Introduction au domaine de recherche : Machine Learning appliqué à la psychiatrie

Introduction

La notion de trouble psychiatrique inclut une large variété de désordres dont on peut trouver des listes et des définitions exhaustives dans des manuels tels que le DSM V ou l'ICM 10. Parmi les plus sévères de ces troubles, en gravité et en incidence, la dépression va atteindre 15 à 20% de la population, la schizophrénie de 0.7 à 1% et les troubles bipolaires vont atteindre 1 à 2.5% de la population (Grande et al., 2016). Ces troubles psychatriques sont la première cause d'handicap dans le monde et induisent une importante charge sociale et financière pour les patients ainsi que pour leur proches.

Aujourd'hui le diagnostic est uniquement basé sur des entretiens entre un psychiatre ou un psychologue et un patient, durant lequel le praticien va tenter de detecter les symptômes d'un trouble psychiatrique au moyen d'un dialogue structuré avec le patient. Cette méthode induit des diagnostics qui dépendent fortement du clinicien et de la méthode d'entretien privilégié. Ces pratiques sont limités pour plusieurs raisons. La première est que ces entretiens dépendent fortement du clinicien, qui induit ses propres biais, et ainsi le diagnostic d'un même patient peut varier selon les clinicien. La seconde raison est dûe au fait que des troubles d'origines différentes vont avoir des symptômes en commun, ce qui rend difficile une identification directe d'un trouble au moyen de ses symptômes. Nous observons notamment un retard moyen de 10 ans dans le diagnostic d'un trouble bipolaire, souvent après une errance médicale et un diagnostic éronné de dépression, ce qui induit un stress sur le patient, une perte du lien et de la confiance clinicien-patient et des traitements inefficaces. Enfin, ces limitations dans le diagnostic induisent une limitation dans notre capacité à classer les troubles et à comprendre leur physio-pathologie (Alda, 2021), comme en témoigne le changement continuel de paradigmes et le fait qu'aucun consensus stable n'ait émergé d'une centaine d'année de recherches en psychiatrie.

La psychatrie a donc besoin de nouveaux biomarqueurs objectifs qui permettent de caracteriser ces troubles et de les discriminer. Ces marqueurs peuvent provenir de différentes sources. Parmi elles, l'IRM et les données génotypiques sont des sources de choix qui ont déjà été l'objet de recherches extensives, avec une littérature varié déjà disponible. Ces modalitées créent des ensembles de données de haute dimension, avec plusieurs dizaines de milliers de features : Les voxels des images IRM 3D et les mutation des séquences ADNs des sujets étudiés. Ces données de haute dimension, et le fait que les différences entre sujet sains et sujets malades ne sont pas visibles à l'oeil nu en regardant les images IRMs font des outils d'apprentissage statistique les outils idéaux pour étudier ces données et pour identifier des bio marqueurs naturels de différents troubles psychiatriques à partir de ces données.

Ainsi, depuis une quinzaine d'années, le domaine de l'apprentissage statistique a été étudié sur de nombreux troubles en commençant par les plus graves, tels que la dépression, la schizophrénie ou le trouble bipolaire, mais aussi sur des troubles tels que l'anxiété, les comportements addictifs ou l'hyperactivité. Dans cette introduction au domaine de recherche, nous nous concentrerons sur des études appliquées aux troubles bipolaires et à la schizophrénie qui utilisent des données de neuroimagerie, et nous présenterons des résultats donnant une intuition quand aux pistes méthodologiques à privilégier dans ce domaine.

Machine Learning

Dans cette section, nous expliquerons les outils principaux que la littérature utilise en machine learning appliqué à la psychiatrie. Le but de tous les modèles que nous allons présenter va être d'associer un diagnostic (sujet sain, patient bipolaire, patient schizophrène), représenté par un chiffre, à un sujet, représenté par un vecteur dans un espace de features. Selon les modèles que nous étudierons le label sera

donné comme entrée du problème (apprentissage supervisé) ou sera déduit par le modèle sans information de notre part (apprentissage non supervisé).

Modèles de classification

Dans cette sous partie nous nous interesserons au cas de l'apprentissage supervisé. Le formalisme est le suivant :

- Un sujet i est représenté par un vecteur $x_i \in \mathbb{R}^p$ où p est le nombre de features du sujet. Une feature peut être un voxel d'une image d'IRM, la présence ou non d'une mutation d'un gène, ou encore une donnée clinique (âge, sexe).
- Le label y_i du sujet i est un nombre entier $y_i \in \mathcal{C}$ où $K = |C|$ est le nombre de classes que le modèle introduit et $\mathcal{C}$ l'ensemble des classes. Le plus souvent, dans la litterature, nous considérons $y_i = 0, 1$ où 0 représente un sujet sain et 1 représente un sujet atteint d'un trouble donné.
- Le cadre probabiliste représente les couples (x_i, y_i) comme des tirages indépendants d'un couple de loi aléatoires (X, Y) de loi jointe $p_{X,Y}$ inconnue.
- Le choix d'un modèle va consister à choisir une fonction de décision :

  $$f : \mathbb{R}^p \longrightarrow 1, \dots, K$$

  Cette fonction associera à un sujet x_i un label $f(x_i)$. Nous la choisirons parmi une famille $\mathcal{F}$ de fonctions de $\mathbb{R}^p \longrightarrow 1, \dots, K$. Dans tous nos modèles, cette famille de fonctions de décision sera paramétré par un vecteur $\beta \in \mathbb{R}^d$, $d \in \mathbb{N}$ et ainsi

  $$\mathcal{F} = \{f_\beta, \beta \in \mathbf{R}^d\}$$

- Le choix du modèle sera un problème d'optimisation, et consistera à choisir les paramètres minimisant une fonction de coût $l : \mathcal{C} \times \mathcal{C} \longrightarrow \mathbb{R}$ propre au modèle étudié qui représente le cout d'une mauvaise classification. Dans le cadre probabiliste, le choix du modèle sera donc la résolution du problème d'optimisation suivant :

