# Rapport de stage Technique: Neurospin

Christophe Launay, Promo 2015, M1, EFREI

24 septembre 2014

# Todo list

# Table des matières

# Introduction

Ici je dois écrire mon introduction

# Chapitre 1

# L'entreprise Neurospin et son environnement

Bla bla.

## 1.1 Neuropsin

Neurospin est un centre de recherche qui allie la neuroscience a la physique pour l'IRM(Imagerie par Résonance Magnétique). Cette alliance des deux en fait toute l'originalité de ce centre de recherche. Basé sur les plateaux de Saclay, à 25 km de Paris. Le site a ouvert ses portes le 1er janvier 2007 et il réunit les plus grands des domaines de la neurologie, du traitement de signal ou de la physique magnétique. Dirigé par le Dr Denis le Bihan, membre de l'Académie des Sciences et de l'Académie des Technologies.

## 1.2 La hiérarchie

présentation de la hiérarchie Neurospin

# Chapitre 2

# Le projet

Étude du trouble bipolaire par l'imagerie de diffusion. Le projet consiste à déterminer si des patients sont atteints de bipolarité ou non suite à un apprentissage de la machine. Les images sur lesquelles nous avons travaillé ont été acquises selon la technique de *Diffusion Weighted Imaging (DWI)*. DWI est une technique d'imagerie IRM qui sert principalement à imager les flux des molécules d'eau a l'intérieur du cerveau. Part cette technique, nous obtenons des mesures du cerveau dépendant de la matière blanche.

## 2.1 Les troubles bipolaires : explication

Pour bien comprendre la nature de la recherche, il faut savoir ce que sont les troubles bipolaires. Une bipolarité est un trouble mental qui influe sur l'humeur, oscillant entre des périodes de dépression et d'hypomanie (élévation de l'humeur) avec entre les deux, des périodes d'humeur "normale" (euthymie).

## 2.2 DWI : diffusion Weighted Imaging

L'IRM de diffusion est une technique basée sur l'imagerie par résonance magnétique (IRM). Elle permet de calculer en chaque point de l'image la distribution des directions de diffusion des molécules d'eau. Étant contrainte par les tissus environnants, cette modalité d'imagerie permet d'obtenir indirectement la position, l'orientation et l'anisotropie des structures fibreuses, notamment les faisceaux de matière blanche du cerveau. L'hypothèse derrière une diffusion pondérée est que cela peut indiquer des changements pathologique.

## 2.3 Présentation du projet

Le projet consiste à étudier une population de 200 sujets, une partie est atteinte de trouble bipolaire, l'autre est constitué de témoins. Le but étant d'ap-

prendre a la machine à reconnaitre les sujets bipolaires des sains. Le principe est de créer deux groupes de cette populations, un qui servira pour l'apprentissage de la machine, et l'autre de test (voir figure 1, a). Elle apprendra une carte des poids des voxels (les pixels de l'image IRM) selon le score (si oui ou non le sujet étudié est bipolaire) ainsi que des hyperplans de prédictions qui correspondent aux coordonnées dans l'espace des voxels de poids les plus forts (voir figure 1, b). grâce a cet apprentissage, la machine calculera sur l'échantillon de test un score de prédiction qui sera comparé aux vraies résultats cliniques des sujets (voir figure 1, c, d). Le sujet porte donc sur une classification des voxels. Pour celle-ci, un vecteur de poids via une régression logistique sera calculé qui va servir ensuite pour les prédictions afin de classés nos résultats grâce à une validation croisée.

Ainsi, nous avons le plan d'ensemble du projet, procédons maintenant par étape en commençant par présenter le cahier des charges.

## 2.4 Le cahier des charges

Une présentation des différentes étapes qui vont menés à la prédiction des résultats :

- Construire notre population : vérifier que notre liste de sujets correspond bien aux images que nous avons, vérifier qu'ils sont normalisés, que les données cliniques sont présentes, etc...
- Traitement des images et création des matrices
- Lancer les calculs de prédiction sur le cluster

Pour continuer sur la mise en œuvre de ce projet, un peu de théorie concernant les méthodes utilisés.

