Régression logistique à base de splines adaptatives avec un réseau de neurones ReLU

Marie GUYOMARD[1], Susana BARBOSA[2], Lionel FILLATRE[1],

[1]Université Côte d'Azur, I3S, France

[2]Université Côte d'Azur, IPMC, France

guyomard@i3s.unice.fr, sudocarmo@gmail.com, lionel.fillatre@i3s.unice.fr

Résumé – Cet article s'intéresse à la classification binaire à l'aide d'une régression logistique non-linéaire. Les modèles linéaires, simples et interprétables, sont très appréciés dans le domaine médical mais leurs performances restent très limitées lorsque les données sont complexes. Nous proposons de remplacer la fonction linéaire de la régression logistique par une fonction linéaire par morceau modélisée par des fonctions splines. L'innovation majeure de cet article consiste à construire un réseau de neurones qui réalise entièrement cette régression logistique non-linéaire. L'architecture particulière de ce réseau automatise la segmentation des variables explicatives, permet une optimisation efficace des paramètres du réseau et garantit l'explicabilité des prédictions calculées.

Abstract – This paper deals with binary classification using a non-linear logistic regression. Linear models, simple and explainable, are very popular in the medical field but their performances remain very limited when data are complex. We propose to replace the linear function of the logistic regression by a piecewise linear function modeled by spline functions. The main novelty of this paper consists in building a neural network which entirely performs this non-linear logistic regression. Our network architecture automates the segmentation of the explanatory variables, allows an efficient optimization of the network parameters and guarantees the explanability of the calculated predictions.

1 Introduction

L'utilisation de l'Intelligence Artificielle dans le domaine médical ne cesse de progresser. Les modèles de "machine learning" pour la classification permettent dans de nombreux cas pratiques de s'affranchir de méthodes invasives, telles que des biopsies, pour fournir un diagnostic précis. L'utilisation de la Régression Logistique (RL) est très répandue. Par exemple, la RL est utilisée pour prédire le développement de la cirrhose non-alcoolique [1]. Contrairement à des méthodes telles que le boosting, les forêts aléatoires et les réseaux de neurones (RN), le modèle estimé par RL est facilement interprétable puisqu'il dépend d'une combinaison linéaire de variables explicatives.

Néanmoins, les médecins sont convaincus qu'intégrer dans la modélisation des phénomènes non-linéaires, tels que des effets de seuils sur certaines variables, augmenterait la performance prédictive. Par exemple, un petit taux de cholestérol pourrait avoir un effet protecteur contre une pathologie alors qu'un taux significativement élevé serait facteur de développement de la maladie. Dans ce but, une piste prometteuse est l'utilisation d'une RL qui exploite un modèle non-linéaire à base de splines linéaires par morceau. Ces splines découpent le domaine de définition des variables explicatives en plusieurs segments et, sur chaque segment, effectuent une approximation linéaire.

La difficulté majeure de l'utilisation d'une approximation avec des splines réside dans le choix des bornes de chaque segment, appelées noeuds des splines. Il a été démontré dans [2, 3] qu'optimiser conjointement les noeuds et l'approximation linéaire associés à chaque segment est difficile. La RL à base de splines fixes [4] s'affranchit de ces problèmes d'optimisation en fixant a priori le nombre ainsi que la valeur des noeuds. Le choix de noeuds est alors plutôt arbitraire et le modèle devient figé. Une alternative est le modèle MARS (Multivariate Adaptative Regression Splines) [5] qui propose une méthode adaptative pour calculer les noeuds. Ceux-ci sont calculés récursivement de façon à améliorer progressivement les performances du modèle. La performance globale de la méthode n'est pas directement optimisée mais de façon gloutone et sous-optimale. Récemment, une troisième approche [6] a émergé : elle établit un pont rigoureux entre les réseaux de neurones (RN) profonds et la théorie des fonctions splines linéaires par morceau. Contrairement au modèle MARS, l'apprentissage de la segmentation dans les RN obéit à la minimisation d'un critère global. Malheureusement, la segmentation produite par un RN est très complexe et il devient donc impossible d'interpréter facilement l'impact des variables explicatives dans la prédiction.

