Régression Logistique One-hot pour la Classification

Baptiste Schall [1] & Rodolphe Anty [2] & Lionel Fillatre [3]

[1] *Université Côte d'Azur, France, bschall@i3s.unice.fr*
[2] *CHU Nice, Unité d'hépatologie, France, anty.r@chu-nice.fr*
[3] *Université Côte d'Azur, France, lionel.fillatre@i3s.unice.fr*

Résumé

Le classifieur de Bayes est très utilisé pour le traitement statistique des données. Lorsqu'on étudie des algorithmes d'apprentissage automatique tels que les réseaux de neurones, le classifieur de Bayes est souvent approché par une régression logistique, mais l'efficacité de cette approximation reste en partie incomprise. Dans cet article, nous proposons une régression logistique non-linéaire basée sur la discrétisation de descripteurs. Nous montrons que cette régression logistique peut parfaitement approcher le classifieur de Bayes naïf à condition d'appliquer un prétraitement spécifique détaillé dans cet article. En outre, grâce à cette méthode, chaque descripteur est associée à une fonction descriptive univariée dont les variations sont apprises de façon unique. Cette fonction descriptive nous permet d'interpréter la contribution d'un descripteur dans le processus de décision. Nous illustrons nos résultats théoriques à l'aide de données radiomiques.

Mots-clés. Classification binaire, Régression Logistique, Modèle additif, Optimalité Bayésienne, Discrétisation de descripteurs

Abstract

The Bayes classifier is widely used for statistical data analysis. When we exploit machine learning algorithms like neural networks, the Bayes classifier is often approximated with a logistic regression but the efficiency of this approximation is still questionable. In this paper, we propose a non-linear logistic regression based on features binning. We show that this logistic regression can perfectly approximate the naive Bayes classifier provided that the features are encoded with a specific preprocessing derived in this paper. Also, using this method each feature is associated with a feature function whose variations are learned uniquely. This feature function allows us to interpret the importance of the feature. We illustrate our theoretical results with radiomics data.

Keywords. Binary classification, Logistic Regression, Additive model, Bayes Optimality, Feature binning

1 Introduction

L'intelligence artificielle, dont les réseaux de neurones, est couramment utilisée dans le domaine du traitement du signal et des images [1]. Les méthodes d'apprentissage profond offrent de bonnes performances mais sont des "boîtes noires" dont les résultats sont souvent très difficiles à interpréter. La Régression Logistique (RL) est massivement utilisée en apprentissage profond en plus d'être très populaire dans le domaine médical.

Dans cet article, nous étudions la RL comme un modèle additif [2] avec décomposition sur une base de fonctions, i.e., la fonction de score de la RL est une somme de fonctions descriptives univariées constantes par morceaux. Chacune de ces fonctions nous permet d'interpréter l'impact de chaque descripteur sur la règle de décision. Cette modélisation est équivalente à la discrétisation des descripteurs et à leur encodage avec le très populaire encodage one-hot [3]. Pour garantir les performances de ce modèle, nous le comparons au Classifier Naïf discret de Bayes Discret (CNBD) [4–7]. Plusieurs articles comparent déjà le CNBD et la RL [8–11] en essayant notamment de créer des méthodes hybrides, par exemple en changeant la méthode d'apprentissage. Notre approche est semblable à [12] qui s'intéresse aux similitudes structurelles entre les modèles. En introduisant l'encodage one-hot [13] explicitement, nous simplifions l'utilisation de cette RL non-linéaire en plus d'un apport notable en terme d'interprétabilité. Nous explorons les deux aspects principaux de la comparaison entre RL et CNBD : l'erreur d'approximation et l'erreur d'estimation [3]. D'une part, nous montrons que l'erreur d'approximation entre la RL et le CNBD est nulle. D'autre part, nous montrons que l'erreur d'estimation liée à l'entraînement du RL nous empêche d'estimer l'ensemble des fonctions descriptives, mais que nous pouvons estimer avec précision leurs tendances.

