Ingénierie de la fouille et de la visualisation de données massives (RCP216)

Aspects éthiques dans la fouille de données

Michel Crucianu
(prenom.nom@cnam.fr)
http://cedric.cnam.fr/vertigo/Cours/RCP216/

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

29 mai 2024

Plan du cours

Quelques définitions

- Éthique : « Science qui traite des principes régulateurs de l'action et de la conduite morale. » (déf. CNRTL)
- Moral : « Qui concerne les règles ou principes de conduite, la recherche d'un bien idéal, individuel ou collectif, dans une société donnée. » (déf. CNRTL)
- Équité : « (Principe impliquant l') appréciation juste, (le) respect absolu de ce qui est dû à chacun. » (déf. CNRTL)
- Discrimination : « Traitement différencié, inégalitaire, appliqué à des personnes sur la base de critères variables. » (déf. CNRTL)
- Loyauté : « [En parlant d'une personne] Fidélité manifestée par la conduite aux engagements pris, au respect des règles de l'honneur et de la probité. » (déf. CNRTL)

Quelques travers (1)

- COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) est utilisé aux États-Unis par des juges pour évaluer le risque de récidive et prendre des décisions de libération conditionnelle. Une étude a trouvé que l'outil défavorise les afro-américains par rapport aux caucasiens de profil similaire.
- Pour un algorithme qui affiche des annonces d'emploi en sciences, technologie, ingénierie et mathématiques, conçu pour ignorer le genre de la personne qui regarde, une étude a trouvé que les publicités étaient montrées plus rarement aux femmes qu'aux hommes pour une raison de coût, le public féminin étant une cible publicitaire plus chère que le public masculin.
- Un établissement bancaire emploie un outil d'aide à la décision pour accorder ou refuser un crédit. Cet outil exclut toute variable interdite mais exploite un grand nombre d'autres variables. Parmi celles-ci, le code postal (entre autres) est très corrélé à la « race » [1] et assez corrélé à la religion.

($\rightarrow$ équité : absence de discrimination, voir aussi [5],[4])

[1] Le consensus scientifique rejette la légitimité de la notion de « race » pour l'espèce humaine.

Quelques travers (2)

- Une société qui a mis en place un réseau social utilise les informations gratuitement fournies par ses utilisateurs, ainsi que leurs échanges sur le réseau, pour optimiser la vente d'espaces publicitaires. Elle vend également à des tiers des informations concernant ses utilisateurs. En situation de quasi-monopole, cette société peut maximiser ses gains tout en offrant un service basique à la grande majorité des utilisateurs.
- Une société de vente de services en ligne adapte au client la tarification de son offre avec pour seul objectif la maximisation de ses bénéfices. Pour cela, elle utilise des données collectées gratuitement auprès de ses clients, ainsi que des données achetées à des *data brokers*.

($\rightarrow$ équité : juste partage des « gains »)

Quelques travers (3)

- Un utilisateur s'informe grâce à un média en ligne. Le site choisit les messages proposés à chaque utilisateur de façon à maximiser le temps de consultation ($\sim$ ses recettes publicitaires), ce qui passe par la maximisation de l'adhésion (et non de la diversité, ni de la véracité). En conséquence, l'utilisateur peut être aliéné.
- Un réseau social vend des informations sensibles concernant les utilisateurs (et les non utilisateurs mentionnés dans les posts des utilisateurs !) à des employeurs actuels ou potentiels, sans que cela soit explicitement mentionné dans les conditions d'utilisation du réseau. Cela a des conséquences sur l'emploi et/ou sur les demandes d'emploi de ces personnes.
- Un comparateur de produits et de services modifie l'ordre de présentation des résultats pour privilégier les fournisseurs avec lesquels il a des liens capitalistiques ou, plus largement, des liens d'affaires.

($\rightarrow$ « loyauté » : respect de ce qui est explicitement ou implicitement promis)

Plan du cours

Est-ce moral ou non ?

