14/12/2022

Fairness in Machine Learning II

Ruta Binkyte

aivancity
SCHOOL FOR
TECHNOLOGY, BUSINESS & SOCIETY
PARIS-CACHAN

A bit of House keeping

Ruta Binkyte – *December 2022*

aivancity
PARIS-CACHAN

Part 2
12/14

Achieving Fairness:
Measuring and mitigating bias

Theory

- Fairness metrics: Group fairness, Individual fairness
- Mitigating bias: Pre-processing, In-processing, Post processing
- Fairness in the context of Trustworthy Machine Learning: synergies and tensions

Practice

- Fairness in business practice (guest speaker)
- Measuring bias practical exercises
- Mitigating bias practical exercises

Expected Outcome

Know the fairness metrics, mitigating approaches. Be able to choose and implement the most plausible technique for a given scenario.

Ruta Binkyte – *December 2022*

Fairness Metrics

Ruta Binkyte – *December 2022*

aivancity
PARIS-CACHAN

A reminder on features used for ML models

Y - The label in the data (Considered Binary $Y \in \{0,1\}$)
$\hat{Y}$ - The prediction (Considered Binary $\hat{Y} \in \{0,1\}$)
$X_1 \ldots X_n$ - The attributes (Features)
S - The sensitive attribute (also known as A) such as gender, race, age, sexual orientation etc. (Considered Binary $S \in \{0,1\}$)
A Proxy - An attribute that correlates with other feature we want to use or predict. When it is correlated with the sensitive attribute and used in the prediction we call it proxy discrimination. For example, medical spending correlates with race.

Statistical Parity Difference (Total Variation)

$$\boxed{P(\hat{Y} = 1 \mid S = 0) - P(\hat{Y} = 1 \mid S = 1)}$$

Where $\hat{Y} = 1$ is a desirable outcome prediction,
$S = 1$ is the privileged group of the Sensitive Attribute and $S = 0$ is the unprivileged group of the Sensitive Attribute.

- Negative value indicate discrimination towards the unprivileged group and positive value indicates the discrimination towards the privileged group.
- Value equal to zero indicates Fairness
- Can be used on the data (with Y) or the predictions (with $\hat{Y}$)
- A group fairness notion

Example: Statistical Parity Difference

Gender (S)	Decision (Y)	Prediction ($\hat{Y}$)
M	1	1
F	1	0
M	0	1
F	0	0
M	1	1
F	0	0

$$P(\hat{Y}=1 \mid S=0) - P(\hat{Y}=1 \mid S=1)$$

= 0/3 - 3/3 = -1

$$P(Y=1 \mid S=0) - P(Y=1 \mid S=1)$$

= 1/3 - 2/3 = - 0.33

aivancity
PARIS-CACHAN

Conditional Statistical Parity Difference

$$P(\hat{Y} = 1 \,|\, S = 0, E = e) - P(\hat{Y} = 1 \,|\, S = 1, E = e)$$

Where $\hat{Y} = 1$ is a desirable outcome prediction,
$S = 1$ is the privileged group of the Sensitive Attribute and $S = 0$ is the unprivileged group of the Sensitive Attribute and E=e is an explanatory attribute.

- Negative value indicate discrimination towards the unprivileged group and positive value indicates the discrimination towards the privileged group.
- Value equal to zero indicates Fairness
- Can be used on the data (with Y) or the predictions (with $\hat{Y}$)
- A group fairness notion

Example: Conditional Statistical Parity Difference

Gender (S)	Merit (E)	Decision (Y)	Prediction $\hat{Y}$
M	1	1	1
F	1	1	0
M	0	0	1
F	0	0	0
M	1	1	1
F	0	0	0

$$P(\hat{Y} = 1 \mid S = 0, E = 1) - P(\hat{Y} = 1 \mid S = 1, E = 1)$$

= 0/1 - 2/2 = - 2/2 = - 1

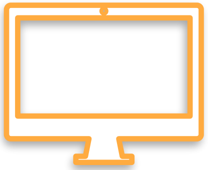
$$P(Y = 1 \mid S = 0, E = 1) - P(Y = 1 \mid S = 1, E = 1)$$

= 1/1 - 2/2 = 0

aivancity PARIS-CACHAN

Statistical Parity or Conditional Statistical Parity?