Les contributions de cet article sont les suivantes. Tout d'abord, nous montrons qu'une RL dont les descripteurs sont one-hot encodés est équivalente à une RL dont la fonction de score est constante par morceaux ; chaque descripteur contribue à la fonction de score de façon non-linéaire via une fonction descriptive spécifique. Cette modélisation rend la RL one-hot encodée, appelée RLO, plus flexible. Ensuite, nous montrons que la RLO est équivalente au CNBD appliqué à des descripteurs discrétisés. Cette équivalence montre que la RLO est presque optimale sous les hypothèses habituelles requises par le CNBD. Troisièmement, nous montrons que les variations des fonctions descriptives peuvent être estimées de manière unique. Cette estimation est possible grâce à l'introduction d'un encodage spécifique, l'encodage one-hot suffisant. Enfin, nos expériences numériques établissent que la RLO est un bon compromis entre complexité et performance. Nous la comparons à des classifieurs plus complexes tels que le “Gradient-Boosting” et la “Random Forest” [2, 3].

La structure de l'article est la suivante. La section 2 étudie l'erreur d'approximation de la RLO. La section 3 introduit l'encodage suffisant et étudie l'estimation des fonctions descriptives. La section 4 illustre, à l'aide de données réelles, la pertinence de nos résultats théoriques. La section 5 conclut l'article.

2 Erreur d'approximation de la RL non-linéaire

Un échantillon $\underline{x}$ contient d descripteurs tel que $\underline{x} = (x_1, \ldots, x_d)$. Les données sont normalisées telles que $x_i \in \Omega_i = \left[-1, 1\right]$ pour tout $i = 1, \ldots, d$. Le vecteur $\underline{x}$ est la réalisation du vecteur aléatoire $\underline{X} = (X_1, \ldots, X_d)$ composé de d variables X_i.

La RL est un classifieur linéaire qui a comme objectif de modéliser la probabilité a posteriori $p(c|\underline{x})$ à l'aide de la fonction sigmoïde, $\sigma(t) = 1/(1+\exp(-t))$, appliquée à une fonction de score linéaire :

$$f_{\underline{\beta}}^{RL}(\underline{x}) = \sigma\left(\beta_0 + \beta_1 x_1 + ... + \beta_d x_d\right) = \sigma(h_{\underline{\beta}}^{RL}(\underline{x})). \tag{1}$$

Il est bien connu qu'étant donné une observation $\underline{x}$, la règle de décision Maximum A Posteriori (MAP) optimale [3] choisit la classe $c^*(\underline{x})$ qui maximise $p(c|\underline{x})$:

$$c^*(\underline{x}) = \underset{c \in \{0,1\}}{\arg\max}\, p(c|\underline{x}). \tag{2}$$

La RL est un modèle additif dans la mesure où la fonction de score $h^{RL}(\underline{x})$ dans (1) se réécrit

$$h_{\underline{\beta}}^{RL}(\underline{x}) = \beta_0 + f_{\beta_1}^{RL}(x_1) + ... + f_{\beta_d}^{RL}(x_d), \quad \text{où} \quad f_{\beta_i}^{RL} : x_i \mapsto \beta_i x_i \tag{3}$$

avec $\beta_i \in \mathbb{R}$. Nous proposons une variante de la RL, appelé la RLO ou RL avec encodage one-hot, avec des fonctions constantes par morceaux. Pour obtenir ce modèle, on partitionne le domaine de définition Ω_i des descripteurs x_i en m_i intervalles disjoints $I_{i,j}$ tels que $\Omega_i = \cup_{j=1}^{m_i} I_{i,j}$. Ainsi, la fonction constante par morceaux pour le descripteur x_i est donnée par :

$$f_{\underline{\beta}_i}^{RLO} : x_i \mapsto \sum_{j=1}^{m_i} \beta_{i,j} \mathbb{1}\left\{x_i \in I_{i,j}\right\}, \tag{4}$$

où $\underline{\beta}_i = \left[\beta_{i,1}, ..., \beta_{i,m_i}\right] \in \mathbb{R}^{m_i}$ et $\mathbb{1}\{A\}$ est la fonction indicatrice qui est égale à 1 si l'événement A est vrai et 0 sinon. De ce fait, la fonction descriptive $f_{\underline{\beta}_i}^{RLO}(x_i)$ sur le $j$ème intervalle $I_{i,j}$ du $i$ème descripteur est égale à $\beta_{i,j}$. Nous montrons dans la sous-section suivante que ce modèle constant par morceaux permet à la RLO d'approximer de manière précise la règle de décision optimale de Bayes naïf. Un descripteur x_i one-hot encodée sera notée $\tilde{\underline{x}}_i \in \{0,1\}^{m_i}$ tel que :

$$\tilde{\underline{x}}_i = \left[\mathbb{1}\left\{x_i \in I_{i,1}\right\}, ..., \mathbb{1}\left\{x_i \in I_{i,m_i}\right\}\right]. \tag{5}$$

