Jean-Marie John-Mathews

Institut Mines-Télécom Business School

Good In Tech

FORMALISER L'ÉQUITÉ EN ML

Revue des méthodes de « fairness » en apprentissage supervisé

Les décisions issues des algorithmes d'apprentissage supervisé s'adaptent à partir d'un historique d'exemples. Un des problèmes éthiques majeurs posés par les algorithmes du *Machine Learning* est celui de l'équite de la décision vis-à-vis de certains groupes de la population.

Les sources de biais

- Biais statistiques :

Algorithme entraîné sur des données statistiquement biaisées (représentativité des données, etc)

- Modèles biaisés :

Un choix de variables trop restreint, des proxys trop approximatifs ou comportant des corrélations indirectes de nature discriminatoire peut entraîner une décision biaisée

- Biais intentionnels :

Le caractère discriminatoire de la décision algorithmique peut aussi provenir d'un comportement intentionnel du concepteur qui cherche délibérément à biaiser les données via la configuration du modèle

Notations

Considérons un problème de classification binaire possédant une variable sensible à l'origine du biais. On note

- $X \in \mathbb{R}^d$ les features
- $S \in \{0,1\}$ la variable sensible (appartenance ethnique, sexe, etc)
- $Y \in \{0,1\}$ la variable target
- $\hat{Y}(X,S) \in \{0,1\}$ le classifieur (ex: 1 si embauche, 0 sinon)

Formalisation de l'équité

- Unawareness :

Il s'agit d'omettre délibérément la variable codant la catégorie préjudiciable (S) au sein des données d'apprentissage. L'algorithme aveugle à la catégorie (*color-blind)* est supposé alors ne pas discriminer la catégorie préjudiciable. Soit

$$\hat{Y}(X,S) = \hat{Y}(X)$$

Les méthodes de type *Unawaress* ont l'inconvénient de reproduire les discriminations en reconstituant la catégorie préjudiciable à partir de variables tierces qui y sont corrélées.

- Group fairness :

Les méthodes d'équité de type *Group fairness* comparent après et avant redressement une mesure statistique d'équité définie au niveau des groupes préjudiciables. On en distingue plusieurs types

- Parité démographique (ou statistique)

Quelque soit le groupe préjudiciable, la probabilité de classification est la même. Autrement dit,

$$\hat{Y} \perp S$$

Cette définition ignore les possibles corrélations en Y et S ce qui dégrade la performance de l'algorithme. A long terme, les algorithmes à parité démographique peuvent être efficaces pour lutter contre les discriminations structurelles.

- Equalized odd :

Se définit par l'égalité d'indépendance conditionnelle suivante :

$$\forall y \quad \hat{Y} \mid Y = y \perp S \mid Y = y$$

Une notion plus faible de l'Equalized odd est la *Parité de précision* définit simplement par :

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

Ou encore *l'égalité d'opportunité* définie par

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

- Predictive Rate Parity :

Symétriquement à l'équité de type *Equalized odd,* on la définit par

$$\forall \hat{y} \quad Y \mid \hat{Y} = \hat{y} \perp S \mid \hat{Y} = \hat{y}$$

que l'on peut aussi décliner par deux notions plus faibles, à savoir *Positive Predictive Parity et Negative Positive Parity.*

- Individual Fairness :

L'équité individuelle ne se calcule pas en fonction de l'appartenance à la catégorie préjudiciable à l'instar de l'équite de type *groupe fairness*. Deux individus dont la distance est faible doivent obtenir des décisions similaires par le classifieur, c'est la distance inter-individuelle qui est calculée. La difficulté provient alors de la définition de l'espace permettant de calculer cette distance inter-individuelle.

Les méthodes de redressement

On peut distinguer trois types de procédure de redressement : redressement par modification des données d'apprentissage en entrée (pre-processing), redressement de la procédure d'apprentissage elle-même (in-processing) ou bien redressement de la prédiction (post-processing). En voici quelques familles :

- Repondération : méthode pre-processing où il s'agit de pondérer les individus de la base d'apprentissage dans chaque configuration de sorte à assurer l'équité avant même l'apprentissage. Cela revient à rajouter ou enlever des individus afin d'assurer l'équité.

- L'équité comme contrainte dans un programme d'optimisation : méthode in-processing où la minimisation du risque empirique s'établit sous contrainte d'égalité d'une des définitions formelles de l'équité.

- Modification des décisions du classifieur a posteriori (Hardt et al, 2016) : méthode de post-processing où, l'on fixe un seuil différent de classification pour chaque groupe sensible. Cela revient donc, pour les classifications les plus incertaines, à donner des décisions plus favorables aux catégories discriminées et moins favorables aux catégories non discriminées.

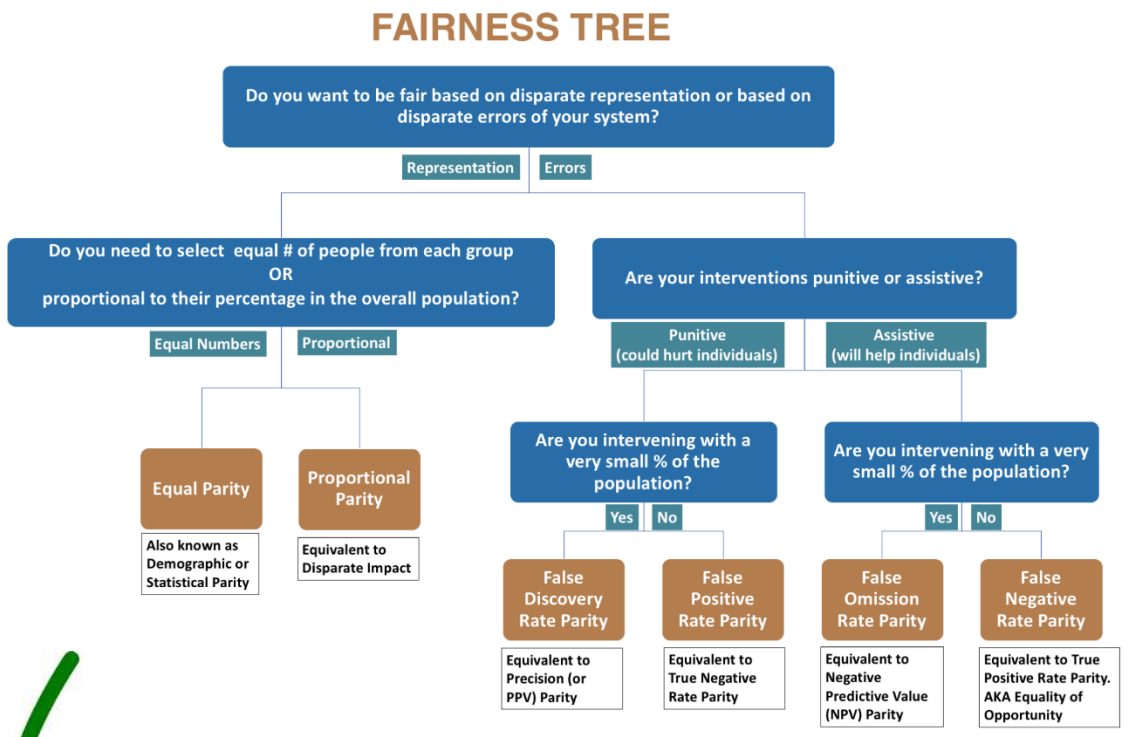

Schéma extrait de Aequitas: A Bias and Fairness Audit Toolkit

ITEM