AALTO UNIVERSITY, SCHOOL OF SCIENCE

Review of Naive Bayes approaches for discrimination-free classification

Sergio Isidoro

November 25, 2015

1 PROBLEM OVERVIEW

The paper being analysed, calders10 [1], proposes 3 naive bayes methods for discrimination free classification. A data set provided by UCI Machine learning repository [5] on the Census [4] data is used to test the methods, trying to predict if an individual has high or low income (>=50 or <50), while removing discrimination on the gender of the individuals being evaluated.
This work aims to review and replicate some of the results and methods

2 THE DISCRIMINATION MEASURE

The discrimination measure used in calders10, and in the algorithms later on it based, are slightly flawed. The discrimination measure is described as:

> A simple solution is the discrimination score, which we define as the difference between the probability of a male and a female of being in the high-income class

First of all, this leads to an asymetric concept of discrimination (we assume a priory knowledge that the discrimination hapens in the class *female*). In the first algorithm we see that the optimization criteria is *while discrimination > 0*, leading to the possibility of reaching a discrimination that is negative, ie. where the class previously privileged will be slightly discriminated against.
Also, this discrimination measure is sensitive to population size and bias. Let's say that all men access credit, but only the most successful women do. This measure assumes that the

probabilty of a man and women being given credit should be the same. This will still discriminate women that are successfull and trying to access credit, since a less successfull man would access credit due to the bias of the population
The first problem could be solved by using a absolute discrimination measure, and make necessary ajustments to the used algorithms. Anoter solution would be to optimize to reach the closest number to zero, instead of trying to reach a disctimination < 50
The second problem is more difficult to tackle, but other works have presented some interesting metrics. Ruggieri10 [3] compares the confidence of an association rule with and without a parameter, resulting in a fairly good measure of direct discrimination independent of the population of each side of the sensitve parameter. Zemel13 [2] uses a concept of Individual Fairness in k-nearest neighbours, comparing how similar samples (except the sensitive parameter) are classified (eg. 2 samples that are exacly similar in every way except gender, are classified differently).

3 METHODS

The methods used were as closely replicated to the methods in calders10 [1] as possible. For discretization of variables, values were ranked into 4 bins (analogous to very low, low, high, very high).
No other action was taken to remove rows with missing atrributes. The missing attibutes were, instead, used as attributes (with value "?").
The methods implemented and presented here are the Modifed Bayes model, which balances the classification post trainig, optimizing the model to a discrimination closer to zero as possible, while maintaining the global probability. The other, 2M Model, breaks the classifier into 2, one for each value of the sensitive parameter. The results follow:

4 RESULTS

Method	**Accuracy**	**Discrimination measure**
Bayes	78.4166	0.3985
Modified Bayes Model	75.9413	-0.010
2M Bayes Model	78.5333	0.1653

Table 4.1: Test results (single run)

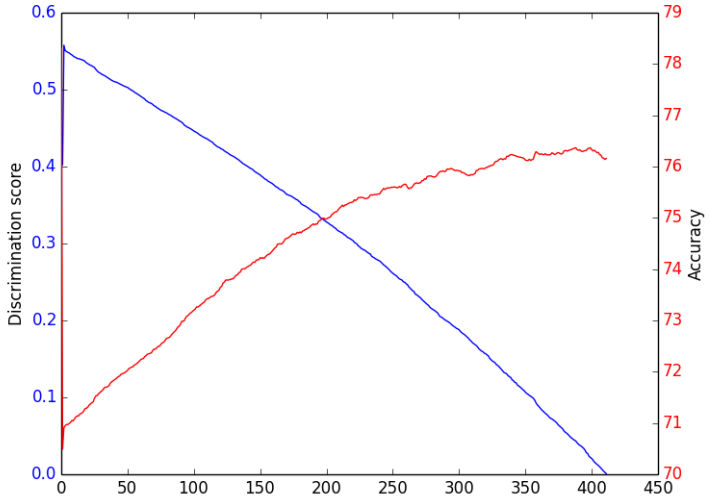

Figure 4.1: Optimization progression, with Accuracy Vs. Discrimination score on Modified Bayes Model

REFERENCES

[1] *Three naive Bayes approaches for discrimination-free classification*, Toon Calders et al., 2015.

[2] *Learning Fair Representations*, Richard Zemel et al., 2013

[3] *Data Mining for Discrimination Discovery*, Salvatore Ruggeri et al., 2010

[4] *UCI Census Income Data Set*, https://archive.ics.uci.edu/ml/datasets/Census+Income

[5] *UCI Machine Learning Repository* M. Lichman, 2013, http://archive.ics.uci.edu/ml , University of California, Irvine, School of Information and Computer Sciences