Sur l'équivalence entre la régression logistique à base de splines et l'apprentissage profond

Marie Guyomard[1], Cyprien Gilet[1], Susana Barbosa[2], Lionel Fillatre[1]
[1] *Université Côte d'Azur, CNRS, I3S*
[2] *Université Côte d'Azur, CNRS, IPMC*

Email : guyomard@i3s.unice.fr

Mots Clés : Classification, Apprentissage statistique, Spline, Réseau de neurones.

Biographie – Marie Guyomard a obtenu un diplôme d'ingénieur en Data Science de l'École Nationale de la Statistique et de l'Analyse de l'Information (Ensai, Rennes). Elle a commencé sa thèse en Machine Learning sous la supervision de Lionel Fillatre (laboratoire I3S) et Nicolas Glaichenhaus (laboratoire IPMC) en Octobre 2020. La thèse est financée par l'école universitaire de recherche *Digital Systems for Humans* (DS4H) de l'Université Côte d'Azur et a pour but de développer des algorithmes afin de prédire l'évolution d'une pathologie et adapter un traitement au profil de chaque patient.

Resumé :

Beaucoup de méthodes de classification en *Machine Learning* exploitent des modèles exclusivement linéaires. Cependant, dans de nombreux domaines d'application réels, des relations non-linéaires interviennent dans le modèle qui relie les variables explicatives à la variable à prédire. Dans le domaine biomédical par exemple, expliquer le développement d'une pathologie par des combinaisons linéaires de variables explicatives s'avère être trop restrictif.

De ce fait, ces dernières années de nombreux travaux ont porté sur l'estimation de modèles de classification non-linéaires par morceau. Autrement dit, en introduisant un changement de comportement du modèle pour les variables en des valeurs précises, les auteurs espèrent capter des effets qui ne sont pas accessibles aux modèles linéaires et ainsi gagner en performance prédictive. Dans ce but, beaucoup d'auteurs ont travaillé sur la régression logistique segmentée. Dans ce modèle, chaque variable explicative, prise individuellement, voit son domaine de définition partitionné en plusieurs régions ou intervalles ; la relation linéaire entre la variable explicative et la variable à prédire dépend alors de la région. L'impact d'une variable explicative sur la probabilité estimée d'appartenance à une classe ne sera plus linéaire mais linéaire par morceau puisqu'un coefficient spécifique est attribué à chaque intervalle de valeurs. Pour ces méthodes, la maximisation standard de la vraisemblance ne peut pas être utilisée. En effet, lorsque les points de changement doivent être estimés, la log-vraisemblance du problème devient différentiable par morceau, les conditions de régularité ne sont alors pas réunies [5], [13], [12]. D'une manière similaire, la régression logistique à base de splines [4] opère une discrétisation des variables explicatives. Dans un premier temps, des noeuds sont fixés afin de découper en différents intervalles les domaines de définition des variables. Ensuite, sur chaque intervalle, la fonction spline coïncide avec une fonction polynomiale spécifique à l'intervalle.

La performance des modèles de classification segmentés étant fortement impactée par la définition des différentes régions, beaucoup d'auteurs se sont intéressés à la problématique du choix du nombre mais aussi des valeurs des noeuds. Un grand nombre de travaux de recherche [14], [6], [10] préconisent l'utilisation d'une méthode coûteuse numériquement, le *Grid Search*, afin de déterminer les régions optimales. Des méthodes Bayésiennes ont aussi été développées afin de détecter le nombre de points de changements [9], [11] ainsi que leurs valeurs optimales [3], [2]. Des algorithmes s'appuyant sur l'estimation du maximum de vraisemblance sont proposés dans la littérature imposant des contraintes non négligeables telle que la continuité de la fonction dans [7] et [15] ou alors l'existence d'une unique région dans [8] et [7].

Afin de limiter le coût numérique et d'éviter la malédiction de la dimension lors de la phase d'apprentissage, les méthodes précédemment décrites segmentent chaque variable explicative de façon individuelle. Les régions multidimensionnelles restent très complexes à construire et à étudier. Par exemple, la construction d'une approximation multidimensionnelle par morceau à base de splines s'avère très difficile lorsque le partitionnement sur l'espace des variables explicatives induit par les splines est très complexe. La pertinence du partitionnement peut également être difficile à étudier. Récemment, il a été montré dans [1] que les réseaux de neurones de profonds permettaient de construire des fonctions multidimensionnelles par morceau. En particulier, la théorie des fonctions multidimensionnelles splines permet de construire un pont rigoureux entre les réseaux profonds et la théorie de l'approximation de fonctions multidimensionnelles. Les auteurs prouvent notamment que des réseaux de neurones peuvent être interprétés comme des opérateurs de splines, en particulier lorsque ces réseaux utilisent des fonctions d'activation convexes par morceau telles que la fonction linéaire rectifiée ReLU. Les réseaux peuvent alors s'exprimer comme étant une composition d'opérateurs *max-affine spline* (MASO) qui introduisent un partitionnement spécifique de l'espace des variables explicatives. Plus il y a d'opérateurs MASO dans le réseau de neurones, plus la subdivision de l'espace d'entrée devient fine. La partition finale correspond à l'intersection de toutes les partitions intermédiaires.