  $$\min_{\beta \in \mathbb{R}^d} \mathbb{E}_{p_{X,Y}}[l(Y, f_\beta(X))]$$

  Comme nous n'avons pas accès à la distribution $p_{X,Y}$, nous chercherons à résoudre ce problème en ne considérant qu'un échantillon $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ de la loi ĵointe (X, Y). Ainsi, nous cherchons à minimiser la fonction

  $$\hat{L}(\beta, \mathcal{D}_n) = \sum_{i=1}^{n} l(y_i, f_\beta(x_i))$$

Machine à vecteur de support

Les machines à vecteur de support (SVM), sont des modèles de classification binaires linéaires. Le principe est le suivant : Étant donné un ensemble d'entrainement $\mathcal{D}_n$ de n sujets, avec $y_i \in \{-1, 1\}$ (problème de classification binaire), nous cherchons un hyperplan affine de l'espace $\mathbb{R}^p$ séparant les sujets x_i tels que $y_i = -1$ et les sujets x_j tels que $y_j = 1$.

Cet hyperplan n'existe pas toujours, nous commencerons donc par le cas particulier d'un échantillon linéairement séparable, c'est un échantillon pour lequel il existe un hyperplan séparateur, c'est à dire un hyperplan H d'équation $\beta \cdot x + \beta_0 = 0$ tel que pour tout $(x_i, y_i) \in \mathcal{D}_n$ tel que $y_i = -1$, $\beta \cdot x_i + \beta_0 < 0$ et pour tout $(x_i, y_i) \in \mathcal{D}_n$ tel que $y_i = 1$, $\beta \cdot x_i + \beta_0 > 0$

Dans la suite, nous ferons l'abus de notation suivant : $\beta = (\beta_0, \beta)$ et $x_i = (1, x_i)$, afin de simplifier les équations précédantes par $\beta \cdot x_i < 0$ pour $y_i = -1$ et $\beta \cdot x_i > 0$ pour $y_i = 1$. Ainsi, un hyperplan séparateur est un hyperplan tel que :

$$\forall\, 1 \leq i \leq n,\ y_i \beta_i \cdot x_i > 0$$

Dans le cadre d'échantillons linéairement séparable, cet hyperplan n'est pas unique, il en existe même une infinité. Afin de restreindre les hyperplans que nous considérerons, nous choisirons celui qui 'sépare le mieux' les points, c'est à dire celui maximisant la distance à l'hyperplan des points les plus proches de l'hyperplan. Pour $1 \leq i \leq n$, comme $y_i \beta \cdot x_i > 0$, la distance de x_i à l'hyperplan est le nombre

$$d(x_i, H) = y_i \frac{x_i \cdot \beta}{\|\beta\|_2}$$

En introduisant par le réel $M > 0$ la marge à l'hyperplan, le problème d'optimisation à résoudre est donc le suivant :

$$\begin{aligned} &\max_{\beta, M} \ M \\ &\text{avec} \ \ y_i \frac{x_i \cdot \beta}{\|\beta\|_2} \geq M, i = 1, \dots, n \end{aligned}$$

En faisant le changement de variable $\tau = M\|\beta\|_2$, on obtient le problème équivalent :

$$\begin{aligned} &\max_{\beta, \tau} \ \frac{\tau}{\|\beta\|_2} \\ &\text{avec} \ \ y_i x_i \cdot \beta \geq \tau, i = 1, \dots, n \end{aligned}$$

Sous cette forme le problème est mal posé, car en multipliant β et τ par le même scalaire positif, nous obtenons une autre solution définissant le même hyperplan. Ainsi, nous fixons par convention $\tau = 1$. Le problème est donc équivalent au problème suivant :

$$\begin{aligned} &\min_{\beta} \ \|\beta\|_2^2 \\ &\text{avec} \ \ y_i x_i \cdot \beta \geq 1, i = 1, \dots, n \end{aligned}$$

C'est un problème d'optimisation quadratique (donc convexe) sous contraintes linéaires et admet une solution, par séparabilité de l'échantillon : Il admet donc une unique solution. Nous avons donc le résultat suivant :

Définition (SVM linéaire à données séparables) : *Soit $\mathcal{D}_n$ un échantillon de n vecteur $x_i \in \mathbb{R}^p$ labelisé par des $y_i \in \{-1, 1\}$. Nous les supposons linéairement séparables. Une machine à vecteur de support linéaire est un discriminateur linéaire de la forme $f_{\beta, \beta_0}(x) = \text{signe}(\beta \cdot x + \beta_0)$, avec β, β_0 les uniques solutions du problème de minimisation sous contrainte :*

$$\begin{aligned} &\min_{\beta} \|\beta\|_2^2 + \beta_0^2 \\ &\text{avec } y_i(x_i \cdot \beta + \beta_0) \geq 1, i = 1, \dots, n \end{aligned}$$

Bien sûr, en général $\mathcal{D}_n$ n'est pas linéairement séparable. Dans ce cas, nous introduisons une pénalité dans le cas ou les x_i franchissent la marge. Pour chaque $1 \leq i \leq n$, nous posons :

$$\xi_i = \max(0, 1 - y_i(x_i \cdot \beta + \beta_0))$$

ξ_i vérifie l'inégalité

$$y_i(x_i \cdot \beta + \beta_0) \geq 1 - \xi_i$$

Et quantifie à quel point x_i franchit la marge. Nous cherchons donc à minimiser les ξ_i. Le problème du SVM non séparable consistera alors à minimiser, pour un $C > 0$ et un $d = 1, 2$ la quantité

$$\hat{L}(\beta, \mathcal{D}_n) = \beta_0^2 + \sum_{i=1}^{n} \beta_i^2 + C \max(0, 1 - y_i(x_i \cdot \beta + \beta_0))^d$$

C sert à quantifier à quel point nous pénalisons les dépassements de marge et les mauvaises classifications. Il est usuellement déterminé par une procédure de gridsearch.