Back to running examples



- Who should get the job?

  - Statistical Parity?
  - Conditional Statistical Parity?
  - Example of Explanatory variable

- Who should get medical priority?

  - Statistical Parity?
  - Conditional Statistical Parity?
  - Example of Explanatory variable

- Who should be recognised?
  Eg. Smart Phone screen lock

  - Statistical Parity?
  - Conditional Statistical Parity?
  - Example of Explanatory variable

Disparate Impact

$$\boxed{\frac{P(\hat{Y}=1 \mid S=0)}{P(\hat{Y}=1 \mid S=1)}}$$

Where $\hat{Y}=1$ is a desirable outcome prediction,
$S=1$ is the privileged group of the Sensitive Attribute and $S=0$ is the unprivileged group of the Sensitive Attribute.

- Smaller value indicate discrimination towards the unprivileged group and larger value indicates the discrimination towards the privileged group.
- Value equal to one indicates Fairness
- Can be used on the data (with Y) or the predictions (with $\hat{Y}$)
- A group fairness notion

Example: Disparate Impact

Gender (S)	Decision (Y)	Prediction ($\hat{Y}$)
M	1	1
F	1	0
M	0	1
F	0	0
M	1	1
F	0	0

$$\frac{P(\hat{Y}=1 \mid S=0)}{P(\hat{Y}=1 \mid S=1)} = 0/ 1= 0$$

$$\frac{P(Y=1 \mid S=0)}{P(Y=1 \mid S=1)} = 0.33/ 0.67 = 0.49$$

aivancity
PARIS-CACHAN

Equal Opportunity Difference

$$P(\hat{Y} = 1 \mid S = 0, Y = 1) - P(\hat{Y} = 1 \mid S = 1, Y = 1)$$

Where $\hat{Y} = 1$ is a desirable outcome prediction, Y= 1 is desirable outcome label (ground truth), $S = 1$ is the privileged group of the Sensitive Attribute and $S = 0$ is the unprivileged group of the Sensitive Attribute.

- Negative value indicate discrimination towards the unprivileged group and positive value indicates the discrimination towards the privileged group.
- Value equal to zero indicates Fairness
- Used with ground truth label (Y) and the predictions $(\hat{Y})$
- A group fairness notion

Example: Equal Opportunity Difference

Gender (S)	Decision (Y)	Prediction ($\hat{Y}$)
M	1	1
F	1	0
M	0	1
F	0	0
M	1	1
F	0	0

$$P(\hat{Y} = 1 \mid S = 0, Y = 1) - P(\hat{Y} = 1 \mid S = 1, Y = 1)$$

= 0/1 - 2/2 = - 1

aivancity
PARIS-CACHAN

Confusion Matrix Metrics

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

https://en.wikipedia.org/wiki/Confusion_matrix

Example: Confusion Matrix

Gender (S)	Decision (Y)	Prediction ($\hat{Y}$)
M	1	1
F	1	0
M	0	1
F	0	0
M	1	1
F	0	0

TP = 2/6

TN = 2/6

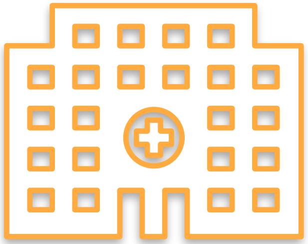
FP = 1/6

FN = 1/6

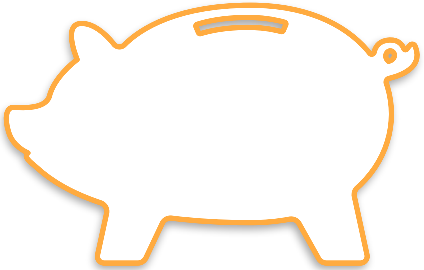
aivancity
PARIS-CACHAN

False Positive or False Negative?