La contribution principale de cet article consiste à développer un RN qui s'inspire du modèle MARS afin de combiner les avantages de MARS et des RN : la minimisation d'un cri-

Ce travail a bénéficié d'une aide du gouvernement français, gérée par l'Agence Nationale de la Recherche au titre du projet Investissements d'Avenir UCA DS4H portant la référence n° ANR-17-IDEX-0004.

tère global pour obtenir un RN avec une architecture facilement explicable. Une seconde contribution consiste à proposer un algorithme explicite pour entraîner ce RN. En effet, [7] démontre qu'il est possible d'approcher les modèles MARS par des RN mais cette démonstration, totalement théorique, ne propose aucun algorithme pour entraîner le RN. Enfin, nous comparons notre RN aux méthodes citées précédemment sur des données simulées et des données réelles. Notre RN présente des performances en prédiction comparables ou supérieures aux autres approches tout en étant tout à fait explicable.

Cet article est structuré sous la forme suivante. La section 2 présente le problème de prédiction étudié. La section 3 décrit notre architecture RN. La section 4 présente les résultats expérimentaux. Enfin, la section 5 conclut l'article.

2 Positionnement du problème

Nous disposons de N couples indépendants et identiquement distribués $(x^{(i)}, y^{(i)})$ où $x^{(i)} \in \mathbb{R}^d$ est le vecteur des variables explicatives et $y^{(i)} \in \{0, 1\}$ est l'étiquette binaire à prédire. La notation (X, Y) désigne le couple de variables aléatoires dont sont issus les couples $(x^{(i)}, y^{(i)})$. Ces données serviront à entraîner et tester tous les modèles utilisés dans l'article.

2.1 Classification binaire Bayésienne

Un classifieur Bayésien du Maximum a Posteriori (MAP) attribue une étiquette y à un échantillon $x = [x_1, \ldots, x_d]$ en fonction de la règle de décision $\delta : \mathbb{R}^d \mapsto [0, 1]$ définie par $\delta(x) = \hat{\mathbb{P}}(Y = y|X = x)$ où $\hat{\mathbb{P}}(Y|X)$ est une estimation de la probabilité conditionnelle a posteriori. La RL est la règle de décision la plus utilisée pour ce type de problème dans le domaine médical. Elle s'écrit sous la forme

$$\delta^{\text{RL}}(x) = \sigma(\psi(x)) = \frac{1}{1 + \exp(-\psi(x))}, \tag{1}$$

où $\sigma(\cdot)$ est la fonction logistique et $\psi(x)$, appelée la fonction de score, est une fonction linéaire $\psi(x) = \beta^\top x$ où $\beta = [\beta_1, \ldots, \beta_d] \in \mathbb{R}^d$ est un vecteur de coefficient et $\beta^\top$ désigne le vecteur β transposé. Chaque coefficient β_i permet de quantifier l'impact de la composante i du vecteur x sur la probabilité de choisir la classe $y = 1$. Ce modèle est très apprécié pour sa simplicité et son explicabilité.

2.2 Fonction de score MARS d'ordre 1

Afin d'obtenir une fonction de score non-linéaire mais encore explicable, une modélisation pertinente est apportée par l'approche MARS [5]. Ce modèle s'appuie sur une approximation avec des splines adaptives de la fonction de score :

$$\psi^{\text{MARS}}(x) = \sum_{m=1}^{M} \beta_m h_m(x), \tag{2}$$

où $h_m(x)$ est une fonction spline de la forme

$$h_m(x) = [s_m(x_{v(m)} - b_m)]_+ \tag{3}$$

$$= \begin{cases} \max\{0, x_{v(m)} - b_m\}, & \text{si } s_m = 1, \\ \max\{0, b_m - x_{v(m)}\}, & \text{si } s_m = -1. \end{cases} \tag{4}$$