Quand $x_i \in I_{i,j}$, tous les éléments dans le vecteur $\tilde{\underline{x}}_i$ sont des zéros sauf le $j$ème On peut donc réécrire avec l'encodage one-hot :

$$f_{\underline{\beta}_i}^{RLO}(x_i) = \underline{\beta}_i^T \tilde{\underline{x}}_i, \tag{6}$$

où $\underline{x}^T$ est la transposée de $\underline{x}$.

2.1 Optimalité Bayésienne

Avec (6), la fonction de score de la RLO devient :

$$h_{\underline{\beta}}^{RLO}(\underline{x}) = \beta_0 + \sum_{i=1}^{d} \underline{\beta}_i^T \underline{\tilde{x}}_i. \tag{7}$$

Grâce à cette nouvelle écriture de la fonction de score de la RLO, nous pouvons montrer qu'elle correspond au Classificateur Naïf de Bayes (CNB). Il est bien connu que le CNB est le classificateur optimal lorsque les variables X_i sont indépendantes conditionnellement à la classe c [3]. La fonction de décision du CNB $f^{CNB}(\underline{x}) \in \{0, 1\}$ est

$$f^{CNB} = \underset{c\in\{0,1\}}{\arg\max}\, \mathbb{P}(C = c) \prod_{i=1}^{d} \mathbb{P}_c(X_i = x_i), \tag{8}$$

où $\mathbb{P}(C = c)$ est la probabilité a priori de la classe c et $\mathbb{P}_c(X_i = x_i)$ est la probabilité conditionnelle de X_i étant donné la classe c. Pour établir l'équivalence avec la RLO, nous devons approximer la probabilité $\mathbb{P}_c(X_i = x_i)$ sur la partition $\Omega_i = \cup_{j=1}^{m_i} I_{i,j}$. En d'autres termes, nous devons considérer le CNB discrétisé (CNBD) f^{CNBD} donné par

$$f^{CNBD} = \underset{c\in\{0,1\}}{\arg\max}\, \mathbb{P}(C = c) \prod_{i=1}^{d} \prod_{j=1}^{m_i} \mathbb{P}_c(X_i \in I_{i,j})^{\mathbb{1}\{x_i \in I_{i,j}\}}. \tag{9}$$

Lorsque la taille maximale des intervalles $I_{i,j}$ est faible, l'écart entre le CNB et le CNBD est presque négligeable. On peut réécrire le CNB comme $f^{CNBD}(\underline{x}) = \mathbb{1}\{h^{CNBD}(\underline{x}) > 0\}$ où

$$h^{CNBD}(\underline{x}) = \alpha_0 + \sum_{i=1}^{d} \sum_{j=1}^{m_i} \alpha_{i,j} \mathbb{1}\{x_i \in I_{i,j}\}. \tag{10}$$

Les coefficients α_0 et $\alpha_{i,j}$ sont donnés par :

$$\alpha_0 = \ln \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)}, \quad \alpha_{i,j} = \ln \frac{\mathbb{P}_1(X_i \in I_{i,j})}{\mathbb{P}_0(X_i \in I_{i,j})}. \tag{11}$$