Back to running examples

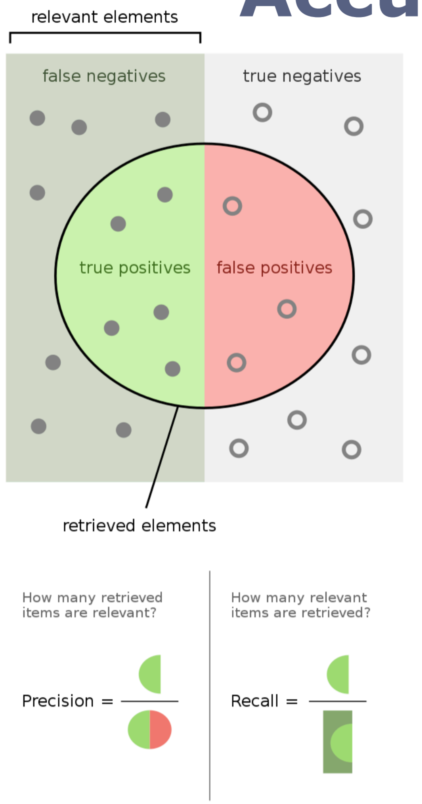


- **Predicting the risk for cancer**
  - False Negative means that someone is sick and not diagnosed. The patient is not informed or risk.
  - False Positive means that someone is not sick and predicted as sick. The patient needs to undergo further examination.



- **Predicting the ability to repay the loan**
  - False Negative means that someone is eligible for the loan, but does not get it. The client may need to reapply after increasing income or savings.
  - False Positive means that someone who is not able to repay is given a loan. The bank may loose money and the client may get indebted.

Accuracy, Precision and Recall



https://en.wikipedia.org/wiki/Precision_and_recall

$$Precision(TruePositiveRate) = \frac{TP}{TP + FP}$$

$$TrueNegativeRate = \frac{TN}{TN + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Possible Fairness Metrics based on Confusion Matrix

$$Precision_{S=0} - Precision_{S=1}$$

$$TrueNegativeRate_{S=0} - TrueNegativeRate_{S=1}$$

$$Recall_{S=0} - Recall_{S=1}$$

$$Accuracy_{S=0} - Accuracy_{S=1}$$

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Average Odds Difference

$$\frac{(FPR_{S=0} - FPR_{S=1}) + (TPR_{S=0} - TPR_{S=1})}{2}$$

Where FPR is false positive rate, TPR is true positive rate $S = 1$ is the privileged group of the Sensitive Attribute and $S = 0$ is the unprivileged group of the Sensitive Attribute.

- Value equal to zero indicates Fairness
- Requires both Y and $\hat{Y}$
- A group fairness notion

Individual Fairness Through Unawareness

$$\boxed{\hat{Y} = f(X)}$$

Where $\hat{Y}$ is a prediction and X is the set of features NOT including the sensitive attribute

- Requires to remove S from the training data
- Does not account for proxy variables
- Problematic to measure
- An individual fairness notion

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 1-21.

Individual Fairness Through Awareness

$$dist_Y(\hat{y}_i, \hat{y}_j) < L \times dist_{\tilde{X}}(\tilde{x}_i, \tilde{x}_j),$$

Where $dist_Y(\hat{y}_i, \hat{y}_j)$ is distance in prediction space, $dist_{\tilde{X}}(\tilde{x}_i, \tilde{x}_j)$ is distance in feature space and L is constant. Intuitively similar individuals to should have similar outcomes.

- Tricky to define the distance metric
- The similarity must include only relevant features
- Does not account for influence of S on values of X
- An individual fairness notion

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 1-21.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Fairness Metrics: *Summary*

- Group - Statistical Parity Difference, Disparate Impact, Conditional Statistical Parity Difference, Equal Opportunity Difference.
- Confusion Matrix - eg. True Positive Rate Difference, False Positive Rate Difference, Average Odds Difference.

