Institut des ACTUAIRES

Bienvenue à toutes et à tous,
nous allons commencer dans un instant.

Institut A
4, rue Chauveau-Lagarde, 75008 paris - tél. 01 44 51 72 72
www.institutdesactuaires.com

Contributions à l'apprentissage semi-supervisé : équité et étiquetage dans les problèmes à classes multiples

François HU

07 février 2023

Institut des ACTUAIRES
4, rue Chauveau-Lagarde – 75008 Paris – tél. 01 44 51 72 72
www.institutdesactuaires.co
m

Remerciements

Caroline Hillairet,
Professeure à l'ENSAE

Romuald Elie,
Professeur à l'Université Gustave Eiffel

Marc Juillard,
Directeur du Data Hub de Société Générale Assurances

Sommaire

- Introduction
- **Défi 1** : étiquetage dynamique
- **Défi 2** : équité dans la classification multi-classes
- Conclusion + extensions

Mots-clés : apprentissage statistique, contrôle stochastique, optimisation.

Contexte général : processus d'apprentissage statistique

Sous ces conditions …

Données **massives**

+

Données **de bonnes qualités**

Données **labellisées**

… nous pouvons avoir :

Prédictions **précises**

Prédictions **robustes**

Contexte général : processus d'apprentissage statistique

Dataset | **Modèle d'apprentissage** | **Prédiction**

Sous ces conditions …

Données **massives**

+

Données **de bonnes qualités**

Données **labellisées**

Données **non biaisées**

… nous pouvons avoir :

Prédictions **précises**

Prédictions **robustes**

Prédictions **non biaisées**

Contexte général : processus d'apprentissage statistique

Dataset → **Modèle d'apprentissage** → **Prédiction**

Sous ces conditions …

Données **massives**

+

Données **de bonnes qualités**

Données **labellisées**

Données **non biaisées**

… nous pouvons avoir :

Prédictions **précises**

Prédictions **robustes**

Prédictions **non biaisées**

Quelques exemples en assurance :

Catégorisation des photos de **voitures endommagées**

Analyse des **données télématiques**

Détection de la conformité **GDPR**

Contexte général : processus d'apprentissage statistique

Dataset → **Modèle d'apprentissage** → **Prédiction**

Sous ces conditions …

Données **massives**

+

Données **de bonnes qualités**

Données **labellisées**

Données **non biaisées**

… nous pouvons avoir :

Prédictions **précises**

Prédictions **robustes**

Prédictions **non biaisées**

Quelques exemples en assurance :

Catégorisation des photos de **voitures endommagées**

Problème d'**étiquetage** :
besoin d'assureurs experts

Analyse des **données télématiques**

Problème **confidentialité** :
déduire certaines var. sensibles

Détection de la conformité **GDPR**

Problèmes d'**étiquetage** et d'**équité** :
besoin de juristes

Défi 1 : problèmes **d'étiquetage** **Défi 2 :** problème **d'équité**

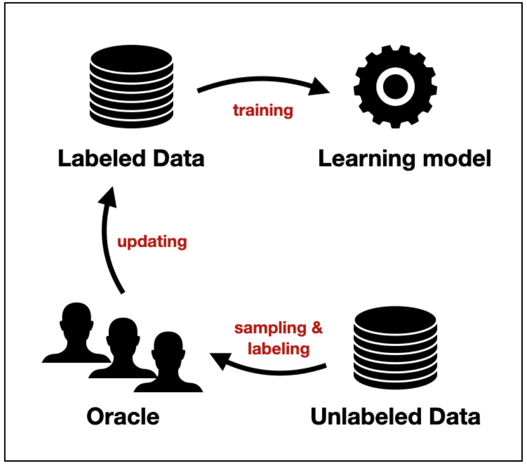
Défis et contributions

Défi 1 : entraîner un modèle ML avec un budget d'étiquetage limité

Défi 2 : garantir l'équité algorithmique dans les problèmes multi-classes

Défi 1 : quelques idées et leurs limites

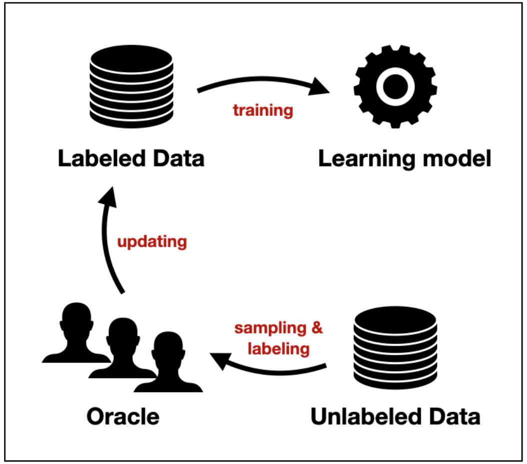
Étiquetage en parallèle



- ✗ coûteux et long
- ✓ génère des labels

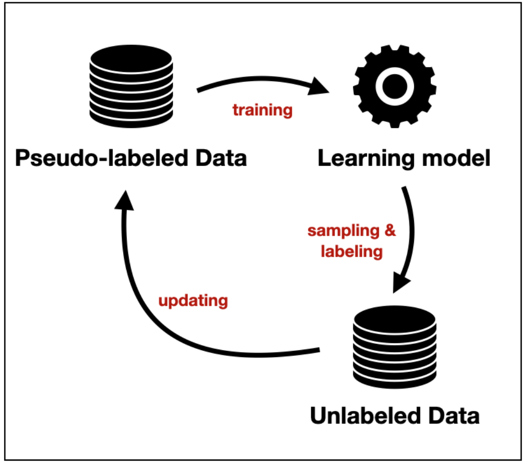
Défi 1 : quelques idées et leurs limites

Étiquetage en parallèle



- ✗ coûteux et long
- ✓ génère des labels

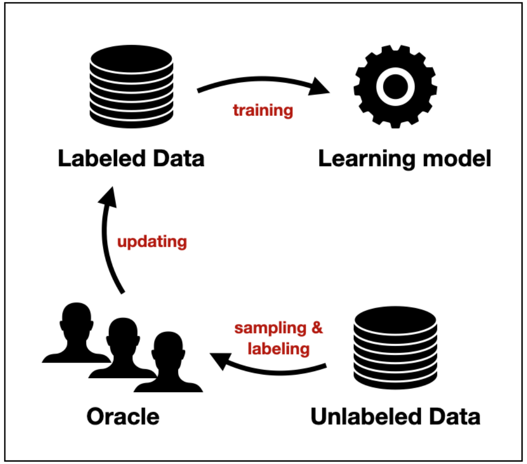
Apprentissage semi-supervisé



- ✓ moins coûteux, moins long
- ✗ génère des **pseudo**-labels

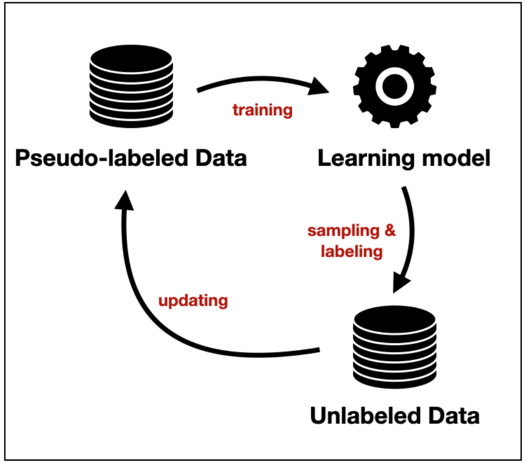
Défi 1 : Active Learning (AL)