Nos travaux de recherche s'intéressent à la formalisation et à l'exploration de la partition induite par le réseau de neurones ReLU sur l'espace des variables explicatives. Nous avons en particulier réaliser des simulations numériques qui mettent en évidence comment la partition induite par le réseau de neurones s'adapte au problème de classification considéré. Notre étude est construite de la façon suivante. Tout d'abord, nous avons comparé la performance de la régression logistique non-linéaire par morceau à base de splines selon différents partitionnements de l'espace. Dans le cas où les variables explicatives suivent une distribution normale multivariée, nous sommes en mesure de conclure que la régression logistique à base de splines naturelles cubiques univariées et multivariées [4] convergent vers le classifieur optimal de Bayes lorsque le nombre d'échantillon d'apprentissage est important. Ensuite, nous avons étudier les réseaux de neurones ReLU. Un réseau de neurones est une fonction $f_\theta : \mathbb{R}^D \longrightarrow \{0, 1\}$ prenant en entrée des variables explicatives $x \in \mathbb{R}^D$ et réalisant une prédiction binaire $f_\theta(x) \in \{0, 1\}$. Nous pouvons le définir comme étant une composition de $L \in \mathbb{N}^*$ fonctions, ou couches, $f_\theta(x) = f^{(L)}_{\theta^{(L)}} \circ \cdots \circ f^{(1)}_{\theta^{(1)}}(x)$ dépendant d'un ensemble de paramètres $\theta = \left(\theta^{(1)}, \ldots, \theta^{(L)}\right)$. Une couche au niveau ℓ est une fonction $f^{(\ell)}_{\theta^{(\ell)}}$, le plus généralement non-linéaire, prenant en entrée un signal $z^{(\ell-1)}(x) \in \mathbb{R}^{D_{\ell-1}}$, et produisant $z^{(\ell)}(x) \in \mathbb{R}^{D_\ell}$, où D_ℓ est le nombre de neurones de la couche ℓ. Les réseaux de neurones profonds peuvent alors être considérés comme une régression logistique non-linéaire lorsque nous appliquons une fonction d'activation sigmoïde sur la couche de sortie du réseau. À l'aide d'un exemple théorique, nous illustrons le partitionnement induit par le réseau de neurones sur l'espace des variables d'entrées. Enfin, nous montrons comment ce partitionnement peut faciliter le fonctionnement et l'estimation de la régression logistique, notamment lorsque le nombre de données d'apprentissage est limité.

Nos prochains travaux s'intéresseront à l'apprentissage explicite d'un réseau de neurones en s'appuyant sur la théorie des splines. Il s'agira en particulier d'étudier s'il est possible d'éviter que le réseau de neurones ne soit perçu comme une boîte noire, notamment dans le domaine biomédical.

Références

[1] Randall Balestriero et al. A spline theory of deep learning. *International Conference on Machine Learning*, pages 374–383, 2018.

[2] Cathy WS Chen, Jennifer SK Chan, Richard H Gerlach, et al. A comparison of estimators for regression models with change points. *Available at SSRN 1579184*, 2010.

[3] Paul Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[5] Douglas M Hawkins. On the choice of segments in piecewise approximation. *IMA Journal of Applied Mathematics (Institute of Mathematics and Its Applications)*, 9(2):250–250, 1972.

[6] Helmut Küchenhoff. An exact algorithm for estimating breakpoints in segmented generalized linear models. *Sonderforschungsbereich*, 386(27), 1996.

[7] Vito MR Muggeo. Estimating regression models with unknown break-points. *Statistics in medicine*, 22(19):3055–3071, 2003.

[8] Roberto Pastor and Eliseo Guallar. Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *American journal of epidemiology*, 148(7):631–642, 1998.

[9] AFM Smith and DG Cook. Straight lines with a change-point: A bayesian analysis of some renal transplant data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):180–189, 1980.

[10] DM Stasinopoulos and RA Rigby. Detecting break points in generalised linear models. *Computational Statistics & Data Analysis*, 13(4):461–471, 1992.

[11] DA Stephens. Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, pages 159–178, 1994.

[12] Asher Tishler and Israel Zang. A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association*, 76(376):980–987, 1981.

[13] Asher Tishler and Isreal Zang. A maximum likelihood method for piecewise regression models with a continuous dependent variable. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 30(2):116–124, 1981.

[14] K Ulm. A statistical method for assessing a threshold in epidemiological studies. *Statistics in medicine*, 10(3):341–349, 1991.

[15] Hongling Zhou and Kung-Yee Liang. On estimating the change point in generalized linear models. *arXiv e-prints*, pages arXiv–0805, 2008.