La notation $[t]_+ = \max\{0, t\}$ désigne la fonction ReLU. La fonction $h_m(x)$ dépend de la composante $x_{v(m)}$ du vecteur x. Le réel b_m est le noeud de la spline. L'entier $s_m \in \{-1, 1\}$ utilisé conjointement avec la fonction ReLU permet d'annuler la partie gauche où la partie droite de $h_m(x)$ comme explicité dans (4).

L'approche MARS apprend les fonctions $h_m(x)$ de façon séquentielle. À chaque itération $m \in \{1, \ldots, M\}$, la fonction spline $h_m(\cdot)$ qui réduit le mieux l'erreur d'apprentissage est ajoutée. La segmentation récursive et adaptative de l'approche MARS est donc similaire à celle des arbres de décision. Si $v(m) \neq k$ pour tout m, alors la composante k de x ne sera jamais incluse dans le modèle. L'approche MARS s'appuie sur un algorithme d'optimisation glouton dont l'optimalité globale n'est pas établie. La récursivité du modèle rend incontrôlable la segmentation des variables. Il se peut qu'une même variable soit segmentée un grand nombre de fois. Or, nous savons par nos échanges avec les médecins et les biologistes que trop segmenter une variable médicale est peu pertinent.

2.3 Réseaux de neurones ReLU

L'approche non-linéaire très répandue actuellement s'appuie sur une fonction de score produite par un RN ReLU [8]. Dans cet article, nous considérons uniquement les RN à une couche cachée qui s'écrivent sous la forme

$$\psi^{\text{RN}}(x) = \beta_0 + \beta^\top [Wx + b]_+, \tag{5}$$

où $\beta \in \mathbb{R}^p$, $W \in \mathbb{R}^{p \times d}$ est la matrice des poids, $b \in \mathbb{R}^p$ est le vecteur des biais, et $[z]_+$ désigne le vecteur z où la fonction $[\cdot]_+$ a été appliquée à chaque composante. La couche cachée comporte p neurones. D'après [6], les réseaux profonds avec la fonction d'activation ReLU introduisent un partitionnement de $\mathbb{R}^d$ équivalent à une approximation avec des splines multidimensionnelles. Toutefois, ce partitionnement est très complexe et très souvent totalement inexplicable. La figure 2-e illustre ce partitionnement lorsque $d = 2$ avec $p = 30$ neurones. Les droites, quasiment toujours obliques, s'entrecroisent et découpent l'espace en polyèdres aux formes géométriques très diverses. Ce découpage explique la flexibilité d'un RN mais aussi pourquoi un RN est considéré comme une "boîte noire".

3 Réseaux de neurones RN-MARS

Afin de bénéficier de l'avantage de l'entraînement d'un réseau de neurones (minimisation d'un critère global avec une descente de gradient) et de conserver une explicabilité proche du modèle MARS, nous proposons dans cette section une fonction de score continue par morceau modélisée avec un réseau

de neurones. Notre modèle $\psi^{\text{RN-MARS}}(x)$ s'écrit

$$\psi^{\text{RN-MARS}}(x) = \beta_0 + \sum_{j=1}^{d} g_j(x_j), \tag{6}$$

$$g_j(t) = \beta_{j1}[b_{j1} - t]_+ + \beta_{j2}[t - b_{j2}]_+, \quad t \in \mathbb{R}. \tag{7}$$

