

Méthode pre-processing et post-processing de mitigation des biais

Session 3:
Mitiger les biais avec des méthodes de type pre-processing et post-processing

Alice HELIOU and Vincent THOUVENOT
Laboratoire de Data Science de ThereSIS



Content

1. Introduction

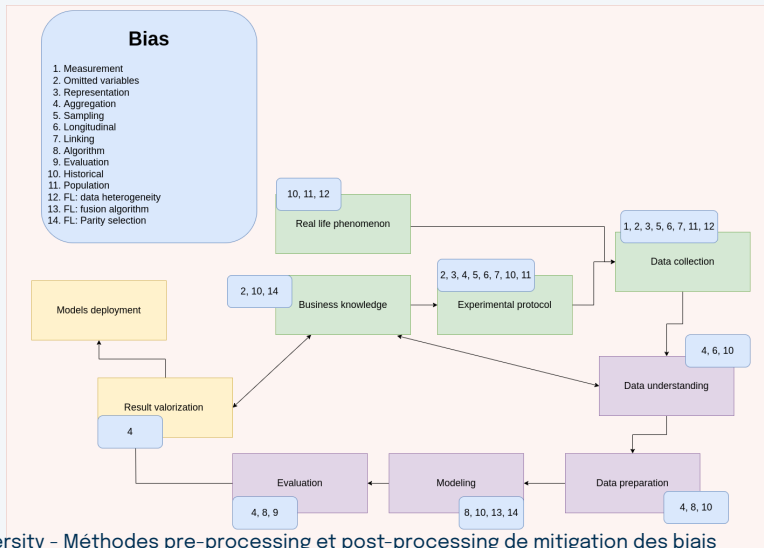
2. Méthodes de pré-processing

3. Méthodes de post-processing

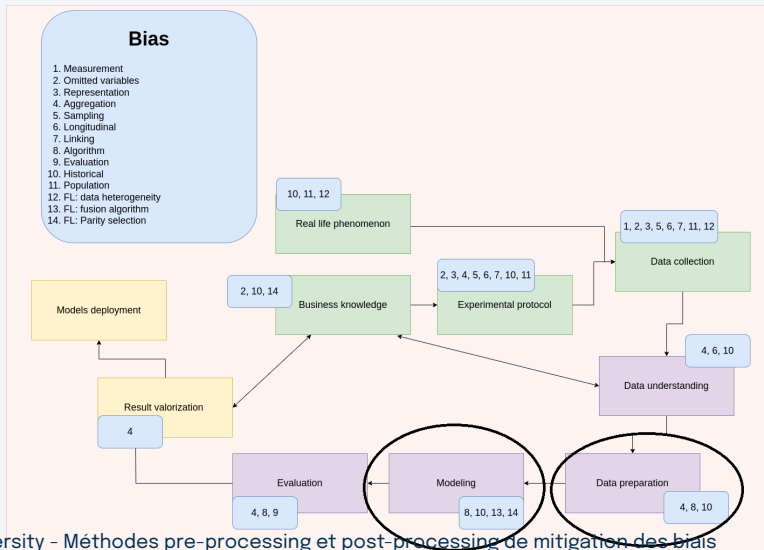
4. AIF360

Introduction

Position du cours



Position du cours



Juste retirer l'attribut sensible?

Exemple: une banque cherche une méthode pour savoir à qui faire un prêt de manière juste

- Objectif de la banque: déterminer qui va rembourser le prêt;
- Le prêt est destiné à deux groupes disjoints de personnes;
- Juste retirer du modèle l'information à propos de l'attribut sensible peut être inefficace à cause des interactions indirectes avec d'autres attributs.

Juste retirer l'attribut sensible n'est pas la solution optimale: les garder peut être plus efficace !

Méthodes de correction des biais

- **Pre-processing:** Modification des bases de données d'apprentissage
- **In-processing:** Correction des biais apprise en même temps que le modèle de machine learning
- **Post-processing:** Modification des prédictions des modèles de machine learning

Méthodes de correction des biais

- **Pre-processing: Modification des bases de données d'apprentissage**
- **In-processing:** Correction des biais apprise en même temps que le modèle de machine learning
- **Post-processing: Modification des prédictions des modèles de machine learning**

Méthodes de pré-processing

Massaging

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

- S attributs sensibles prenant b (groupe discriminé) et w (groupe favorisé);
- Label prend $+$ et $-$ comme valeur;
- Changer les labels de certaines observations:
 - Pour certaines observations telles que $S = b$, changer le label $-$ à $+$;
 - Pour certaines observations telles que $S = w$, changer le label $+$ à $-$.

