THALES
Building a future we can all trust

Unbiased and private machine learning

Session 4: Bias mitigation with pre-processing and post-processing

Alice HELIOU and Vincent THOUVENOT
Laboratoire Cyber de cortAIx Labs

Disclaimer

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of Thales. Thales cannot be held responsible for them.

Content

1. Lessons journey
2. Pre-processing methods
3. Post-processing methods
4. AIF360

Lessons journey

Program

- **05/01 - Section 1:** Introduction to bias, fairness and data analysis
- **12/01 - Section 2:** Data analysis for bias detection
- **19/01 - Mi-Project début**
- **26/01 - Section 3:** Model auditing, explainability and interpretability
- **02/02 - Section 4: Bias mitigation with pre-processing and post-processing** methods
- **09/02 - Mi-Project fin**
- **16/02 - Section 5:** Bias mitigation with in-processing methods
- **09/03 - Project début**
- **16/03 - Section 6:** Data privacy, metrics and countermeasures
- **23/03 - Project fin**
- **30/03 - Section 7:** Perspectives with unlearning
- **06/04 - Soutenance**

Just Remove the Sensitive Attribute?

Example: A bank is looking for a fair method to decide who to grant a loan to

- Bank's goal: determine who will repay the loan;
- The loan is intended for two distinct groups of people;
- Simply removing the sensitive attribute from the model can be ineffective due to indirect interactions with other attributes.

Just removing the sensitive attribute is not the optimal solution: keeping it can be more effective!

Bias Correction Methods

- **Pre-processing:** Modification of training databases
- **In-processing:** Bias correction learned simultaneously with the machine learning model
- **Post-processing:** Modification of machine learning model predictions

Bias Correction Methods

- **Pre-processing: Modification of training databases**
- **In-processing:** Bias correction learned simultaneously with the machine learning model
- **Post-processing: Modification of machine learning model predictions**

Pre-processing methods

Massaging

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

- S sensitive attributes taking values b (disadvantaged group) and w (privileged group);
- Label takes $+$ and $-$ as values;
- Change the labels of certain observations:
  - For some observations where $S = b$, change the label from $-$ to $+$;
  - For some observations where $S = w$, change the label from $+$ to $-$.

Massaging

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

Algorithm 2: *Rank*

Input: Labeled dataset D, Sensitive attribute and value S, b, desired class $+$
Output: Ordered promotion list pr and demotion list dem
1: Learn a ranker R for prediction $+$ using D as training data
2: $pr := \{X \in D \mid X(S) = b, X(Class) = -\}$
3: $dem := \{X \in D \mid X(S) = w, X(Class) = +\}$
4: Order pr descending w.r.t. the scores by R
5: Order dem ascending w.r.t. the scores by R
6: **return** (pr, dem)

Data Reweighting

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

Algorithm 3: *Reweighing*

Input: $(D, S, Class)$
Output: Classifier learned on reweighed D
1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: $D_W := \{\}$
7: **for** X in D **do**
8: Add $(X, W(X(S), X(Class)))$ to D_W
9: **end for**
10: Train a classifier C on training set D_W, taking onto account the weights
11: **return** Classifier C

Data Reweighting

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

Sex	Ethnicity	Highest degree	Job type	Class
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	–
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Education	–
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Board	+

- What weight is given to a woman in the positive class?

Data Reweighting

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

Sex	Ethnicity	Highest degree	Job type	Class
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	–
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Education	–
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Board	+

- What weight is given to a woman in the positive class?
- $\frac{0.5 \times 0.6}{0.2} = 1.5$

Data Reweighting

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

Sex	Ethnicity	Highest degree	Job type	Class
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	–
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Education	–
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Board	+

- Weight given to a woman in the positive class: 1.5
- Weight given to a woman in the negative class: 0.67
- Weight given to a man in the positive class: 0.75
- Weight given to a man in the negative class: 2

Uniform Sampling

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

Algorithm 4: *Uniform Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D
1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: Sample uniformly $W(b, +) \times |DP|$ objects from DP;
7: Sample uniformly $W(w, +) \times |FP|$ objects from FP;
8: Sample uniformly $W(b, -) \times |DN|$ objects from DN;
9: Sample uniformly $W(w, -) \times |FN|$ objects from FN;
10: Let D_{US} be the bag of all samples generated in steps 6 to 9
11: **return** Classifier C learned on D_{US}