Dans (6), la fonction réelle non-linéaire $g_j(\cdot)$ est appliquée à la composante x_j du vecteur $x \in \mathbb{R}^d$. La fonction $g_j(t)$ correspond à une paire de neurones fonctionnant ensemble : le premier neurone est une spline non-nulle avant la valeur de noeud b_{j1} et le second neurone est une spline non-nulle après la valeur de noeud b_{j2}. De ce fait, la fonction $g_j(t)$ modélise des fonctions avec un profil composé de 3 segments linéaires comme illustré dans l'encadré gris sur la Figure 1. Dans cette figure, la variable X_1 peut par exemple représenter le poids d'un patient. Être en sous-poids ($X_1 < b_{11}$) ou en sur-poids ($X_1 > b_{12}$) augmente la probabilité de développer la pathologie. En revanche, entre ces deux intervalles, l'impact du poids sur la maladie est négligeable. La segmentation du réseau est donc très contrôlée : au plus 3 segments sont créés pour chaque variable descriptive.

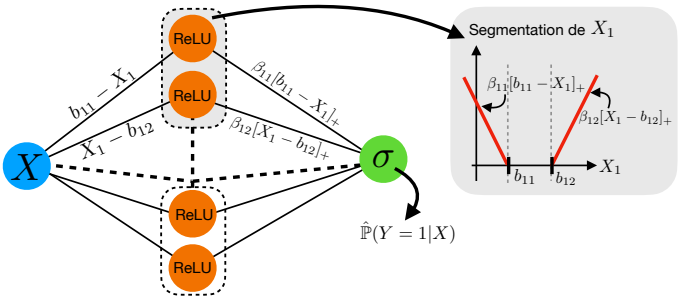

FIGURE 1 – Architecture de RN-MARS : les entrées en bleu, la couche cachée en orange et l'étiquette estimée en vert.

Par ailleurs, la nature de la segmentation opérée par le RN-MARS est également controlée. Contrairement aux RN ReLU classiques qui créent des régions obliques en combinant linéairement les composantes de x, le RN-MARS découpe les composantes de x de façon indépendante. Il s'appuie sur des hyperplans orthogonaux à la base canonique de l'espace $\mathbb{R}^d$, tout comme le font les arbres de décisions ou le modèle MARS, comme illustré sur la Figure 2-f. Le découpage de l'espace $\mathbb{R}^d$ s'effectue avec des hypercubes et non des polyèdres aux formes complexes. La règle de décision obtenue est facilement interprétable : la fonction de score est linéaire sur chaque hypercube. En pratique, cela revient à effectuer une RL localement en seuillant les composantes du vecteur x. La fonction $\psi^{\text{RN-MARS}}(x)$ modélise l'impact de chaque composante x_i avec un profil non-linéaire spécifique comme illustré dans la Figure 3. Globalement, le RN-MARS est composé de $2d$ neurones cachés. La figure 1 montre que les neurones cachés fonctionnent par paire. L'entraînement de RN-MARS se fait avec une descente de gradient ordinaire en utilisant l'entropie croisée comme fonction de perte.

4 Expériences

Modèles testés : Nous comparons les performances et l'explicabilité de RN-MARS aux arbres de décision (DT), à la RL Splines Cubiques Naturelles (RL SCN) [4, section 5.2] dont les noeuds sont fixés à l'aide de quantiles uniformes, à MARS et aux RN ReLU classiques.

Données simulées : Nous avons simulé des données $x \in \mathbb{R}^2$ afin de pouvoir visualiser les frontières de décision estimées. Nous cherchons à prédire la probabilité de développer une pathologie en fonction du taux de cholestérol (x_1) et du poids (x_2). Lorsque le taux de cholestérol est faible, le patient est davantage protégé contre la maladie. En revanche, lorsqu'il est élevé, le patient est plus à risque. Enfin, être en sous-poids ou sur-poids augmente la probabilité de tomber malade. Nous avons calculé les étiquettes telles que si la probabilité estimée de développer la pathologie est supérieure à 0.5 alors le patient appartient à la classe des malades ($y = 1$).