Avec l'encodage one-hot $\underline{\tilde{x}}_i$, on peut réécrire h^{CNBD} tel que :

$$h^{CNBD}(\underline{x}) = \alpha_0 + \sum_{i=1}^{d} \underline{\alpha}_i^T \underline{\tilde{x}}_i, \tag{12}$$

où $\underline{\alpha}_i = \left[\alpha_{i,1}, ..., \alpha_{i,m_i}\right]$. Nous pouvons remarquer que (7) coïncide avec (12) à l'exception de leurs coefficients qui sont différents : les coefficients CNBD sont déduits des distributions de probabilité discrétisées, tandis que les coefficients RLO sont appris à partir de l'ensemble de données d'apprentissage. Par conséquent, le modèle RLO est structurellement optimal par rapport au modèle CNBD. Ses coefficients peuvent être interprétés comme des rapports de vraisemblance optimaux (11). L'erreur d'approximation est donc nulle. Toutefois, comme le montre la section suivante, l'erreur d'estimation n'est pas négligeable. Il est essentiel de s'assurer que les paramètres du RLO peuvent être estimés de manière fiable.

3 Erreur d'estimation

3.1 Non unicité de l'estimation

Le modèle RLO est estimé avec l'entropie croisée binaire habituelle $\mathcal{L}(\underline{\beta})$:

$$\mathcal{L}(\underline{\beta}) = -\sum_{i=1}^{N} c_i \ln\big(\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\big) + (1-c_i)\ln\big(1-\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\big) \tag{13}$$

où $\underline{\tilde{x}}_i$ désigne l'encodage (5) de $\underline{x}_i$. Soit $\underline{\hat{\beta}}$ un estimateur qui minimise $\mathcal{L}(\underline{\beta})$. Nous nous attendons bien sûr à ce que $\underline{\hat{\beta}}$ soit proche du vecteur de paramètres inconnu $\underline{\alpha}$. La RLO (12) est entièrement caractérisée par l'ensemble fini des scores $\sigma(\underline{\hat{\beta}}^\top \underline{\tilde{x}}_i)$ calculés sur l'ensemble du jeu de données $\mathcal{D}$. Par conséquent, la RLO est définie de manière unique par le vecteur de score $\tilde{X}\underline{\hat{\beta}}$ où $\tilde{X}$ est la matrice des données d'entrée (un échantillon par ligne) :

$$\tilde{X} = \left[\underline{\tilde{x}}_1^\top, \ldots, \underline{\tilde{x}}_N^\top\right]^\top \in \mathbb{R}^{N\times m}. \tag{14}$$

Le hessien $\nabla^2\mathcal{L}(\underline{\beta})$ de $\mathcal{L}(\underline{\beta})$ peut être réécrit :

$$\nabla^2\mathcal{L}(\underline{\beta}) = \tilde{X}^\top D(\underline{\beta})\tilde{X}, \tag{15}$$

$$D(\underline{\beta}) = \operatorname{diag}\left(\left[\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\left(1-\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\right)\right]\right), \tag{16}$$

où $\operatorname{diag}(\underline{u})$ désigne la matrice diagonale avec comme diagonale le vecteur $\underline{u}$. Puisque $D(\underline{\beta})$ est diagonale avec tous les éléments diagonaux non nuls, le rang de $\nabla^2\mathcal{L}(\underline{\beta})$ est égal à celui de la matrice de données $\tilde{X}$. On en déduit le lemme suivant.

Lemme 1. *La hessienne $\nabla^2\mathcal{L}(\underline{\beta})$ est semi-définie positive de rang $m-d$. La fonction $\mathcal{L}(\underline{\beta})$ n'est pas strictement convexe et l'estimateur $\underline{\hat{\beta}}$, s'il existe, n'est pas unique.*

3.2 Encodage Suffisant

Afin de garantir l'unicité de l'estimation remise en cause par le lemme 1, la proposition suivante introduit un nouveau encodage, appelé encodage suffisant, qui garantit la convexité stricte de $\mathcal{L}(\underline{\beta})$.

