Thesis Review for *Interpretability and Fairness in Machine Learning: A Formal Methods Approach*

This thesis offers substantial contributions to the field of trustworthy machine learning, systematically addressing issues of interpretability, fairness, and scalability in machine learning models.

The author starts by motivating the need for trustworthy machine learning and setting up the relevant technical background. Subsequent chapters describe technical contributions, e.g., IMLI, an incremental MaxSAT-based approach for learning interpretable classifiers; a combinatorial learning framework (CRR) for learning relaxed CNF rules; Justicia, an SSAT approach to fairness metric verification; and FVGM, a method for handling feature correlations in fairness verification. The thesis also introduces Fairness Influence Functions (FIFs), which aim to determine individual and intersectional influence of features on fairness. The thesis concludes by summarizing its contributions and outlining promising directions for future research.

While certain sections could benefit from improved placement and elaboration, my overall impression of the thesis is positive. The following provides chapter-specific comments.

Chapter 1: Introduction

This chapter provides sufficient motivation for the contributions in trustworthy machine learning. The key contributions of the thesis are well summarized, along with relevant prior work and existing challenges. Overall, the chapter is well-written and I do not have major comments.

Chapter 2: Preliminaries

This chapter sets up the notation used and provides relevant background on formal methods, interpretable ML, and fairness in ML. The technical background is clear and provides clarification for the operations used (e.g., inner-products for booleans and reals).

Subsection 2.3.3 on Bayesian Networks seems out of place (placed under Fairness in ML). BNs are a general modeling framework. Potentially, you may want to shift it to the previous section on interpretable ML.

2.3.5: Is the decomposition across all subsets of Z?

Chapter 3: Scalability via Incremental Learning

Chapter 3 describes IMLI, the first technical contribution of this thesis, which is an incremental MaxSAT-based approach for learning an interpretable classifier. Writing is clear: a problem formulation is given, followed by the method and its extension to incremental learning.

Chapter 4: Expressiveness via Logical Relaxation

This chapter proposes a combinatorial learning framework (CRR) for learning relaxed CNF rules. It formulates a mixed-integer linear program for learning these rules. Experiments

show CRR to achieve good accuracy with a sparser set when using relaxed-CNF rules. Suggestions:

- In Sec. 4.1: useful to give the precise definition of |R| again.
- In the experiments, a general observation is that the rule sets for CRR tend to be smaller, but generally with slightly lower performance. What are the characteristics of the Adult dataset where CRR achieves comparable accuracy with a much smaller rule set compared to RIPPER? Also, there are datasets where CRR's rule set is larger (e.g., WDBC); are there any hypothesized reasons for this?
- Minor quirk; \lambda is often used as the regularization parameter (or the Lagrange multiplier) in ML settings. Perhaps use a different symbol?
- Footnote 1: Unless I am mistaken, it is still considered bad form to cite Wikipedia. I'm sure there are better sources for the majority function.

Chapter 5: Fairness Verification using SSAT

This chapter introduces Justicia, which is a SSAT approach to fairness metric verification. A key idea is to represent the conditional probability computation as a RE-SSAT formula that can be passed to a solver. This enables more efficient computation of the fairness metric. Experiments show the approach to be effective and more efficient computationally compared to FairSquare and Verifair. I don't have major comments for this chapter.

Chapter 6: Handling Features Correlations in Fairness Verification

The chapter aims to address a limitation of Justicia in that it does not consider feature correlations. Focussing on linear classifiers, the authors propose FVGM based on PGMs (specifically, Bayes Nets) to represent conditional independence assertions between random variables (features). Experiments shows FVGM to achieve better results compared to baselines. As stated in the thesis proposal comments, it remains unclear how first learning a Bayes Net outperforms a direct approach.

Chapter 7: Interpreting Fairness: Identifying Sources of Bias

This chapter introduces Fairness Influence Functions (FIFs), where the key aim is to develop a methodology for determining the individual and intersectional influence of features. FIFs leverage ideas from variance-based sensitivity analysis (in Sec 2.3.4) to define the contribution of a subset of features to the variance of the prediction based on the sensitive group. An issue with the experiments is that there is no "ground truth" FIF, but the analyses using fairness attacks and interventions helps to support the method. Some minor suggestions:

- The symmetry property is not apparent to me; some elaboration would be helpful.
- Add the definition statistical parity (it is given in Chp 2.3.2, but it would help to remind readers as to the precise definition).
- Consider moving your related work to after formally defining FIFs.
- Regarding the degenerate cases of perfectly biased/unbiased classifiers, detection of these cases can be performed in a straightforward manner?
- Fig 7.2 is similar to Fig 6.2, but their axes are flipped. I suggest consistency between the figures.

Chapter 8: Conclusions and future work

Chapter concludes this thesis with a summary of contributions and avenues for future work. I don't have major comments for this chapter.

Examiner 2

The student needs to provide a better positioning of his work with respect to the the rich prior work