- Déjà intégré à la réglementation : toute règle de droit doit être connue et respectée
  - « La règle de droit peut trouver sa source dans la morale, mais une fois établie en tant que règle, elle s'en sépare pour intégrer le domaine du droit. » (Espace Éthique APHP)
- Au-delà de ce qui a été intégré à la réglementation en vigueur, la préoccupation morale pour les moyens et les finalités d'une action est naturelle

... par ailleurs, une réglementation n'est pas nécessairement complète

- Si la morale n'a pas un caractère universel et immuable, il y a néanmoins des droits fondamentaux reconnus par l'ONU dans la Déclaration universelle des droits de l'homme

Droits fondamentaux

- Déclaration des droits de l'homme et du citoyen (1789)
- Déclaration universelle des droits de l'homme (ONU, 10/12/1948) :

Art. 1 Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.

Art. 2 Chacun peut se prévaloir de tous les droits et de toutes les libertés proclamés dans la présente Déclaration, sans distinction aucune, notamment de race, de couleur, de sexe, de langue, de religion, d'opinion politique ou de toute autre opinion, d'origine nationale ou sociale, de fortune, de naissance ou de toute autre situation.

- Charte des droits fondamentaux de l'UE (18/12/2000, chap. III, art. 21) :

  Est interdite toute discrimination fondée notamment sur le sexe, la race, la couleur, les origines ethniques ou sociales, les caractéristiques génétiques, la langue, la religion ou les convictions, les opinions politiques ou toute autre opinion, l'appartenance à une minorité nationale, la fortune, la naissance, un handicap, l'âge ou l'orientation sexuelle.

Exemples de réglementation aux États-Unis

- Cadres : *disparate treatment* (utilisation intentionnelle d'une variable protégée), *disparate impact* (accès à une variable protégée à travers des variables *proxy*)
- Exemple : variables protégées par *Fair Housing Act* (FHA) et *Equal Credit Opportunity Act* (ECOA)

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

Nature des problèmes

1. Absence d'équité (présence de discrimination) dans la prise de décision : partialité ou favoritisme envers un individu ou un groupe sur la base de caractéristiques inhérentes ou acquises
   - Présence de réglementation concernant plus largement la prise de décision publique ou privée, reste à la (faire) respecter
   - Principal sujet d'étude concernant l'exploitation des données pour la prise de décision
   - Définitions multiples, dont certaines incompatibles entre elles
2. Absence de loyauté : non respect du « contrat » (explicite ou implicite) fait avec l'utilisateur
   - Prise de conscience en cours, accélérée aussi par des lanceuses et lanceurs d'alerte

   → Explicitation du « contrat » dans l'utilisation des données personnelles (par ex. RGPD)
3. Absence d'équité dans le partage des « gains » ou avantages issus de l'exploitation des données : prise de conscience en cours, accélérée par certains comportements monopolistiques manifestes

- Plusieurs autres aspects sont liés à la prise de décision par des systèmes automatiques (voir par ex. [2]) dans la conduite autonome, la sécurité, etc.

Plan du cours

Équité comme absence de discrimination

- Historiquement, la préoccupation pour la non discrimination s'est manifestée à travers deux notions :
  1. L'équité de groupe : parité statistique entre les décisions pour les différents groupes (différentes valeurs des attributs protégés[2] : sexe, « race », etc.)
  2. L'équité individuelle : décisions similaires pour individus similaires par rapport à la tâche

- Des conflits peuvent apparaître dans la mise en œuvre des deux notions
  → Par ex. un outil automatique d'analyse de CV choisit les candidat.e.s pour des entretiens d'embauche ; l'employeur constate qu'un groupe protégé[3] est désavantagé par l'outil et décide d'équilibrer les taux de sélection pour assurer l'équité de groupe ; un.e candidat.e d'un groupe non protégé conteste sa non sélection, montrant que des candidat.e.s du groupe protégé, ayant les mêmes qualifications, ont pourtant été sélectionné.e.s

[2] Attribut protégé : attribut qui doit être exclu de la prise de décision.
[3] Groupe protégé : groupe typiquement victime de discrimination.