Étiquetage en parallèle



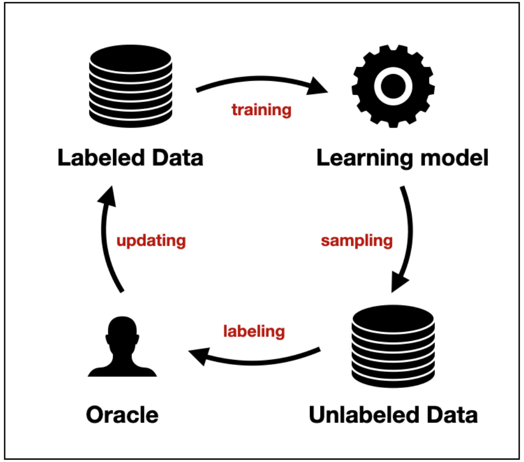
- ✗ coûteux et long
- ✓ génère des labels

Apprentissage semi-supervisé



- ✓ moins coûteux, moins long
- ✗ génère des **pseudo**-labels

Apprentissage actif (*Active Learning*)



- ✓ moins coûteux, moins long
- ✓ génère des labels

Exemple dans le cas multi-classes : entropie de Shannon

Algorithm 1 Outline of active learning (AL) process

Input: h base estimator, train-set $\mathcal{D}^{(train)}$, pool-set $\mathcal{D}_{\mathcal{X}}^{(pool)}$

Step 1. Fit h on the train-set $\mathcal{D}^{(train)}$

Step 2. Given a **score** $I(x, h)$, we sample:

$$\boxed{x^* = \underset{x \in \mathcal{D}_{\mathcal{X}}^{(pool)}}{\operatorname{argmax}} \{I(x, h)\}}$$

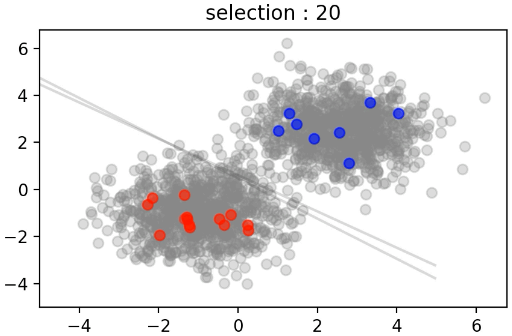
Example: Entropy-based [Sha48]

$$I(x, h) = -\sum_{k=1}^{K} \mathbb{P}(h(x) = k|x) \log \mathbb{P}(h(x) = k|x)$$

Step 3. If y^* is its **label** then we update:

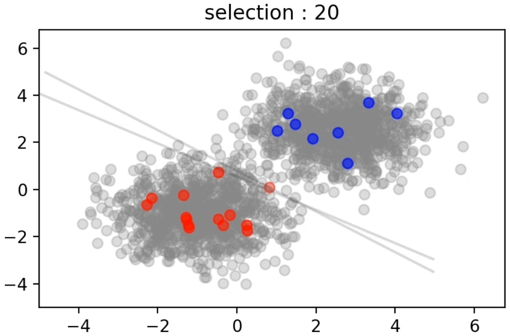
$$\mathcal{D}^{(train)} = \mathcal{D}^{(train)} \cup \{(x^*, y^*)\}$$
$$\mathcal{D}_{\mathcal{X}}^{(pool)} = \mathcal{D}_{\mathcal{X}}^{(pool)} - \{x^*\}$$

Step 4. Return to **step 1** until convergence.



Apprentissage **passif**

(Passive learning)



Apprentissage **actif**

(Active learning)

Exemple : entropie de Shannon

Algorithm 1 Outline of active learning (AL) process

Input: h base estimator, train-set $\mathcal{D}^{(train)}$, pool-set $\mathcal{D}_{\mathcal{X}}^{(pool)}$

Step 1. Fit h on the train-set $\mathcal{D}^{(train)}$

Step 2. Given a **score** $I(x, h)$, we sample:

$$\boxed{x^* = \underset{x \in \mathcal{D}_{\mathcal{X}}^{(pool)}}{\operatorname{argmax}} \{I(x, h)\}}$$

Example: Entropy-based [Sha48]

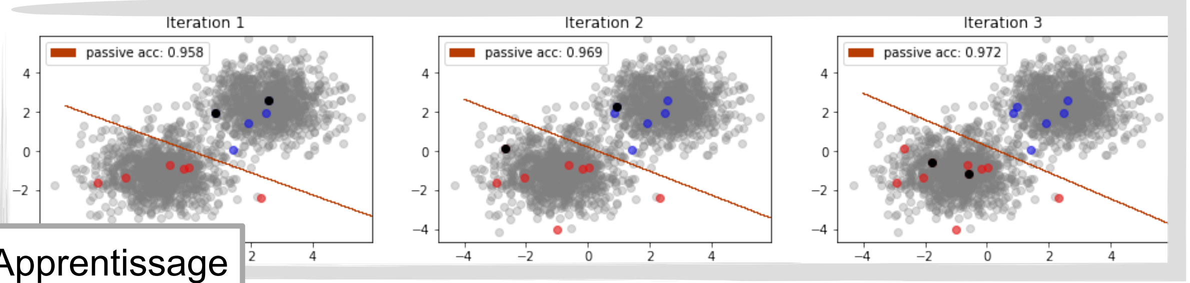
$$I(x, h) = -\sum_{k=1}^{K} \mathbb{P}(h(x) = k|x) \log \mathbb{P}(h(x) = k|x)$$

Step 3. If y^* is its **label** then we update:

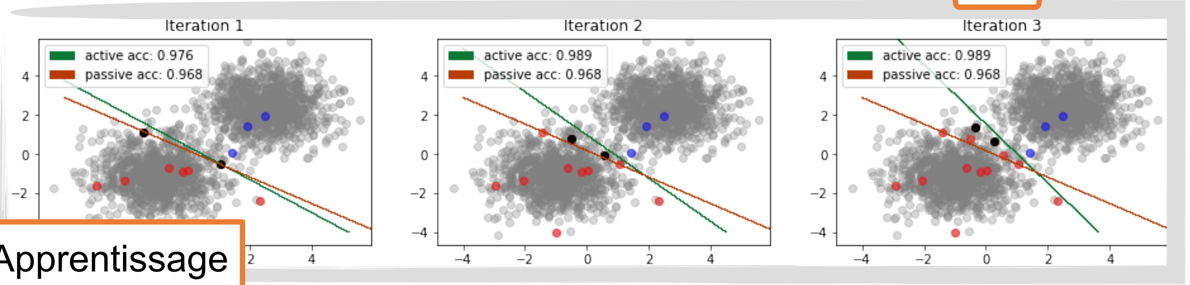
$$\mathcal{D}^{(train)} = \mathcal{D}^{(train)} \cup \{(x^*, y^*)\}$$

$$\mathcal{D}_{\mathcal{X}}^{(pool)} = \mathcal{D}_{\mathcal{X}}^{(pool)} - \{x^*\}$$

Step 4. Return to **step 1** until convergence.



