

## Summary

Classifier construction is one of the most researched topics within the data mining and machine learning communities. Literally thousands of algorithms have been proposed. The quality of the learned models, however, depends critically on the quality of the training data. No matter which classifier inducer is applied, if the training data is incorrect, poor models will result. In this thesis, we study cases in which the input data is discriminatory and we are supposed to learn a classifier that optimizes accuracy, but does not discriminate in its predictions. Such situations occur naturally as artifacts of the data collection process when the training data is collected from different sources with different labeling criteria, when the data is generated by a biased decision process, or when the sensitive attribute, e.g., gender serves as a proxy for unobserved features. In many situations, a classifier that detects and uses the racial or gender discrimination is undesirable for legal reasons. The concept of discrimination is illustrated by the next example: *Throughout the years, an employment bureau recorded various parameters of job candidates. Based on these parameters, the company wants to learn a model for partially automating the matchmaking between a job and a job candidate. A match is labeled as successful if the company hires the applicant. It turns out, however, that the historical data is biased; for higher board functions, Caucasian males are systematically being favored.* A model learned directly on this data will learn this discriminatory behavior and apply it over future predictions. From an ethical and legal point of view it is of course unacceptable that a model discriminating in this way is deployed.

Our proposed solutions to the discrimination problem fall into two broad categories. First, we propose pre-processing methods to remove the discrimination from the training dataset. Second, we propose solutions to the discrimination problem by directly pushing the non-discrimination constraints into classification models and post-processing of built models.

We further studied the discrimination-aware classification paradigm in the presence of explanatory attributes that were correlated with the sensitive attribute, e.g., low income may be explained by the low education level. In such a case, as we show, not all discrimination can be considered bad. Therefore, we introduce a new way of measuring discrimination, by explicitly splitting it up into explainable and bad discrimination and propose methods to remove the bad discrimination only.

We tried our discrimination-aware methods over real world data sets. We observed in our experiments that our methods show promising results and clearly outperform the traditional classification model w.r.t. accuracy discrimination trade-off.

To conclude, we believe that discrimination-aware classification is a new and exciting area of research addressing a societally relevant problem.