Équité comme absence de discrimination : critères observationnels

- Notations :
  - X : variables explicatives disponibles
  - A : variable qui représente explicitement les attributs protégés
  - Y : variable expliquée (ou cible) [exemple : $Y = 1 \Leftrightarrow$ remboursement prêt]
  - R : score (ou éventuellement décision) [exemple : $R = 1 \Leftrightarrow$ prêt accordé]
- Trois critères représentatifs (voir par ex. [1] à https://fairmlbook.org/) :
  1. Indépendance (*demographic parity*) : $R \perp\!\!\!\perp A$. Score indépendant des attributs protégés (par ex., pour R binaire, même « taux d'acceptation » pour différentes valeurs de A)
  2. Séparation (*equalized odds*) : $R \perp\!\!\!\perp A \mid Y$. Score et attributs protégés sont indépendants conditionnellement[4] à la variable cible (pour R et Y binaires, les taux de faux positifs et de faux négatifs ne changent pas pour différentes valeurs de A)
  3. « Suffisance » : $Y \perp\!\!\!\perp A \mid R$. Cible et attributs protégés sont indépendants conditionnellement au score (pour R et Y binaires, les valeurs prédictives[5] positives et les valeurs prédictives négatives ne changent pas pour différentes valeurs de A)

[4] Déf. indépendance conditionnelle $R \perp\!\!\!\perp A \mid Y$: $P(R|A,Y) = P(R|Y)$, ou $P(R,A|Y) = P(R|Y)P(A|Y)$

[5] Valeur prédictive positive $= \frac{\text{Vrais Positifs}}{\text{Prédits Positifs}}$, valeur prédictive négative $= \frac{\text{Vrais Négatifs}}{\text{Prédits Négatifs}}$.

Matrice de confusion et définitions associées

	Positifs (**P** : $Y = 1$)	Négatifs (**N** : $Y = 0$)	
Prédits positifs (**PP** : $R = 1$)	Vrais positifs (**VP**)	Faux positifs (**FP**)	Valeur prédictive positive $= \frac{VP}{PP}$
Prédits négatifs (**PN** : $R = 0$)	Faux négatifs (**FN**)	Vrais négatifs (**VN**)	Valeur prédictive négative $= \frac{VN}{PN}$
	Taux faux négatifs TFN $= \frac{FN}{P}$ (= 1-TVP)	Taux faux positifs TFP $= \frac{FP}{N}$ (= 1-TVN)	

Trois exemples

1. Accorder un prêt : Y correspond au remboursement ($Y = 1$ pour prêt remboursé, $Y = 0$ pour non remboursé), R correspond au score (si continu) ou à la décision (accorder ou non le prêt) si binaire ($R = 1$ pour accordé, $R = 0$ pour non accordé)

2. Embauche : $Y = 1$ si l'employé donne satisfaction, $Y = 0$ sinon. R correspond au score (si continu) ou à la décision (embaucher ou non) si binaire ($R = 1$ pour embauché, $R = 0$ pour non embauché)

3. Libération conditionnelle : Y correspond à la récidive ($Y = 0$ pour récidive, $Y = 1$ pour non récidive), R correspond au score (si continu) ou à la décision (libérer ou non) si binaire ($R = 1$ pour libéré, $R = 0$ pour non libéré)

Critère d'indépendance

- Conditions équivalentes :
  - Cas général : information mutuelle nulle entre R et A : $I(A;R) = 0$
  - Pour R binaire : $P(R = 1|A = a) = P(R = 1|A = b)$ ($R = 1$ désigné par « acceptation »)
- Relaxations de la condition d'indépendance ($\epsilon > 0$ mais proche de 0) :