Apprentissage **passif**



Apprentissage **actif**

Batch Mode Active Learning (BMAL)

Dataset : NPS de Sogécap (scores attribués par les clients sur un produit)

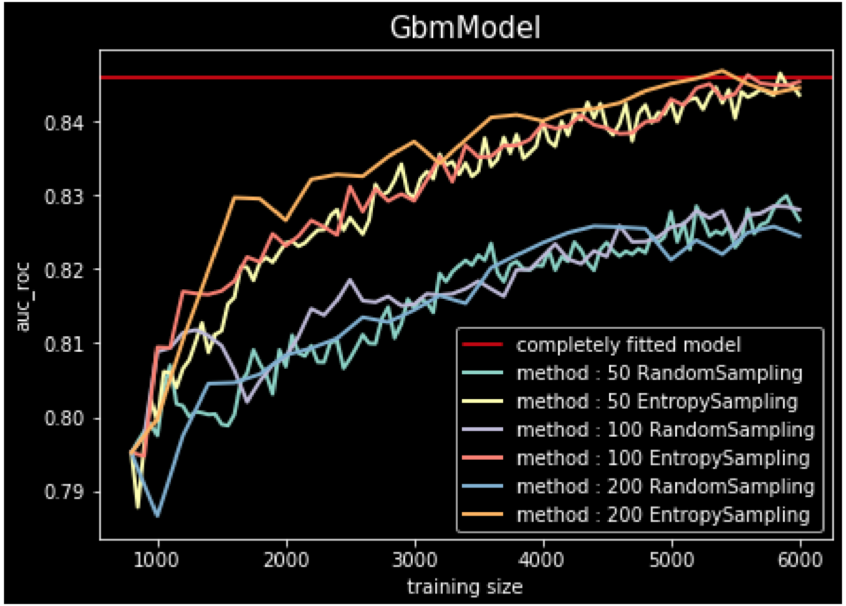
- **Verbatims** ($\mathscr{X}$) explication du score par le client, encodé par doc2vec
- **Analyse des sentiments** $\mathscr{Y}$ = { score $\leq$ T, score $>$ T }, T à déterminer

Processus d'AL : XGBoost + Entropie de **Shannon**

- **BMAL** : à chaque itération AL, échantillonner les b meilleures instances
- $b \in \{50,100,200\}$

Pourquoi utiliser **BMAL** en pratique ?

1. utile pour l'**étiquetage parallèle**
2. éviter le coût des **délais de ré-apprentissage**.



Batch Mode Active Learning (BMAL)

Dataset : **NPS** de Sogécap (scores attribués par les clients sur un produit)

- **Verbatims** ($\mathscr{X}$) explication du score par le client, encodé par doc2vec
- **Analyse des sentiments** $\mathscr{Y}$ = { score $\leq$ T, score $>$ T }, T à déterminer

Processus d'AL : **XGBoost** + Entropie de **Shannon**

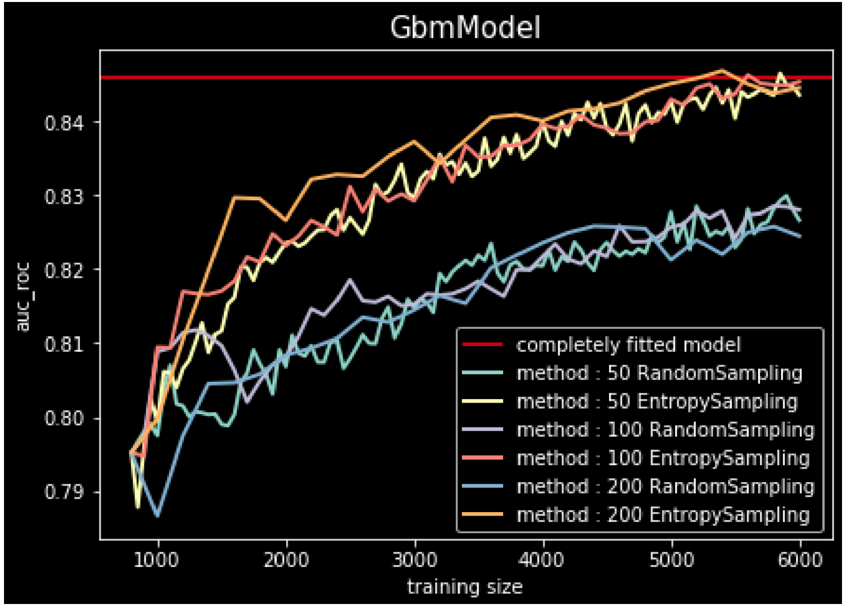
- **BMAL** : à chaque itération AL, échantillonner les b meilleures instances
- $b \in \{50,100,200\}$

Pourquoi utiliser **BMAL** en pratique ?

1. utile pour l'**étiquetage parallèle**
2. éviter le coût des **délais de ré-apprentissage**.

Objectif : trouver une séquence de tailles de lots $(b_t)_t$ avec un compromis entre

- **Maximiser** la performance du modèle
- **Réduire** le nombre d'itérations AL



Contribution : considérer BMAL comme un **Processus de Décision Markovien** (PDM)

Dynamic-Size Batch Mode Active Learning (DS-BMAL)

Dynamique des processus d'état (mouvement brownien W_t) :

(performance/qualité) (taille du train-set)

$$\begin{cases} dQ_t = \mu(B_t, b_t) \cdot Q_t(1 - Q_t) \cdot dt + \sigma(B_t, b_t) \cdot Q_t(1 - Q_t) \cdot dW_t \\ dB_t = b_t \cdot dt \end{cases}$$

Problème d'optimisation

U **fonction d'utilité** qui modélise l'aversion au risque du modèle étiquetage-performance

τ le **temps d'arrêt**

$$V_0 = \sup_b \mathbb{E}\left[U(Q_\tau) - \int_0^\tau C(b_s)ds \right]$$ (Fonction de valeur)

C **coût** supposé convexe p.r. au batch b

Paramètres :

- μ et σ sont déduits par analyse numérique :

$$\mu = \mu(B, b) \propto \frac{b}{B} \text{ et } \sigma = \sigma(B, b) \propto \frac{b}{B}$$

- C et U sont des sont des fonctions de puissance standard :

$$C(b) \propto b^2 \text{ et } U(Q) = Q^p, p \in (0,1).$$

Institut des ACTUAIRES

Dynamic-Size Batch Mode Active Learning (DS-BMAL)

Dynamique des processus d'état (mouvement brownien W_t) :

(performance/qualité) (taille du train-set)
$$\begin{cases} dQ_t = \mu(B_t, b_t) \cdot Q_t(1 - Q_t) \cdot dt + \sigma(B_t, b_t) \cdot Q_t(1 - Q_t) \cdot dW_t \\ dB_t = b_t \cdot dt \end{cases}$$

Problème d'optimisation

U **fonction d'utilité** qui modélise l'aversion au risque du modèle étiquetage-performance

τ le **temps d'arrêt**

$$V_0 = \sup_b \mathbb{E}\left[U(Q_\tau) - \int_0^\tau C(b_s)ds\right]$$ (Fonction de valeur)

C **coût** supposé convexe p.r. au batch b

Paramètres :

- μ et σ sont déduits par analyse numérique :

$$\mu = \mu(B, b) \propto \frac{b}{B} \text{ et } \sigma = \sigma(B, b) \propto \frac{b}{B}$$

- C et U sont des sont des fonctions de puissance standard :

$$C(b) \propto b^2 \text{ et } U(Q) = Q^p, p \in (0,1).$$

Principe de la programmation dynamique

$$V_t := v(Q_t, B_t) = \sup_{b_s,\ s\in[t,\tau]} \mathbb{E}\left[v(Q_{t+h}, B_{t+h}) - \int_t^{t+h} C(b_s)ds \,|\, \mathcal{F}_t^W\right]$$

Hamilton Jacobi Bellman (HJB) equation

Equation de HJB dans $[0,1] \times [0,B_{MAX}]$ nous donne :

$$\sup_{\substack{b \geq 0 \\ \leq B_{MAX} - B}} \underbrace{\left\{\mu Q(1-Q)\frac{\partial V}{\partial Q}(Q,B) + b\frac{\partial V}{\partial B}(Q,B) + \frac{1}{2}\sigma^2 Q^2(1-Q)^2\frac{\partial^2 V}{\partial Q^2}(Q,B) - C(b)\right\}}_{= \quad A(Q,B,b,V)} = 0$$