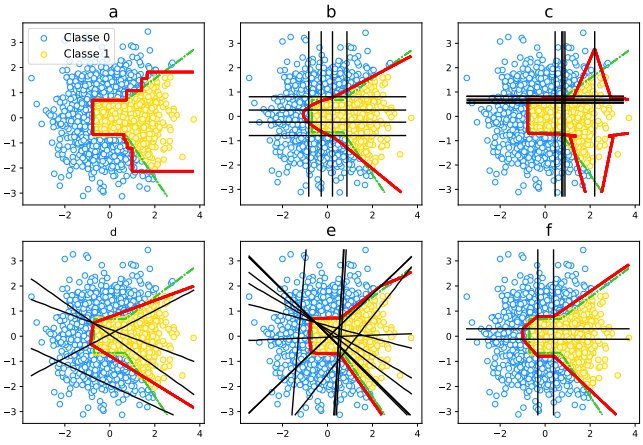

FIGURE 2 – Résultats sur données simulées : a) DT, b) RL SNC, c) MARS avec $M = 10$, d) RN avec $p = 4$ dans (5), e) RN avec $p = 30$ dans (5), f) RN-MARS. Légende : frontière idéale en vert, frontière estimée en rouge, segmentation en noir.

La Figure 2 présente les 6 méthodes comparées. Les deux classes des données d'entraînement sont séparées par une frontière idéale en vert. La frontière en rouge représente la frontière de décision de la méthode testée. Les traits noirs représentent les bordures du partitionnement produit par la méthode testée. Le DT (Fig. 2-a), la RL SCN (Fig. 2-b), MARS (Fig. 2-c) ainsi que RN-MARS (Fig. 2-f) segmentent les variables par des hyperplans. Le modèle MARS avec $M = 13$ définit 8 noeuds pour la variable x_1 et 5 pour x_2 en des valeurs très proches. Le RN-MARS obtient les mêmes performances que la RL SCN avec moins de fonctions splines, d'où l'intérêt d'automatiser la segmentation et de ne pas fixer les noeuds a priori. Les RN ReLU (Fig. 2-d & e) partitionnent l'espace par des frontières obliques rendant difficile l'interprétation de la segmentation. Le RN-MARS est presque aussi performant que les RN ReLU tout en nécessitant moins de paramètres à estimer et en garan-

tissant une bonne explicabilité. Le RN profond (Fig. 2-e) est le modèle le plus performant car sa frontière estimée est la plus proche de celle simulée. Néanmoins, l'interprétation de sa règle de décision est trop complexe car l'espace $\mathbb{R}^2$ est partitionné en de très nombreux polyèdres, dont certains sont finalement peu utiles pour l'approximation de la frontière.

Données réelles : Nous avons comparé les performances de RN-MARS à celles des autres méthodes sur un jeu de données réelles. L'objectif principal de la base de données "Parkinson" [9] est de détecter les personnes atteintes de la maladie de Parkinson à partir d'enregistrements vocaux. Nous avons conservé $d = 16$ mesures biomédicales de la voix, telles que par exemple les fréquences vocales maximales, moyennes et minimales. Le RN-MARS est composé de 32 neurones. L'entraînement des RNs est stoppé lorsque l'erreur sur les données de test ne diminue plus afin d'éviter le sur-apprentissage. Une validation croisée à 5 "folds" a été réalisée pour chacune des méthodes. Les résultats sont détaillés dans le tableau 1.

	Apprentissage		Test	
	Accuracy	AUC	Accuracy	AUC
RL	85 (2)	87 (2)	76 (1)	80 (6)
DT	91 (2)	94 (2)	88 (1)	77 (3)
RL SCN	90 (2)	94 (1)	82 (3)	87 (5)
MARS	90 (3)	91 (6)	82 (4)	89 (4)
RN ($p = 16$)	87 (4)	91 (7)	81 (6)	88 (7)
RN ($p = 70$)	86 (4)	91 (7)	83 (6)	88 (7)
RN-MARS	**87 (1)**	**92 (3)**	**83 (5)**	**91 (5)**

TABLE 1 – Résultats des performances prédictives (en %) sur les données réelles (moyenne et écart-type entre parenthèses) : le DT, la RL SCN, le RN à 16 neurones, le RN à 70 neurones et le RN-MARS (6).