Proposition 1. *Il existe une matrice de permutation Q telle que la matrice $\tilde{X}$ et le vecteur $\underline{\beta}$ satisfassent :*

$$\tilde{X}Q = \left[\tilde{S} \;\middle|\; \tilde{R}\right], \quad Q^\top\underline{\beta} = \begin{bmatrix}\underline{\theta}^S \\ \underline{\theta}^R\end{bmatrix}, \tag{17}$$

où $\tilde{S}$ est de taille $N\times(m-d)$, $\tilde{R}$ est de taille $N\times d$, $\underline{\theta}^S \in \mathbb{R}^{m-d}$ et $\underline{\theta}^R \in \mathbb{R}^d$. Les propriétés suivantes sont vérifiées :

- *$\tilde{S}$ est une matrice de rang plein colonne $m-d$,*

- *Il y a une matrice L de taille $(m-d)\times d$, de rang plein colonne d, telle que $\tilde{R}=\tilde{S}L$.*
- *Les scores $\tilde{X}\underline{\beta}$ sur $\mathcal{D}$ sont préservés, i.e*

$$\tilde{X}\underline{\beta}=\tilde{S}(\underline{\theta}^S+L\underline{\theta}^R)=\tilde{S}\underline{\theta}, \tag{18}$$

avec $\underline{\theta}=\underline{\theta}^S+L\underline{\theta}^R$.

La proposition 1 introduit une nouvelle matrice d'encodage $\tilde{S}$ déduite de l'encodage initial $\tilde{X}$ qui est détaillé dans la sous-section 3.3. Ce nouvel encodage définit une nouvelle RL, appelé RLOS et notée $h_{\underline{\theta}}^{RLOS}(\underline{x})$.

Lemme 2. *Soit $\mathcal{L}(\underline{\theta})$ la fonction de perte à minimiser pour estimer $h_{\underline{\theta}}^{RLOS}(\underline{x})$. Le hessien $\nabla^2\mathcal{L}(\underline{\theta})\in\mathbb{R}^{(m-d)\times(m-d)}$ de $\mathcal{L}(\underline{\theta})$ est défini positif de rang $m-d$. La fonction $\mathcal{L}(\underline{\theta})$ est donc strictement convexe et l'estimateur $\hat{\underline{\theta}}$ qui minimise $\mathcal{L}(\underline{\theta})$, s'il existe, est unique.*

Grâce au lemme 2, nous avons la garantie d'obtenir une estimation unique $\hat{\underline{\theta}}$. Il s'agit d'un avantage significatif par rapport aux méthodes alternatives qui sont souvent utilisées pour obtenir une estimation raisonnable, comme une initialisation aléatoire pertinente de l'optimisation RL ou une régularisation de la fonction de perte (nous obtenons une estimation biaisée). Même si l'estimation de $\underline{\theta}$ est maintenant facilitée, il est important de reconstruire le vecteur $\underline{\beta}$ afin de trouver une interprétation de la RLOS en termes de fonctions descriptives constantes par morceaux (4). La proposition suivante montre qu'il existe un nombre infini de vecteurs $\underline{\beta}$ qui forment un espace linéaire de dimension d.

Proposition 2. *Soit les matrices Q et L définies dans la Proposition 1. Soit $\hat{\underline{\theta}}$ l'estimateur obtenu par le lemme 2. Tout RLO $h_{\underline{\beta}}^{RLO}(\underline{x})$ à coefficients $\underline{\hat{\beta}}=\underline{\hat{\beta}}(\hat{\underline{\theta}},\underline{\theta}^R)$ de la forme*

$$\underline{\hat{\beta}}(\hat{\underline{\theta}},\underline{\theta}^R)=Q\left(\begin{bmatrix}\hat{\underline{\theta}}\\0\end{bmatrix}+\begin{bmatrix}-L\\I_d\end{bmatrix}\underline{\theta}^R\right)=Q\Big(b(\hat{\underline{\theta}})+A\underline{\theta}^R\Big), \tag{19}$$

où $\underline{\theta}^R\in\mathbb{R}^d$ est choisi arbitrairement, satisfait

$$h_{\underline{\hat{\beta}}}^{RLO}(\underline{x})=h_{\hat{\underline{\theta}}}^{RLSO}(\underline{x}),\forall\underline{x}. \tag{20}$$