Avec les **conditions aux bords** :

$$v(0^+, B) = U(0)$$
$$v(1^-, B) = U(1)$$
$$v(Q, B_{MAX}) = U(Q) \quad \text{pour} \quad Q \in (0,1)$$

Institut des ACTUAIRES

Dynamic-Size Batch Mode Active Learning (DS-BMAL)

Dynamique des processus d'état (mouvement brownien W_t) :

$$\begin{cases} dQ_t = \mu(B_t, b_t)\cdot Q_t(1-Q_t)\cdot dt + \sigma(B_t,b_t)\cdot Q_t(1-Q_t)\cdot dW_t & \text{(performance/qualité)} \\ dB_t = b_t \cdot dt & \text{(taille du train-set)} \end{cases}$$

Problème d'optimisation

U **fonction d'utilité** qui modélise l'aversion au risque du modèle étiquetage-performance

τ le **temps d'arrêt**

$$V_0 = \sup_b \mathbb{E}\left[U(Q_\tau) - \int_0^\tau C(b_s)ds\right] \quad \text{(Fonction de valeur)}$$

C **coût** supposé convexe p.r. au batch b

Paramètres :

- μ et σ sont déduits par analyse numérique :

$$\mu = \mu(B,b) \propto \frac{b}{B} \text{ et } \sigma = \sigma(B,b) \propto \frac{b}{B}$$

- C et U sont des sont des fonctions de puissance standard :

$$C(b) \propto b^2 \text{ et } U(Q) = Q^p, p \in (0,1).$$

Principe de la programmation dynamique

$$V_t := v(Q_t, B_t) = \sup_{b_s,\ s\in[t,\tau]} \mathbb{E}\left[v(Q_{t+h}, B_{t+h}) - \int_t^{t+h} C(b_s)ds \,|\, \mathcal{F}_t^W\right]$$

Hamilton Jacobi Bellman (HJB) equation

Equation de HJB dans $[0,1]\times[0,B_{MAX}]$ nous donne :

$$\sup_{\substack{b\geq 0 \\ \leq B_{MAX}-B}} \underbrace{\left\{\mu Q(1-Q)\frac{\partial V}{\partial Q}(Q,B) + b\frac{\partial V}{\partial B}(Q,B) + \frac{1}{2}\sigma^2 Q^2(1-Q)^2\frac{\partial^2 V}{\partial Q^2}(Q,B) - C(b)\right\}}_{=\quad A(Q,B,b,V)} = 0$$

Avec les **conditions aux bords** :

$$v(0^+, B) = U(0)$$
$$v(1^-, B) = U(1)$$
$$v(Q, B_{MAX}) = U(Q) \quad \text{pour} \quad Q \in (0,1)$$

Résolution numérique pour trouver b^* optimal:

- Différences finies + Algorithme de Howard

Résultats empiriques

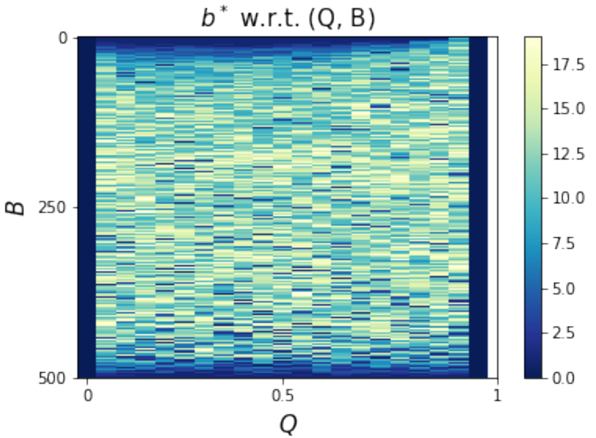


Figure **a** - Heatmap de b^* p.r. au processus (B, Q)

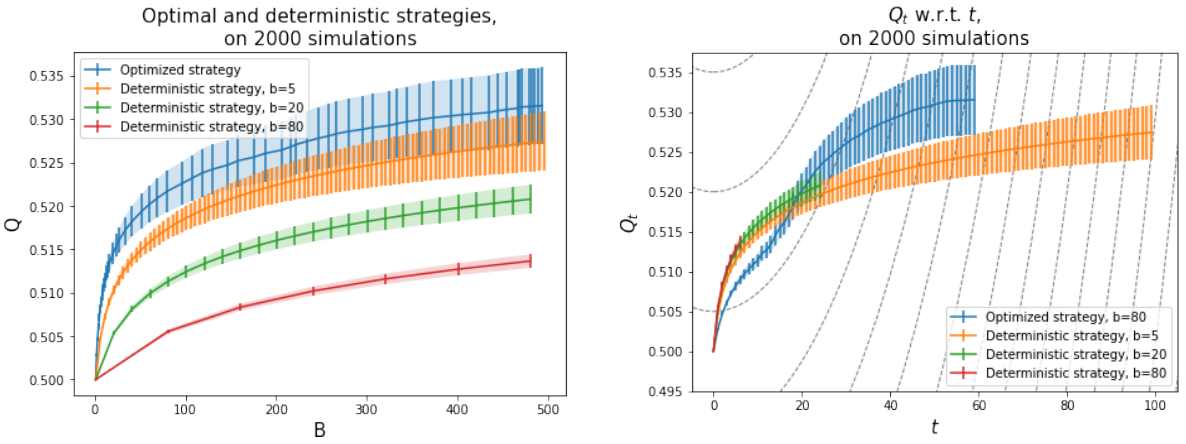


Figure **b et c** - stratégie **optimale** vs **déterministe** avec $b \in \{5{,}20{,}80\}$ et $(B_0, Q_0) = (0{,}0.5)$.

La stratégie optimisée (DS-BMAL) montre de meilleurs résultats que les stratégies déterministes en termes de :

- **qualité du modèle** en étiquetant moins au début du processus
- **réduction des délais de réapprentissage**.
  En effet, cette stratégie dynamique permet de réduire considérablement le nombre d'itérations.

Défis et contributions

Défi 1 : entraîner un modèle ML avec un budget d'étiquetage limité

Contributions : Des analyses numériques présentent DS-BMAL une stratégie optimale d'étiquetage qui réduit considérablement le nombre d'itérations d'AL tout en gardant une bonne performance du modèle. Cela permettrait d'améliorer les conditions d'étiquetage pour les experts humains.

Défi 2 : garantir l'équité algorithmique dans les problèmes multi-classes

Institut des ACTUAIRES

Défi 2 : équité dans la classification multi-classes

Données : $(\underbrace{\text{variables}}_{X}, \underbrace{\text{variable sensible}}_{S}, \underbrace{\text{label-}}_{Y}) \sim \mathbb{P}$ sur $\mathcal{X} \times \mathcal{S} \times [K]$. Considérons $\mathcal{S} = \{+1, -1\}$, **distribution** de S : $(\pi_s)_{s \in \mathcal{S}}$

Défi 2 : équité dans la classification multi-classes

Données : $(\underbrace{\text{variables}}_{X}, \underbrace{\text{variable sensible}}_{S}, \underbrace{\text{label-}}_{Y}) \sim \mathbb{P}$ sur $\mathcal{X} \times \mathcal{S} \times [K]$. Considérons $\mathcal{S} = \{+1, -1\}$ **, distribution** de S : $(\pi_s)_{s \in \mathcal{S}}$

Parité démographique (*Demographic Parity* or DP) : classifier $g \in \mathcal{H}$ est