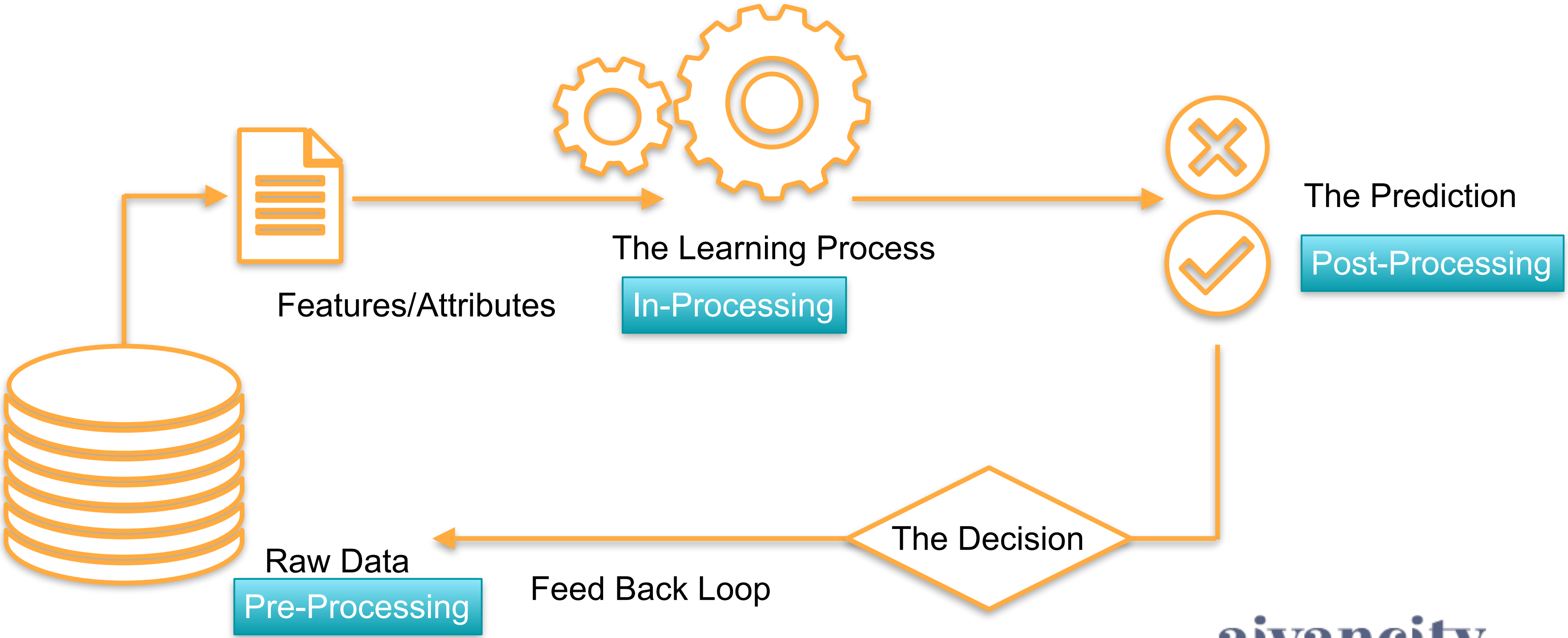
  Individual - Fairness Through Awareness, Fairness Through Unawareness.

- The choice of metric is contextual and depends on situation.
- Group Fairness metrics are more popular, because they are easier to define and implement.

Mitigating Bias

Ruta Binkyte – *December 2022*

Machine Learning Process



Pre-Processing Methods

- **Reweighing Pre-Processing**: *Generates weights for the training samples in each (group, label)* combination differently to ensure fairness before classification. It does not change any feature or label values, so this is ideal if you are unable to make value changes.

- **Optimized Pre-Processing:** *Learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual sample distortion, and data utility constraints and objectives.*

- **Disparate Impact Remover:** *Edits feature values to increase group fairness while preserving rank ordering within groups*.

Mahoney, T., Varshney, K., & Hind, M. (2020). *AI Fairness*. O'Reilly Media, Incorporated.

Pre-Processing Methods: Zoom In on Reweighing

$$Weight_{S=s,Y=y} = \frac{N_{S=s} \times N_{Y=y}}{N_{all} \times N_{S=s,Y=y}}$$

- Can be used with classifiers that can handle row-level weights, otherwise weights can be used for oversampling
- Satisfies Disparate Impact fairness notion
- Applied on data before training process (pre-processing)
- Data with weights can be safely used without explicit Sensitive attribute in the dataset

F. Kamiran and T. Calders, “Data Preprocessing Techniques for Classification without Discrimination,” Knowledge and Information Systems, 2012.

