Ethics in Data Science

Quantititive Methods

Dr. Philippe J.S. De Brouwer
Honorary Consul of Belgium in Kraków
guest professor at the UJ, AGH, UEK and UW
board member of AGH and ISK
SVP at HSBC in Kraków

2024-11-27

AGH University of Krakow

Table of Contents i

Bias in Data and Mathematical Models

- Basic Definitions
- Algorithmic Fairness
- The Confusion Matrix and Related Measures
- Measures for Algorithmic Fairness

Origin and Types of Bias

Conclusions

Case Study

Bias in Data and Mathematical Models

Bias in Data and Mathematical Models

Basic Definitions

What is a Model?

Definition (Mathematical Model)

A mathematical model is a description of a system using mathematical concepts and language.

What is a Model?

Definition (Mathematical Model)

A mathematical model is a description of a system using mathematical concepts and language.

Use of a model in banking is fact-based (data is used to fit the model):

$$\begin{tikzcd} & \textit{data} \arrow[d, Rightarrow, "\textit{fit}"] & \\ \textit{input} \arrow[r] & \boxed{\text{MODEL}} \arrow[r] & \textit{output} \end{tikzcd}$$

What is Bias in Models?

What is Bias in Models?

Definition (Bias in Mathematical Models)

Model Bias (also Algorithmic Bias) refers to the systematic and repeatable error in a Model that creates outcomes that are statistically at odds with the reality of the population whose behaviour it is supposed to reflect.

Bias in Data and Mathematical Models

Algorithmic Fairness

Fairness Assumptions

Assume S the protected feature with an under-privileged group S_u and a privileged group S_p, and predicted outcomes $\hat{Y}$.

1. **Independence**: the probability of being in S_p or S_u has nothing to do with $\hat{Y}$ – $\hat{Y}$ is independent of S.
2. **Separation**: The predicted probability of having a favourable outcome has nothing to do with membership of the protected feature – $\hat{Y}$ is independent of S, given Y.
3. **Sufficiency**: the prediction should not depend on the protected group – Y is independent of S, given $\hat{Y}$.

Bias in Data and Mathematical Models

The Confusion Matrix and Related Measures

Useful Concepts for the Confusion Matrix

The following are useful measures for how good a classification model fits its data:

- *Accuracy:* The proportion of predictions that were correctly identified.
- *Precision* (or positive predictive value): The proportion of positive cases that correct.
- *Negative predictive value:* The proportion of negative cases that were correctly identified.
- *Sensitivity* or Recall: The proportion of actual positive cases which are correctly identified.
- *Specificity:* The proportion of actual negative cases which are correctly identified.

Some Acronyms for the Confusion Matrix

Let us use the following definitions:

- Objective concepts (depends only on the data):
  - *P:* The number of positive observations ($y = 1$)
  - *N:* The number of negative observations ($y = 0$)
- Model dependent definitions:
  - True positive (TP) the positive observations ($y = 1$) that are by the model correctly classified as positive;
  - False positive (FP) the negative observations ($y = 0$) that are by the model incorrectly classified as positive – this is a false alarm (Type I error);
  - True negative (TN) the negative observations ($y = 0$) that are by the model correctly classified as negative;
  - False negative (FN) the positive observations ($y = 1$) that are by the model incorrectly classified as negative – miss (Type II error).

The Definition of the Confusion Matrix

	Observed pos.	**Observed neg.**	
Pred. pos.	TP	FP	$\text{Pos.pred.val} = \frac{TP}{TP+FP}$
Pred. neg.	FN	TN	$\text{Neg.pred.val} = \frac{TN}{FN+TN}$
	Sensitivity $= \frac{TP}{TP+FN}$ $= \frac{TP}{TP+FN}$	Specificity $= \frac{TN}{FP+TN}$ $= \frac{TN}{FP+TN}$	Accuracy $= \frac{TP+TN}{TP+FN+FP+TN}$ $= \frac{TP+TN}{TP+FN+FP+TN}$

Table 1: The confusion matrix, where "pred." refers to the predictions made by the model, "pred." stands for "predicted," and the words "positive" and "negative" are shortened to three letters.