# Chapitre 3

# Méthodes

Ce chapitre fournit des explications sur les méthodes d'évaluation et de calcul utilisés dans le projet

## 3.1 Algorithme d'apprentissage

Le principe de l'apprentissage machine est le suivant : nous avons un estimateur à qui nous envoyons deux jeux de données. Une matrice X qui représente les données à évaluer et une matrice Y qui représente les résultats correspondant aux données. Des hypers-Paramètres lui sont également envoyé afin que le modèle puisse estimer correctement le paramètre $\theta$. (voir Figure 3.1)

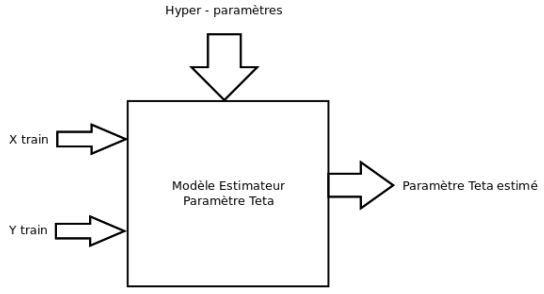


FIGURE 3.1 – Schéma d'apprentissage Machine

Ici, l'estimation du paramètre $\theta$ doit être tel que l'erreur estimé sur les prédictions soit le minimum possible. Cette erreur est estimée selon un modèle de régression linéaire qui doit minimiser selon le paramètre $\theta$ par la méthode des moindres carrés, c'est-à-dire minimiser la formule suivante :

$$erreur = \sum_{i=1}^{n} (|y_i - \hat{y}_i|^2) \tag{3.1}$$

avec $y_i$ les vrais résultats et $\hat{y}_i$ les résultats estimés.

Une fois que l'erreur a été estimé et que les $\theta$ ont été calculé, il est testé sur un échantillon test, suite à cela, la machine calcul une prédiction des résultats qu'on compare aux vrais résultats (voir Figure 3.2).

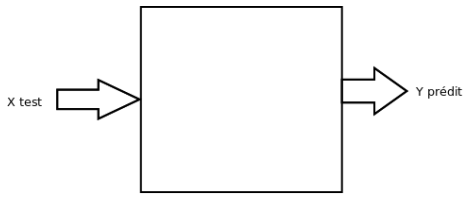


FIGURE 3.2 – Schéma d'apprentissage Machine

expliquer l'apprentissage des k voxels

## 3.2 Estimateur

Notre estimateur consiste à effectuer une régression logistique afin de trouver le vecteur de poids $\beta$ (ici, $\beta = \theta$). Dans un premier temps, une régression linéaire est effectuée et par dessus celle-ci, une fonction logistique est appliquée afin de classé les résultats à 0 ou 1.

### Régression Linéaire

La régression linéaire consiste à calculer une droite qui passe au plus prés de toutes les données en minimisant l'erreur. Cette erreur est représenté par la distance des données à la droite (la ligne verte sur le graphique) (voir Figure 3.3)

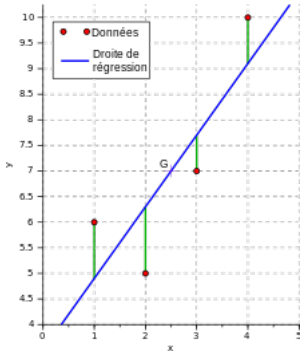


FIGURE 3.3 – Graphique d'une régression linéaire classique

les valeurs Y représente les résultats correspondant aux données de X et $\beta$ Les poids qui permettent de minimiser l'erreur de l'estimation, c'est à dire faire

en sorte que la droite passe au plus prés possible de tous les points. La formule de la régression linéaire est la suivante (à 1 dimension) :

$$Y = X * \beta + \epsilon \tag{3.2}$$

$\epsilon$ est ici un nombre infiniment petit qui représente une variation minime sur tout estimateur car l'erreur est humaine et l'apprentissage aussi. Cette formule nous permet facilement d'isoler $\beta$ afin de le calculer.

En ce qui nous concerne, Y correspond à une matrice de dimension n ligne et une colonne. X représente une matrice de n ligne et p colonne ainsi $\beta$ représente une matrice à 1 ligne et p colonne. Donc notre régression se calcul non pas sur une dimension mais sur n dimension. La formule devient donc :

$$Y = X_1 * \beta_1 + X_2 * \beta_2 + X_3 * \beta_3 + ... + \epsilon \tag{3.3}$$

En nous basant sur cette formule, il est facile après d'isoler les $\beta_i$ et de les calculés.

expliquer le but du choix de ce modele ci ( voir classif envoyé par DOUDOU)

**Régression Logistique**

Une fois nos paramètres estimés, il faut que le Y calculé soit entre 0 ou 1. Or, les résultats donnés par la régression linéaire sont définis sur l'ensemble des réels. On applique donc une fonction logistique afin de classé les résultats.

