Towards a Framework for Bias Analysis in Data

Andrei Olaru

[andrei.olaru@upb.ro]

AI-MAS Group

23.05.2025

Towards a Framework for Bias Analysis in Data

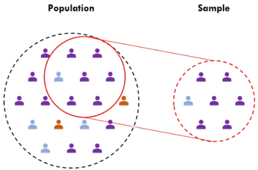
overview

Terminology

Bias in learning

Decision support systems must be interpretable and transparent. They must be fair and not give privilege to one or other category of individuals.

Bias can arise in algorithms or in data, but in ML models, using unbiased data is critical to the results of algorithms.

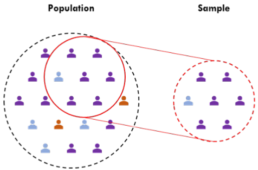


Bias in learning

Decision support systems must be interpretable and transparent. They must be fair and not give privilege to one or other category of individuals.

Bias can arise in algorithms or in data, but in ML models, using unbiased data is critical to the results of algorithms.

Most times, bias in data is linked to a correlation between the output and one protected attributes, such as gender, race, social status, religion, political views, and others.

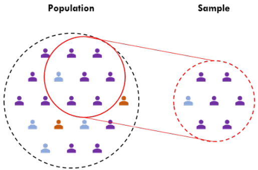


Bias in learning

Decision support systems must be interpretable and transparent. They must be fair and not give privilege to one or other category of individuals.

Bias can arise in algorithms or in data, but in ML models, using unbiased data is critical to the results of algorithms.

Most times, bias in data is linked to a correlation between the output and one or more protected attributes, such as gender, race, social status, religion, political views, and others.

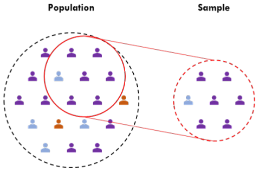


Bias in learning

Decision support systems must be interpretable and transparent. They must be fair and not give privilege to one or other category of individuals.

Bias can arise in algorithms or in data, but in ML models, using unbiased data is critical to the results of algorithms.

Most times, bias in data is linked to a correlation between the output and one or more protected attributes, such as gender, race, social status, religion, political views, and others. This correlation may also arise via other, non-protected attributes.



Bias in learning

Dataset: a series of data *records*

Data record: a series of values for various attributes, all corresponding to the same entity (e.g. a *person*).

Attribute (or feature): a property of entities represented by the dataset (e.g. the person's *salary*, or the person's *gender*).

Class: a property predicted by an algorithm, for a given entity (e.g. the person's *credit score*.)

Bias in learning

Protected attribute: an attribute that is legally or ethically safeguarded against discrimination (e.g. race, gender, age, religion, disability, or sexual orientation).

Bias: a systematic error or prejudice in data or in the decision-making process that leads to unfair, unbalanced, or inaccurate predictions, especially relevant when it favors or disadvantages certain groups, and especially when based on protected attributes like race, gender, or age.

Bias in learning

Explicit bias: a correlation between a protected attribute (or specific values therein) and the predicted class (or specific values therein), or an incorrect distribution of the values in a protected attribute with respect to the real distribution.

Implicit bias: unconscious or unintended bias and are often harder to detect, caused by

- attributes that correlate with protected attributes
- bias in how data is measured, entered, or annotated

Bias in learning

Explicit bias: a correlation between a protected attribute (or specific values therein) and the predicted class (or specific values therein), or an incorrect distribution of the values in a protected attribute with respect to the real distribution.

Implicit bias: unconscious or unintended bias and are often harder to detect, caused by

- attributes that correlate with protected attributes
- bias in how data is measured, entered, or annotated

Group fairness: the principle that the predicted class should have the same distribution over groups with the same protected attributes.

Individual fairness: the principle that similar entities (with similar values for non-protected attributes) should benefit from similar predictions.

Bias in learning

- detection of bias in data and/or in the predictions of an algorithm
- discovering implicit bias
- evaluating individual fairness versus group fairness in a data set
- mitigation of bias, e.g. by resampling and/or by reweighting the records in the dataset

Existing frameworks

What approaches to bias-related support frameworks already exist?

Existing frameworks

IBM AI Fairness 360 (AIF360) [https://github.com/Trusted-AI/AIF360] [Bellamy et al., 2018]

- comprehensive set of metrics and algorithms for evaluating fairness and mitigating bias in data
- available in R and Python

Existing frameworks

IBM AI Fairness 360 (AIF360) [https://github.com/Trusted-AI/AIF360] [Bellamy et al., 2018]

- comprehensive set of metrics and algorithms for evaluating fairness and mitigating bias in data
- available in R and Python
- however, there is no high-level manner in which to connect the different algorithms

Existing frameworks

Risk-of-bias VISualization (robvis) [https://github.com/mcguinlu/robvis] [McGuinness and Higgins, 2020]

- implemented in R
- has a web app

Existing frameworks

Risk-of-bias VISualization (robvis) [https://github.com/mcguinlu/robvis] [McGuinness and Higgins, 2020]

- implemented in R ← difficult to interoperate with ML algorithms in Python
- has a web app
- no flexibility in adding algorithms or metrics

Existing frameworks

fairmodels [https://modeloriented.github.io/fairmodels/] [Wisniewski and Biecek, 2022]

- implemented in R
- features a variety of plots

Existing frameworks

fairmodels [https://modeloriented.github.io/fairmodels/] [Wisniewski and Biecek, 2022]

- implemented in R $\leftarrow$ difficult to interoperate with ML algorithms in Python
- features a variety of plots
- flexibility is limited to the metrics integrated in the framework

Existing frameworks

- little support for a comprehensive, easy-to-use framework in Python
- extensibility is needed as new algorithms are developed
- visualization tools exist (e.g. for producing plots) but there is little support for a visual language for constructing the processing pipeline
- support for non-tabular data (e.g. images) is close to none

Envisioned result

What do we envision as an end result?