  (le groupe $A = b$ est le groupe protégé)
  - $P(R = 1|A = b) \geq P(R = 1|A = a) - \epsilon$
  - Ratio de disparité démographique proche de 1 : $\frac{P(R=1|A=b)}{P(R=1|A=a)} \geq 1 - \epsilon$
  - Information mutuelle en A et R proche de 0 : $I(A;R) \leq \epsilon$
- Insuffisances du critère d'indépendance :
  - Son respect ne suffit pas à garantir la non discrimination
    - Exemple : taux d'acceptation identique dans les groupes a et b mais sélection attentive dans le groupe a contre choix aléatoire dans le groupe b

    $\Rightarrow$ résultats inférieurs pour les accepté.e.s du groupe b

    $\Rightarrow$ justification *a posteriori* pour privilégier le groupe a
  - Si $Y \not\perp\!\!\!\perp A$ (cas habituel) alors le score « parfait » qui serait $R = Y$ ne peut pas satisfaire le critère d'indépendance $R \perp\!\!\!\perp A$

Critère d'indépendance : signification sur les trois exemples

1. La probabilité d'accorder un prêt (ou la proportion de membres du groupe, candidats à un prêt, auxquels le prêt est accordé) est la même pour le groupe favorisé ($A = a$) et pour le groupe défavorisé ($A = b$)
2. La probabilité d'embaucher (ou la proportion de membres du groupe, candidats à l'embauche, qui ont été embauchés) est la même pour le groupe favorisé ($A = a$) et pour le groupe défavorisé ($A = b$)
3. La probabilité de libération conditionnelle (ou la proportion de membres du groupe, candidats à une libération, qui ont été libérés) est la même pour le groupe favorisé ($A = a$) et pour le groupe défavorisé ($A = b$)

Critère de séparation

- Pourquoi « séparation » : dans le modèle graphique associé, Y sépare R de A
- Intuition : l'absence d'indépendance entre R et A peut être acceptable mais pas au-delà de ce qui est justifié par la cible Y
  - Pour R et Y binaires :
    $P(R = 1|Y = 1, A = a) = P(R = 1|Y = 1, A = b)$ (même taux de vrais positifs) et
    $P(R = 1|Y = 0, A = a) = P(R = 1|Y = 0, A = b)$ (même taux de faux positifs)
- Relaxations de la condition de séparation :
  - Contrainte sur le seul taux de faux négatifs quand « positif » signifie opportunité (ex. embauche), donc les faux négatifs sont les opportunités non accordées
  - Contrainte sur le seul taux de faux positifs quand « positif » signifie pénalité (ex. maintien en prison car récidive prédite), donc les faux positifs sont les pénalités injustifiées
  - Seuil > 0 sur l'écart entre le taux pour $A = a$ et celui pour $A = b$
- Difficulté : il faut avoir accès à Y

Critère de séparation : signification sur les trois exemples

1. La probabilité d'obtenir un score qui correspond à accorder un prêt ($R = 1$) pour un membre du groupe favorisé ($A = a$) qui rembourserait le prêt ($Y = 1$) est la même que pour un membre du groupe défavorisé ($A = b$) qui rembourserait le prêt. Ou, parmi ceux qui rembourseraient le prêt aucun groupe n'est favorisé dans la décision d'accorder le prêt.
2. La probabilité d'obtenir un score qui correspond à être embauché ($R = 1$) pour un membre du groupe favorisé ($A = a$) qui donnerait satisfaction ($Y = 1$) est la même que pour un membre du groupe défavorisé ($A = b$) qui donnerait satisfaction. Ou, parmi ceux qui donneraient satisfaction aucun groupe n'est favorisé dans la décision d'embauche.
3. La probabilité d'obtenir un score qui correspond à être libéré ($R = 1$) pour un détenu membre du groupe favorisé ($A = a$) qui ne récidiverait pas ($Y = 1$) est la même que pour un membre du groupe défavorisé ($A = b$) qui ne récidiverait pas. Ou, le taux de « perte de chance » par non libération alors qu'il n'y aurait pas eu récidive est le même dans les deux groupes.