Pre-Processing Methods: Zoom In on Reweighing

Gender (S)	Decision (Y)
M	1
F	1
M	0
F	0
M	1
F	0

$$Weight_{S=s,Y=y} = \frac{N_{S=s} \times N_{Y=y}}{N_{all} \times N_{S=s,Y=y}}$$

$N_{S=0} = 3$ $N_{S=0,Y=0} = 2$ $N_{S=0,Y=1} = 1$

$N_{S=1} = 3$ $N_{S=1,Y=0} = 1$ $N_{S=1,Y=1} = 2$

$N_{all} = 6$ $N_{Y=1} = 3$ $N_{Y=0} = 3$

F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012.

Pre-Processing Methods: Zoom In on Reweighing

Gender (S)	Decision (Y)
M	1
F	1
M	0
F	0
M	1
F	0

$$Weight_{S=0,Y=1} = \frac{3 \times 3}{6 \times 1} = 1.5$$

Higher weight !

$$Weight_{S=1,Y=1} = \frac{3 \times 3}{6 \times 2} = 0.75$$

$N_{S=0} = 3$ $N_{S=0,Y=0} = 2$ $N_{S=0,Y=1} = 1$

$N_{S=1} = 3$ $N_{S=1,Y=0} = 1$ $N_{S=1,Y=1} = 2$

$N_{all} = 6$ $N_{Y=1} = 3$ $N_{Y=0} = 3$

F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012.

In-Processing Methods

- **Adversarial Debiasing**: Learns a classifier to maximize prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier because the predictions can't carry any group discrimination information that the adversary can exploit.
- **Prejudice Remover:** Adds a discrimination-aware regularization term to the learning objective.
- **Meta-Fair Classifier :** Optimises classifier for more than one fairness metric.

Mahoney, T., Varshney, K., & Hind, M. (2020). *AI Fairness*. O'Reilly Media, Incorporated.

In-Processing Methods: Zoom In on Prejudice Remover

$$PI = \sum_{Y,A \in D} P(Y,A) ln(\frac{P(Y,A)}{P(Y)P(A)})$$

$$\min_{f} L(f(X), Y) + \eta PI$$

- Can be used with any discriminative probabilistic classifier
- Added to the optimisation function as part of the learning process (in-processing)
- Measures mutual information between the sensitive attribute and the label and penalises the dependency between the two

Post-Processing Methods

- **Reject Option Classification**: Gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

- **Equalized Odds:** finds a classification threshold with which output labels change to satisfy Equalized Odds.

Mahoney, T., Varshney, K., & Hind, M. (2020). *AI Fairness*. O'Reilly Media, Incorporated.

Bias Mitigation: *Summary*

- Pre-processing can be used with any classifier.
- In-processing make the trade-offs explicit.
- Post-processing is closest to the decision making.
- No one method is proved to perform better than others.
- The choice depends on dataset characteristics.
- A good practice is to try several methods on your dataset to see which one performs better.

Fairness Tensions

Ruta Binkyte – *December 2022*

Fairness and Accuracy

- Often fairness mitigation results in decreased Accuracy

- Mitigating for fairness degrades Accuracy, but it is a fundamental question should we aim for Accuracy towards labels indicating historical biases?

- In some cases Fairness mitigation can actually increase Accuracy (Remember Medical Expenditure and Face Recognition bias examples)

Fairness and Privacy

- No learning algorithm can simultaneously satisfy − differential privacy and guarantee to generate a fair (equal opportunity) classifier which is non-trivial. (*Cummings, R., Gupta, V., Kimpara, D., Morgenstern, J.: On the compatibility of pri-vacy and fairness. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 309–315 (2019)*)

- Some new results show increase in unfair bias when applying differential privacy on gradients (*Esipova, M. S., Ghomi, A. A., Luo, Y., & Cresswell, J. C. (2022). Disparate Impact in Differential Privacy from Gradient Misalignment. arXiv preprint arXiv:2206.07737.*)

A mechanism M is (ε, δ)-differentially
private if for all adjacent databases
5 D, D′ and for every measurable S
⊆ Range(M) holds that:

$$P\,[M(D) \in S\,] \leq e^{\epsilon}\, P\,[M(D') \in S] + \delta$$

In a nutshell, differential privacy ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis.

