Modèle de prédiction des crises d'épilepsie : k plus proches voisins et forêt aléatoire.

Raphael Sitruk

Lo Seen Luke

Rachdi Chems Eddine

Résumé : Nous souhaitons prédire la présence de crises d'épilepsie à partir de l'encéphalographie des patients. Pour ce faire nous utiliserons plusieurs méthodes statistiques telles que les k plus proches voisins et les forêts aléatoires.

Sommaire

Introduction

L'épilepsie est un trouble neurologique qui se caractérise par des crises d'épilepsie où le patient est sujet à vif tremblements et convulsions. De manière plus générale, il s'agit d'une altération de certaines zones neuronales du cerveau due à des décharges épileptiques.

Ces crises, quelques soit leur forme, sont difficilement soignables, malgré les nombreuses recherches sur le sujet. En l'absence de traitement stable, il devient important pour les personnes atteintes d'être capable de prévoir des crises risquant de se dérouler à tout moment et donc pouvant occasionner des accidents mortelles pendant les états de crises (accident de voiture, chute, etc).

L'idée derrière cette prévision serait d'étudier des données cérébrales pouvant être intimement liées à des crises prochaines. C'est ce que nous proposons de faire via un recueil d'informations obtenu sur le sujet.

Nous possédons effectivement 5 dossiers regroupant chacun 100 fichiers, chaque fichier désignant un individu dont l'activité cérébrale est mesurée pendant 23,6 secondes. Cette série temporelle d'analyse cérébrale sont échantillonnées en 4097 temps.

Nous avons, finalement, une masse d'informations conséquente, ainsi qu'une variable décrivant l'état du patient à l'issue des 23,6 secondes où l'information sur une crise d'épilepsie ou non est donnée.

Ainsi, avec ces observations, nous pouvons facilement relier l'activité cérébrale aux crises d'épilepsie et donc espérer créer un modèle qui nous permettrait de déterminer si le patient va faire une crise d'épilepsie en se basant sur les données neuronales possédées.

Notre travail prédictif utilisera deux méthodes :

- La méthode des k plus proches voisins
- Les forêts aléatoires

Avec ces deux outils, nous allons obtenir des résultats que nous comparerons afin de savoir lequel est le meilleur en termes de qualité de prédiction.

Nous utiliserons également la méthode du « parallel computing » afin d'accélérer les processus de calculs sur les ordinateurs.

Ainsi, nous souhaitons au travers de ce travail expliquer pas à pas comment ces outils ont été mis en place malgré les différents problèmes auxquels nous avons dû faire face et quels sont les résultats obtenus.

Dans un premier temps seront détaillées les données en notre possession. C'est une étape primordiale pour la suite de notre travail. Deuxièmement, les méthodes utilisées seront expliquées. En troisième partie, nous présenterons les résultats obtenus. Enfin, la dernière section nous permettra de conclure de manière plus globale sur le sujet.

Base de données

Notre base de données provient du site « kaggle ». Elle contient 179 variables et 11500 observations. Chaque ligne représente une seconde du résultat de l'encéphalographie d'une personne tiré aléatoirement. Les 178 premières colonnes décrivent les caractéristiques de l'encéphalographie alors que la dernière colonne nous donne une information sur la personne. Cette dernière variable prend 5 modalités de 1 à 5 :

1 : L'encéphalographie a été réalisée pendant une crise d'épilepsie

2 : L'encéphalographie a été réalisée sur la zone où est située la tumeur

3 : L'encéphalographie a été réalisée sur la zone saine du cerveau

4 : L'encéphalographie a été réalisée alors que le patient avait les yeux fermés

5 : L'encéphalographie a été réalisée alors que le patient avait les yeux ouverts

Traitement de la base

Afin de rendre plus robuste notre régression nous avons commencé par retravailler la variable y. Pour rappel, cette variable contient une valeur de 1 à 5 décrivant les conditions de l'encéphalographie. Pour les patients ayant un code de 2 à 5 l'encéphalographie n'a pas été réalisée durant une crise d'épilepsie alors que pour les patients ayant la valeur 1 pour cette variable l'encéphalographie a été réalisée durant une crise.