Critère de « suffisance »

- Dans le modèle graphique associé, R sépare Y de A
- Signification : la valeur prédictive ne dépend pas du groupe (valeur de A)
  - Pour R et Y binaires :
    $P(Y = 1|R = 1, A = a) = P(Y = 1|R = 1, A = b)$ (même valeur prédictive positive)
    $P(Y = 0|R = 0, A = a) = P(Y = 0|R = 0, A = b)$ (même valeur prédictive négative)
- Difficulté : comme pour la séparation, il faut avoir accès à Y
- Rapport avec la calibration (permettant de mieux comprendre la suffisance) :
  - Calibration : R est calibrée si, pour toute valeur $r \in [0, 1]$ que peut prendre R, $P(Y = 1|R = r) = r$ (parmi les observations de score r, une fraction égale à r sont positives)
  - Calibration *par groupe* : la condition est satisfaite pour tout groupe défini par une valeur différente de A, $P(Y = 1|R = r, A = a) = P(Y = 1|R = r, A = b) = r$
  - La calibration par groupe implique la suffisance, la suffisance implique la calibration par groupe à une transformation de score près

Critère de suffisance : signification sur les trois exemples

1. La probabilité de rembourser le prêt ($Y = 1$) est la même pour les membres du groupe favorisé ($A = a$) qui ont obtenu un prêt ($R = 1$) que pour les membres du groupe défavorisé ($A = b$) qui ont obtenu un prêt. Ou, parmi ceux à qui le prêt a été accordé la même proportion le remboursera dans le groupe favorisé et dans le groupe défavorisé (R est également prédictif de la cible Y dans les deux groupes).
2. La probabilité de donner satisfaction ($Y = 1$) est la même pour les membres du groupe favorisé ($A = a$) qui ont obtenu le score d'embauche ($R = 1$) que pour les membres du groupe défavorisé ($A = b$) qui ont obtenu le score d'embauche. Ou, parmi ceux qui ont été embauchés la même proportion donnera satisfaction dans le groupe favorisé et dans le groupe défavorisé (R est également prédictif de la cible Y dans les deux groupes).
3. La probabilité de ne pas récidiver ($Y = 1$) est la même pour un détenu membre du groupe favorisé ($A = a$) qui a été libéré ($R = 1$) est la même que pour les membres du groupe défavorisé ($A = b$) qui a été libéré. Ou, parmi ceux qui ont été libérés la même proportion récidivera dans le groupe favorisé et dans le groupe défavorisé (R est également prédictif de la cible Y dans les deux groupes).

Relations entre critères

- Indépendance $\leftrightarrow$ séparation :
  - L'absence d'indépendance entre A et Y est le cas typique : le taux de positifs ($Y = 1$) est différent entre groupes définis par les différentes valeurs de l'attribut protégé A

  $\Rightarrow$ Si A et Y ne sont pas indépendantes et Y est binaire, alors $R \perp\!\!\!\perp A$ et $R \perp\!\!\!\perp A \mid Y$ ne peuvent pas être valables simultanément
- Indépendance $\leftrightarrow$ suffisance :

  $\Rightarrow$ Si A et Y ne sont pas indépendantes, alors $R \perp\!\!\!\perp A$ et $Y \perp\!\!\!\perp A \mid R$ ne peuvent pas être valables simultanément
- Séparation $\leftrightarrow$ suffisance :

  $\Rightarrow$ Si A et Y ne sont pas indépendantes et toutes les probabilités jointes (A, R, Y) sont positives, alors $R \perp\!\!\!\perp A \mid Y$ et $Y \perp\!\!\!\perp A \mid R$ ne peuvent pas être valables simultanément

$\rightarrow$ Plusieurs critères ne peuvent pas être satisfaits simultanément de façon stricte, en revanche il est envisageable de rechercher des compromis grâce aux relaxations

