MaxSAT) queries in mini-batch learning and enables classification on million-size datasets. Although being interpretable, rule-based classifiers often suffer from limited expressiveness, for example, classifiers based on propositional logic. To learn more expressible yet interpretable classification rules, we discuss a relaxation of classifiers based on logical formulas. For learning relaxed rule-based classifiers, we discuss an efficient learning

framework, called CRR, building on incremental learning and mixed integer linear programming (MILP). CRR obtains higher accuracy yet less rule-size than existing interpretable classifiers.

Fairness in machine learning centers on quantifying and mitigating the bias or unfairness of machine learning classifiers. In the presence of multiple fairness metrics for quantifying bias, we discuss a probabilistic fairness verifier, called Justicia, with the goal of formally verifying the bias of a classifier given the probability distribution of features. Building on stochastic satisfiability (SSAT), Justicia improves the scalability of verification; and unlike prior approaches, Justicia verifies compound sensitive groups combining multiple sensitive features. For a more accurate fairness verification, we extend Justicia to consider feature correlations represented as a Bayesian Network, resulting in an accurate verification of fairness.

Fairness metrics globally quantify bias, but do not detect or interpret its sources. To interpret group-based fairness metrics, we discuss fairness influence function (FIF) with an aim of quantifying the influence of individual features and the intersection of multiple features on the bias of a classifier. FIF interprets fairness by revealing potential individual or intersectional features attributing highly to the bias. Building on global sensitivity analysis, we discuss an algorithm, called FairXplainer, for estimating the FIFs of features, resulting in a better approximation of bias based on FIFs and a higher correlation of FIFs with fairness interventions.