Nous avons créé une variable binaire prenant 1 si y égal 1 et 0 sinon.

Comme nous le disions nous avons 178 variables explicatives et 1 variable dépendante binaire.

Le nombre de variables dépendantes étant très élevé nous avons choisi d'utiliser la méthode de l'analyse en composantes principales (PCA). Cette méthode consiste à capturer l'information présente dans des variables corrélées pour créer d'autres variables, moins nombreuses, et non corrélées.

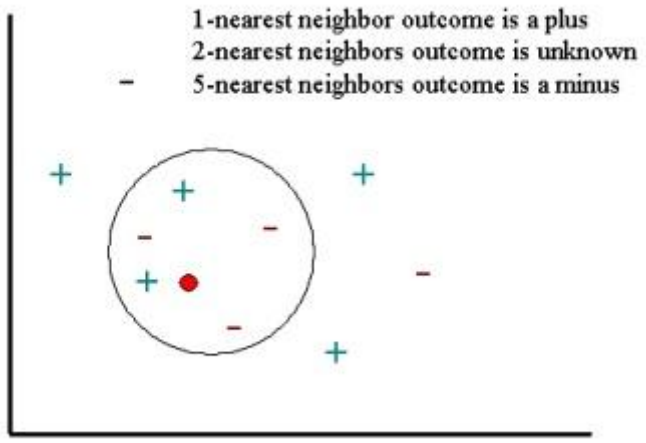
Enfin, afin de pouvoir tester les résultats de notre modèle nous allons diviser notre échantillon en deux sous-échantillons : le premier servant à « entrainer » notre modèle, le second à « tester » notre modèle.

Modèle des k plus proches voisins

La méthode des k plus proches voisins est l'un des plus anciens algorithmes de classification puisqu'il aurait été inventé au Xème siècle par Alhazen Ibn Al Haytham. (Mark Smith, 2001, Alhacen's Theory of Visual Perception. Philadelphia: American Philosophical Society)

La méthode des « k plus proches voisins » est une méthode non paramétrique. C'est-à-dire que nous n'entrainons pas un modèle dans le but d'obtenir des paramètres à l'instar de la plupart des régressions et nous ne faisons pas d'hypothèse sur la distribution des données. A la place, cette méthode possède un paramètre de « réglage » k. Ce paramètre détermine comment sera entrainé le modèle. Notons que les paramètres de réglage ne sont pas exclusifs aux méthodes non paramétriques.

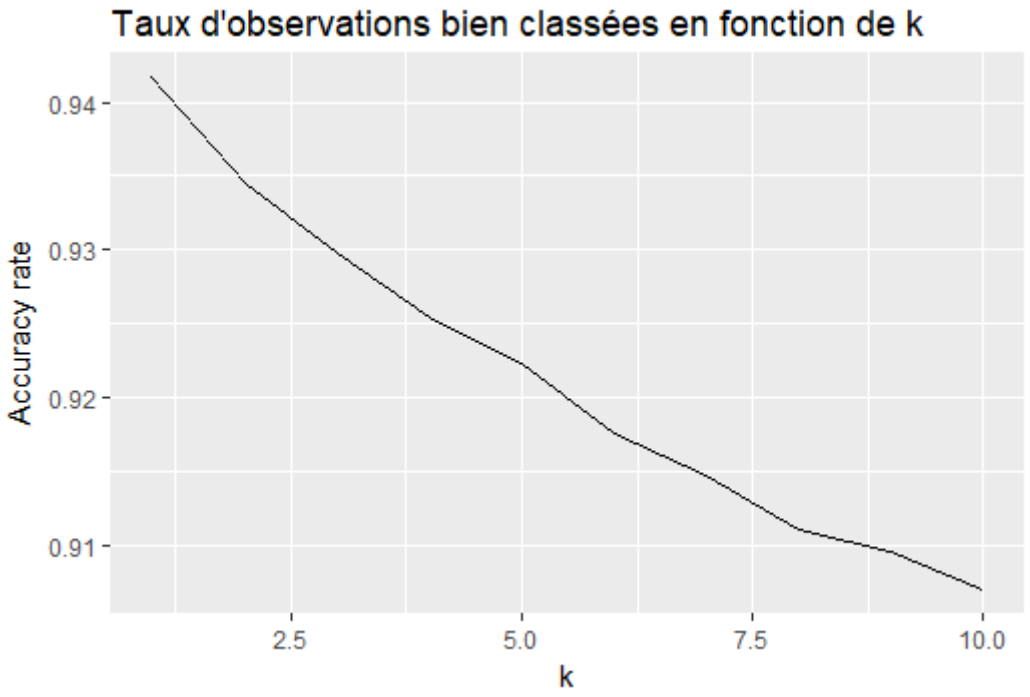
Il existe deux types de méthodes des « k plus proches voisins » : la régression et la classification. Nous nous intéresserons à la deuxième catégorie. Commençons par expliquer comment fonctionne l'algorithme des « k plus proches voisins ».