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)
-

Repondération des données

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

Algorithm 3: *Reweighting*

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

Repondération des données

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	+

- Poids donné à une femme de la classe positive ?

Repondération des données

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	+

- Poids donné à une femme de la classe positive ?
- $\frac{0.5 \times 0.6}{0.2} = 1.5$

Repondération des données

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	+

- Poids donné à une femme de la classe positive: 1.5
- Poids donné à une femme de la classe négative: 0.67
- Poids donné à une homme de la classe positive: 0.75
- Poids donné à une homme de la classe négative: 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;

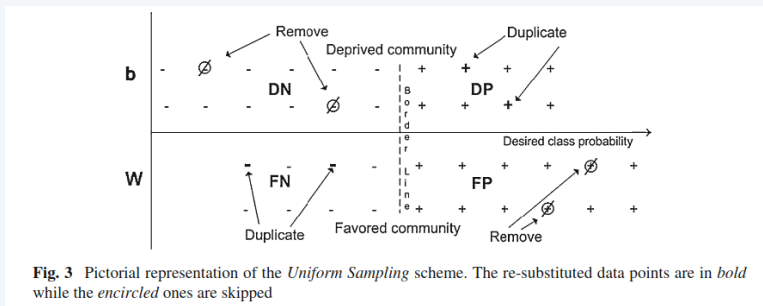
9: Sample uniformly $W(w, -) \times |FN|$ objects from FN;

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.



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 $\lfloor W(b, -) \rfloor \times |DN| \rfloor$ lowest ranked elements of DN to D_{PS}

11: Add $\lfloor W(w, +) \rfloor \times |FP| \rfloor$ highest ranked elements of FP to D_{PS}

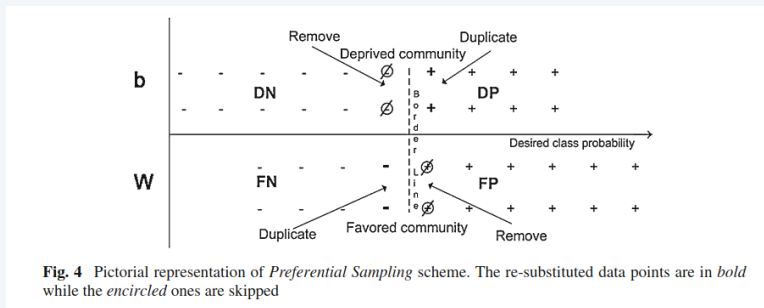
12: Add $\lfloor W(w, -) \rfloor$ copies of FN to D_{PS}

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.



Disparate Impact Remover

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

Definition 4.1 (BER). Let $f : Y \rightarrow X$ 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 $f : Y \rightarrow X$ 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 $f : Y \rightarrow X$

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

with (empirical) probabilities estimated from D .

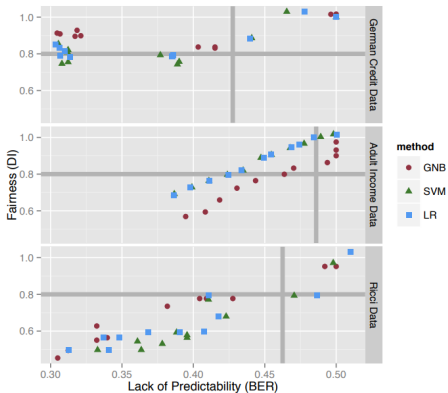
Disparate Impact Remover

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

- Une certification a montré un problème de disparate impact sur D ;
- Recherche d'un nouveau jeu de données $\tilde{D}$ où tous les éléments de D ont été changés pour que $\tilde{D}$ soit ε -fair

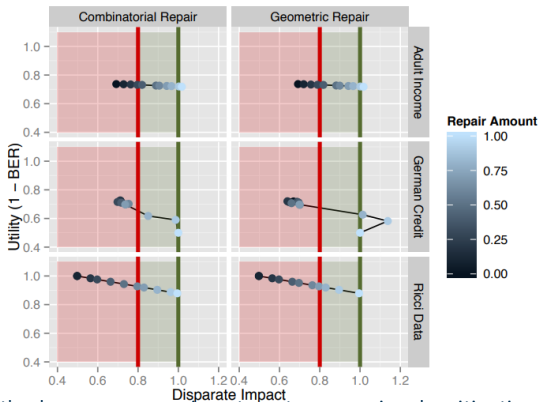
Disparate Impact Remover

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.