Uniform Sampling

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

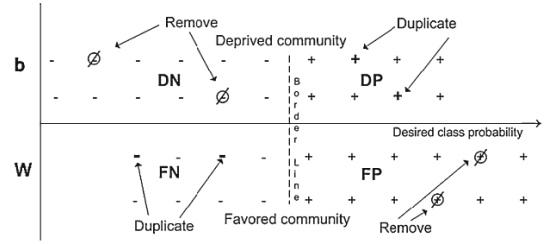


Fig. 3 Pictorial representation of the *Uniform Sampling* scheme. The re-substituted data points are in *bold* while the *encircled* ones are skipped

Preferential Sampling

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

Algorithm 5: *Preferential Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D
1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: Learn a ranker R for predicting $+$ using D as training set
7: $D_{PS} := \{\}$
8: Add $\lfloor W(b, +) \rfloor$ copies of DP to D_{PS}
9: Add $\lfloor W(b, +) - \lfloor W(b, +) \rfloor \times |DP| \rfloor$ lowest ranked elements of DP to D_{PS}
10: Add $\lfloor W(b, -) \times |DN| \rfloor$ lowest ranked elements of DN to D_{PS}
11: Add $\lfloor W(w, +) \times |FP| \rfloor$ highest ranked elements of FP to D_{PS}
12: Add $\lfloor W(w, -) \rfloor$ copies of FN to D_{PS}
13: Add $\lfloor W(w, -) - \lfloor W(b, -) \rfloor \times |FN| \rfloor$ highest ranked elements of FN to D_{PS}
14: **return** Classifier C learned on D_{PS}

Preferential Sampling

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

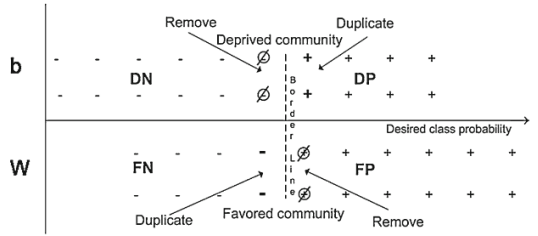


Fig. 4 Pictorial representation of *Preferential Sampling* scheme. The re-substituted data points are in *bold* while the *encircled* ones are skipped

Disparate Impact Remover

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.

Definition 4.1 (BER). *Let $f : Y \to X$ be a predictor of X from Y. The* balanced error rate BER *of f on distribution $\mathcal{D}$ over the pair (X, Y) is defined as the (unweighted) average class-conditioned error of f. In other words,*

$$\text{BER}(f(Y), X) = \frac{\Pr[f(Y) = 0 | X = 1] + \Pr[f(Y) = 1 | X = 0]}{2}$$

Definition 4.2 (Predictability). *X is said to be* ϵ-predictable *from Y if there exists a function $f : Y \to X$ such that*

$$\text{BER}(f(Y), X) \leq \epsilon.$$

This motivates our definition of ϵ-fairness, as a data set that is *not* predictable.

Definition 4.3 (ϵ-fairness). *A data set $D = (X, Y, C)$ is said to be* ϵ-fair *if for any classification algorithm $f : Y \to X$*

$$\text{BER}(f(Y), X) > \epsilon$$

with (empirical) probabilities estimated from D.

Disparate Impact Remover

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.

- A certification showed a disparate impact issue on D;
- Search for a new dataset $\tilde{D}$ where all elements of D have been changed so that $\tilde{D}$ is ε-fair

Disparate Impact Remover

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.

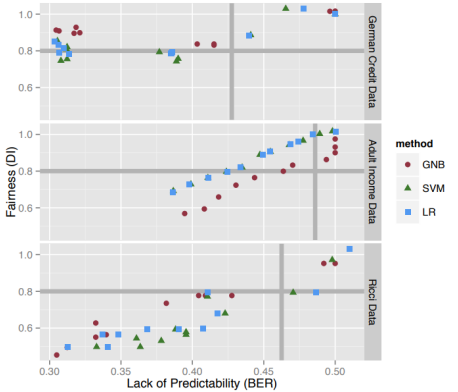
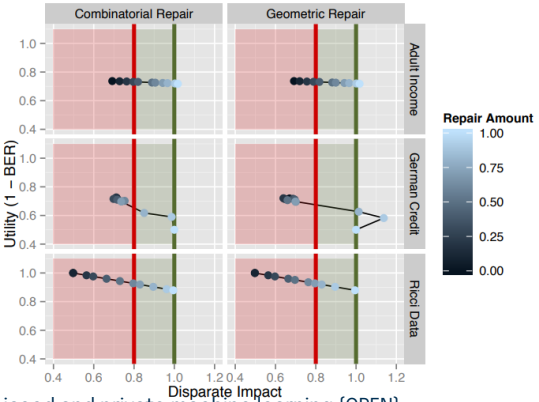