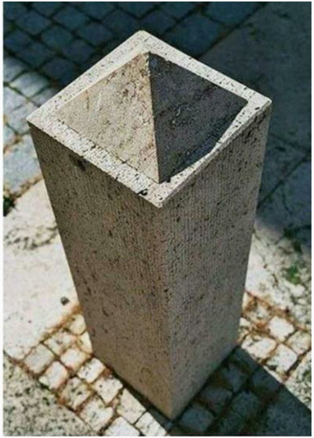
Bias in Data and Mathematical Models

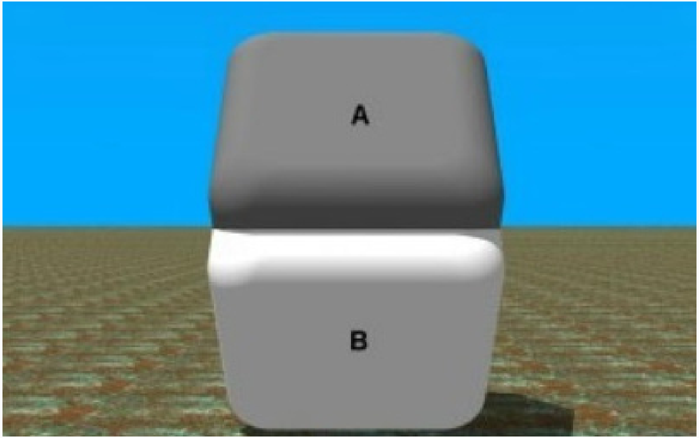
Measures for Algorithmic Fairness

Fairness Metrics

1. **Equal Opportunity**: $FNR_p = FNR_u$ (equivalent with $TPR_p = TPR_u$)
2. **Predictive Equality**: $FPR_p = FPR_u$ (equivalent with $TNR_p = TNR_u$)
3. **Equalised Odds**: both previous
4. **Predictive Parity** or outcome test: $Prec_p = Prec_u$ (equal positive predictive value or precision ($Prec = \frac{TP}{TP+FP}$))
5. **Demographic Parity**: membership of S has no correlation with favourable outcome ($\hat{Y}$).

Visual Biases are systematic miss-interpretations





Origin and Types of Bias

General Causes of Bias in Data

- **Confirmation Bias**: bias governs what we search for and we find what we search for (e.g. belief preservation via social media)

General Causes of Bias in Data

- **Confirmation Bias**: bias governs what we search for and we find what we search for (e.g. belief preservation via social media)
- **Selection Bias**: sample is not representative

General Causes of Bias in Data

- **Confirmation Bias**: bias governs what we search for and we find what we search for (e.g. belief preservation via social media)
- **Selection Bias**: sample is not representative
- **Historical Bias**: socio-cultural prejudices are reflected in data (e.g. Google image search)

General Causes of Bias in Data

- **Confirmation Bias**: bias governs what we search for and we find what we search for (e.g. belief preservation via social media)
- **Selection Bias**: sample is not representative
- **Historical Bias**: socio-cultural prejudices are reflected in data (e.g. Google image search)
- **Survival Bias**: the winner takes it all and losers are forgotten (e.g. hedge fund performance statistics)

General Causes of Bias in Data

- **Confirmation Bias**: bias governs what we search for and we find what we search for (e.g. belief preservation via social media)
- **Selection Bias**: sample is not representative
- **Historical Bias**: socio-cultural prejudices are reflected in data (e.g. Google image search)
- **Survival Bias**: the winner takes it all and losers are forgotten (e.g. hedge fund performance statistics)
- **Availability Bias**: violence has systematically decreased over the millennia, however we might think of modern times to be more violent. (focus on through the cycle data, different perspectives)

General Causes of Bias in Data

- **Confirmation Bias**: bias governs what we search for and we find what we search for (e.g. belief preservation via social media)
- **Selection Bias**: sample is not representative
- **Historical Bias**: socio-cultural prejudices are reflected in data (e.g. Google image search)
- **Survival Bias**: the winner takes it all and losers are forgotten (e.g. hedge fund performance statistics)
- **Availability Bias**: violence has systematically decreased over the millennia, however we might think of modern times to be more violent. (focus on through the cycle data, different perspectives)
- **Outlier Bias**: thinking about a startup one things of Google, Amazon, Facebook, etc. These are the outliers. (use median instead of average, identify outliers, etc.)

Process Related Bias

- **Reporting bias**: selective reporting of some data: citation bias, language bias (ignore reports in other languages), duplicate publication bias (copied data found twice), location bias (some studies are hard to find), publication bias (secret or not popular data is hard to find), outcome reporting bias (e.g. company does not report with much bravo when results are bad), time lag bias
- **Automation bias**: we prefer automated systems to provide data
- **Selection bias**: data is not representative (e.g. sampling bias, convergence bias (only people who got a loan are in our database), participation bias, etc.)
- **Over-generalisation bias**: no black swans in the data
- **Group Attribution bias**: generalisation of stereotypes, in-group bias (preference for members in the group), out-group bias (stereotype for other groups), etc.
- **Implicit bias**: search for conforming information

What causes Bias in Models?

Bias can appear due to

What causes Bias in Models?

Bias can appear due to

1. biases in the data

What causes Bias in Models?

Bias can appear due to

1. biases in the data
    - less representative group

What causes Bias in Models?

Bias can appear due to

1. biases in the data
    - less representative group
    - the data is not a good representation of the reality (e.g. survivor bias, previous societal bias or limitations, etc.)

What causes Bias in Models?

Bias can appear due to

1. biases in the data
    - less representative group
    - the data is not a good representation of the reality (e.g. survivor bias, previous societal bias or limitations, etc.)
2. the data processing pipeline (extraction, cleaning, transformation, binning)

What causes Bias in Models?

Bias can appear due to

1. biases in the data
   - less representative group
   - the data is not a good representation of the reality (e.g. survivor bias, previous societal bias or limitations, etc.)
2. the data processing pipeline (extraction, cleaning, transformation, binning)
3. the model development process (including selection of algorithm, variables, etc.)

What causes Bias in Models?

Bias can appear due to

1. biases in the data
   - less representative group
   - the data is not a good representation of the reality (e.g. survivor bias, previous societal bias or limitations, etc.)
2. the data processing pipeline (extraction, cleaning, transformation, binning)
3. the model development process (including selection of algorithm, variables, etc.)
4. the particulars of model implementation

Conclusions

Fairness is a matter of perspective

Figure 1: Fairness is a matter of perspective

Fairness is a matter of perspective

Figure 1: Fairness is a matter of perspective

Bias is learnt

Case Study

Case study on addressing bias: car insurance