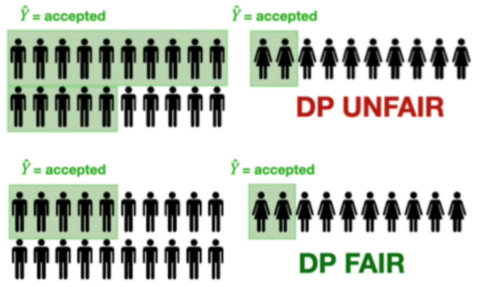
- **Exactement équitable** si

$$\mathbb{P}\left(g(X,S) = k \,|\, S = 1\right) = \mathbb{P}\left(g(X,S) = k \,|\, S = -1\right) \quad \forall k \in [K].$$

- **Approximativement** (ou ε-) **équitable** si pour $\varepsilon \geq 0$,

$$\left|\mathbb{P}\left(g(X,S) = k \,|\, S = 1\right) - \mathbb{P}\left(g(X,S) = k \,|\, S = -1\right)\right| \leq \varepsilon \quad \forall k \in [K].$$

Exemple : entretien d'embauche



Défi 2 : équité dans la classification multi-classes

Données : $(\underbrace{\text{variables},}_{X}\ \underbrace{\text{variable sensible},}_{S}\ \underbrace{\text{label-}}_{Y}) \sim \mathbb{P}$ sur $\mathcal{X} \times \mathcal{S} \times [K]$. Considérons $\mathcal{S} = \{+1, -1\}$ **, distribution** de S : $(\pi_s)_{s \in \mathcal{S}}$

Parité démographique (*Demographic Parity* or DP) : classifier $g \in \mathcal{H}$ est

- **Exactement équitable** si

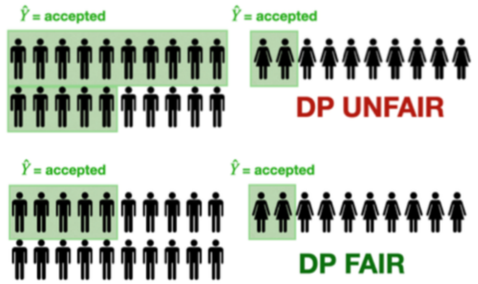
$$\mathbb{P}\left(g(X,S) = k \,|\, S = 1\right) = \mathbb{P}\left(g(X,S) = k \,|\, S = -1\right) \quad \forall k \in [K].$$

- **Approximativement** (ou ε-) **équitable** si pour $\varepsilon \geq 0$,

$$\boxed{\left|\mathbb{P}\left(g(X,S) = k \,|\, S = 1\right) - \mathbb{P}\left(g(X,S) = k \,|\, S = -1\right)\right| \leq \varepsilon \quad \forall k \in [K].}$$

pour cette présentation !

Exemple : entretien d'embauche



Q : où réduire les biais ? **R** : réduire les biais …

dans les données (**pre-processing**)

lors de l'entraînement ML (**in-processing**)

après l'entraînement ML (**post-processing**)

Défi 2 : équité dans la classification multi-classes

Données : $(\underbrace{\text{variables},}_{X}\ \underbrace{\text{variable sensible},}_{S}\ \underbrace{\text{label-}}_{Y}) \sim \mathbb{P}$ sur $\mathcal{X} \times \mathcal{S} \times [K]$. Considérons $\mathcal{S} = \{+1, -1\}$ **, distribution** de S : $(\pi_s)_{s \in \mathcal{S}}$

Parité démographique (*Demographic Parity* or DP) : classifier $g \in \mathcal{H}$ est

- **Exactement équitable** si

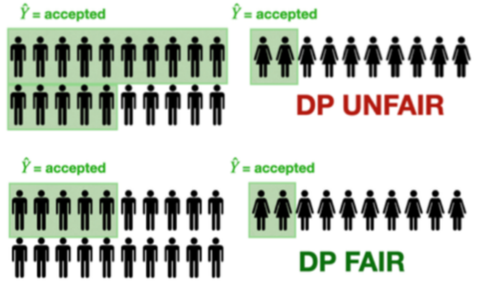
$$\mathbb{P}\left(g(X,S) = k \,|\, S = 1\right) = \mathbb{P}\left(g(X,S) = k \,|\, S = -1\right) \quad \forall k \in [K].$$

- **Approximativement** (ou ε-) **équitable** si pour $\varepsilon \geq 0$,

$$\boxed{\left|\mathbb{P}\left(g(X,S) = k \,|\, S = 1\right) - \mathbb{P}\left(g(X,S) = k \,|\, S = -1\right)\right| \leq \varepsilon \quad \forall k \in [K].}$$

pour cette présentation !

Exemple : entretien d'embauche



Q : où réduire les biais ? **R** : réduire les biais …

Contribution :

- La plupart des travaux sur l'équité algorithmique : tâches **binaires** ou de **régression !**
- Cette thèse propose de renforcer l'équité algorithmique en **multi-classes** avec une stratégie **post-processing**

après l'entraînement ML
(**post-processing**)

Institut des ACTUAIRES

Equité approximative : notations et objectif

Risque (de mauvaise classification) pour un classifieur $g : \mathcal{X} \times \{-1,1\} \to [K]$

$$\mathcal{R}(g) := \mathbb{P}\left(g(X,S) \neq Y\right).$$

Scores : pour $k \in [K]$, on note

$$p_k(X,S) := \mathbb{P}\left(Y = k \,|\, X,S\right).$$

Classifieur de Bayes minimise le risque : $\forall (x,s) \in \mathcal{X} \times \mathcal{S}$

$$g^*(x,s) \in \arg\max_k p_k(x,s).$$

Equité approximative : notations et objectif

Risque (de mauvaise classification) pour un classifieur $g : \mathscr{X} \times \{-1,1\} \to [K]$

$$\mathscr{R}(g) := \mathbb{P}\left(g(X,S) \neq Y\right).$$

Scores : pour $k \in [K]$, on note

$$p_k(X,S) := \mathbb{P}\left(Y = k \,|\, X,S\right).$$

Classifieur de Bayes minimise le risque : $\forall (x,s) \in \mathscr{X} \times \mathscr{S}$

$$g^*(x,s) \in \arg\max_k p_k(x,s).$$

Objectif

- **Minimiser le risque** : $g^* \in \arg\min_g \mathscr{R}(g)$
- **Garantir l'équité** : pour $\varepsilon \geq 0$, $\max_{k\in[K]} \left|\mathbb{P}\left(g^*(X,S) = k \,|\, S = 1\right) - \mathbb{P}\left(g^*(X,S) = k \,|\, S = -1\right)\right| \leq \varepsilon$. (dénoté $g^* \in \mathscr{G}_{\varepsilon-\text{fair}}$)

Considérons son **Lagrangien** et introduisons pour $\lambda^{(1)} = (\lambda_1^{(1)}, \ldots, \lambda_K^{(1)}) \in \mathbb{R}_+^K$ et $\lambda^{(2)} = (\lambda_1^{(2)}, \ldots, \lambda_K^{(2)}) \in \mathbb{R}_+^K$,

$$\mathcal{R}_{\lambda^{(1)},\lambda^{(2)}}(g) := \mathcal{R}(g) + \sum_{k=1}^{K} \lambda_k^{(1)} \left[\mathbb{P}\left(g(X,S) = k|S = 1\right) - \mathbb{P}\left(g(X,S) = k|S = -1\right) - \varepsilon\right] \qquad \text{(risque-équitable)}$$

$$+ \sum_{k=1}^{K} \lambda_k^{(2)} \left[\mathbb{P}\left(g(X,S) = k|S = -1\right) - \mathbb{P}\left(g(X,S) = k|S = 1\right) - \varepsilon\right].$$

Equité approximative : prédicteur équitable optimal

Hypothèse de continuité : $t \mapsto \mathbb{P}(p_k(X,S) - p_j(X,S) \leq t \,|\, S = s)$ considéré comme **continu** pour tout $k, j \in [K]$ et $s \in \mathcal{S}$.