Fig. 2: Lack of predictability (BER) of the protected attributes on the German Credit Adult

Disparate Impact Remover

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.



Disparate Impact Remover

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.

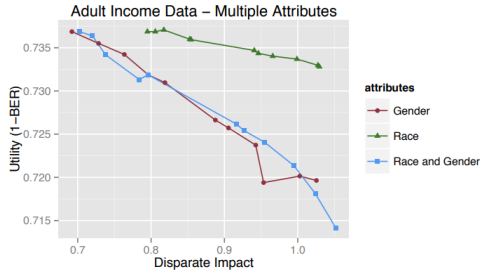
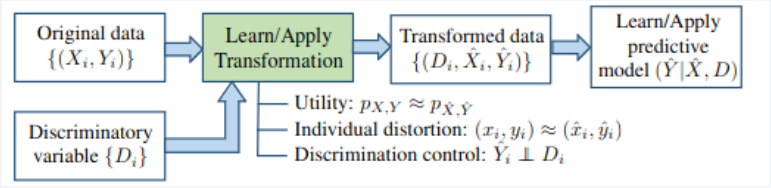


Fig. 4: Disparate impact (DI) vs. utility (1-BER) from our combinatorial and geometric partial repair processes using the SVM as the classifier. For clarity in the figure, only the combinatorial repairs are shown, though the geometric repairs follow the same pattern.

Optimized Pre-processing

F. P. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. "Optimized Pre-Processing for Discrimination Prevention." Conference on Neural Information Processing Systems, 2017.



Solving an optimization problem under three constraints:

- Preserve data utility;
- Limit data distortion;
- Control discrimination.

Fair Latent Representation Learning

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations." International Conference on Machine Learning, 2013.

- Learn a representation that can be used for making accurate predictions without bias from sensitive information
- Approaches
  - With Adversarial Learning
  - Without Adversarial Learning
    - Clustering methods with probabilistic mapping
    - Variational Fair AutoEncoder with Maximum Mean Discrepancy measure
    - Orthogonal Disentangled Fair Representation with orthogonal priors to enforce an orthogonality constraint between sensitive and non sensitive representation

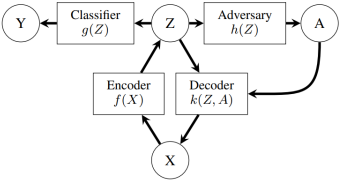


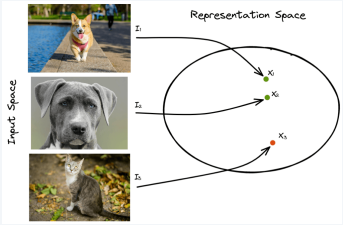
Figure 1. Model for learning adversarially fair representations. The variables are data X, latent representations Z, sensitive attributes A, and labels Y. The encoder f maps X (and possibly A - not shown) to Z, the decoder k reconstructs X from (Z, A), the classifier g predicts Y from Z, and the adversary h predicts A from Z (and possibly Y - not shown).

Fair Latent Representation Learning

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations." International Conference on Machine Learning, 2013.