Expliquer aussi la pénalité ds la régression linéaire

## 3.3 Validation Croisée

La validation croisée est une méthode d'évaluation qui consiste a moyenné les scores calculés sur plusieurs jeux de données pour un seul tuple de paramètre. (voir Figure 3.4)

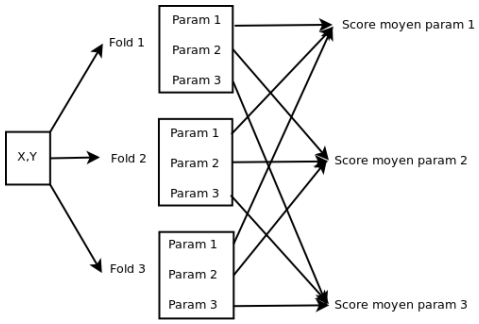


FIGURE 3.4 – Schéma de fonctionnement de la validation croisée

Le principe de la validation croisée est de séparer un ensemble de donnée en plusieurs groupes de tailles équivalentes. Sur le schéma de la Figure 3.5, on peut observer un ensemble de donné séparé en 3 avec $X_1$, $X_2$, $X_3$. Chacun à tour de

rôle sera utilisé pour l'apprentissage et le test. Sur les deux schémas, on peut lire "score", cela représente le résultat estimé par la régression linéaire. C'est ce même score qui sera comparé aux vrais score pour savoir si notre estimation est correcte et les résultats prédits bons. Il va sans dire que

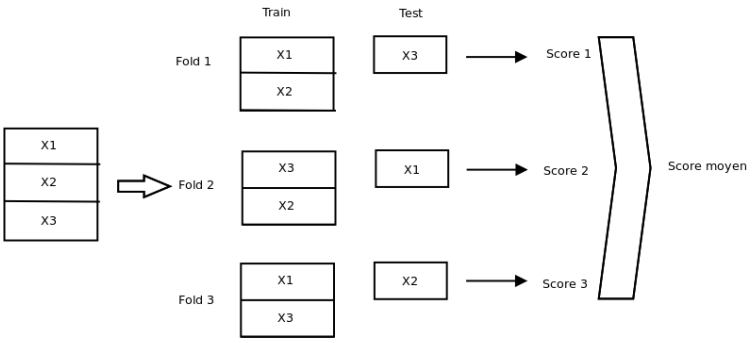


FIGURE 3.5 – Schéma de calcul d'un score par validation croisée

Toutes les méthodes décrites dans cette partie représente enormément de temps de calculs. Pour réduire ce temps la, le module MapReduce va être utilisé.

## 3.4 MapReduce

MapReduce est un patron d'architecture de développement informatique, inventé par Google, dans lequel sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses. Ce module consiste en deux étapes :

- map
- reduce

### Map

Dans l'étape Map le nœud analyse un problème, le découpe en sous-problèmes, et les délègue à d'autres nœuds (qui peuvent en faire de même récursivement). Les sous-problèmes sont ensuite traités par les différents nœuds à l'aide de la fonction Reduce qui à un couple (clé, valeur) associe un ensemble de nouveaux couples (clé, valeur) :

$$map(key1, value1) \rightarrow list(key2, value2) \tag{3.4}$$

### Reduce

l'étape Reduce, où les nœuds les plus bas font remonter leurs résultats au nœud parent qui les avait sollicités. Celui-ci calcule un résultat partiel à l'aide de la fonction Reduce (réduction) qui associe toutes les valeurs correspondantes à la même clé à une unique paire (clé, valeur). Puis il remonte l'information à

son tour. À la fin du processus, le nœud d'origine peut recomposer une réponse au problème qui lui avait été soumis :

$$reduce(key2, list(value2)) \rightarrow list(value2) \tag{3.5}$$

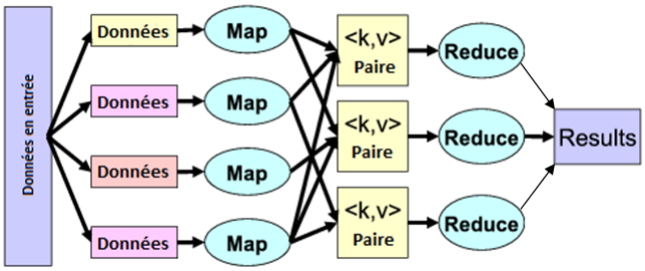


FIGURE 3.6 – Schéma de fonctionnement du MapReduce (source wikipedia)