La RL est moins performante (76% d'accuracy sur l'échantillon de test) que les autres méthodes, d'où l'importance d'introduire une modélisation non-linéaire. Parmi toutes les méthodes non-linéaires testées, le RN-MARS obtient la meilleure AUC (Area Under the Curve) sur l'échantillon de test (91%) qui est un critère apprécié par les médecins. L'automatisation de la segmentation des modèles MARS et des RN explique leurs AUCs plus élevées que celle de la RL SCN. Les écarts-types montrent que le RN-MARS est plus stable que les RN classiques.

La Figure 3 illustre les splines estimées pour 3 variables prédictives. Les noeuds estimés sont représentés par les points sur les courbes. Les différentes méthodes trouvent des profils similaires pour les variables mais tout de même assez dissemblables. Par ailleurs, cette figure met en avant les limites de l'apprentissage glouton de la méthode MARS. Seulement 5 variables ont été segmentées, dont une 5 fois (Fig. 3-a). Le modèle MARS n'est pas en mesure d'augmenter sa performance prédictive en ajoutant une nouvelle fonction spline, alors qu'il est possible de trouver une règle de classification plus performante, comme les RN le démontrent dans la Table 1.

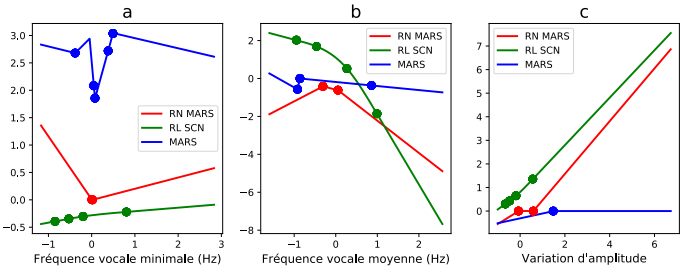

FIGURE 3 – Splines estimées sur les données réelles : a) Fréquence vocale minimale, b) Fréquence vocale moyenne, c) Variation d'amplitude. Légende : RN-MARS en rouge, MARS en bleu, RL SCN en vert.

5 Conclusion

Cet article développe un réseau de neurones avec une architecture explicable pour la classification binaire non-linéaire. L'architecture particulière de ce réseau permet d'une part de contrôler la segmentation des variables, mais aussi de produire une règle de décision interprétable. Ainsi, ce modèle est adapté aux problématiques médicales et aux attentes des spécialistes du domaine. Dans de futurs travaux, il serait souhaitable d'inclure des variables catégorielles et des interactions entre les variables afin de gagner davantage en performance prédictive.

Références

[1] M. Guyomard *et al.*, "Diagnostic non-invasif de la nash fibrosante à l'aide de l'intelligence artificielle," *AFEF (Société Française d'Hépatologie)*, 2020.

[2] D. M. Hawkins, "On the choice of segments in piecewise approximation," *IMA Journal of Applied Mathematics*, vol. 9, no. 2, pp. 250–256, 1972.

[3] A. Tishler *et al.*, "A new maximum likelihood algorithm for piecewise regression," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 980–987, 1981.

[4] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*, vol. 2. Springer, 2009.

[5] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[6] R. Balestriero *et al.*, "A spline theory of deep learning," in *International Conference on Machine Learning*, pp. 374–383, PMLR, 2018.

[7] K. Eckle *et al.*, "A comparison of deep networks with relu activation function and linear spline-type methods," *Neural Networks*, vol. 110, pp. 232–242, 2019.

[8] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA : MIT Press, 2016.

[9] M. Little *et al.*, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, 2008.