Biais dans la prise de décision

- Biais : « Déformation, travers » (déf. CNRTL)
- Significations diverses :
  - Statistique : écart systématique entre l'espérance de la valeur estimée et la vraie valeur
  - *Machine learning* : erreur systématique souvent attribuée à des hypothèses erronées (par ex. utilisation d'une famille de modèles de capacité insuffisante)
  - Ici : disparités démographiques répréhensibles ou du moins discutables dans la prise de décision basée sur des algorithmes
- Quelques sources de disparités :
  - Disparités démographiques présentes dans la société
    $\rightarrow$ disparités dans les données recueillies
    $\rightarrow$ disparités dans les modèles décisionnels (aussi bien algorithmiques qu'humains) dans la mesure où ils sont développés à partir de données
    - Boucle : certaines décisions ont un impact sociétal et peuvent contribuer à perpétuer les disparités existantes
    - Biais délibérés : introduire (ou renforcer) délibérément des biais dans les données d'apprentissage pour modèles réputés opaques (forêts d'arbres de décision, réseaux de neurones profonds) afin de rendre ces biais moins visibles

Biais dans la prise de décision (2)

- Quelques sources de disparités (suite) :
  - La définition et la mesure des variables sont soumises à des disparités
    - Impact plus fort lorsqu'il s'agit des variables cibles (variables expliquées)
    - Exemple : la variable *récidive* ne pouvant être mesurée directement, c'est la variable *arrestation* qui s'y substitue, or celle-ci peut être soumise à des disparités démographiques
  - Constitution de bases de données par échantillonnage de convenance : peut renforcer des disparités présentes dans la population
    - Exemple : pour minimiser le coût d'un sondage, on l'applique là où il est plus facile d'avoir des réponses en grand nombre, c'est à dire là où sont présentes les catégories majoritaires
- Mécanismes d'action des disparités :
  - Action directe : inclusion de variables protégées dans les variables explicatives employées ($\rightarrow$ *disparate treatment*)
  - Action par l'intermédiaire de *proxy* : variable de substitution qui remplace une variable critique non observable ou non mesurable ou protégée ($\rightarrow$ *disparate impact*)
    - Exemple classique (États-Unis) : *zip code* comme proxy de la « race »
    - Absence de *proxy* clair $\not\Rightarrow$ absence de disparités : cumuler un nombre suffisant de variables « anodines » (peu corrélées à la variable protégée) peut suffire pour estimer la variable protégée
  - Action par l'intermédiaire de disparités dans les données qui engendrent des corrélations entre variables explicatives « anodines » et la variable protégée

Biais dans la prise de décision (3)