Fairness and … Fairness

- Some Fairness notions are fundamentally incompatible because they represent different worldviews and values (Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.)

- There is no one Fairness Notion that can be applicable every time and in every situation.

SP and EP Difference

Gender (S)	Decision (Y)	Prediction ($\hat{Y}$)
M	1	1
F	1	0
M	0	1
F	0	0
M	1	1
F	0	0

Statistical Parity

$$P(\hat{Y} = 1 \mid S = 0) - P(\hat{Y} = 1 \mid S = 1)$$

= 0/3 - 3/3 = -1

Equal Opportunity Rate

$$P(\hat{Y} = 1 \mid S = 0, Y = 1) - P(\hat{Y} = 1 \mid S = 1, Y = 1)$$

= 0/1 - 2/2 = - 1

aivancity
PARIS-CACHAN

SP and EP Difference

Gender (S)	Decision (Y)	Prediction ($\hat{Y}$)
M	1	1
F	1	1
M	0	1
F	0	0
M	1	1
F	0	0

Statistical Parity

$$P(\hat{Y} = 1 \mid S = 0) - P(\hat{Y} = 1 \mid S = 1)$$

= 1/3- 3/3 = - 0.67

Equal Opportunity Rate

$$P(\hat{Y} = 1 \mid S = 0, Y = 1) - P(\hat{Y} = 1 \mid S = 1, Y = 1)$$

= 1/1 - 2/2 = 0

aivancity
PARIS-CACHAN

Example: CSP and SP Difference

Gender (S)	Merit (E)	Decision (Y)	Prediction $\hat{Y}$
M	1	1	1
F	1	1	0
M	0	0	1
F	0	0	0
M	1	1	1
F	0	0	0

Conditional Statistical Parity

$$P(\hat{Y}=1 \mid S=0, E=1) - P(\hat{Y}=1 \mid S=1, E=1)$$

= 0/1 - 2/2 = - 2/2 = - 1

$$P(Y=1 \mid S=0, E=1) - P(Y=1 \mid S=1, E=1)$$

= 1/1 - 2/2 = 0

Statistical Parity

$$P(\hat{Y}=1 \mid S=0) - P(\hat{Y}=1 \mid S=1)$$

= 0/3 - 3/3 = -1

$$P(Y=1 \mid S=0) - P(Y=1 \mid S=1)$$

= 1/3 - 2/3 = -0.33

aivancity PARIS-CACHAN

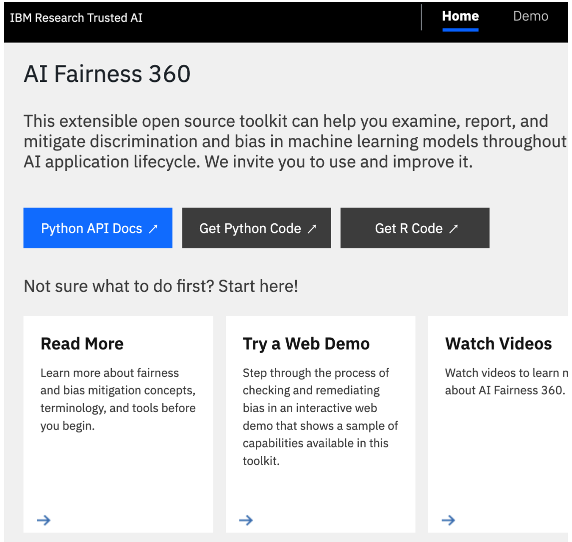
Practical Exercises

Ruta Binkyte – *December 2022*

Bias Measurement and Mitigation

The libraries

- **AI Fairness 360 (IBM) https://aif360.mybluemix.net/**
- **Fair-learn (Microsoft) https://github.com/fairlearn/fairlearn**
- **What-If-Tool (Google) https://github.com/PAIR-code/what-if-tool**