Nous souhaitons connaitre la classe du point rouge. Si nous choisissons un paramètre k = 1, nous classerons le point rouge comme un « plus » car le plus proche voisin du point rouge est un « plus ». Dans le cas où k vaut 2, l'algorithme échouera à classer le point rouge puisqu'il possède un voisin « plus » et un voisin « moins ». Dans ce cas-là, l'algorithme choisit aléatoirement la classe. Sur le graphique ci-dessus k = 5 et le point rouge sera classé en « moins » car il possède 3 voisins de type « moins » contre 2 voisins de type « plus ». Les voisins les plus proches sont ceux avec lesquelles la distance Euclidienne est la plus petite.

La classification des « k plus proches voisins » requiert de n'avoir que des variables numériques. Si certaines variables possèdent des nombres vraiment plus grands d'autres variables, elles domineront en distances de mesure. Mais cela ne signifie pas forcément que cette variable est plus importante. Une pratique courante est de modifier la variable afin d'avoir une moyenne nulle et une variance unitaire. Nous n'avons cependant pas rencontré ce type de problème dans notre base de données.

Comment choisir le paramètre de réglage ? Il y a une méthode similaire à la courbe ROC pour les régressions logistiques. Il s'agit de la courbe de « cross-validation ». Cette courbe nous donne le pourcentage d'observations correctement classées pour différentes valeurs de k.



Si k est trop faible nous risquons de « surentrainer » le modèle et son pouvoir prédictif en pâtira. Si k est trop élevé, le modèle sera biaisé. La valeur optimale de k est celle qui minimise l'erreur de validation. En testant plusieurs valeurs de k (de 1 à 10) on remarque que la valeur qui maximise « l'accuracy » (c'est-à-dire la proportion d'observations bien classées) est k = 1.

Afin de mesurer la qualité de notre système de classification nous allons utiliser une matrice de confusion.

		Valeurs observées	
		Absence de crise	Crise
Valeurs prédites	Absence de crise	4528	321
	Crise	11	831

Cette matrice nous donne plusieurs informations. En plus du taux d'accuracy de 94%, nous obtenons le taux de rappel (vrais positifs / vrais positifs + faux positifs) de 98% et le taux de précision (vrais positifs / vrais positifs + faux négatifs) de 72%.

Modèle des forêts aléatoires

Le modèle, développé par Léo Breiman et Adèle Cutier en 2001 [Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.], est une amélioration de l'algorithme CART (i.e. Classification And Regression Tree) [Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press]. La Forêt Aléatoire, plus robuste face à l'*overfitting*, dispose d'un meilleur pouvoir prédictif qu'un simple arbre de décision. En contrepartie, la visualisation des résultats n'est pas aussi agréable.