- **Remarque** : en pratique cette hypothèse est satisfaite en donnant une petite perturbation uniforme.

Soit $H : \mathbb{R}_+^{2K} \to \mathbb{R}$ la fonction $H(\lambda^{(1)}, \lambda^{(2)}) = \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_k \left(\pi_s p_k(X,s) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right)\right] + \varepsilon \sum_{k=1}^{K} (\lambda_k^{(1)} + \lambda_k^{(2)})$.

Proposition

- Sous l'*hypothèse de continuité*, nous définissons $\lambda^{*(1)}, \lambda^{*(2)} \in \mathbb{R}_+^{2K}$ par

$$(\lambda^{*(1)}, \lambda^{*(2)}) \in \arg \min_{(\lambda^{(1)}, \lambda^{(2)}) \in \mathbb{R}_+^{2K}} H(\lambda^{(1)}, \lambda^{(2)}) .$$

Alors, $g^*_{\varepsilon-\mathrm{fair}} \in \arg \min_{g \in \mathcal{G}_{\varepsilon-\mathrm{fair}}} \mathcal{R}(g)$ **ssi** $g^*_{\varepsilon-\mathrm{fair}} \in \arg\min_{g \in \mathcal{G}} \mathcal{R}_{\lambda^{*(1)}, \lambda^{*(2)}}(g)$.

- De plus $\forall (x, s) \in \mathcal{X} \times \mathcal{S}$,

$$g^*_{\varepsilon-\mathrm{fair}}(x,s) = \arg\max_{k \in [K]} \left(\pi_s p_k(x,s) - s(\lambda_k^{*(1)} - \lambda_k^{*(2)})\right) .$$

Equité approximative : prédicteur équitable optimal + estimateur

Hypothèse de continuité : $t \mapsto \mathbb{P}(p_k(X,S) - p_j(X,S) \leq t \,|\, S = s)$ considéré comme **continu** pour tout $k, j \in [K]$ et $s \in \mathcal{S}$.

- **Remarque** : en pratique cette hypothèse est satisfaite en donnant une petite perturbation uniforme.

Soit $H : \mathbb{R}_+^{2K} \to \mathbb{R}$ la fonction $H(\lambda^{(1)}, \lambda^{(2)}) = \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_k \left(\pi_s p_k(X,s) - s(\lambda_k^{(1)} - \lambda_k^{(2)})\right)\right] + \varepsilon \sum_{k=1}^{K} (\lambda_k^{(1)} + \lambda_k^{(2)})$.

Proposition

- Sous l'*hypothèse de continuité*, nous définissons $\lambda^{*(1)}, \lambda^{*(2)} \in \mathbb{R}_+^{2K}$ par

$$(\lambda^{*(1)}, \lambda^{*(2)}) \in \arg \min_{(\lambda^{(1)},\lambda^{(2)})\in\mathbb{R}_+^{2K}} H(\lambda^{(1)}, \lambda^{(2)}) .$$

Alors, $g^*_{\varepsilon-\text{fair}} \in \arg \min_{g\in\mathcal{G}_{\varepsilon-\text{fair}}} \mathcal{R}(g)$ **ssi** $g^*_{\varepsilon-\text{fair}} \in \arg\min_{g\in\mathcal{G}} \mathcal{R}_{\lambda^{*(1)},\lambda^{*(2)}}(g)$.

- De plus $\forall (x,s) \in \mathcal{X} \times \mathcal{S}$,

$$g^*_{\varepsilon-\text{fair}}(x,s) = \arg\max_{k\in[K]} \left(\pi_s p_k(x,s) - s(\lambda_k^{*(1)} - \lambda_k^{*(2)})\right) .$$

Approche semi-supervisée :

- **Données étiquetées** : $\mathcal{D}_n = (X_i, S_i, Y_i)_{i=1,\ldots,n}$

Entraîner les estimateurs $(\hat{p}_k)_k$ (ex. RF, SVM, …)

Randomization trick $\bar{p}_k(X,S,\zeta_k) = \hat{p}_k(X,S) + \zeta_k$

- **Données non-étiquetées** : $X_1^s, \ldots, X_{N_s}^s \overset{\text{iid}}{\sim} \mathbb{P}_{X|S=s}$

Fréquences empiriques $(\hat{\pi}_s)_s$ comme estimateurs de $(\pi_s)_s$

Estimateur post-processing :

Si $(\hat{\lambda}^{(1)}, \hat{\lambda}^{(2)}) \in \arg \min_{(\lambda^{(1)},\lambda^{(2)})\in\mathbb{R}_+^{2K}} \hat{H}(\lambda^{(1)}, \lambda^{(2)})$

$$\boxed{\hat{g}_\varepsilon(x,s) = \arg\max_{k\in[K]} \left(\hat{\pi}_s \bar{p}_k(x,s,\zeta_k) - s(\hat{\lambda}_k^{(1)} - \hat{\lambda}_k^{(2)})\right)}$$

Equité approximative : garanties théoriques

Mesure d'iniquité (*Unfairness measure*). $\mathscr{U}(g) := \max_{k \in [K]} \left| \mathbb{P}\left(g(X,S)=k \,|\, S=1\right) - \mathbb{P}\left(g(X,S)=k \,|\, S=-1\right) \right|$

Théorème

Garantie universelle d'équité. Il existe une constante $C > 0$ dépendant uniquement de K et de $\min_{s \in S} \pi_s$, tel que, pour tout estimateur $\hat{p}_k$

$$|\mathbb{E}[\mathscr{U}(\hat{g}_\varepsilon)] - \varepsilon| \leq \frac{C}{\sqrt{N}}$$

Consistance. Si l'estimateur est consistant avec la norme L_1 alors,

$$\mathbb{E}\left[\mathscr{R}(\hat{g}_\varepsilon)\right] \to \mathscr{R}(g^*_{\varepsilon-\mathrm{fair}}) \text{ quand } n, N \to \infty$$

n = # données étiquetées

N = # données non-étiquetées

$\hat{g}_\varepsilon$ asymptotiquement aussi performant que g^* en termes **d'équité** et de **précision**

Résultats sur des données synthétiques

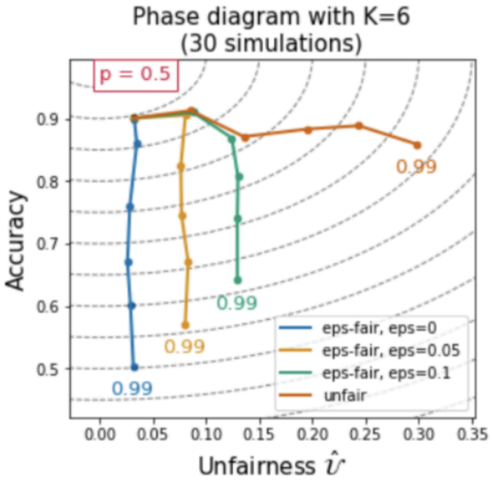
Dataset :

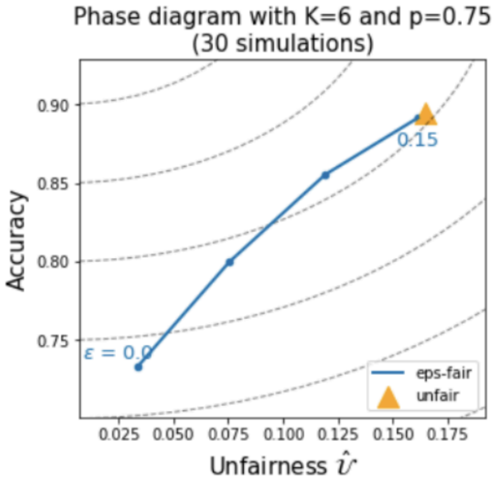
données **synthétiques,** pour $k \in [K]$

$(X \mid Y = k) \sim$ mélange Gaussien
$(S \mid Y = k) \sim 2 \cdot \mathscr{B}(p) - 1$, si $k \leq \lfloor K/2 \rfloor$,
$(S \mid Y = k) \sim 2 \cdot \mathscr{B}(1-p) - 1$, si $k > \lfloor K/2 \rfloor$.