- Learning low-dimensional representations of data by contrasting between similar and dissimilar samples
- Contrastive Loss
  - Pair of observations (I_i, I_j), a label $Y = 0$ if samples are similar, 1 otherwise, f is CNN that encodes input I_i and I_j into an embedding space such that $x_i = f(I_i)$ and $x_j = f(I_j)$
    $(1 - Y) \times ||x_i - x_j||^2 + Y \times \max(0, m - ||x_i - x_j||^2)$
  - Triplet Loss
    - Anchor sample I, positive sample I^+, negative sample I^-

$$\max(0, ||x - x^+||^2 - ||x - x^-||^2 + m)$$



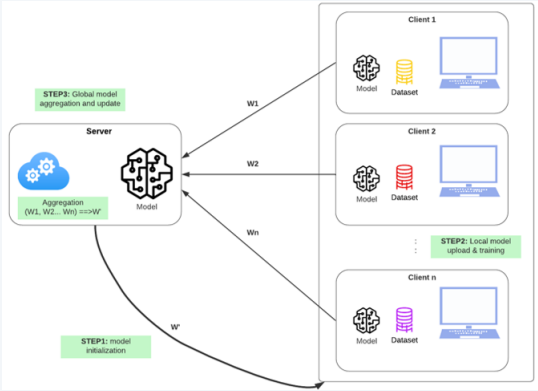
Federated Learning Case

- Case where several data owners (e.g. hospitals) want to collaborate to learn a Machine Learning model together...
- ...but cannot share the data with each other.

Federated Learning Case

	Distributed Learning	Federated Learning
Data	I.I.D. data, potentially accessible to the server	Non-I.I.D. data. Sensitive and local data. The distribution cannot be chosen by the server. Potentially, many data sources
Objectives	Scaling up, accelerating computations without performance loss	Collaborating without leaking sensitive information. Trade-off between security/privacy and model performance

Federated Learning Case



1. The central server sends the latest model update to the participants;
2. Each participant updates the model weights with their local data;
3. The updated weights are sent by all participants to the central server;
4. The central server aggregates the locally updated weights.

Federated Learning Case

Internship by Zachary Fakir under the supervision of Alice Héliou

Several sources of bias in federated learning:

- Bias in the selection of federated learning participants;
- Heterogeneous data;
- Aggregation algorithm.

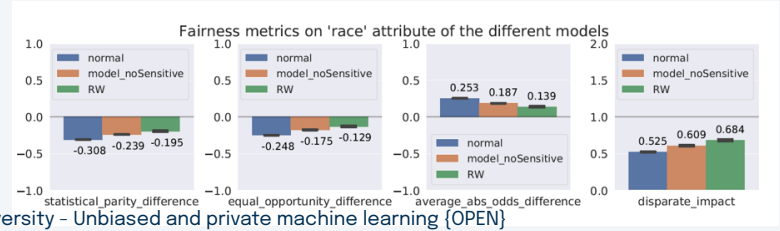
Methods	Privacy	Additional Communication	Hyperparameters
Local Reweighing	Same as plain FL	0	none
Global Reweighing with Differential Privacy	ϵ-Differential privacy	1.5	ϵ
Federated Prejudice Removal	Same as plain FL	0	η

- Local reweighting produces fair models without sacrificing privacy nor model accuracy...
- ...But with lack of stability

Federated Learning Case

Internship by Zachary Fakir under the supervision of Alice Héliou

- Combining Federated Learning and Fairness approaches to protect sensitive features:
  - Use cases: a set of sensitive attributes for which we observe biases in the dataset
  - Example: ProPublica Compass dataset containing information on 7,215 people arrested. The sensitive and biased attribute considered is race.
- The privacy of sensitive attributes can be improved by a preprocessing Fairness approach (Local Reweighting)
  - Naively removing the sensitive attribute is not enough