Ainsi la définition du SVM est la suivante :

Définition (SVM) : *Soit $\mathcal{D}_n$ un échantillon de n vecteur $x_i \in \mathbb{R}^p$ labelisé par des $y_i \in \{-1, 1\}$. Une machine à vecteur de support linéaire L^d de coefficient $C > 0$ est un discriminateur linéaire de la forme $f_{\beta,\beta_0}(x) = \text{signe}(\beta \cdot x + \beta_0)$, avec β, β_0 les uniques solutions du problème de minimisation :*

$$\min_{\beta,\beta_0} \hat{L}(\beta, \mathcal{D}_n) = \beta_0^2 + \sum_{i=1}^{n} \beta_i^2 + C \max(0, 1 - y_i(x_i \cdot \beta + \beta_0))^d$$

Regression logistique

La régression logistique est un autre modèle de classification binaire. Ici les labels de l'ensemble d'entrainement $\mathcal{D}_n$ seront des $y_i \in \{0, 1\}$.

Le principe général est le suivant : Le modèle de regression logistique va associer à chaque sujet x_i une loi binomiale $b_i \sim B(p_\beta(x_i))$, qui représente la prédiction que le modèle va faire, ainsi que sa confiance en sa prédiction : Si $p_\beta(x_i) = 0,7$, le modèle pense avec une "probabilité de $0,7$" que x_i appartient à la classe 1. Cela se formalise avec la loi b_i. Demander une classification au modèle revient à tirer la loi b_i : Il classifiera x_i comme appartenant à la classe 1 à une fréquence $p_\beta(x_i)$.

Nous cherchons alors la fonction $p_\beta : \mathbb{R}^d \longrightarrow [0, 1]$ maximisant la vraisemblance du modèle de régression logistique, c'est à dire la fonction qui maximise la probabilité qu'à le modèle de donner une bonne classification pour tous les éléments de l'ensemble d'entrainement en même temps (en supposant les classification, donc les b_i comme étant indépendantes). Soit alors $1 \leq i \leq n$. La probabilité que b_i donne une bonne classification est :

$$\mathbb{P}(b_i = y_i) = p_\beta(x_i)^{y_i}(1 - p_\beta)^{1-y_i}$$

Et donc la vraisemblance du modèle est :

$$L(\beta, \mathcal{D}_n) = \mathbb{P}(b_i = y_i, \forall 1 \leq i \leq n) = \prod_{i=1}^{n} p_\beta(x_i)^{y_i}(1 - p_\beta)^{1-y_i}$$

En pratique nous maximisons plus la log-vraisemblance, le logarithme de la vraisemblance, donné par :

$$l(\beta, \mathcal{D}_n) = \ln(L) = \sum_{i=1}^{n} y_i \ln(p_\beta(x_i)) + (1 - y_i) \ln(1 - p_\beta(x_i))$$

Nous cherchons maintenant la fonction p_β parmi une famille de fonctions paramétrées par un vecteur β . Comme dans le cas d'une régression linéaire classique, nous voudrions une fonction p_β de la forme $p_\beta(x) = \beta \cdot x + \beta_0$. Cela donne une fonction linéaire, simple à utiliser, dont on peut aisément interpréter les poids. Cependant, cette fonction n'est pas à valeurs dans $[0, 1]$. Afin de la restreindre à $[0, 1]$, nous allons la composer par une fonction dite sigmoïde, de la forme suivante :

$$f : x \longrightarrow \frac{1}{1 + e^{-x}}$$

C'est une fonction de $\mathbb{R} \to [0, 1]$, dérivable. Ainsi la forme finale de p_β sera donc :

$$p_\beta(x) = \frac{1}{1 + e^{-\beta \cdot x - \beta_0}}$$

Finalement, la définition du modèle de regression logistique linéaire est la suivante :

Définition (Régression logistique) : *Soit $\mathcal{D}_n$ un échantillon de n vecteur $x_i \in \mathbb{R}^p$ labelisé par des $y_i \in \{0, 1\}$. Un modèle de régression logistique discriminant les x_i est un discriminateur de la forme $f_{\beta,\beta_0}(x) = \text{signe}(p_\beta(x) - 1/2)$, avec*

$$p_\beta(x) = \frac{1}{1 + e^{-\beta \cdot x - \beta_0}}$$

et où les β, β_0 sont donnés par l'unique solution du problème de minimisation :

$$\min_{\beta,\beta_0} l(\beta, \mathcal{D}_n) = \sum_{i=1}^{n} y_i \ln(p_\beta(x_i)) + (1 - y_i) \ln(1 - p_\beta(x_i))$$

Méthodes de régularisation et sélection de features

Un des problèmes majeurs en machine learning appliqué à la neuroimagerie est la situation dite d'overfitting, ou de sur-apprentissage. Cette situation est dûe au fait que l'apprentissage du modèle est très sensible à $\mathcal{D}_n$. Le modèle peut alors être estimé et décrire précisément $\mathcal{D}_n$ sans avoir de pouvoir prédictif sur d'autres enchantillons. Il aura alors 'appris' le bruit aléatoire propre à l'échantillon $\mathcal{D}_n$ et non une bonne approximation de la loi $p_{X,Y}$. Ce phénomène arrive quand le nombre de paramètres du modèle est trop grand par rapport au nombre de sujets de l'échantillon ou quand le modèle est trop complexe par rapport à la loi que nous essayons d'estimer.

En neuroimagerie, et spécifiquement en psychiatrie, le nombre de features, qui est de l'ordre de 10^5 voir 10^6 pour une image d'IRM, excède de beaucoup le nombre de sujets, qui est de l'ordre de 10^2 en général, avec peu cohortes qui arrivent à un ordre de grandeur de 10^3 sujets. Ainsi, ce phénomène est extrêmement présent et de nombreux efforts sont réalisés afin de réussir à contrer cet effet. Cela passe par de l'ingénieurie de caractéristiques, c'est à dire à transformer les features initiales en un nombre plus petit de features représentant aussi bien le problème que nous cherchons à étudier, mais aussi par un choix d'une méthode de sélection de features au sein même du modèle ou par une méthode de régularisation qui forcera le modèle à être 'plus simple'.

Régularisation L^1-L^2

Dans le cas des modèles linéaires, l'overfitting se manifeste en général par un modèle dont les poids (un poids est un paramètre associé à une feature) vont être de grande ampleur et tous non nuls.

