Conception et implémentation d'un algorithme d'apprentissage machine produisant des modèles précis, équitables et interprétables

Julien FERRY - Informatique et Réseaux - Stage réalisé au LAAS-CNRS et supervisé par Marie-José Huguet et Mohamed Siala

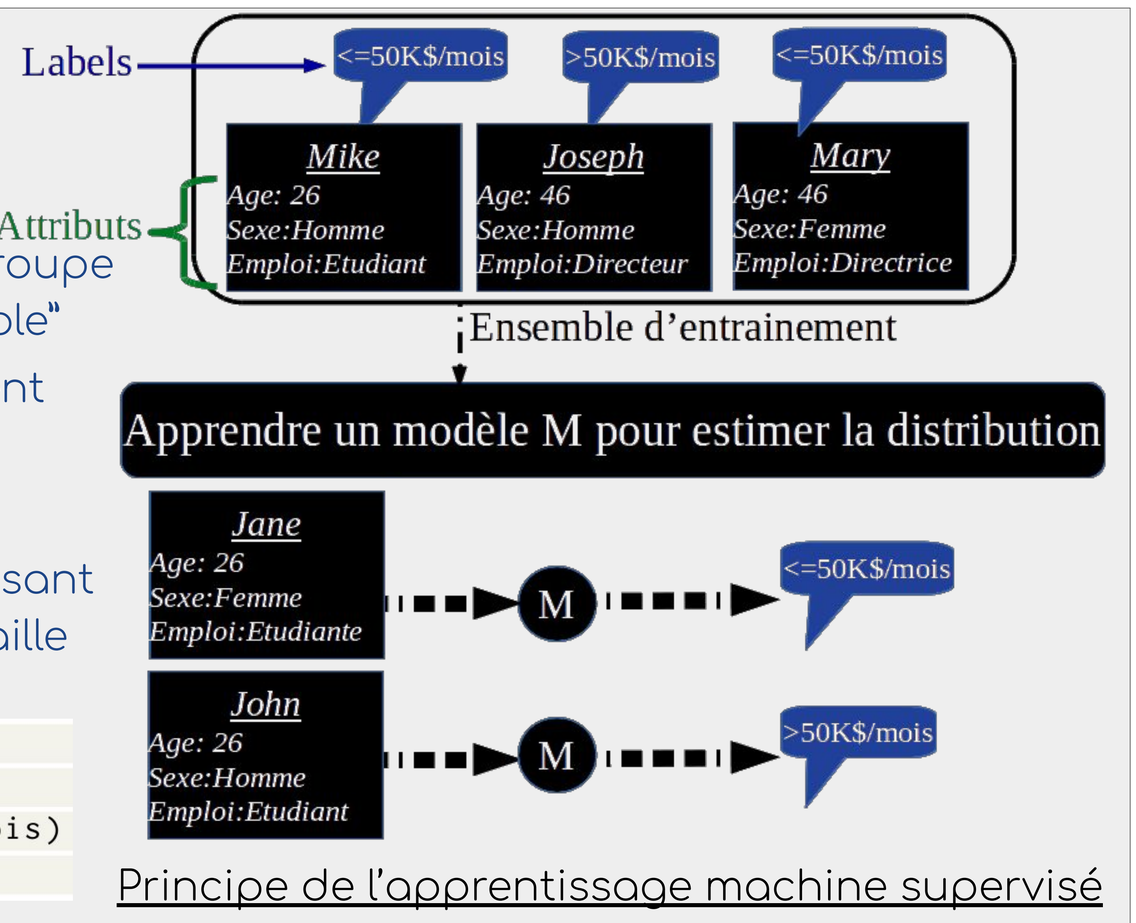
INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Contexte

- Collaboration entre deux équipes :
  - L'**équipe ROC**[1] du **LAAS-CNRS**[2] :
    - Recherche opérationnelle
    - Intelligence artificielle/Programmation par contraintes
  - La **Chaire de recherche du Canada en analyse respectueuse de la vie privée et éthique des données massives (UQAM**[3]**)**
- Stage au LAAS-CNRS (Toulouse) et mission de deux semaines à l'UQAM (Montréal, Canada)