Federated Learning Case

Internship by Zachary Fakir under the supervision of Alice Héliou

Division			AAOD (10^{-2})		Accuracy (%)		Attack precision (%)	
			RW	noRW	RW	noRW	RW	noRW
race	60-40	P0 50C-33AF (1989)	9.5 ± 4.0	36 ± 10	66 ± 1.4	66 ± 0.5	$\mathbf{66 \pm 1.7}$	71 ± 2.5
		P1 33C-50AF (1989)	13 ± 7.0	27 ± 1.9	66 ± 2.0	68 ± 2.0	$\mathbf{65 \pm 1.4}$	68 ± 0.7
		FL 41C-41AF (3978)	$\mathbf{8.4 \pm 1.9}$	30 ± 0.3	$\mathbf{69 \pm 0.1}$	68 ± 0.1	$\mathbf{67 \pm 0.6}$	68 ± 0.1
	90-10	P0 74C-8AF (1989)	18 ± 4.4	31 ± 7.3	65 ± 0.5	67 ± 1.0	$\mathbf{66 \pm 1.9}$	69 ± 2.0
		P1 8C-74AF (1989)	19 ± 5.2	27 ± 5.8	67 ± 0.8	67 ± 0.8	$\mathbf{65 \pm 1.4}$	68 ± 0.8
		FL 41C-41AF (3978)	$\mathbf{12 \pm 2.1}$	28 ± 1.4	$\mathbf{69 \pm 0.1}$	68 ± 0.3	$\mathbf{66 \pm 0.5}$	67 ± 0.9
city		P0 58C-23AF (588)	18 ± 9.1	27 ± 9.0	64 ± 1.5	65 ± 0.2	$\mathbf{67 \pm 2.0}$	68 ± 3.2
		P1 45C-38AF (1275)	15 ± 4.2	33 ± 9.2	66 ± 1.0	67 ± 1.2	$\mathbf{64 \pm 0.6}$	70 ± 2.6
		P2 26C-60AF (2779)	14 ± 3.6	28 ± 4.7	66 ± 0.1	$\mathbf{68 \pm 0.2}$	$\mathbf{64 \pm 0.5}$	69 ± 1.2
		FL 35C-44AF (4642)	$\mathbf{12 \pm 1.1}$	27 ± 1.7	$\mathbf{68 \pm 0.1}$	$\mathbf{68 \pm 0.3}$	$\mathbf{66 \pm 0.5}$	67 ± 1.0

Post-processing methods

Reject-option Classification

F. Kamiran, A. Karim, and X. Zhang, "Decision Theory for Discrimination-Aware Classification," IEEE International Conference on Data Mining, 2012

Algorithm 1: Reject Option based Classification (ROC)

Input: $\{\mathcal{F}_k\}_{k=1}^{K}$ ($K \geq 1$ probabilistic classifiers trained on $\mathcal{D}$), $\mathcal{X}$ (test set), $\mathcal{X}^d$ (deprived group), θ
Output: $\{C_i\}_{i=1}^{M}$ (labels for instances in $\mathcal{X}$)
** Critical region **
$\forall X_i \in \{Z | Z \in \mathcal{X}, \max[p(C^+|Z), 1 - p(C^+|Z)] < \theta\}$
  If $X \in \mathcal{X}^d$ **then** $C_i = C^+$
  If $X \notin \mathcal{X}^d$ **then** $C_i = C^-$
** Standard decision rule **
$\forall X_i \in \{Z | Z \in \mathcal{X}, \max[p(C^+|Z), 1 - p(C^+|Z)] \geq \theta\}$
  $C_i = \text{argmax}_{\{C^+, C^-\}}[p(C^+|X_i), p(C^-|X_i)]$

Equalized-odds Post-processing

M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Conference on Neural Information Processing Systems, 2016.

- Equalized odds: $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}$;
- Equal opportunity: $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$

Equalized-odds Post-processing

M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Conference on Neural Information Processing Systems, 2016.

$$\gamma_a(\widehat{Y}) \stackrel{\text{def}}{=} \left(\Pr\left\{\widehat{Y} = 1 \mid A = a, Y = 0\right\}, \Pr\left\{\widehat{Y} = 1 \mid A = a, Y = 1\right\}\right). \tag{4.1}$$

The first component of $\gamma_a(\widehat{Y})$ is the *false positive rate* of $\widehat{Y}$ within the demographic satisfying $A = a$. Similarly, the second component is the *true positive rate* of $\widehat{Y}$ within $A = a$. Observe that we can calculate $\gamma_a(\widehat{Y})$ given the joint distribution of $(\widehat{Y}, A, Y)$. The definitions of equalized odds and equal opportunity can be expressed in terms of $\gamma(\widehat{Y})$, as formalized in the following straight-forward Lemma:

Lemma 4.2. *A predictor $\widehat{Y}$ satisfies:*

1. *equalized odds if and only if $\gamma_0(\widehat{Y}) = \gamma_1(\widehat{Y})$, and*
2. *equal opportunity if and only if $\gamma_0(\widehat{Y})$ and $\gamma_1(\widehat{Y})$ agree in the second component, i.e., $\gamma_0(\widehat{Y})_2 = \gamma_1(\widehat{Y})_2$.*