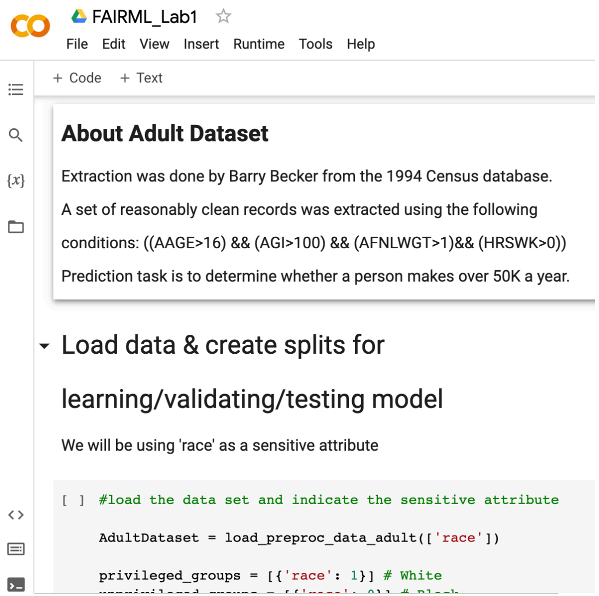


Bias Measurement and Mitigation

The exercise

Access the Colab FAIRML_Lab1Notebook here

- Make a copy of FAIRML_Lab1 on your drive
- Run the examples
- Implement exercises and answer the questions
- Upload your notebook with the output to Blackboard

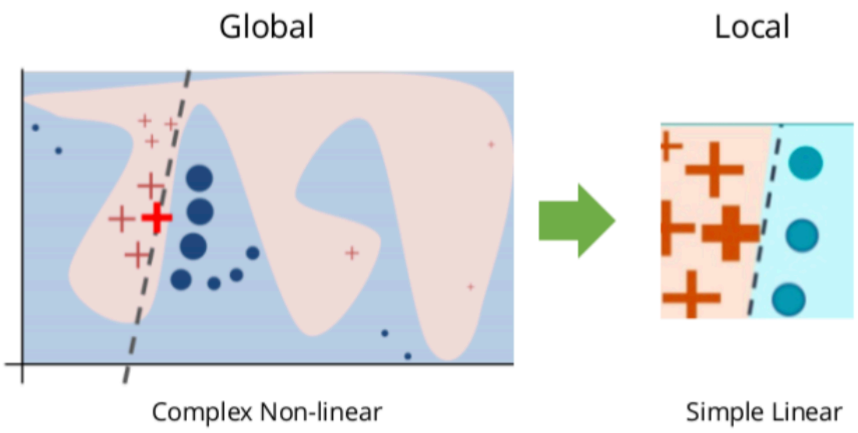


Ruta Binkyte – *December 2022*

Short Introduction to LIME

Local Interpretable Model-Agnostic Explanations

- Provide Local Explanations.
- Approximates black box model locally by an interpretable model.
- Model Agnostic



https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

Reading Homework

Machine learning fairness notions: Bridging the gap with real-world applications

arXiv > cs > arXiv:2006.16745

Searc

Computer Science > Machine Learning

[Submitted on 30 Jun 2020 (v1), last revised 7 Jun 2022 (this version, v5)]

Machine learning fairness notions: Bridging the gap with real-world applications

Karima Makhlouf, Sami Zhioua, Catuscia Palamidessi

Fairness emerged as an important requirement to guarantee that Machine Learning (ML) predictive systems do not discriminate against specific individuals or e
populations, in particular, minorities. Given the inherent subjectivity of viewing the concept of fairness, several notions of fairness have been introduced in the
This paper is a survey that illustrates the subtleties between fairness notions through a large number of examples and scenarios. In addition, unlike other surv
literature, it addresses the question of: which notion of fairness is most suited to a given real-world scenario and why? Our attempt to answer this question co
identifying the set of fairness-related characteristics of the real-world scenario at hand, (2) analyzing the behavior of each fairness notion, and then (3) fitting
elements to recommend the most suitable fairness notion in every specific setup. The results are summarized in a decision diagram that can be used by practi
policymakers to navigate the relatively large catalog of ML.