Disparate Impact Remover

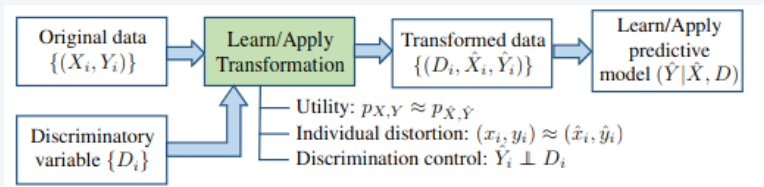
Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.



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.

Pre-processing optimisé

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.



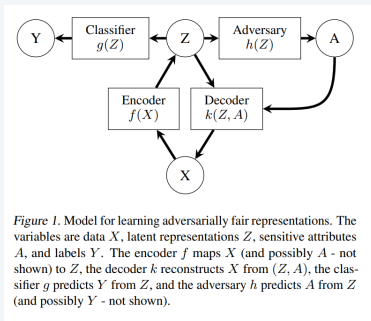
Résolution d'un problème d'optimisation sous trois contraintes:

- Préserver l'utilité des données;
- Limiter la distorsion des données;
- Contrôler la discrimination.

Apprentissage de représentation latente fair

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



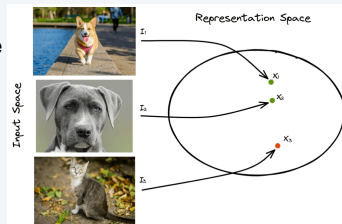
Apprentissage de représentation latente fair

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 $x_j = f(I_j)$
- Triplet Loss
 - Anchor sample I , positive sample I^+ , negative sample I^-

$$(1 - Y) \times ||x_i - x_j||^2 + Y \times \max(0, m - ||x_i - x_j||^2)$$

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



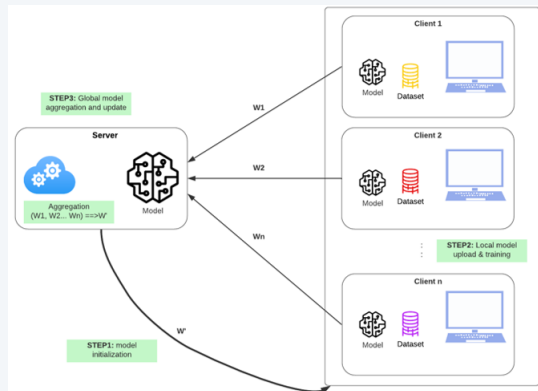
Cas de l'apprentissage fédéré

- Cas où plusieurs propriétaires de données (e.g. hôpitaux) veulent collaborer ensemble pour apprendre un modèle de Machine Learning...
- ...mais ne peuvent pas se partager les données.

Cas de l'apprentissage fédéré

	Apprentissage distribué	Apprentissage fédéré
Données	Données i.i.d., potentiellement accessibles au serveur	Données non i.i.d. Données sensibles et locales. La distribution ne peut pas être choisie par le serveur. Potentiellement, de nombreuses sources de données
Objectifs	Passage à l'échelle, accélérer les calculs sans perte de performance	Collaborer sans faire fuiter des informations sensibles. Compromis entre la sécurité/privacy et les performances du modèle

Cas de l'apprentissage fédéré



1. Le serveur central envoie la mise à jour la plus récente du modèle aux participants;
2. Chaque participant met à jour les poids du modèle avec ses données locales;
3. Les poids mis à jour sont envoyés par tous les participants au serveur central;
4. Le serveur central agrège les poids mis à jour localement.