For $a \in \{0, 1\}$, consider the two-dimensional convex polytope defined as the convex hull of four vertices:

$$P_a(\widehat{Y}) \stackrel{\text{def}}{=} \text{convhull}\left\{(0,0), \gamma_a(\widehat{Y}), \gamma_a(1-\widehat{Y}), (1,1)\right\} \tag{4.2}$$

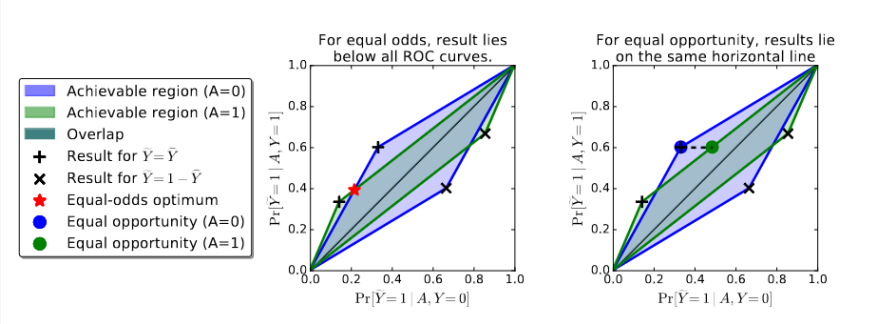
Equalized-odds Post-processing

M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Conference on Neural Information Processing Systems, 2016.

$$\begin{aligned} \min_{\widetilde{Y}} \quad & \mathbb{E}\ell(\widetilde{Y}, Y) \\ \text{s.t.} \quad & \forall a \in \{0,1\} : \gamma_a(\widetilde{Y}) \in P_a(\widehat{Y}) \\ & \gamma_0(\widetilde{Y}) = \gamma_1(\widetilde{Y}) \end{aligned}$$

Equalized-odds Post-processing

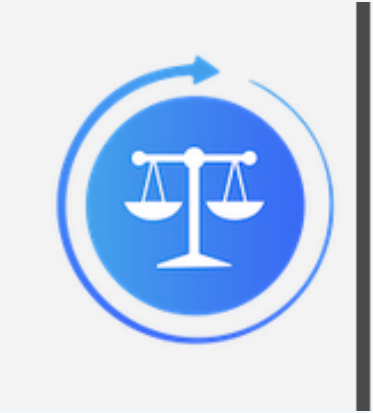
M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Conference on Neural Information Processing Systems, 2016.



AIF360

Overview

- Proposed by IBM;
- Python and R;
- Apache 2.0 license;
- Fairness evaluation metrics and mitigation tools.



AIF360 Methods

- **Pre-processing:**
  - Reweighting;
  - Uniform and Preferential sampling;
  - Disparate Impact Remover;
  - Optimized Pre-processing;
  - Fair Latent Representation Learning.
- **Post-processing:**
  - Reject Option Classification;
  - Equalized-odds Post-processing;
  - Calibrated Equalized-odds Post-processing.
- **In-processing:** See the next lecture.

Documentation Example

aif360.algorithms.preprocessing.DisparateImpactRemover

class aif360.algorithms.preprocessing.DisparateImpactRemover(repair_level=1.0, sensitive_attribute='') [source]

Disparate impact remover is a preprocessing technique that edits feature values increase group fairness while preserving rank-ordering within groups [1].

References

[1] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

Parameters:
- repair_level (float) – Repair amount. 0.0 is no repair while 1.0 is full repair.
- sensitive_attribute (str) – Single protected attribute with which to do repair.

Methods

fit	Train a model on the input.
fit_predict	Train a model on the input and predict the labels.
fit_transform	Run a repairer on the non-protected features and return the transformed dataset.
predict	Return a new dataset with labels predicted by running this Transformer on the input.
transform	Return a new dataset generated by running this Transformer on the input.

__init__(repair_level=1.0, sensitive_attribute='') [source]

Parameters:
- repair_level (float) – Repair amount. 0.0 is no repair while 1.0 is full repair.
- sensitive_attribute (str) – Single protected attribute with which to do repair.

fit_transform(dataset) [source]

Run a repairer on the non-protected features and return the transformed dataset.

Parameters: dataset (BinaryLabelDataset) – Dataset that needs repair.
Returns: dataset (BinaryLabelDataset) – Transformed Dataset.

Note

In order to transform test data in the same manner as training data, the distributions of attributes conditioned on the protected attribute must be the same.