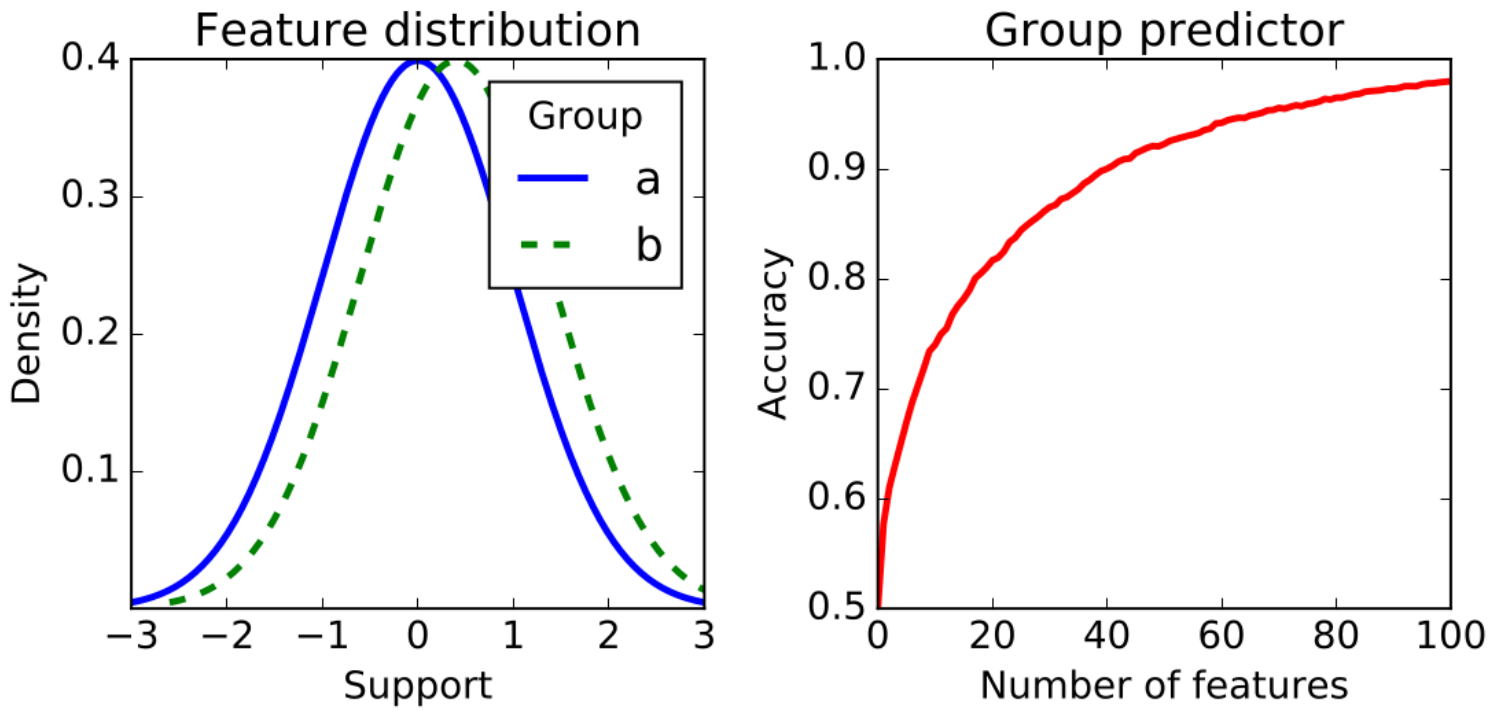

FIG. – A partir d'un nombre suffisant de variables légèrement corrélées à l'attribut protégé (distributions légèrement différentes pour les valeurs différentes de l'attribut protégé) il est possible de (très) bien prédire la valeur de l'attribut protégé (donc de l'imputer si absente). Illustration issue de [1].

Détecter l'iniquité

Techniques utilisées déjà pour décision humaine :

- Audits : expérimentation en environnement réel, pratiquement des tests d'aveuglement (de non prise en compte des attributs protégés) ; imposent le contrôle des autres variables (tests avec différentes valeurs pour l'attribut protégé « toutes choses égales par ailleurs »), ce qui ne correspond pas aux situations réelles
- *Blinding* : cacher les attributs protégés (et autres attributs qui permettraient de les deviner de façon trop évidente) sans contrôle des autres variables
- Révéler l'arbitraire dans les décisions : procédure non systématique, relations entre décision (ou score) et facteurs arbitraires (qui ne devraient pas intervenir)

Détecter l'iniquité (2)

- Tests basés sur les issues (*outcomes*) Y si disponibles (par ex. Y : défaut ou non du remboursement d'un prêt, R : score sur la base duquel le prêt est accordé ou non) :
  - Test de suffisance ($Y \perp\!\!\!\perp A \mid R$) : pour candidats dont le score R est dans un intervalle étroit, vérifier si même taux de $Y = 1$ pour chaque valeur de A (chaque groupe)
  - Sans accès au score R, test de parité de prédiction : vérifier si même taux de $Y = 1$ pour les candidats classés favorablement ($\hat{Y} = 1$) dans chaque groupe
  - Test de séparation ($R \perp\!\!\!\perp A \mid Y$) : difficile car en général Y est observé seulement pour $\hat{Y} = 1$; possible seulement si on peut évaluer Y aussi pour $\hat{Y} = 0$

Détecter l'iniquité (3)

