

Fairness Definitions Explained

Sahil Verma, IIT Kanpur, India

Julia Rubin, University of British Columbia, Canada

May 29th, 2018

Our Goals

- Collect definitions of fairness for the *algorithmic classification* problem
- Explain the *rationale* behind each definitions
- Demonstrate each definition on a single *unifying case-study*:
German Credit Dataset

German Credit Dataset

- Contains 1000 records from 1994
- Each record has ~20 attributes, such as, credit amount, duration, employment, age, marital status and gender
- Popular in fairness literature
- Ground truth: good or bad credit score

Question: Will a classifier trained on this data discriminate by gender?

Methodology

- Trained Logistic Regression classifier (Python) on 90% of the data; tested on 10% of the data (repeated for 10 folds)
- The data set does not contain single women
 - considered whether married/divorced men are treated similarly to married/divorced women

Coefficient Analysis

Attribute	Coefficient
Personal status and gender: single male	0.16
Personal status and gender: married male	-0.04
Personal status and gender: married/divorced female	-0.08
Personal status and gender: divorced male	-0.14

Considered Definitions

From NIPS, Big Data, AAAI, FATML, ICML, KDD, online reports

Statistical
Similarity-
Based
Causal Reasoning

	Definition	Paper	Citation #
3.1.1	Group fairness or statistical parity	[12]	208
3.1.2	Conditional statistical parity	[11]	29
3.2.1	Predictive parity	[10]	57
3.2.2	False positive error rate balance	[10]	57
3.2.3	False negative error rate balance	[10]	57
3.2.4	Equalised odds	[14]	106
3.2.5	Conditional use accuracy equality	[8]	18
3.2.6	Overall accuracy equality	[8]	18
3.2.7	Treatment equality	[8]	18
3.3.1	Test-fairness or calibration	[10]	57
3.3.2	Well calibration	[16]	81
3.3.3	Balance for positive class	[16]	81
3.3.4	Balance for negative class	[16]	81
4.1	Causal discrimination	[13]	1
4.2	Fairness through unawareness	[17]	14
4.3	Fairness through awareness	[12]	208
5.1	Counterfactual fairness	[17]	14
5.2	No unresolved discrimination	[15]	14
5.3	No proxy discrimination	[15]	14
5.4	Fair inference	[19]	6

Statistical Measures

- Vertically:
 - The ratio of “good” applicants who were assigned a good predicted credit score
 - The ratio of “good” applicants who were assigned a bad predicted credit score
 - The ratio of “bad” applicants who were assigned a good predicted credit score
 - ...

- Horizontally:
 - The ratio of applicants with good predicted score who actually have a good score
 - ...





	Actual – Positive	Actual – Negative
Predicted – Positive	True Positive (TP) $PPV = \frac{TP}{TP+FP}$ $TPR = \frac{TP}{TP+FN}$	False Positive (FP) $FDR = \frac{FP}{TP+FP}$ $FPR = \frac{FP}{FP+TN}$
Predicted – Negative	False Negative (FN) $FOR = \frac{FN}{TN+FN}$ $FNR = \frac{FN}{TP+FN}$	True Negative (TN) $NPV = \frac{TN}{TN+FN}$ $TNR = \frac{TN}{TN+FP}$

What is really fair?

- The ratio of “good” applicants who get the loan is the same for both males and females [equal opportunity]
- The ratio of “bad” applicants who do not get the loan is the same for both males and females [predictive equality]
- The same ratio of male and female applicants receives the loan [group fairness]
- The ratio of “good” applicants within the loan recipients is the same for both males and females [predictive parity]
- Anything else?

Mathematically, a classifier cannot satisfy all definitions at the same time when the base rates for a good credit score are different (72% and 65% for males and females in our case)

Experiments

-  The ratio of “good” applicants who get the loan is the same for males and females (86% for both)
-  The ratio of truly “good” males and females within those who got the loan is same (73% for males and 74% for females)
-  The ratio of male and female applicants who get the loan is not the same (81% for males and 75% for females)
-  “Bad” male applicants are more likely to be assigned with a good predicted credit score (70% for males and 55% for females)

Question ...

- Suppose we believe in group fairness: the same ratio of male and female applicants receives the loan.
- Are we happy?

Similarity-based Measures

- Fairness through unawareness:
 - Individuals that only differ in the sensitive attributes should get a similar classification.
 - No sensitive attributes are explicitly used in the decision-making process.
- Fairness through awareness
 - The similarity of individuals is defined via a distance metric.
 - The distance between the distributions of outputs for individuals should be at most the distance between the individuals.

Experiments

- For 8.8% “generated” identical applicants, the output classification is not the same
- Becomes “fair” when the gender attribute is excluded
- Distance metric affects the outcomes

Age difference	k	Avg. D	% violating cases
5	0.09	0.02	0.0
10	0.18	0.05	0.5
15	0.27	0.10	1.8
20	0.36	0.2	4.5
25	0.45	0.3	6.7

Conclusions

- Tens of definitions, some are satisfied and some are not
- Statistical definitions are easy to compute
 - But some rely on the availability of the actual outcome
- Similarity-based definitions are sensitive to the distance metric
- Understanding which definition is appropriate to a particular situation is challenging

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓
3.3.2	Well calibration	[16]	81	✓
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–