Dans notre cas, l'étape Map ne découpe pas nos données en plusieurs groupes, cependant elle appelle plusieurs fois la fonction de calcul des prédictions sur plusieurs clés de paramètres différents tous indépendants les uns des autres. C'est pourquoi MapReduce est utilisé afin de lancer les calculs (qui sont importants et qui prennent du temps) en parallèle sur le cluster.

# Chapitre 4

# La mise en œuvre

## 4.1 Les sujets

Cette étape va concerner l'initialisation de nos sujets. Comment ils ont été choisi, quels sont les spécificités, etc...

### 4.1.1 Le choix des participants

Les participants a ce projet ont été recruté dans trois sites différents :
- Assistance Publique-Hôpitaux de Paris Hôpital Henri Mondor-Albert Chenevier à Créteil, Fernand Widal-Lariboisière à Paris, France
- Western Psychiatric Institute and Clinic in Pittsburgh, Pennsylvania
- Central Institute for Mental Health in Mannheim, Germany, une fondation publique associée a l'université de Heidelberg.

les sujets témoins ont été recrutés parmi des contrôles, des annonces médiatiques ou encore dans des bureaux d'enregistrement parmi les trois sites. Ils ont tous été soumis à des tests reconnus pour les troubles mentaux et non aucun membre de leur famille sujet a ces mêmes troubles. Un autre critère de sélection est qu'aucun des participants n'a subi de traumatisme neurologique, n'ai une contre indication pour l'IRM.

### 4.1.2 L'acquisition de donnée

Toute les données ont été acquises par le même logiciel d'acquisition avec une machine IRM 3T qui ont toutes été paramétrés de la même manière sur les 3 sites. Par la suite, les images ont été normalisés et corrigés si besoin à l'aide de logiciel libre.

inclure un petit schéma concernant le prétraitement des données

### 4.1.3 L'initialisation de la population

Au commencement, nous avons 3 fichiers :

– Le fichier qui contient les imageries de diffusion : *all_FA.nii.gz*
– Un autre contenant les identifiants de chaque sujet qui correspondent aux images : *ID.tbss*
– Un dernier fichier qui correspond aux données cliniques de chaque sujet : *BD_clinic.xlsx*

Avant tout, il faut s'assurer que les identifiants contenus dans le ID.tbss soient les mêmes que dans notre fichier clinique. Une fois cela fait, nous sélectionnons les données cliniques qui nous intéressent pour l'étude en question. C'est à dire les données suivantes :

– L'identifiant du sujet
– l'état du patient : si il est bipolaire ou non
– l'age auquel l'IRM a été effectué
– le sex du sujet
– le site dans lequel l'acquisition a été fait.

A la fin de toutes ces étapes, ces données sont enregistrés dans un nouveau fichier que nous appelons *population.csv* que nous utiliserons tout du long du projet.

Une fois toutes ces données récupéré, les images seront traités.

## 4.2 Traitement des images et création de nos matrices

Ici vous sera expliqué comment les données ont été traités informatiquement pour la suite de notre projet. Plusieurs hypothèses ont été évaluées :

1. les images initiales et une matrice avec le sexe et l'age auquel l'image a été prise l'IRM sans intercept.
2. les images initiales avec une matrice au mêmes covariables avec l'intercept.
3. ces deux hypothèses mais avec des images squeletonisées.
4. Les images de bases avec une covariable en plus : les sites ont été prises les IRM ainsi que l'intercept.
5. un masque tronqué d'une partie du cerveau afin de focaliser notre analyse sur des régions précises du cerveau.