- Différences entre décision humaine et décision algorithmique
  - Décision humaine : peut facilement être non aveugle et/ou non systématique (arbitraire), en revanche ne peut pas traiter un grand nombre de variables
  - Décision algorithmique : systématique, on peut exclure *a priori* les attributs protégés, mais elle peut les inférer à partir des autres variables
- Conséquences pour systèmes algorithmiques :
  - Tests d'aveuglement peu pratiqués car les attributs protégés sont exclus *a priori* ; *blinding* peu utile car attributs protégés peuvent souvent être inférés des autres variables individuellement anodines
  - Révéler l'arbitraire : procédure algorithmique en principe systématique, mais un excès de variables explicatives peut permettre de trouver prédictifs des facteurs arbitraires
  - Critères observationnels souvent facilement applicables : parité démographique, parité des taux d'erreur, calibration ; à considérer avec précaution, utiles pour motiver une analyse plus approfondie des mécanismes d'apparition d'iniquités

Détecter l'iniquité : interprétation des résultats de tests

- Important de comprendre le mécanisme exact qui mène au résultat d'un test dans chaque situation, par ex.
  - La précision de classement de visages par genre diffère suivant la couleur de peau : bases d'apprentissage effectivement biaisées en faveur d'une couleur de peau, mais aussi contenant surtout des photos de célébrités où les femmes emploient plus de maquillage que dans la population générale
  - La personnalisation des résultats de moteurs de recherche (des recherches équivalentes donnent des résultats qui dépendent des recherches antérieures, renforçant ainsi les croyances de la personne qui cherche) : les requêtes « équivalentes » ne sont pas identiques, chaque personne emploie des termes spécifiques suivant ses croyances, ces termes orientent les résultats différemment
  - Un site de vente en ligne qui a aussi des magasins en dur propose des réductions plus fortes pour certains *zip codes* que pour d'autres (*zip code* est un *proxy* pour « race ») : la réduction augmente avec la concurrence locale autour du *zip code*, concurrence qui n'est pas la même pour tous les *zip codes* (elle est plus forte là où le pouvoir d'achat local est supérieur)

Supprimer ou réduire l'iniquité : approches[6]

1. Prétraitement : transformer la représentation des données pour la rendre indépendante des attributs protégés
   - Exige l'accès complet aux données de départ et à l'attribut protégé
   - Valable quelles que soient les méthodes de modélisation appliquées en aval
2. Inclure le respect de la contrainte dans le processus d'optimisation qui permet d'obtenir le modèle décisionnel
   - Exige l'accès aux données, à l'attribut protégé et à la procédure d'apprentissage
   - Contraintes sur les familles de modèles utilisées et sur les procédures d'optimisation
3. Post-traitement : ajuster le modèle décisionnel obtenu pour que ses décisions respectent le critère
   - Ne nécessite pas l'accès aux données ou à la procédure d'apprentissage
   - Peut s'appliquer à tout modèle (y compris « boîte noire »)

[6] Certaines formulations sont adaptées au critère d'indépendance mais les mêmes approches s'appliquent à d'autres critères.

Supprimer ou réduire l'iniquité : exemples de méthodes

1. Prétraitement
   - Compléter l'information : si possible, compléter la collecte de données en suivant des plans d'expérience
   - Redressement de l'échantillon (par ex. augmentation des pondérations des individus dans les populations sous-représentées) ou ré-échantillonnage (sous/sur-échantillonnage, génération par interpolation)
2. Inclure le respect de la contrainte dans le processus d'apprentissage
   - Introduction de termes de « coût d'iniquité » dans la fonction globale de coût à minimiser
3. Post-traitement
   - Randomisation partielle des prédictions pour respecter un critère observationnel d'équité, voir par ex. [3] (approche mise en œuvre dans cette séance de travaux pratiques)
   - Modification sélective des seuils de décision (voir également [3])

Références I

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan.
Fairness and Machine Learning Limitations and Opportunities.
2018.

[2] CERNA Collectif.
Research Ethics in Machine Learning.
Technical report.

[3] Moritz Hardt, Eric Price, and Nati Srebro.
Equality of opportunity in supervised learning.
In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[4] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan.
A survey on bias and fairness in machine learning, 2019.

[5] Cathy O'Neil.
Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy.
Crown Publishing Group, USA, 2016.