

Résumé du mémoire de maîtrise de Florian TRAMÈR

François BIDEET

24 avril 2017

1 Présentation

Ce document regroupe des définitions et considérations extraites de mémoire de master “Algorithmic Fairness Revisited” de Florian Tramèr.

2 Notations

Notation. *Un algorithme est noté \mathcal{A} .*

Notation. *On note N la taille du dataset.*

Notation. *Les données collectées d’un utilisateur sont notées $X \in \mathcal{X}$.*

Notation. *L’ensemble des résultats d’un algorithme est noté \mathcal{O} et une sortie particulière est notée $o \in \mathcal{O}$.*

Notation. *Les attributs protégés (ou sensibles) sont notés S .*

Si S est binaire, on appelle s^+ la valeur pour laquelle l’algorithme favorise et s^- celle pour laquelle il défavorise.

Notation. *Une estimation par rapport aux données d’une probabilité P est notée \hat{P} .*

Notation. *La fonction d’utilité d’un utilisateur est notée $U \in \mathcal{U}$.*

Notation. *On note les attributs non protégés nécessaires à un travail (business) $B \in \mathcal{B}$. On note les classes d’utilisateurs nécessaires à un travail $K \in \mathcal{K}$. Un business fournit une fonction d’association $h : \mathcal{B}^n \rightarrow \mathcal{K}^n$ pour tout $n \geq 1$.*

3 Mesures d’équité

Definition 1. *Un algorithme est juste/équitable si sa sortie est indépendante des attributs protégés.*

Definition 2. Pour un attribut protégé $S \in \{s^+, s^-\}$, respectivement favorisé et défavorisé pour la même sortie $o \in \mathcal{O}$, on définit la mesure “selection-lift” :

$$slift(s^+; o) = \frac{\hat{P}(o|s^+)}{\hat{P}(o|s^-)}$$

Definition 3. Pour un attribut protégé $S \in \{s^+, s^-\}$, respectivement favorisé et défavorisé pour la même sortie $o \in \mathcal{O}$, on définit la mesure “difference-based selection-lift” :

$$slift_d(s^+; o) = \hat{P}(o|s^+) - \hat{P}(o|s^-)$$

Definition 4 (*a*-protection). Pour une mesure d’équité $f(\cdot)$ et un seuil fixé $a \in \mathbb{R}$, on dit que \mathcal{A} est “*a*-protecteur” par rapport à $f(\cdot)$, $s^+ \in \mathcal{S}$ et $o \in \mathcal{O}$ si $f(s^+, o) < a$. Sinon, \mathcal{A} est dit “*a*-discriminant” ou “*a*-discriminatoire”.

Soit $[L_1, L_2]$ l’intervalle de confiance de $f(s^+, o)$ au niveau de signification β , alors :

- \mathcal{A} est *a*-protecteur au niveau de signification β si $L_2 < a$
- \mathcal{A} est *a*-discriminant au niveau de signification β si $L_1 \geq a$

Definition 5 (Mesures théoriques de l’information). La divergence de Kullback-Leibler entre $\hat{P}(S)$ et $\hat{P}(S|O = o)$ est définie par

$$D_{KL}(\hat{P}(S|O = o) || \hat{P}(S)) = \sum_s \hat{P}(s|o) \ln \left(\frac{\hat{P}(s|o)}{\hat{P}(s)} \right)$$

Definition 6 (Parité statistique). Un algorithme statisfait empiriquement la parité statistique jusqu’au biais ϵ si

$$D_{TV}(\hat{P}(O|S = s^+), \hat{P}(O|S = s^-)) = \frac{1}{2} \sum_{o \in \mathcal{O}} \left| \hat{P}(o|s^+) - \hat{P}(o|s^-) \right| \leq \epsilon$$

où $D_{TV}(P, Q)$ est la distance de variation totale entre les distributions P et Q .

Definition 7 (Information mutuelle (MI)). L’information mutuelle empirique entre S et O est définie par

$$\hat{I}(S; O) = \mathbb{E}_O \left[D_{KL}(\hat{P}(S|O = o) || \hat{P}(S)) \right] = \sum_{s \in \mathcal{S}} \sum_{o \in \mathcal{O}} \hat{P}(s, o) \ln \left(\frac{\hat{P}(s, o)}{\hat{P}(s)\hat{P}(o)} \right)$$

On définit de même l’information mutuelle empirique normalisée entre S et O par

$$\hat{I}_{norm}(S; O) = \frac{\hat{I}(S; O)}{\min\{\hat{H}(S), \hat{H}(O)\}} \in [0, 1]$$

où $\hat{H}(Y)$ est l’entropie empirique d’une variable aléatoire Y :

$$\hat{H}(Y) = - \sum_y \hat{P}(Y = y) \ln \hat{P}(Y = y)$$

Definition 8 (MI-équité). *Un algorithme satisfait l'équité ϵ -MI par rapport à l'attribut protégé S si $\hat{I}(S; O) \leq \epsilon$. De même, un algorithme satisfait l'équité ϵ -MI normalisée par rapport à l'attribut protégé S si $\hat{I}_{norm}(S; O) \leq \epsilon$.*