- p mesure le biais historique dans le dataset
- K représente le nombre de classes

$$\max_{k \in [K]} \left| \mathbb{P}\left(Y = k \mid S = 1\right) - \mathbb{P}\left(Y = k \mid S = -1\right) \right| = \left| \frac{2}{K} \cdot (2p - 1) \right| .$$





Modèle d'apprentissage :

Random Forest (**RF**)

l'équité de g est mesurée par la version empirique de la mesure d'iniquité $\mathscr{U}(g)$

Lien vers code/expérimentations (Github)

SCAN ME

Résultats sur des données réelles : cas binaire (1/2)

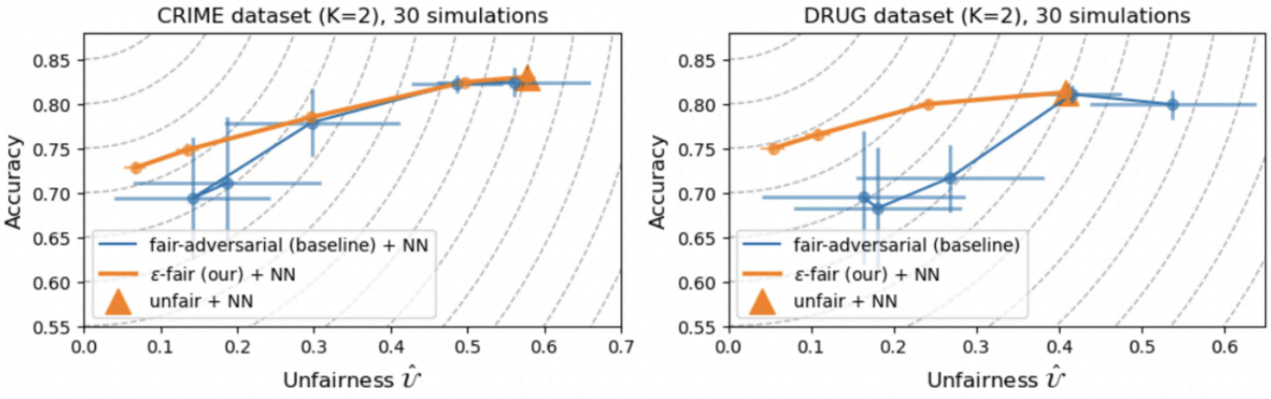
Dataset : DRUG, CRIME

Modèle d'apprentissage :

Réseaux de Neurones (**NN**)

Benchmarks :

Fair-adversarial[1]



Lien vers code/expérimentations (Github)

SCAN ME

1: approche **in-processing** https://aif360.readthedocs.io/en/stable/modules/generated/aif360.sklearn.inprocessing.AdversarialDebiasing.html

Résultats sur des données réelles : cas binaire (2/2)

Dataset : **DRUG, CRIME**

Modèle d'apprentissage :

Régression Logistique (**reglog**),

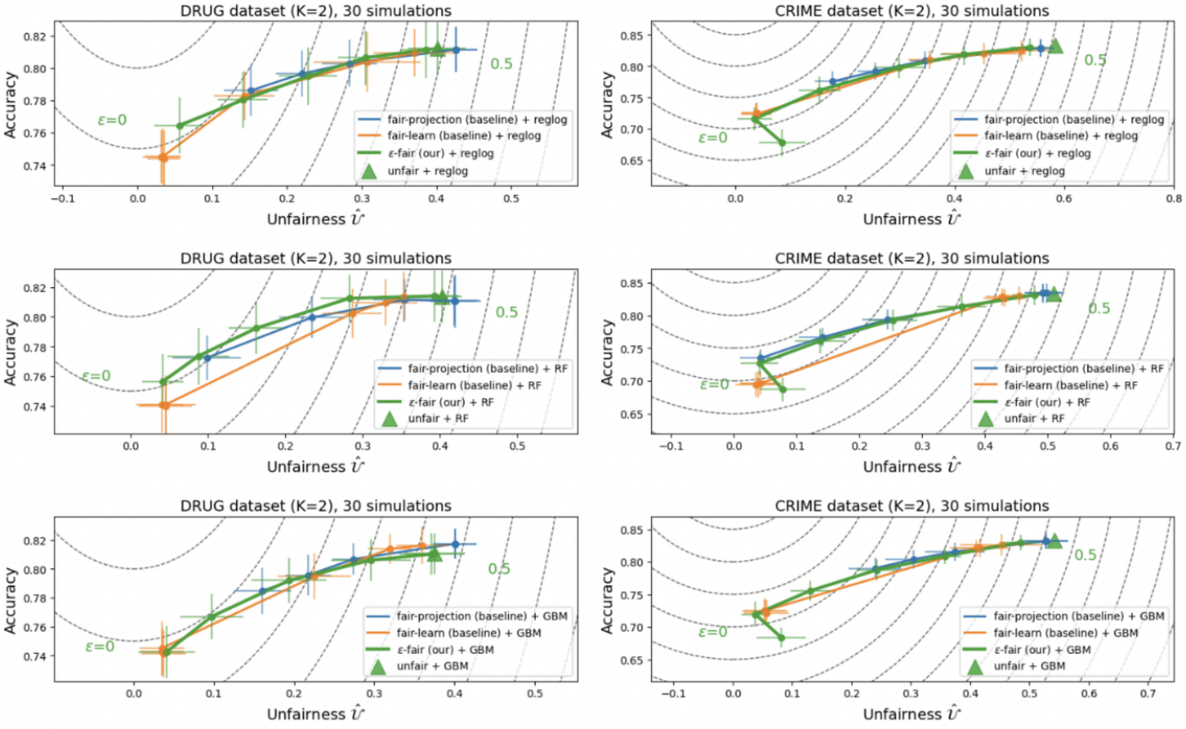
Random Forest (**RF**),

XGBoost (**GBM**)

Benchmarks :

Fair-learn[1]

Fair-projection[2]



1: approche **in-processing** pour binaire et régression https://fairlearn.org/ (Agarwal et al., ICLM 2018)
2: approche **post-processing** pour multi-classes https://github.com/HsiangHsu/Fair-Projection (NeurIPS 2022)

SCAN ME

Résultats sur des données réelles : cas multi-classes

Dataset : **DRUG, CRIME**

Modèle d'apprentissage :