Une manière de contrer cet effet est de contraindre le modèle à adopter des coefficients plus petits et en plus petit nombre. Cela passe par une pénalisation des modèles ayant un grand nombre de poids avec de grandes valeurs. Ainsi lors de l'optimisation nous introduisons un terme de pénalisation, souvent paramétré par un réel positif λ qui marquera l'importance que nous voulons donner au terme de régularisation. Le problème d'optimisation revient alors au problème :

$$\min_{\beta} \hat{L}(\beta, \mathcal{D}_n) + \lambda \cdot \text{pen}(\beta)$$

De là nous avons différentes manières de paramétrer la pénalité :

- Régularisation L^2 : La régularisation L^2, ou régularisation Ridge, est la norme L^2 des poids :

  $$\text{pen}(\beta) = \|\beta\|_2^2$$

  Elle permet de pénaliser les poids de façon homogène, sans forcer certains poids à zéro tout en minimisant leur valeur totale
- Régularisation L^1 : Cette régularisation, aussi appelé Régularisation Lasso, est la norme L^1 des poids :

  $$\text{pen}(\beta) = \|\beta\|_1$$

  Elle difère de la régularisation L^2 en ce qu'elle favorise un modèle sparse, c'est à dire avec plus de poids à zéro, car à norme L^2 égale, le minimum de la norme L^1 est atteint pour des vecteurs avec des composantes nulles, ie des poids à zéro. C'est un avantage pour l'interpretabilité des modèles (voir plus bas), mais cette régularisation est plus instable (Si deux features sont importantes mais correlées, cette régularisation va en choisir une arbitrairement) et non différentiable : Il faut des algorithmes spéciaux pour l'optimisation.
- ElasticNet : Cette régularisation est une combinaison de la régularisation L^1 et L^2, parametré par un réel $\alpha \in [0, 1]$:

  $$\text{pen}(\beta) = \alpha\|\beta\|_1 + (1 - \alpha)\|\beta\|_2^2$$

Elle combine les avantages de la régularisation L^1 et L^2 en favorisant des régularisations sparses et stables en cas de variables correlées.

Ces méthodes de régularisation des modèles peut aussi bien s'appliquer aux modèles de regression logistique que les SVMs. Il est intéressant de noter que le paramètre λ représente une manière détournée de faire de la sélection de feature dans le cas de la régularisation L^1 et ElasticNet : Le plus il sera grand, le moins de features seront non nulles.

Sélection de features

La régularisation notamment la régularisation L^1, sont des méthodes indirectes de sélection de features. D'autres méthodes sont possible, nous en donneront deux exemples ici :

- Filtres univariés : Cette méthode consiste tout simplement à appliquer un test multiple (t-test, ANOVA, tests non paramétriques) en amont du modèle d'apprentissage et de sélectionner les k features séparant le mieux les deux ensembles de sujets que nous cherchons à classer.
- Elimination de features récursive (RFE) : Cette méthode consiste à éliminer une feature à la fois en ré-entrainant le modèle et en sélectionnant la feature participant le moins à la classification (Dans les modèles linéaires, nous pouvons considérer le poids, sinon des méthodes telles que la permutation de features peuvent donner une mesure de l'importance de chaque feature).

Afin d'éviter une fuite de données, chacun de ces sélections doivent être réalisées sur l'ensemble d'entrainement uniquement, en laissant l'ensemble de test à l'écart.

État de l'art

Comparaison des méthodes de selection de feature et des modèles non-linéaires :

Au vu des différentes méthodes de sélection de features et des modèles qui nous sont proposés, nous pouvons nous demander laquelle nous donne les meilleures performances. Cette section suivra les résultats présentés par (Duchesnay, n.d.), qui utilise plusieurs ensembles de données pour comparer différents modèles et en tirer des conduites à tenir lors du choix de modèle. Il faut toutefois garder en tête que chaque problème de classification en neuroimagerie est unique et que les résultats sur les comparaisons de méthodes obtenus sur un ensemble de données ne vont pas forcément se généraliser à un autre ensemble de données. Ainsi, même si les résultats que je vais présenter peuvent donner des intuitions sur le comportement des modèles en général, il faut rester critique quant à leur généralisation.

Enfin, les modèles présents dans la suite de cette section seront des modèles simples choisis pour leur robustesse par rapport à une situation d'overfitting, et présentent en général, pour les tailles d'échantillons actuelles, de meilleurs résultats et une meilleur capacité de généralisation que des RandomForest, XGBoost, LDA... et que des réseaux de neurones qui sur-apprennent.

Comparaison d'un algorithme linéaire contre un algorithme non linéaire

Le premier exemple est celui de la prédiction du sexe de 151 sujets (65 femmes pour 86 hommes) à partir de la mesure de plissements corticaux (116 régions d'intérêt) venant de la base de donnée ICBM. Les images ont subit une normalisation affine avant la mesure des plissements corticaux afin d'éviter d'apprendre sur les différences globales de volumes entre les hommes et les femmes.
Une des comparaison faite consiste en deux modèles :

- Modèle 1 : Filtre univarié combiné avec un SVM avec une régularisation L^2
- Modèle 2 : Filtre univarié combiné avec un SVM-RBF non linéaire

Nous comparons les modèles en utilisant une validation croisée (10-fold cross-validation) avec le filtre univarié dans la boucle de validation croisée afin d'éviter une fuite de données et donc des prédictions trop

optimistes. Nous obtenons le graphe de la balanced accuracy (précision équilibré vis à vis de la différence du nombre de sujets dans chaque classe) en fonction du nombre de features sélectionnées par le filtre univarié :