Le Random Forest procède comme ceci :

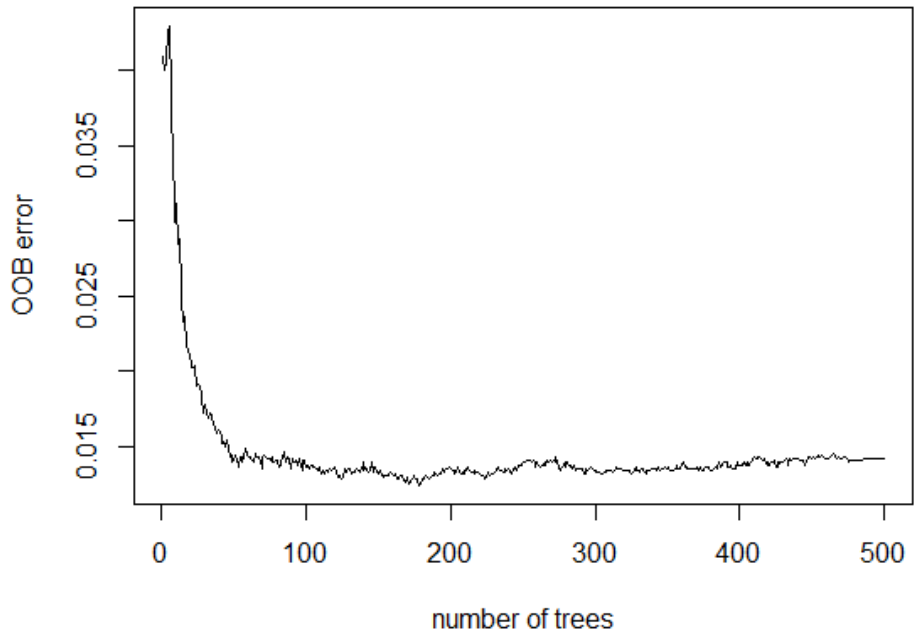
1. Le modèle génère B échantillons de mêmes tailles à partir de l'échantillon original par *bootsraping* (i.e. tirage aléatoire avec remise des observations de l'échantillon original, jusqu'à obtenir un nouvel échantillon de la taille voulue).

2. Sur chaque échantillon, on entraîne un arbre en prenant un sous-ensemble aléatoire de taille m de l'ensemble des prédicteurs.

3. La valeur de la variable à prédire pour une observation donnée est obtenue par vote majoritaire en agrégeant les prédictions de l'ensemble des arbres (notons qu'il est possible de modifier la « proportion critique » où la valeur de la prédiction finale change).

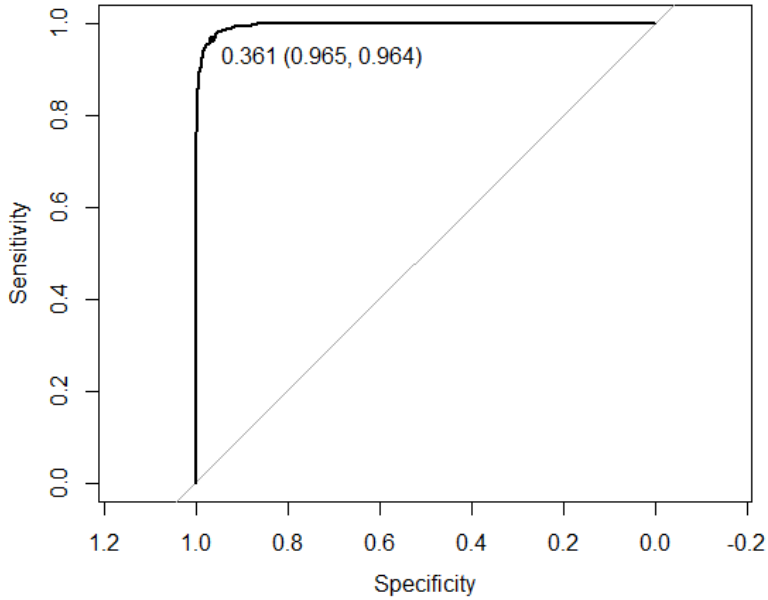
Nous avons ici fixé B=500 et p=59 (i.e. la partie entière du tiers du nombre de prédicteurs)