Cas de l'apprentissage fédéré

Stage de Zachary Fakir sous la supervision d'Alice Héliou

Plusieurs causes de biais en apprentissage fédéré:

- Biais dans la sélection des participants de l'apprentissage fédéré;
- Données hétérogènes;
- Algorithme d'agrégation.

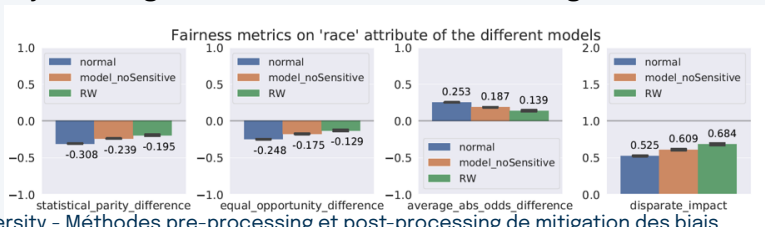
METHODS	PRIVACY	ADDITIONAL COMMUNICATION	HYPERPARAMETERS
LOCAL REWEIGHING	SAME AS PLAIN FL	0	NONE
GLOBAL REWEIGHING WITH DIFFERENTIAL PRIVACY	ϵ -DIFFERENTIAL PRIVACY	1.5	ϵ
FEDERATED PREJUDICE REMOVAL	SAME AS PLAIN FL	0	η

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

Cas de l'apprentissage fédéré

Stage de Zachary Fakir sous la supervision d'Alice Héliou

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



Cas de l'apprentissage fédéré

Stage de Zachary Fakir sous la supervision d'Alice Héliou

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

Méthodes de post-processing

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 ****

$\forall X_i \in \{Z | Z \in \mathcal{X}, \max[p(C^+|Z), 1 - p(C^+|Z)] \geq \theta\}$

$C_i = \operatorname{argmax}_{\{C^+, C^-\}} [p(C^+|X_i), p(C^-|X_i)]$

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.

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

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:

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

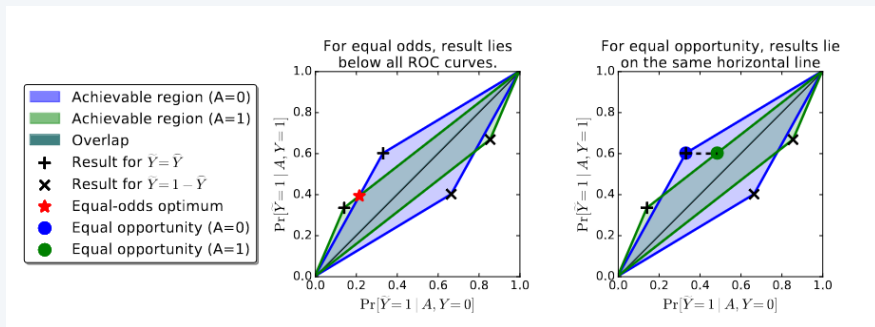
Equalized-odds post-processing

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

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

Equalized-odds post-processing

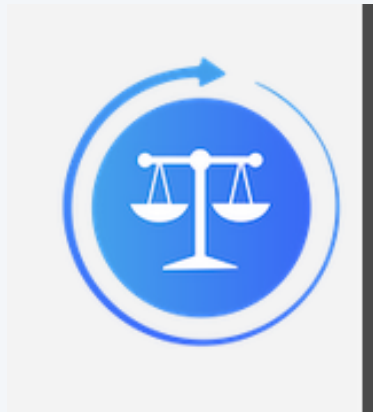
M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning,"
Conference on Neural Information Processing Systems, 2016.



AIF360

Vue d'ensemble

- Proposé par IBM;
- Python et R;
- Licence Apache 2.0.
- Métriques d'évaluation de la fairness et de mitigation



Méthodes de AIF360

- **Pre-processing:**
 - Repondération;
 - Uniform et Preferetial sampling;
 - Disparate Impact Remover;
 - Pre-processing optimisé;
 - Apprentissage de représentation latente fair.
- **Post-processing:**
 - Reject option classification;
 - Equalized-odds post-processing;
 - Calibrated Equalized-odds post-processing.
- **In-processing:** Voir le prochain cours.

Exemple de documentation

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.