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Computers and Society (cs.CY); Machine Learning (stat.ML)
Cite as: arXiv:2006.16745 **[cs.LG]**
(or arXiv:2006.16745v5 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2006.16745
Journal reference: Information Processing and Management, 58(5). pp. 107-132 (2021)
Related DOI: https://doi.org/10.1016/j.ipm.2021.102642

https://arxiv.org/abs/2006.16745

Ruta Binkyte – *December 2022*

aivancity
PARIS-CACHAN

The Fairness Guidelines

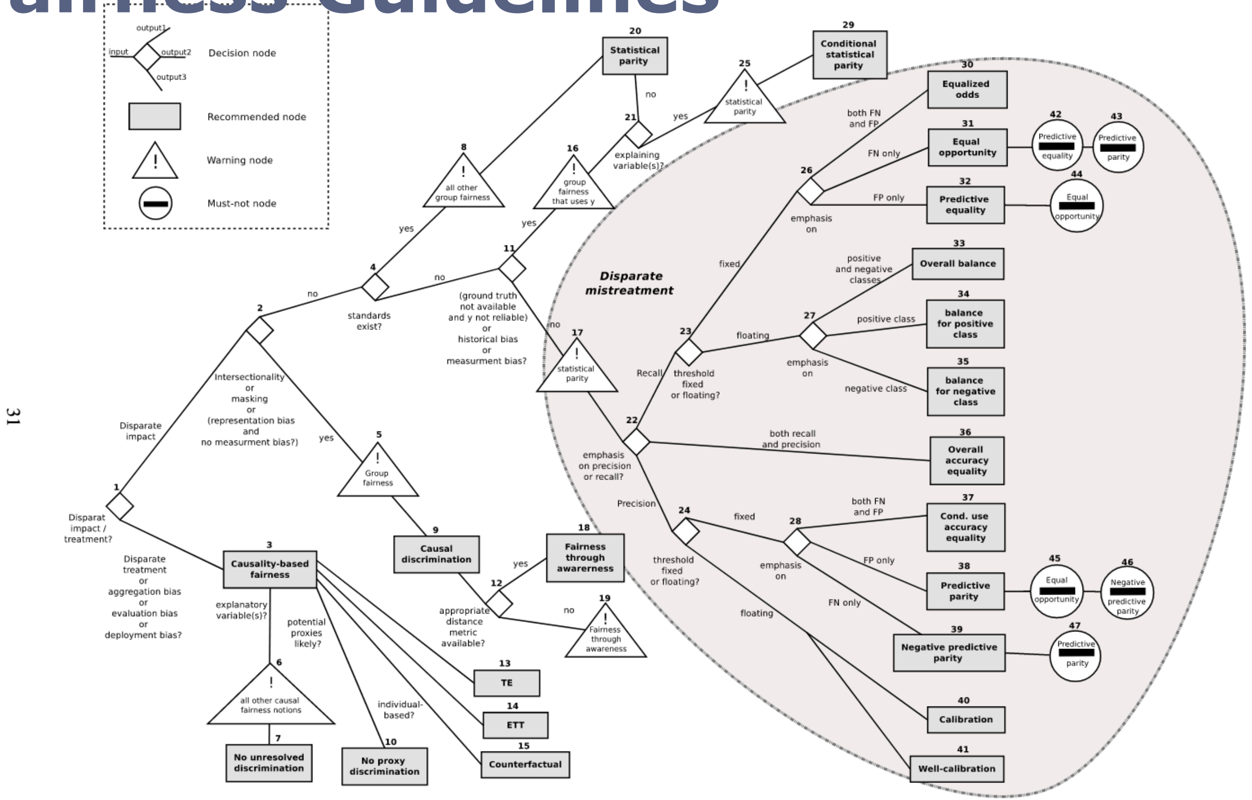


Figure 6: Fairness notions applicability decision diagram.

https://arxiv.org/abs/2006.16745

aivancity

PARIS-CACHAN

advancing education
in artificial intelligence