Biais et ethique en apprentissage statistique

Nicolas Leroy, Ema Cerezo, Axel de Montgolfier

2023-10-16

Enjeux

- Comment catégoriser une discrimination ?
- Comment catégoriser les source de discrimination ?
- Quelles méthodes pour minimiser de telles sources de discriminations ?

Cas d'application

- COMPAS : Estimateur de la capacité d'un criminel à récidiver adopté en 2016 par l'etat du Wisconsin.
- Gender Shades project : Tentative en 2018 de rendre plus éthique les système de reconnaissance faciale par la repondération d'échantillon.

Deux définitions d'un modèle éthique

Equalized Odds (Égalité des chances) : Veille à ce que les groupes bénéficient de taux de faux positifs et de vrais positifs égaux, indépendamment de la variable sensible.

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

Demographic Parity (Parité démographique) : Garantit que les individus ont des chances égales d'avoir un résultat favorable quel que soit leur groupe.

$$\mathbb{P}(\hat{Y} = y|S = 0) = \mathbb{P}(\hat{Y} = y|S = 1)$$

Catégories de biais

- Biais de collection de données
- Biais de modèlisation
- Biais d'utilisation

Comment remedier aux discriminations

- Correction de données (pre-processing)
- Modèle resilient (in-processing)
- Correction de resultat (post-processing)