Definition 9 (équité statistique générique). *Soit p la p -value obtenue par un test statistique pour l'hypothèse nulle $S \perp O$. Alors \mathcal{A} est statistiquement juste/équitable par rapport à S au niveau de signification β si $p \leq \beta$.*

Definition 10 ("User-Utilitarian Fairness"). *\mathcal{A} est juste par rapport à un attribut protégé S et à une fonction d'utilité U si et seulement si $U \perp S$.*

Definition 11 ("Business-Utilitarian Fairness"). *\mathcal{A} est juste par rapport à un attribut protégé S et à une classe K si et seulement si*

1. *Il existe un ensemble d'attributs \mathcal{B} et une association $h : \mathcal{B}^n \rightarrow \mathcal{K}^n$ qui sont une représentation valide des nécessités d'un travail*
2. *$O \perp S | K$*

Definition 12 (Information mutuelle conditionnelle). *L'information mutuelle conditionnelle entre S et O sachant K est définie par*

$$\hat{I}(S; O | K) = \mathbb{E}_K[\hat{I}(S; O) | K] = \sum_{k \in \mathcal{K}} \hat{P}(k) \sum_{s \in \mathcal{S}} \sum_{o \in \mathcal{O}} \hat{P}(s, o | k) \ln \left(\frac{\hat{P}(s, o | k)}{\hat{P}(s | k) \hat{P}(o | k)} \right)$$

La mesure normalisée $\hat{I}_{norm}(S; O | K)$ est donnée par $\frac{\hat{I}(S; O | K)}{\min\{\hat{H}(S | K), \hat{H}(O | K)\}}$.

La définition du "G-test" est une définition qui n'est pas créée par l'auteur. "G-test" est une évaluation statistique populaire de la qualité d'une hypothèse.

Definition 13 (G-test). *Pour chaque paire $(s, o) \in \mathcal{S} \times \mathcal{O}$, on note le nombre d'observation $f_{s,o}$. Alors le G-test est donné par*

$$G = 2 \sum_{s \in \mathcal{S}} \sum_{o \in \mathcal{O}} f_{s,o} \ln \left(\frac{f_{s,o}}{E_{s,o}} \right)$$

où $E_{s,o}$ est la fréquence d'observation théorique si l'hypothèse nulle est valide.

Definition 14 (G-test conditionnel). *Avec $f_{s,o,k}$ le nombre d'observation du triplet (s, o, k) et $E_{s,o,k}$ le nombre théorique si l'hypothèse nulle est valide, le G-test est donné par*

$$G_K = 2 \sum_{s \in \mathcal{S}} \sum_{o \in \mathcal{O}} \sum_{k \in \mathcal{K}} f_{s,o,k} \ln \left(\frac{f_{s,o,k}}{E_{s,o,k}} \right)$$

4 Remarques dans le document

4.1 Remarques sur les mesures

Remarque 1. Pour un attribut protégé binaire $S \in s^+, s^-$, on a :

$$\begin{aligned}\hat{P}(S, O) = \hat{P}(S)\hat{P}(O) &\iff \text{slift}(s^+; o) = 1, \forall o \in \mathcal{O} \\ &\iff \text{slift}_d(s^+; o) = 0, \forall o \in \mathcal{O} \\ &\iff D_{KL}(\hat{P}(S|O=o) || \hat{P}(S)) = 0, \forall o \in \mathcal{O}\end{aligned}$$

Remarque 2 (Asymétrie). Les mesures de “selection-lift” ne sont pas symétriques. En général

$$\text{slift}(s^+; o) \neq \text{slift}(s^-; o)$$

et

$$\text{slift}_d(s^+; o) \neq \text{slift}_d(s^-; o)$$

Remarque 3 (lien entre G-test et information mutuelle). Le G-test vérifie

$$G = 2 \cdot N \cdot \hat{I}(S; O)$$

Remarque 4 (lien entre G-test et information mutuelle conditionnelle). Le G-test vérifie

$$G_K = 2 \cdot N \cdot \hat{I}(S; O|K)$$

avec $E_{s,o,k} = N \cdot \hat{P}(k)\hat{P}(s|k)\hat{P}(o|k)$

Remarque 5. Le G-test dépend de la taille du dataset : plus le dataset est grand, plus l’amplitude du G-test peut être importante.

4.2 Considérations

Remarque 6. Un certain nombre de mesures d’équité ne prennent pas en compte la taille des catégories, ce qui peut mener à des contradiction : un résultat peut être annoncé comme peu biaisé alors qu’il y a un énorme biais sur une très petite population.