Envisioned result

A tool to support bias detection and mitigation

- a visual application
- plug-in building blocks
  - bias detection / mitigation algorithms

Envisioned result

A tool to support bias detection and mitigation

- a visual application
- plug-in building blocks
  - bias detection / mitigation algorithms
  - attribute selection
  - dataset filtering
  - running model predictions on dataset
  - training model on dataset

Envisioned result

A tool to support bias detection and mitigation

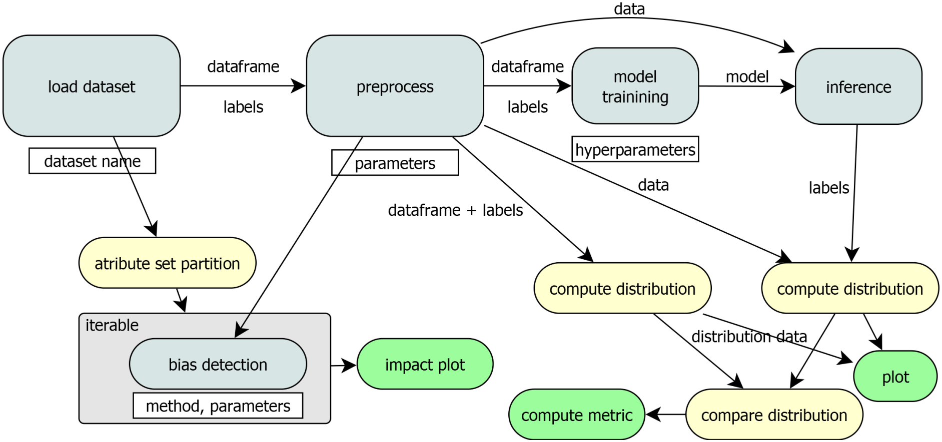
- a visual application
- plug-in building blocks
  - bias detection / mitigation algorithms
  - attribute selection
  - dataset filtering
  - running model predictions on dataset
  - training model on dataset
  - metric evaluation
  - plots on metrics

Envisioned result

A tool to support bias detection and mitigation

- a visual application
- plug-in building blocks
  - bias detection / mitigation algorithms
  - attribute selection
  - dataset filtering
  - running model predictions on dataset
  - training model on dataset
  - metric evaluation
  - plots on metrics
- a visual language for assembling building blocks into processing pipelines

Envisioned result



Architecture

Objects

- Data frames / Data sets
- Attribute set
- Ground truth labels
- Predicted labels
- Record selection
- Prediction models
- Distribution data
- Metrics

Architecture

Processes

- Algorithms
  pre-processing · post-processing
- Model training
- Model inference
- Computation of distributions and statistical metrics

Objects

- Data frames / Data sets
- Attribute set
- Ground truth labels
- Predicted labels
- Record selection
- Prediction models
- Distribution data
- Metrics

Architecture

Objects

- Data frames / Data sets
- Attribute set
- Ground truth labels
- Predicted labels
- Record selection
- Prediction models
- Distribution data
- Metrics

Processes

- Algorithms
  pre-processing · post-processing
- Model training
- Model inference
- Computation of distributions and statistical metrics

Control

- Iterables
- Aggregators
- Object identifiers
- Object flow

Architecture

Engineering + usability

- package management
- dynamic algorithm and model loading
- data types across implementations

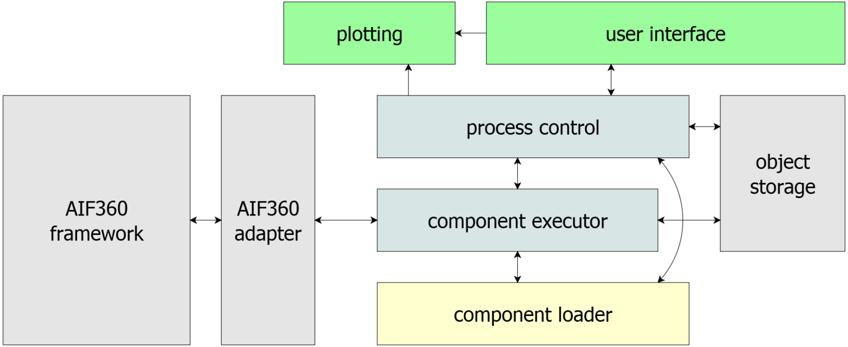
Architecture

Engineering + usability

- package management
- dynamic algorithm and model loading
- data types across implementations
- visual interface
- dimensionality challenges
- programming related elements – identifiers and iterative structures

Architecture

Preliminary high-level architecture



Future work

- assisting users in bias detection via automatic variation of parameters
- addition of new features based on existing data, to help in detecting biases
- interoperation with other frameworks providing algorithms
- couple the visual language with an LLM to operate and *inter*-operate the building blocks.

Thanks to Isel Grau Garcia (TU/e) for contributions to this work.

Thank You!

Questions are welcome!

[andrei.olaru@upb.ro]

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018).
AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

McGuinness, L. A. and Higgins, J. P. T. (2020).
Risk-of-bias visualization (robvis): An r package and shiny web app for visualizing risk-of-bias assessments.
Research Synthesis Methods, 11(1):40–46.

Wisniewski, J. and Biecek, P. (2022).
fairmodels: A flexible tool for bias detection, visualization, and mitigation in binary classification models.
The R Journal, 14(1):227–243.

Thank You!

Questions are welcome!

[andrei.olaru@upb.ro]