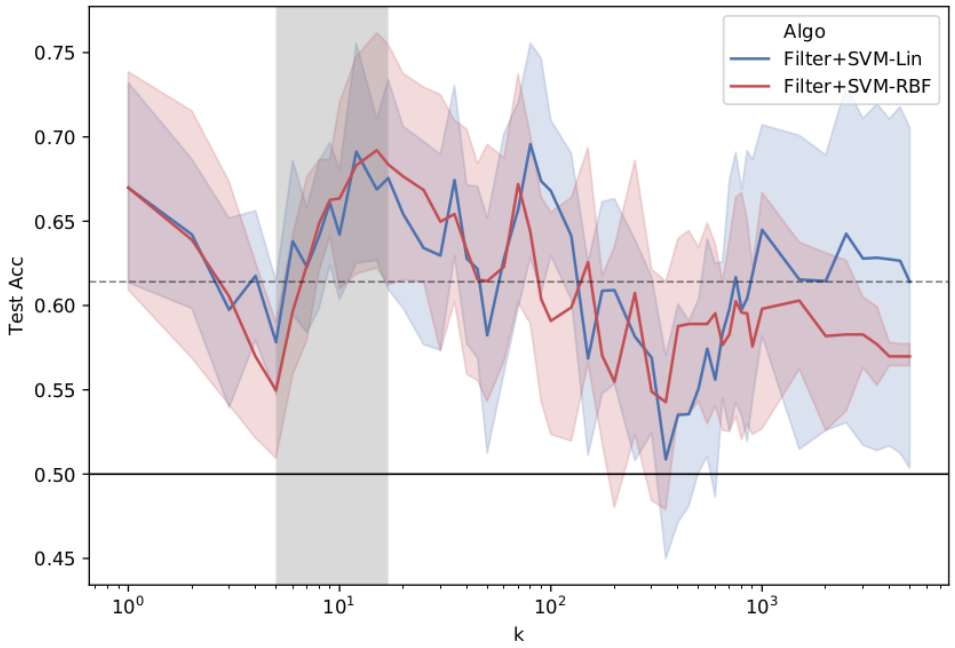


Figure 1 : Accuracy de l'ensemble de test en fonction du nombre de features sélectionnées. Les zones bleutées ou rougeatres représentent les intervalles de confiances vis à vis de la variation de l'accuracy en fonction des folds. La zone grisatre représente la zone où l'addition de features améliore l'accuracy.

Cet exemple montre que la sélection de feature, dans ce cas particulier, augmente bel et bien l'accuracy de façon consistante. Cependant, et c'était le but du test, nous remarquons que le SVM non-linéaire n'améliore pas l'accuracy de façon significative par rapport au SVM linéaire. Cela pourrait être dû a un mauvais calibrage des paramètres du SVM non-linéaire. En effet, comparé à un SVM linéaire qui est moins sensible au paramètrage de la pénalisation, un SVM non linéaire sera plus sensible à son paramétrage. Une autre explication est que les performances du modèle sont surtout dûes aux features que l'on met en input, et que le choix du modèle à partir de là n'aura qu'un effet marginal.

Comparaison des différentes méthodes de selection de feature

Pour cette comparaison, deux ensembles de données sont utilisés. Le premier est celui d'une étude visant à classifier des patients schizophrènes (SCZ) contre des contrôles sains (CTL). Elle comporte 330 controles pour 275 patients atteint de schizophrénie. Le second est celui d'une étude visant à classifier des patients bipolaires (BD) contre des controles (CTL). Elle comporte 356 controles contre 306 patients bipolaires.

Dans ces études, la sélection de feature est nécessaire car nous donnons en entrée le volume de matière grise contenue dans $\sim 360,000$ voxels qui englobent le cerveau. Nous comparons les régularisation L^1, L^2 et ElasticNet appliquées à une régression logistique, et deux méthodes de sélections de features, en premier un filtre univarié, en second une RFE. Enfin, tous ces résultats sont comparés à une baseline composée d'une régularisation L^2 avec le paramètre $C = 1$.

Les résultats sont les suivants :