\mathcal{A}	Homme	Femme	\mathcal{A}	Homme	Femme
Président	20	5	Président	20	20
Manager	9 980	9 995	Manager	9 970	9 995
Employé	10 000	10 000	Employé	10 010	9 985

$$D_{TV} = 7.50 \cdot 10^{-4}$$

$$\hat{I}_{norm} = 1.74 \cdot 10^{-4}$$

$$D_{TV} = 1.25 \cdot 10^{-3}$$

$$\hat{I}_{norm} = 1.13 \cdot 10^{-6}$$

Remarque 7. On ne peut pas observer un biais sur des catégories en convertissant en donnée binaire. Exemple, la discrimination envers les personnes noires n’apparaît pas :

S					
Black		White		Hispanic	
Applicants	Hired	Applicants	Hired	Applicants	Hired
100	50%	100	80%	100	20%

S'			
Black		Not Black	
Applicants	Hired	Applicants	Hired
100	50%	200	50%

Le biais envers les personnes noires n'apparaît pas car la favorisation des personnes blanches est compensée par la discrimination envers les personnes hispaniques.

Remarque 8 (Justification “business-utilitarian fairness”). Considérons le cas d'un recrutement de 120 employés avec les candidats répartis comme suit :

	Male	Female
PhD	60	24
Master	240	156
Bachelor	150	270

1. On peut considérer que le niveau Master est le niveau minimal à avoir pour pouvoir être embauché. On définit alors

$$\mathcal{K} = \{qualified, non-qualified\}$$

$$B \in \{PhD, Master, Bachelor\}$$

et

$$h(B) = qualified \iff B \in \{PhD, Master\}$$

On a alors 120 postes pour 480 candidats qualifiés, ce qui fait 25% des candidats à embaucher :

	Male		Female	
	Applicants	Hired	Applicants	Hired
PhD	60	25%	24	25%
Master	240	25%	156	25%
Bachelor	150	0%	270	0%
Total	450	16.7%	450	10%

2. On peut considérer que l'on désire les employés les plus qualifiés. On définit alors

$$\mathcal{K} = \{most-qualified, qualified, non-qualified\}$$

On embauche alors l'ensemble des candidats de niveau doctorat et 36 des 396 candidats de niveau Master, avec le cas favorable où 21 des 240

hommes et 15 des 156 femmes de niveau Master sont embauchés :

	Male		Female	
	Applicants	Hired	Applicants	Hired
PhD	60	100%	24	100%
Master	240	9%	156	9%
Bachelor	150	0%	270	0%
Total	450	18%	450	8.7%

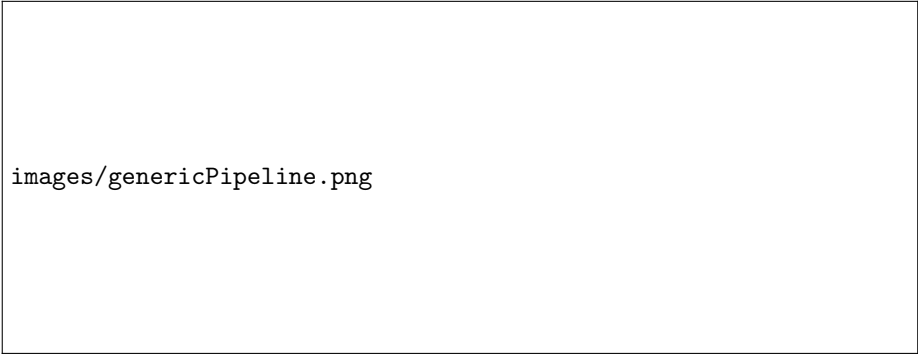
Les résultats présentent donc une discrimination d’après la définition 1 mais sont “justes” d’après la définition 11 si les ensembles \mathcal{K} sont acceptés.

Remarque 9 (Classifieur à deux étages). Pour effectuer une classification qui est juste par rapport à des exigences de l’entreprise, on peut concevoir un classifieur en deux niveaux comme suit :



On définit une première fonction h qui associe un profil d’attributs B , nécessaires à l’entreprise, à une classe K puis une deuxième fonction g effectue la classification en fonction de la classe K et du profil privé des attributs B . C’est cette dernière fonction qui doit être juste pour chaque classe K .

Remarque 10 (Pipeline générique pour tests statistiques d’équité).



`images/genericPipeline.png`

On a quatre étapes qui peuvent en théorie être effectuées indépendamment :

1. Acquisition de donnée

étape consistant à mettre en forme le dataset, à le séparer en ensemble d'apprentissage et ensemble de test, à les traiter pour enlever les attributs protégés, en les sauvegardant dans un ensemble auxiliaire afin d'étudier ultérieurement les biais.

2. Partionnement de données et apprentissage

étape consistant à entraîner un classifieur approximant la décision prise sur l'ensemble d'entraînement pour pouvoir par la suite interpréter.

3. Génération d'hypothèse interprétable

étape consistant à formuler l'hypothèse nulle, dans notre cas qu'il n'y a pas de biais dans les groupes observés.

4. Tests statistiques

étape consistant à effectuer des mesures d'équité en fonction des hypothèses précédemment établies et à calculer les p-valeurs correspondantes.

Après ces quatre étapes, on doit rajouter une étape d'études des résultats statistiques pour déterminer s'il y a vraiment un biais et si on doit réitérer en modifiant le dataset.