3.3 Fonction descriptive constante par morceaux

D'après le lemme 1, l'estimation de la RLO n'est pas unique. Par conséquent, nous devons d'abord apprendre la RLOS. Pour apprendre une RLOS, nous devons encoder chaque variable avec l'encodage spécifique, appelé encodage suffisant, donné par la matrice $\tilde{S}$ dans la proposition 1. Cet encodage plus compact ne conserve que $m-d$ bits de l'encodage original $\underline{\tilde{x}}_i$. Les bits supprimés sont choisis arbitrairement mais, selon (18), cela n'affecte pas le score final. Il existe un moyen simple d'obtenir un encodage suffisant. Soit $\underline{\tilde{x}}_i^s$ les descripteurs codés tels que

$$\underline{\tilde{x}}_i^s=\left[\mathbb{1}\left\{x_i\in I_{i,1}\right\},...,\mathbb{1}\left\{x_i\in I_{i,m_i-1}\right\}\right]. \tag{21}$$

Le vecteur $\tilde{\underline{x}}_i^s$ est le vecteur $\tilde{\underline{x}}_i$ sans sa dernière composante. Nous pouvons alors construire $\tilde{S}$ dont les lignes sont $\tilde{\underline{x}}_i^s$. De cet encodage, nous pouvons déduire les matrices Q, $\tilde{R}$ et L dans la Proposition 1. Ensuite, nous pouvons apprendre les coefficients uniques de la RLOS. Comme indiqué dans la proposition 2, étant donné la RLOS, nous pouvons construire un nombre infini de RLO en choisissant différents $\underline{\theta}_R$. Toutes ces RLO ont le même score mais pas les mêmes fonctions descriptives. Choisissons deux vecteurs arbitraires $\underline{\theta}^{R_1}$ et $\underline{\theta}^{R_2}$ et calculons $\underline{\hat{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_1})$ and $\underline{\hat{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_2})$. Avec (19), on a

$$\underline{\hat{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_1}) - \underline{\hat{\beta}}(\hat{\underline{\theta}}, \underline{\theta}^{R_2}) = Q \begin{bmatrix} -L \\ I_d \end{bmatrix} \left(\underline{\theta}^{R_1} - \underline{\theta}^{R_2}\right). \tag{22}$$

Ensuite, en utilisant les valeurs appropriées de Q et L, un bref calcul montre que

$$\underline{\hat{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_1}) - \underline{\hat{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_2}) = \underline{\theta}_i^{R_1} - \underline{\theta}_i^{R_2} = \delta_i, \tag{23}$$

en notant $\underline{\theta}^{R_k} = \left[\underline{\theta}_1^{R_k}, \ldots, \underline{\theta}_d^{R_k}\right]$. Par conséquent, on obtient immédiatement que

$$f^{RLO}_{\underline{\hat{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_1})}(x_i) - f^{RLO}_{\underline{\hat{\beta}}_i(\hat{\underline{\theta}}, \underline{\theta}^{R_2})}(x_i) = \delta_i, \ \forall x_i \in \Omega_i, \tag{24}$$

à partir de la définition des fonctions descriptives dans (4). Par conséquent, bien que les fonctions descriptives ne soient pas uniques, elles sont simplement liées les unes aux autres par un décalage constant δ_i comme nous pouvons le voir sur la figure 1. Ce décalage peut être différent pour chaque descripteur. Par conséquent, l'interprétation des coefficients RLO est relative et non absolue : on interprète les variations des fonctions descriptives pour évaluer l'impact d'une variable d'entrée. En outre, dans la figure 1, nous pouvons noter la non-linéarité des fonctions descriptives RLO par rapport aux fonctions descriptives linéaires d'une RL conventionnelle.

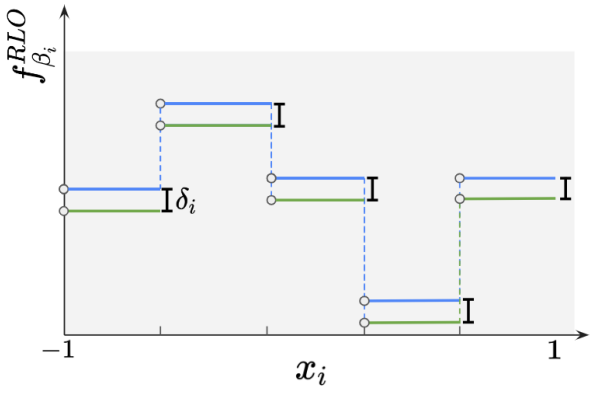

Figure 1: Exemple de fonctions descriptives de deux RLO, en vert et en bleu, provenant d'une même RLOS. Elles présentent les mêmes variations mais il y a un décalage vertical.

