Bias Simulation Study

Responsible Data Science

Gesa Götte, Marcel Öfele, Viviane Wolters

01.10.2019

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INF

FAKULTÄT FÜR
INFORMATIK

Contents

The team: Biaspects

- **B.Sc. Gesa Götte**
  Statistik (Master: 4.Semester)
  - Hypotheses evaluation
  - Code monitoring
  - Weekly reports
- **B.Eng. Marcel Öfele**
  Digital Engineering (Master: 3. Semester)
  - Coding
  - Weekly reports
- **B.Eng. Viviane Lisa Wolters**
  Digital Engineering (Master: 4.Semester)
  - Code monitoring
  - Visualisation & presentation preparation
  - Weekly reports

Idea

- Create simple data sets
- Provoke in different ways a bias in the data sets
- Train selected classifiers for samples of the created data set
- Run the classification models on data test sets
- Compare the true-positive / true-negative rates of the different models

Motivation

- There occur performance biases in machine learning algorithms
- The bias may have different causes (unbalanced samples, real differences in the dependencies)

$\rightarrow$ **Experimental exploration of potential bias sources and their interactions**

GOALS:

- Find sources for performance bias
- Find adjustments that minimize the bias
- In general: Find settings for ML-Algorithms that minimize the occurrence of performance biases

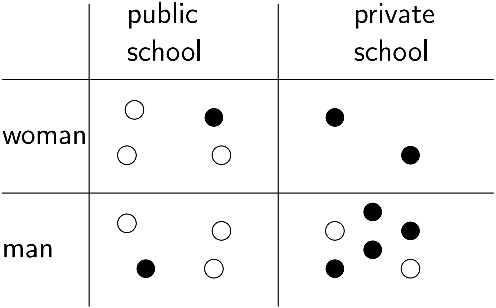
State of the art

- A lot of research for methods how to prevent biased ML models
- Some theoretical works on how biased datasets affect different algorithms
- Little to no work on empirical investigations to this kind of topic

Hypotheses

Hypothesis 1:

If there is no real difference between populations regarding the underlying label distribution, there will be no bias in a classifiers performance in any direction.



● = CEO / ○ = not a CEO

Hypotheses

Hypothesis 2:

If each population is represented equally in a learning data set of a (binary) classifier, there will be no bias, independent of probable differences between the populations regarding the underlying label distribution.

	public school	private school
woman	○ ○ ○ ● ○	● ● ○
man	● ● ○ ○	○ ● ● ●

● = CEO / ○ = not a CEO

Hypotheses

Hypothesis 3:

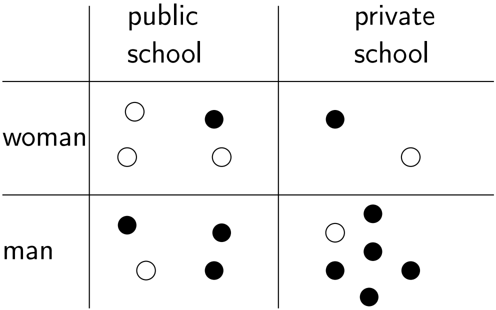
If one population is overrepresented in a learning data set of a (binary) classifier, its underlying distribution of the label will impact the performance of the classifier on the whole learning data set. The sensitivity or specificity for the underrepresented population will be worse.

	public school	private school
woman	○ ● ○ ○	● ○
man	● ○ ○ ○	● ○ ● ● ○ ●

● = CEO / ○ = not a CEO

Hypotheses

Hypothesis 3.1:
The greater the shift between the populations regarding the underlying label distribution, the higher the loss of sensitivity/specificity for the underrepresented one.



● = CEO / ○ = not a CEO

Hypotheses

Hypothesis 3.2:

The greater the overrepresentation of one population, the higher the bias

	public school	private school
woman	○ ● ○ ○	● ○
man	● ● ○ ●	● ○ ● ● ● ● ● ○

• = CEO / ○ = not a CEO

Hypotheses

Hypothesis 4: Exclusion of the sensitive variable will not reduce the bias.

Hypothesis 5: Different classification algorithms are differently vulnerable for bias.

Methods: Study1

Data set:
Sensitive variable: 'woman' or 'man'
Feature: 'private school' or 'public school'
Label: 'CEO' or 'not a CEO'

- Factor 1: Sample distribution of the sensitive variable
- Factor 2: Label distribution in the men's population
- Factor 3: Label distribution shifts between the men and women (indirect: Label distribution in the women's population)

Methods: Study1

Evaluation:

Measurement of bias:

- difference in TPR for men and women
- difference in TNR for men and women

Averaged over multiple data sets (for each factor combination)

Methods: Study2

Data set:
Create data sets as in study 1 but only this one's which resulted in a significant bias

- Factor1: inclusion/exclusion of the sensitive variable
- Factor2: subsequent adaptation of the population distribution (50%/50%)

Evaluation:

- Measurement of bias according to study 1
- Look at bias reduction

Current status & next steps

KW38	Research, topic concretization
KW39	Work up hypotheses, start coding
$\rightarrow$ KW40 $\leftarrow$	Kickoff presentation, hypotheses operationalization
KW41	Finish coding
KW42 + KW43	Evaluation & adjustments
KW44 + KW45	Final evaluation & visualisation
KW46	Preparing presentation/poster, code cleaning
KW47	Final & public presentation