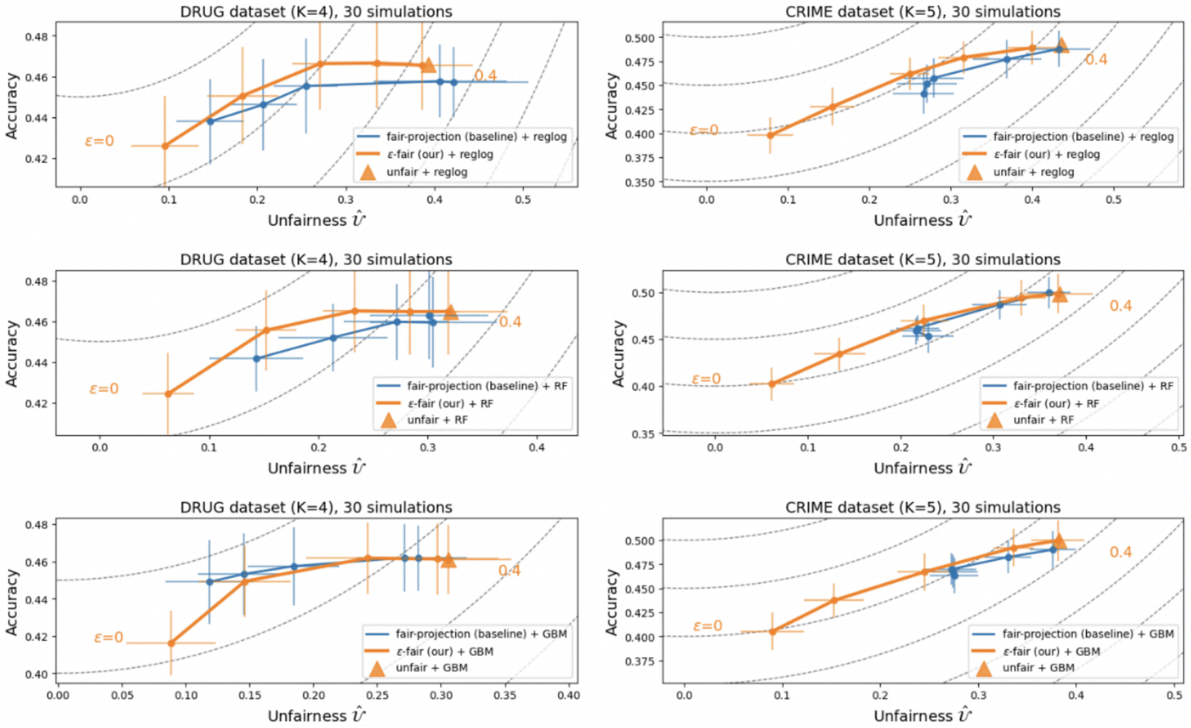
Régression Logistique (**reglog**),

Random Forest (**RF**),

XGBoost (**GBM**)

Benchmarks :

Fair-projection[1]



1: approche **post-processing** pour multi-classes https://github.com/HsiangHsu/Fair-Projection (NeurIPS 2022)

SCAN ME

Défis et contributions

Défi 1 : entraîner un modèle ML avec un budget d'étiquetage limité

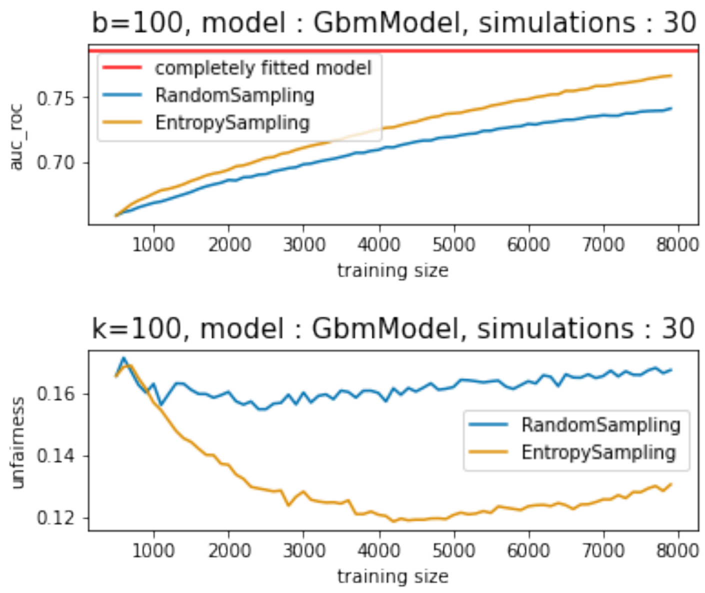
Contributions : Des analyses numériques présentent DS-BMAL une stratégie optimale d'étiquetage qui réduit considérablement le nombre d'itérations d'AL tout en gardant une bonne performance du modèle. Cela permettrait d'améliorer les conditions d'étiquetage pour les experts humains.

Défi 2 : garantir l'équité algorithmique dans les problèmes multi-classes

Contributions : Dans le cadre de la classification multi-classes, nous fournissons une règle de classification équitable optimale sous la contrainte DP. Nous traitons l'équité exacte et approximative et montrons que notre approche obtient des résultats remarquables sur divers ensembles de données synthétiques et réels. En particulier, notre algorithme est efficace pour faire respecter un niveau d'équité pré-spécifié.

Conclusion

Défi 1+2 : *fair active classifier* sur données d'assurance



Perspective

- Extension en **classification multi-labels :**

Au lieu de considérer **le risque de mauvaise classification**

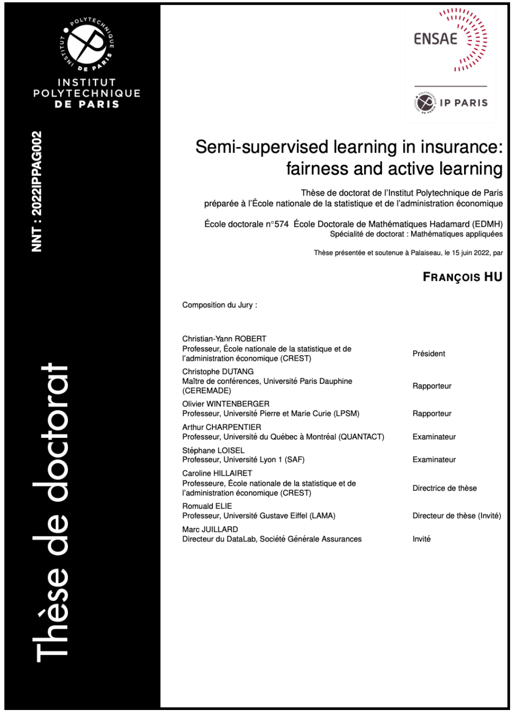
$$R(g) = \mathbb{P}(g(X,S) \neq Y)$$

Nous considérons **le risque** L_2

$$R_2(g) = \mathbb{E}\left[\sum_{k=1}^{K}\left(1_{Y=k} - g_k(X,S)\right)^2\right]$$

Bientôt un **package** sur Python pour l'équité algorithmique en multi-classes

Références



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAG002

Thèse de doctorat

ENSAE

IP PARIS

Semi-supervised learning in insurance: fairness and active learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 15 juin 2022, par

FRANÇOIS HU

Composition du Jury :

Christian-Yann ROBERT Professeur, École nationale de la statistique et de l'administration économique (CREST)	Président
Christophe DUTANG Maître de conférences, Université Paris Dauphine (CEREMADE)	Rapporteur
Olivier WINTENBERGER Professeur, Université Pierre et Marie Curie (LPSM)	Rapporteur
Arthur CHARPENTIER Professeur, Université du Québec à Montréal (QUANTACT)	Examinateur
Stéphane LOISEL Professeur, Université Lyon 1 (SAF)	Examinateur
Caroline HILLAIRET Professeure, École nationale de la statistique et de l'administration économique (CREST)	Directrice de thèse
Romuald ELIE Professeur, Université Gustave Eiffel (LAMA)	Directeur de thèse (Invité)
Marc JUILLARD Directeur du DataLab, Société Générale Assurances	Invité

PhD THESIS

PhD DEFENSE

PAGE PERSO

PhD THESIS : https://curiousml.github.io/research/TheseFrancoisHUSubmitted.pdf
PhD DEFENSE : https://curiousml.github.io/talks/SLIDES_PhD_defense_HU_VF.pdf
PAGE PERSO : https://curiousml.github.io/

Merci pour votre attention