4 Expériences numériques

Nous avons testé notre méthode avec deux ensembles de données radiomiques créés à l'aide de la librairie PyRadiomic. La radiomique peut être définie comme un processus complexe visant à extraire des données numériques exploitables à partir d'images tomographiques. Les deux jeux de données proviennent du Mathematical Oncology Laboratory qui a publié en 2023 un jeu de données radiomiques massif [14] regroupant différents types de pathologie. Nous avons décidé de concentrer notre étude sur deux sous-ensembles de données qui en sont extraits. Le premier, RadioNsclc, traite les patients atteints d'un cancer du poumon non à petites cellules (Nsclc) et le second, RadioBreast, traite les patients atteints d'un cancer du sein. Pour les deux sous-ensembles de données, notre objectif est de prédire le temps de survie après la détection de la tumeur, qui est ici binaire (0 temps de survie court, 1 temps de survie long), déterminé en utilisant le temps de survie médian comme point critique. Ces ensembles de données contiennent respectivement 621 et 316 d'échantillons.

Il est commun d'utiliser des techniques de sélection de descripteurs en radiomique. Le problème de ces techniques est qu'elles introduisent un biais ; elles peuvent favoriser leur modèle sous-jacent. Par exemple, la sélection de descripteurs basée sur la RL avec une méthode LASSO [2] peut favoriser la sélection de variables qui interviennent de façon linéaire dans le modèle (au détriment de variables plus informatives qui interviendraient de façon non-linéaire). Par souci de simplicité, nous avons sélectionné manuellement les descripteurs pertinents pour mettre en évidence la non-linéarité en radiomique. Pour les deux ensembles de données, nous conservons une dizaine de descripteurs sur la centaine initiale.

Nous comparons d'abord la RL conventionnelle et la RLOS. Chaque modèle est entraîné à l'aide d'une validation croisée à 5 blocs. L'ensemble de données est discrétisé avec des intervalles de discrétisation variant de 5 à 10 en fonction du descripteur. La taille des intervalles est uniforme et déterminée par la règle de Sturges [15]. Les résultats sont présentés dans le tableau 1. La RL non-linéaire (RLOS) obtient de meilleures performances (+5-6%), ce qui prouve l'importance de la modélisation des effets non-linéaires. La figure 2 affiche la fonction descriptive RLO (4) pour le descripteur "Maximum 3D Diameter". Ce descripteur correspond à la plus grande distance euclidienne entre deux sommets du maillage représentant la tumeur en 3D. Comme indiqué dans la sous-section 3.3, ce sont les variations de la fonction descriptive RLO qui sont informatives, d'où la pertinence de n'étudier qu'une seule RLO. Des interprétations médicales pourraient découler de cette fonction descriptive.

	Train	**Test**	**Train**	**Test**
Méthodes	**Précision moyenne (écart type)**			
RL	**69 (1)**	**64 (5)**	**68 (1)**	**64 (3)**
RLOS	**77 (2)**	**70 (4)**	**75 (2)**	**69 (3)**
Données	**RadioBreast**		**RadioNsclc**	

Table 1: Tableau des précisions et des écart-types sur l'ensemble d'entraînement "Train" et l'ensemble de test "Test" pour les ensembles de données RadioBreast et RadioNsclc pour différents modèles RL (la meilleure précision est surlignée).

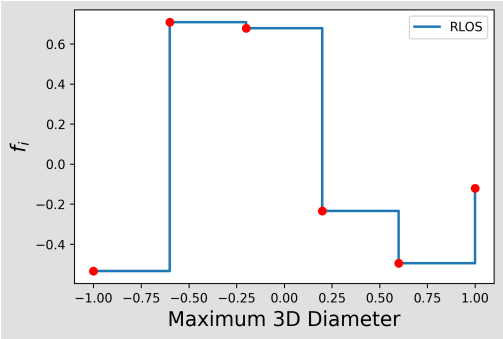

Figure 2: Fonction descriptive RLO du "Maximum 3D Diameter" pour RadioNsclc.