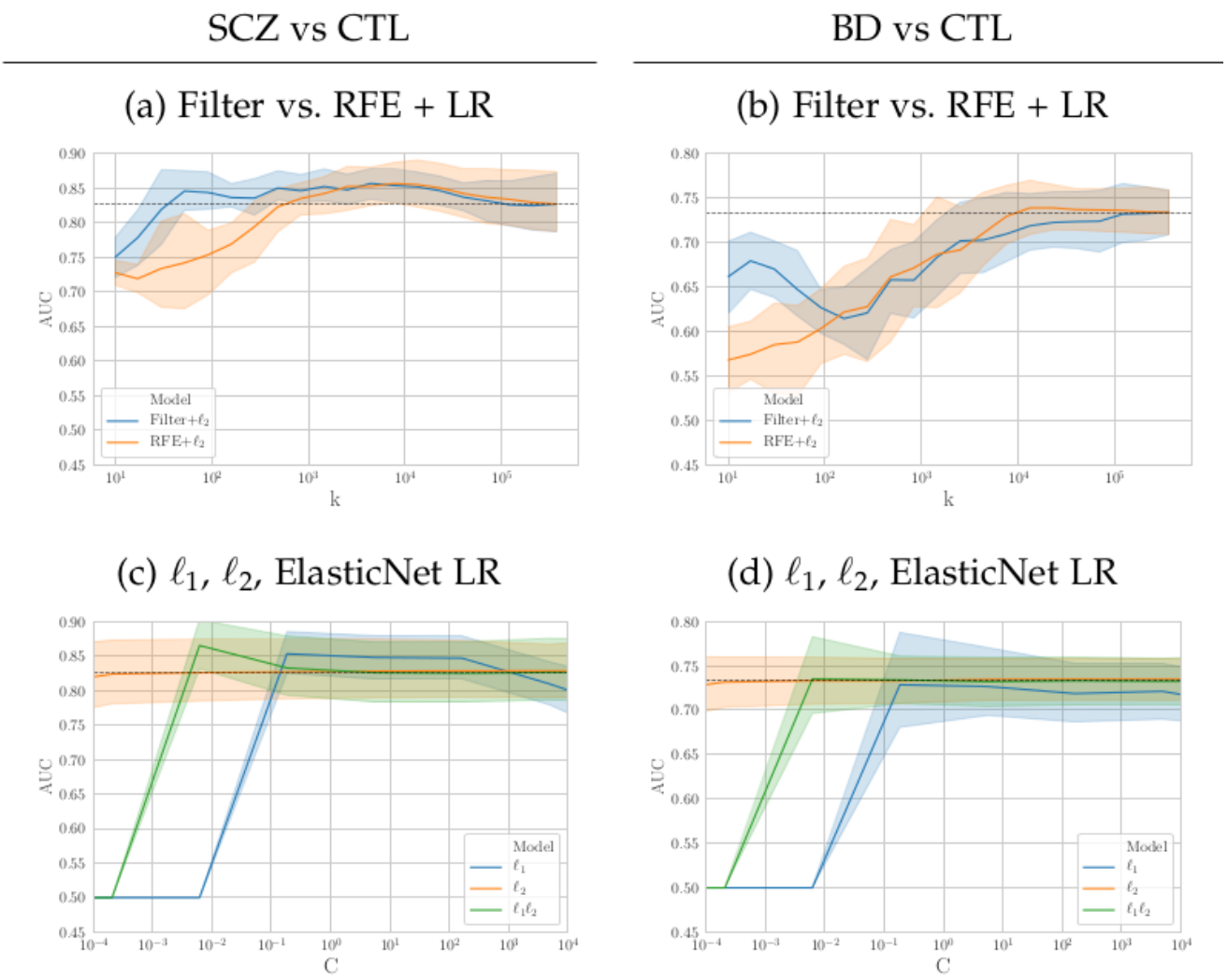


Figure 2 : Comparaison des différentes méthodes régularisation/sélection de feature. La ligne horizontale pointillé correspond à la baseline. Les zones colorées sont les intervalles de confiances de l'AUC issus de la validation croisée

En premier lieu, les exemples (a) et (b) montrent que la méthode de sélection de feature, que ce soit une RFE ou un filtre univarié, n'importe peu en terme de résultats. Sur la figure (a) nous observons que le filtre univarié obtient de meilleures performances (visibles par rapport à l'intervalle de confiance) pour un k entre 10^3 et 10^4 que la régularisation L^2 avec un $C = 1$. Cependant, nous ne pouvons rien déduire de ces meilleures performances car elles sont fines et peut être dépendantes de l'ensemble d'entrainement.

Pour la comparaison de régularisation L^1, L^2 et ElasticNet, nous observons que les régularisation L^1 et ElasticNet obtiennent de meilleures performances que la régularisation L^2 dans le cas (c) mais pas dans le (d), et que ces différences sont peu distinguables par rapport à la baseline.

Conclusions

Nous observons un faible impact des méthodes de selection de feature et des algorithmes choisis dans les performances des modèles. Ces résultats indiquent qu'il est probable que nous obtenons les performances maximales possibles compte tenu de la taille des échantillons et des features choisies avec des modèles linéaires simples et régularisés. Partant de ce constat, et à échantillons égaux, deux directions de recherches sont envisageables. Nous pouvons augmenter le nombre de modalités choisies et faire un travail de feature engieneering en amont afin de maximiser l'information que nous pouvons extraire des données en entrée. En second lieu, nous pouvons choisir les modèles possédant la meilleure interprétabilité afin de tirer le maximum de contenu sémantique des paramètres et peut être d'inférer des hypothèses physiopathologiques des résultats du modèle. Ainsi, le modèle ElasticNet semble intéressant car combinant la stabilité d'une régularisation L^2 et d'une solution sparse.

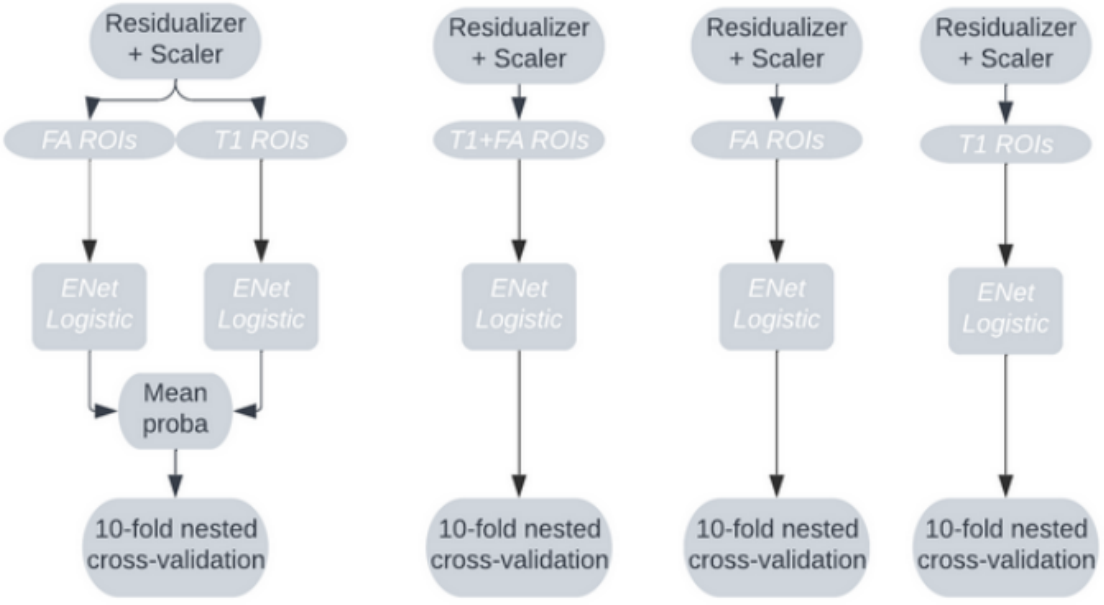
Exemple d'une étude : IRM multimodal pour la caractérisation du trouble bipolaire

Dans le cadre de mon Master 2, j'ai pu participer à une étude sur le trouble bipolaire. L'objectif de l'étude était multiple. En premier lieu, (Favre et al., 2019) a montré l'intérêt de l'IRM de diffusion dans le cadre de l'étude du trouble bipolaire. En second lieu, (Schulz et al., 2022) ont montré que la combinaison de plusieurs modalités d'IRM pouvaient améliorer les performances des algorithmes de classification. Dans cette étude, nous avons combiné deux modalités d'IRM différentes afin de voir si cela améliorait les performances de classification.