Problématique et notions clés

- **Interprétabilité** : Explication des prédictions d'un modèle
- **Équité (fairness)** :
  - ***Statistique*** : équilibrer une certaine métrique entre un groupe "sensible" (instances discriminées) et un autre "non sensible"
  - ***Individuelle*** : deux instances aux attributs proches doivent avoir des prédictions proches
- **Rule list**[4] : type de modèle interprétable
- **CORELS**[5] : algorithme de type "branch and bound" produisant des rule lists certifiées optimales en termes de précision/taille



Principe de l'apprentissage machine supervisé

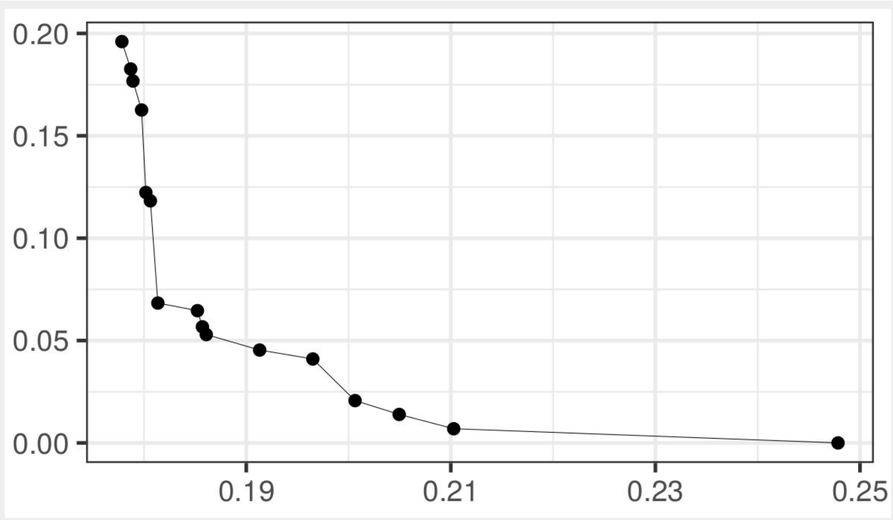
Exemple de rule list

```
if [sexe:Femme] then (<=50K$/mois)
else if [age:<22] then (<=50K$/mois)
else if [emploi:Directeur] then (>50K$/mois)
else (<=50K$/mois)
```

Objectif du stage: Conception et implémentation d'un algorithme d'apprentissage machine supervisé produisant des modèles interprétables, réalisant un compromis entre précision et équité

Contributions

- Etat de l'art des mesures d'équité en apprentissage machine
- Adaptation de **CORELS** en **fairCORELS** pour prendre en compte des contraintes d'équité lors de l'apprentissage
- Ajout de bornes pour plusieurs métriques d'équité , et de nouvelles stratégies d'exploration pour améliorer l'**efficacité** de fairCORELS
- Utilisation de **fairCORELS** pour générer un ensemble de compromis équité/précision (**front de Pareto**) avec différentes méthodes d'optimisation multi-objectif



Exemple de front de Pareto construit par fairCorels (abscisse : (1-précision) et ordonnée : (1-équité))

Conclusions scientifiques et apports

- **fairCORELS** :
  - Modèles interprétables proposant de **meilleurs compromis précision/équité** que la littérature
  - **Facilement paramétrable et intégrable** dans différents frameworks d'apprentissage (ensemble learning, etc.)
- Méthode et résultats présentés dans le **papier "Learning fair rule lists"**, soumis à la conférence ACM FAT* et publié sur ArXiv
- **Module Pypi** : "faircorels" (écrit en C++, Cython et Python)

- Acquis personnels :
  - **Scientifiques** (machine learning, optimisation combinatoire)
  - **Méthodologiques** (LateX, rédaction, présentation, évaluation expérimentale d'algorithmes)
  - **Techniques** (C++, Python, Cython, Pypi)
  - **Professionnels** (Découverte du milieu de la recherche, dans deux pays différents, à travers l'expérience de la co-écriture d'un papier)

(1) Recherche Opérationnelle, Optimisation Combinatoire et Contraintes, (2) Laboratoire d'Analyse et d'Architecture des Systèmes du CNRS, (3) Université du Québec à Montréal, (4) Règles de décision, en français, (5) Papier d'Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, et CynthiaRudin. 2017. "Learning certifiably optimal rule lists."