	Train	**Test**	**Train**	**Test**
Méthodes	**Précision moyenne (écart type)**			
RF	**85 (1)**	**77 (3)**	**80 (1)**	**72 (3)**
GB	**84 (2)**	**76 (4)**	**86 (1)**	**76 (2)**
Données	**RadioBreast**		**RadioNsclc**	

Table 2: Tableau des précisions et des écart-types sur l'ensemble d'entraînement "Train" et l'ensemble de test "Test" pour les ensembles de données RadioBreast et RadioNsclc (la meilleure précision est surlignée).

Maintenant que nous avons vu que la RLOS est plus performante que la RL conventionnelle, nous pouvons comparer cette méthode avec des méthodes de classification avancées telles que le Gradient Boosting (GB) et les Random Forest (RF). Comme nous pouvons le voir dans le tableau 2, les modèles plus complexes obtiennent de meilleurs résultats. Cependant, ces méthodes souffrent d'un manque d'interprétabilité car leurs règles de décision sont très difficiles à comprendre. Par exemple, GB est basé sur un système de vote, dans notre cas un vote entre cent arbres sous-jacents, qui rend la règle de décision obscure. Le but de la RLOS n'est pas de surpasser ces méthodes mais de montrer qu'un modèle additif plutôt simple peut obtenir des résultats satisfaisants en plus d'un gain massif en interprétabilité. Enfin, la RLOS est très facile à configurer car les seuls hyper-paramètres sont liés à l'étape de discrétisation (nombre et largeur des intervalles). Pour entraîner les GB et RF, nous avons utilisé un algorithme dit "gridsearch" qui est à la fois fastidieux et chronophage.

5 Conclusion

Cet article montre que nous pouvons déduire une RL non-linéaire additive avec une performance et une interprétabilité accrues en représentant les descripteurs avec des vecteurs one-hot. Nous avons prouvé que cette RL one-hot est quasi-optimale au sens de Bayes. En outre, nous avons introduit l'encodage suffisant qui établit l'unicité des variations des fonctions descriptives associées aux variables d'entrées. Nos futurs travaux porteront sur la manière de prendre en compte les interactions entre les descripteurs.

References

[1] Erik Meijering et al., "Deep learning in biological image and signal processing," *IEEE Signal Processing Magazine*, vol. 39, no. 2, pp. 24–26, 2022.

[2] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[3] Kevin P Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.

[4] Gherardo Varando et al., "Decision boundary for discrete bayesian network classifiers," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2725–2749, 2015.

[5] Daniel Berend and Aryeh Kontorovich, "A finite sample analysis of the naive bayes classifier," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1519–1545, jan 2015.

[6] Concha Bielza and Pedro Larranaga, "Discrete bayesian network classifiers: A survey," *ACM Computing Surveys*, vol. 47, pp. 1–43, 07 2014.

[7] Harry Zhang, "The optimality of naive bayes," in *Proceedings of the Seventeenth International FLAIRS Conference*, 2004.

[8] Andrew Ng and Michael Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *NIPS*, vol. 14, 2001.

[9] Chenyu Zheng et al., "Revisiting discriminative vs. generative classifiers: Theory and implications," *arXiv preprint arXiv:2302.02334*, 2023.

[10] P. Charan Kumar and B. T. Geetha, "Efficient removal of real time rain streaks from a image using novel naive bayes (NB) compare over linear regression (LR) with improved accuracy," in *International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 2023, pp. 1–6.

[11] Tapan Kumar Bhowmik, "Naive bayes vs logistic regression: Theory, implementation and experimental validation," *Inteligencia Artificial*, vol. 18, no. 56, pp. 14–30, Dec. 2015.

[12] Teemu Roos et al., "On discriminative bayesian network classifiers and logistic regression," *Machine Learning*, vol. 59, pp. 267–296, 2005.

[13] Samet Oymak, Mehrdad Mahdavi, and Jiasi Chen, "Learning feature nonlinearities with regularized binned regression," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1452–1456.

[14] Beatriz Ocaña-Tienda et al., "A comprehensive dataset of annotated brain metastasis MR images with clinical and radiomic data," *Scientific Data*, vol. 10, no. 1, pp. 208, 2023.

[15] David W Scott, "Sturges' rule," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, pp. 303–306, 2009.