Les deux modalités utilisées sont donc :

- IRM structurelle : À partir d'images d'IRM structurelle, nous avons utilisé le logiciel CAT12 afin d'extraire les volume de matière grise de chaque voxels, que nous avons par la suite regroupé en régions d'intérêts et obtenus 141 régions d'intérêt.
- IRM de diffusion : Cette modalité d'IRM mesure la manière dont l'eau diffuse dans le cerveau. La Fraction Anisotropique (FA) en particulier, qui mesure l'anisotropie de la diffusion de l'eau, est un scalaire qui est utilisé dans de nombreuses études en imagerie. Cette valeur est considéré comme un proxy de la qualité et du nombre d'axones présent dans des directions données. L'idée étant que plus l'eau va diffuser dans une direction privilégiée, plus les axones seront gros et nombreux dans cette direction. Ainsi, en utilisant des pipelines préconisées par le groupe de travail ENIGMA, qui utilise des logiciels de prétraitement FSL et TBSS, nous avons extrait la FA moyenne de 48 régions d'intérets.

Nous avons donc comparé les prédictions de quatres modèles sur 490 sujets (227 patients bipolaires pour 263 contrôles). La démographie précise est présente dans la Table 1 du rapport de stage. Nous avons utilisé une méthode de Gridsearch dans une boucle de 10-fold cross validation pour estimer les paramètres de régularisation d'un modèle ElasticNet.



balanced accuracy	0.652(0.050)	0.639(0.070)	0.628(0.058)	0.588(0.087)
AUC	0.702(0.062)	0.700(0.071)	0.672(0.079)	0.629(0.097)

Figure 3 : Résultats de 4 modèles d'apprentissage statistique

Nous observons ici donc l'intéret de la combinaison de modalités qui améliore la prédiction des modèles. Pour le moment, nous n'avons pas de résultats significatifs, cette étude sur 490 sujets n'étant qu'une étude exploratoire. Durant ce stage de M2, nous avons aussi participé au traitement de 390 nouveaux sujets qui serviront dans la suite à obtenir des résultats statistiquement significatifs pour confirmer l'intuition que donne cette étude exploratoire.

En parallèle de ces résultats, nous avons pu tirer profit de l'interprétabilité de nos modèles et d'identifier des régions d'intérêts qui participent particulièrement à la prédiction. Notamment, nous avons observé que la combinaison de la moyenne de FA dans le cingulum gauche et dans colonne rayonnante droite obtenait une balanced accuracy de 0.617(0.057).

Domaines de recherche

Apprentissage profond

L'arrivée de l'apprentissage profond a fait exploser la recherche en apprentissage statistique. Les performances impressionnantes de ces modèles pour la reconnaissance d'images ont incité les chercheurs du monde de l'imagerie biomédicale à utiliser ces algorithmes afin de prédire des diagnostic cliniques.

Dans le cadre de la recherche en psychiatrie particulièrement, plusieurs limitations importantes paraissent immédiatement. La première, et la plus importante, est l'interprétabilité des ces modèles. Les maladies psychiatriques sont des troubles lourds, potentiellement mortels et en être atteint est extrêmement stigmatisé. Il est donc très important pour un clinicien qui utilise un modèle d'apprentissage profond pour un diagnostic de comprendre en premier lieu la signification exacte de la sortie de l'algorithme (est-ce un vrai diagnostic ou seulement l'expression d'un risque de développer la maladie au vu de la structure cérebrale ou de la génétique ?) et de comprendre les raisons de la décision de cet algorithme afin de voir si cela peut être dû à une erreur. Dans le cas d'analyse d'images directes, des méthodes d'interprétabilités telles que Grad-CAM permettent de montrer les régions d'intérêts pour l'algorithme, ce qui pourrait permettre une interprétabilité de ceux-ci. Toutefois, il faut garder en tête le fait que les algorithmes d'apprentissage profonds ne raisonnent pas comme des humains. En témoigne l'exemple de (Szegedy et al., 2014) où les auteurs ont pu changer la prédiction d'un modèle de deep learning en appliquant une modification invisible à l'oeil nu à certains pixels d'une image. Ainsi, il nous faut mieux comprendre comment ces algorithmes prennent leur décision avant de les utiliser dans le domaine médical.

La seconde limitation est le nombre de sujets des bases de données actuelles. Les bases de données sont limitées par la difficulté de recrutement de patients en psychiatrie. Ainsi, nous parvenons difficilement à obtenir des bases de données de plus de la centaine de patients sur un site, notamment pour les schizophrénie ou le trouble bipolaire. Cela a causé l'emergence d'études intersites et de programmes internationaux comme le groupe de travail ENIGMA, qui a permi des études avec un très grand nombre de sujet (Nunes et al., 2020). Bien que ces initiatives augmentent de façon considérable le nombre de patients, elles apportent leurs propres limitations.

Toutefois, ces initiatives obtiennent déjà des résultats en population générale. Nous voyons ici l'exemple de l'UK-Biobank, une base de données qui regroupe plusieurs dizaines milliers de sujets. Une étude de Valki and al. (Vakli et al., 2020) a donc utilisé cette base de données (leur étude a regroupé 9518 femmes et 8420 hommes) pour prédire l'indice de masse corporelle (IMC) à partir d'une unique image d'IRM structurelle. Leur résultats ont donné une erreur moyenne de prédiction de 2.36 kg/m 2 (l'indice de masse corporelle allait de 13,39 à 58,70 kg/m2 dans cette étude). Ils ont de plus utilisé Grad-CAM afin de repérer les régions qui contribuent le plus à la prédiction de l'IMC, et ont entre autres identifié l'Amygdale, zone cérébrale impliqué dans la régulation de la faim. Ces résultats montrent qu'il est possible, à partir d'un ensemble de donné suffisamment grand, de prédire une variable clinique non visible à l'oeil nu à partir de données d'imagerie brutes à l'aide de réseaux de neurones (CNN).