Afin de savoir si le nombre d'arbres que nous avons fixé est suffisant, nous pouvons regarder la proportion d'erreur *Out Of Bag* en fonction du nombre d'arbres (représenté ci-dessous). L'erreur *Out Of Bag* (ou OOB) est la moyenne pour l'ensemble des observations x_i de la proportion d'erreur de classification uniquement chez les arbres dont l'échantillon ne contient pas x_i.



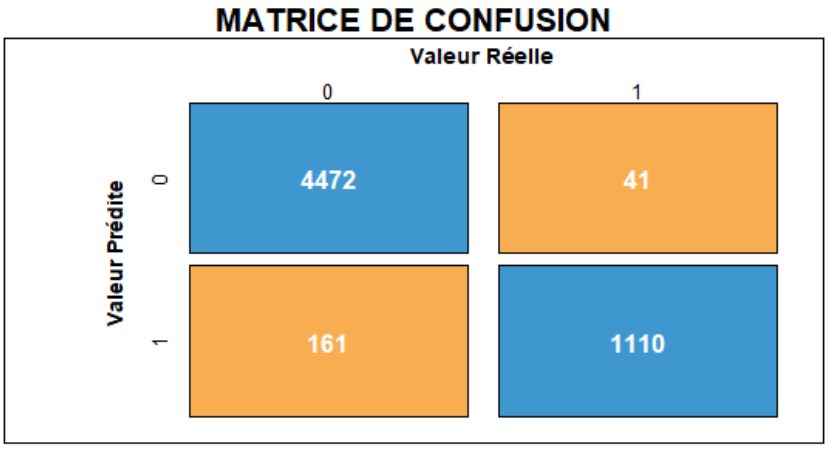
Si la proportion d'erreur OOB est stabilisée, alors le nombre d'arbres choisi est suffisant. On constate que c'est le cas ici, et qu'il aurait même été possible de fixer un nombre d'arbres B bien inférieur. Aussi, il semblerait que nous ayons un taux d'erreur OOB d'environ 1%, ce qui est très bas. C'est donc une bonne nouvelle en ce qui concerne le pouvoir prédictif du modèle.

Nous allons maintenant chercher à déterminer le « seuil critique » de décision optimal. La courbe ROC ci-dessous représente la sensibilité (i.e. la fraction de positifs classés positifs) et la spécificité (i.e. la fraction de négatifs classés négatifs) en fonction du seuil de proportion à partir duquel la forêt prédit une valeur positive.



Ainsi, on peut voir que la proportion qui optimise la sensibilité et la spécificité est 0.361. Par ailleurs, nous obtenons d'excellentes valeurs (0.965 pour la sensibilité et 0.964 pour la spécificité) pour cette proportion.

En nous basant sur la courbe ROC, nous fixons donc la « proportion critique » à 0.361. Nous obtenons la matrice de confusion suivante :



On constate les choses suivantes :

- La forêt possède un taux de classifications correctes de 97%
- La forêt possède un taux de précision de 87%
- La forêt possède un taux de rappel de 96%

Comparé à la méthode des k plus proches voisins, nous obtenons un taux de classifications correctes légèrement plus élevé (97% pour la forêt contre 94% pour les plus proches voisins), un taux de rappel plus faible (87% pour la forêt contre 98% pour les plus proches voisins) mais surtout un taux de précision bien plus élevé (96% pour l'arbre contre 72% pour les k plus proches voisins).

Dans notre cas, étant donné que l'on tente de prédire des crises d'épilepsie, le taux de rappel est crucial puisque l'on considère qu'un faux négatif (i.e. une vraie crise non prédite) est plus grave qu'un faux positif (i.e. une crise prédite pour un patient sain). À cet égard, les performances de la méthode des k plus proches voisins semblent meilleures que celles de la forêt aléatoire.