### 4.2.1 présentation des images

A chaque sujet correspond une image de dimension (182, 218, 182). Ce sont donc des images 3D qui sont regroupés en un seul volume. Le fichier image correspond donc a un volume 4D qui correspond aux dimensions d'une image plus le nombre de sujet.

inclure des images d'une imagerie de diffusion + skeletonised

expliquer le but du masque dans le cas du traitement des images

### 4.2.2 Création du masque

Chaque image fais environ deux millions de voxels (pixel du cerveau). Nous allons appliquer un masque afin de réduire le nombre de voxel et ainsi focaliser notre analyse sur une plus petite partie du cerveau. Cela nous permettra d'éliminer une grande partie des voxels qui ne nous intéressent pas. Ce masque sera fait selon les images de base avec un certain seuil suivi d'un paufinage basé sur un atlas qui va nous permettre de sélectionner des régions du cerveau qui ne sont pas révélateurs dans notre cas. (voir figure 2)

Dans le cas du masque tronqué, nous suivons les étapes du dessus, suivi d'élimination des zones du cerveau que nous ne voulons pas.

Dans le cas des images skeletonisés, nous créons le masque sur la base de ces images.

inclure des images du masque, troncés et skeletonised

### 4.2.3 Application du masque

Chaque image 3D sera transformé en un vecteur ligne. Celui-ci, que nous appellerons *vecteur d'image* se verra appliqué le masque crée plutôt après avoir été transformé en vecteur ligne. Ainsi, nous obtenons une matrice X de vecteur ou chaque ligne correspond un sujet et le nombre de colonne aux voxels. Cette matrice sera ensuite centrée et réduite afin d'être normalisé. Cette démarche est commune a toutes les hypothèses.

### 4.2.4 Ajout des covariables

Suite à la création de cette matrice, nous y ajoutons des colonnes de covariables. Ces variables sont des données cliniques qui ne seront pas pénalisés lors du calcul de prédiction. Les covariables sont le sexe des sujets et leur âge à auquel a été pris l'image. L'âge sera centrée et réduit mais pas le sexe qui sera codé en binaire (1, -1) ou le 1 correspond aux femmes et -1 les femmes. Ces deux colonnes sont ensuite ajoutées a la matrice X.

Dans le cas de l'hypothèse 2, une colonne de 1 a été ajouté au tout début de la matrice.Il s'agit de la matrice d'intercept qui va diminuer l'effet de biais(équilibrage de la classification des sujets).

Dans le cas de l'hypothèse 3, une covariable va être ajouté, celle des sites auxquelles a été prises l'IRM. Il s'agit de *Dummy Coding*, c'est à dire transformé une colonne contenant plusieurs informations qualitatives en n colonnes, une pour chaque valeur différente.

Une autre matrice est crée, celle des Y qui correspond a l'état clinique des sujets (sain ou malade). Cette matrice est commune a toutes les hypothèses.

Ces deux matrices seront enregistrées dans deux fichiers différents et seront utilisés lors du calcul d'apprentissage et de prédiction.

Suite cela, nous lançons les calculs de prédiction sur un cluster.

## 4.3 Le calcul de prédiction

expliquer l'ordre dans lequel sont effecués les différentes étapes du calcul de prédiction (shéma maybe ?)

**Sensibilité et spécificité : explication**

Au dessus, on a mentionné la sensibilité et la spécificité. mots d'une grande importance qui doivent être explicité afin de bien comprendre les résultats qui vous seront présenté aprés. Ces termes vient d'une technique d'analyse statistique : *Receiver Operating Characteristic curve*. Cette technique permet de classé des résultats binaires (0 ou 1) en quatre groupes sous-jacent :

- vrai positif
- faux positif
- vrai négatif
- faux négatif

Il s'agit donc de classer des résultats entre 0 ou 1 en comparant les scores obtenus avec les vrais. Ceci forme une courbe (voir figure X) qui représente la valeur de seuil selon la spécificité et la sensibilité. Ces termes ont pour historique la détection sur des radars. Les radars sont dis sensibles si ils détectent correctement les événements importants parmi les événements qu'il a détecté. En revanche, un radar est spécifique si il ne détecte que des événements importants même si il n'en détecte pas beaucoup. Dans notre cas, nous nous dirons sensibles si parmi tous les sujets classés malades, tous le sont, au contraire, nous serons spécifiques si nous avons classés peu de sujet malade mais ceux qui sont classés sont vraiment malade. La sensibilité correspond donc au taux de vrais positifs bien classés alors que la spécificité correspond au taux de vrais négatifs bien classés.

Maintenant que nous savons comment sont effectués les calculs et notre classification, passons maintenant a l'analyse de nos résultats.

# Chapitre 5

# chapitre_5

Ici commence le dernier chapitre de mon rapport

Présentation des résultats

expliquer la sensibilité, la spécificité, la ROC curve, et ce que l'on voudrait obtenir,