Ainsi, plusieurs directions s'offrent à la recherche en apprentissage profond en psychiatrie. La première direction, la plus évidente, est de rassembler des bases de données toujours plus conséquentes de

patients en psychiatrie, afin de pouvoir entrainer sans sur-apprentissage. Cela impliquerait des moyens colossaux et une cooperation nationale, voir internationale, et semble peu réaliste à court terme. La seconde, plus envisageable, est d'utiliser des méthodes de transfert learning afin d'utiliser des réseaux de neurones convolutionnels en population générale, pour prédire des varaiables cliniques telles que l'age ou le poids, puis de transposer cet entrainement à des patients en psychiatrie pour prédire des diagnostic cliniques.

Méthodes de clustering et modèles semi-supervisés

Un aspect de l'apprentissage statistique que nous n'avons pas évoqué jusqu'ici est l'apprentissage non supervisé. Cette branche est consacrée à la classification de données dans le cas où elles ne sont pas étiquettées a priori. Cela permettrait d'obtenir des classifications 'naturelles', basées sur l'anatomie et la physiologie des données de neuroimagerie. Nous pourrions alors nous servir de ces résultats afin d'inférer de nouvelles classification des troubles psychiatriques. En effet, actuellement les classifications basées sur la symptomatologie sont partielles et ne prédisent pas toujours l'évolution ni la réponse au traitement car elles n'ont pas de bases biologiques.

Cependant, comme remarqué précédement, la présence ou non de trouble psychiatrique n'est pas la première source de variation d'une image IRM. D'autres facteurs sont susceptibles de modifier la structure du cerveau comme l'age, le poids, le sexe. Il est alors nécessaire de faire ressortir les paramètres intéressants pour un clinicien avant de pratiquer la classification. C'est ce que proposent des modèles d'apprentissages semi-supervisés. (Schulz et al., 2020) proposent en premier lieu d'entrainer un algorithme supervisé pour la prédiction contrôle vs malade, puis d'en déduire un 'explanation space', mettant en avant les features utilisées pour la selection et de proposer un clustering non supervisé dans cet espace afin d'en déduire des sous types de maladie. D'autres modèles semi-supervisés ont aussi été développé, comme UCSL (Louiset et al., 2021). Cet algorithme propose une boucle de rétro-action où la classification supervisé et le clustering s'influencent jusqu'à trouver un clustering stable de sous types de patients malades.

Conclusion

En conclusion, le domaine du machine learning appliqué à la psychiatrie est un domaine de recherche en plein développement qui profite autant du bond de la recherche en machine learning que du rassemblement de bases de données de taille de plus en plus importantes. Cependant, malgré cet important engouement, de nombreuses limites méthodologiques se posent, telles que le bruit, la généralisation entre différents ensembles de données, les différences méthodologiques entre plusieurs sites d'acquisitions, les comorbidités qui peuvent fausser les prédictions (les patients atteints de troubles psychiatriques ont plus de chance d'être en surpoids, un modèle peut donc se retrouver à prédire le surpoids plutôt que le trouble). Ces limites ralentissent fortement les progrès dans la compréhension de ces maladies et ainsi de nouveaux outils doivent être développés afin de pouvoir continuer à mieux comprendre les troubles psychiatriques au moyen de l'imagerie.

Biblographie :

Grande, I., Berk, M., Birmaher, B., & Vieta, E. (2016). Bipolar disorder. *The Lancet*, *387*(10027), 1561–1572. https://doi.org/10.1016/S0140-6736(15)00241-X

Alda, M. (2021). The moving target of psychiatric diagnosis. *Journal of Psychiatry and Neuroscience*, *46*(3), E415–E417. https://doi.org/10.1503/jpn.210098

Duchesnay, E. (n.d.). *Neuroimaging signatures of brain disorders: Fighting overfitting in predictive models*. 145.

Favre, P., Pauling, M., Stout, J., Hozer, F., Sarrazin, S., Abé, C., Alda, M., Alloza, C., Alonso-Lana, S., Andreassen, O. A., Baune, B. T., Benedetti, F., Busatto, G. F., Canales-Rodríguez, E. J., Caseras, X., Chaim-Avancini, T. M., Ching, C. R. K., Dannlowski, U., … Houenou, J. (2019). Widespread white matter microstructural abnormalities in bipolar disorder: Evidence from mega- and meta-analyses across 3033

individuals. *Neuropsychopharmacology*, *44*(13), 2285–2293. https://doi.org/10.1038/s41386-019-0485-6
Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D., & Ritter, K. (2022). *Performance reserves in brain-imaging-based phenotype prediction* [Preprint]. Neuroscience. https://doi.org/10.1101/2022.02.23.481601
Vakli, P., Deák-Meszlényi, R. J., Auer, T., & Vidnyánszky, Z. (2020). Predicting Body Mass Index From Structural MRI Brain Images Using a Deep Convolutional Neural Network. *Frontiers in Neuroinformatics*, *14*, 10. https://doi.org/10.3389/fninf.2020.00010
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks* (arXiv:1312.6199). arXiv. http://arxiv.org/abs/1312.6199
Nunes, A., Schnack, H. G., Ching, C. R. K., Agartz, I., Akudjedu, T. N., Alda, M., Alnæs, D., Alonso-Lana, S., Bauer, J., Baune, B. T., Bøen, E., Bonnin, C. del M., Busatto, G. F., Canales-Rodríguez, E. J., Cannon, D. M., Caseras, X., Chaim-Avancini, T. M., Dannlowski, U., … Hajek, T. (2020). Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group. *Molecular Psychiatry*, *25*(9), 2130–2143. https://doi.org/10.1038/s41380-018-0228-9
Schulz, M.-A., Chapman-Rounds, M., Verma, M., Bzdok, D., & Georgatzis, K. (2020). Inferring disease subtypes from clusters in explanation space. *Scientific Reports*, *10*(1), 12900. https://doi.org/10.1038/s41598-020-68858-7
Louiset, R., Gori, P., Dufumier, B., Houenou, J., Grigis, A., & Duchesnay, E. (2021). UCSL: A Machine Learning Expectation-Maximization framework for Unsupervised Clustering driven by Supervised Learning. *ArXiv:2107.01988 [Cs, Stat]*. http://arxiv.org/abs/2107.01988