D'ordinaire, nous testons également la significativité d'une variable v_j d'une Forêt Aléatoire en redistribuant aléatoirement les valeurs de la colonne v_j sur l'échantillon d'entraînement, puis en mesurant la différence entre la proportion d'erreur OOB ainsi obtenue et celle du modèle original. Cependant, nous ne l'avons pas fait ici car nos observations correspondent à une seconde tirée aléatoirement de l'encéphalogramme d'un sujet. Chaque colonne ne correspond pas forcément au même instant pour chaque observation donc nous avons considéré que cela n'avait pas beaucoup de sens de le faire ici.

Conclusion

En conclusion, ce travail nous permet, à l'aide d'un jeu de données conséquent sur l'activité cérébrale, de mettre en place des méthodes de prédiction des crises d'épilepsie à venir. Ces deux méthodes sont celles des « k-plus proches voisins » et des « forêts aléatoires ».

En ce qui concerne la méthode des k-plus proches voisins, il est nécessaire, dans un premier temps de trouver la valeur de paramètre k. Pour cela, nous avons pu utiliser la courbe de « cross-validation » qui donne le taux de classifications correctes pour différentes valeurs de k et qui nous a amené à choisir k = 1.

D'un autre côté, nous avons utilisé le modèle de forêts aléatoires qui reposent sur la création de multiples arbres de classification via la formation de multiples échantillons d'entraînements issus de nos données initiales. La valeur de la variable à prédire ressort de l'agrégation des prédictions de tous nos arbres. Une fois ces étapes effectuées, nous avons pu tester la fiabilité de cette méthode au travers de divers tests.

Notons que pour obtenir des résultats plus rapidement à l'aide de ces deux outils, nous avons également utilisé du parallel computing qui nous permet d'augmenter le nombre de cœurs du processeur en charge des calculs.

Les résultats obtenus par ces deux méthodes sont de bonnes qualités avec des résultats très proches en termes de valeurs de taux de classifications correctes (94% pour les k-plus proches voisins et 97% pour les forêts aléatoires). En revanche, les valeurs diffèrent largement lorsqu'il s'agit des taux de rappel et de précision avec des taux très proches dans le cas des forêts aléatoire, tout en étant élevés. Ce n'est pas le cas avec la méthode des k plus proches voisins et dans un tel cas, les résultats sont difficilement exploitables car le nombre de résultats erronés renvoyés seraient assez élevé.

Finalement, nous avons pu conclure que la méthode la plus précise était celle des forêts aléatoires. Néanmoins, sachant que nous cherchons ici à prédire les crises d'épilepsies à venir, avoir un taux de rappel plus élevé peut s'avérer plus intéressant car cela signifie que le nombre de faux négatifs est réduit. Or, dans les tests médicaux, il vaut mieux limiter les erreurs de résultats impliquant la non-détection de maladie. Et ce même, si cela doit être fait en dépit d'une précision plus grande

et d'un nombre plus grand de faux positifs. Dans ce cas précis, la méthode des k-plus proches voisins est à privilégier.

En ce qui concerne la qualité de notre travail de manière générale, nous avons pu déplorer l'absence d'égalité dans la taille des échantillons ce qui nous amènent à avoir des résultats légèrement biaisés. Idéalement, il nous aurait fallu avoir un nombre équivalent d'individu n'ayant pas fait de crise que de personnes épileptiques. Cela peut être amélioré avec de *l'oversampling* ou de *l'undersampling* mais un biais subsistera malgré tout.

Il n'en reste pas moins que nous possédons ici des méthodes de prévision des crises d'épilepsie de bonnes qualités qui pourraient être utile pour le corps médical et qui pourraient être améliorées en ajoutant des variables autres que les mesures d'activités cérébrales (tension artérielle, rythme cardiaque, etc).