Équivalence entre la régression logistique et le classifieur naïf de Bayes

BAPTISTE SCHALL[1] LIONEL FILLATRE[1] RODOLPHE ANTY[2]

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

[2]CHU de Nice, Hôpital de l'Archet 2, Unité d'hépatologie, pôle DIGi-TUNED, Nice

Résumé – La régression logistique est très appréciée en apprentissage automatique mais, lorsque les données ne sont pas linéairement séparables, son optimalité et son interprétabilité ne sont pas encore établies. Dans le cas de données discrètes (ou discrétisées) indépendantes, nous montrons que la régression logistique peut parfaitement approximer le test de Bayes naïf avec un encodage approprié des données. Toutefois, elle n'est pas équivalente au test de Bayes naïf. Nous montrons qu'une régression logistique entraînée correspond à une classe de modèles probabilistes Bayésien paramétrée par un espace affine. Des simulations numériques illustrent les résultats théoriques.

Abstract – The logistic regression is well established in machine learning but, when data are not linearly separable, its optimality and interpretability have not been established yet. When data are discrete (or discretized) and independent, we show that the logistic regression can perfectly approximate the Naive Bayes Test with a specific encoding. However, it is not equivalent to the Naive Bayes Test. We show that a trained logistic regression is associated to a class of Bayesian probabilistic models parameterized by an affine space. Numericals simulations illustrate our theoretical results.

1 Introduction

L'intelligence artificielle est couramment utilisée dans le domaine médical [11, 6]. Dans le domaine de la classification, il existe deux types de modèles [7] : i) les modèles génératifs qui à partir des données $x \in \mathbb{R}^d$ et des étiquettes $c \in \{0, 1\}$ modélisent la probabilité conjointe $\mathbb{P}(x, c)$, puis en déduisent $\mathbb{P}(c|x)$ pour estimer l'étiquette la plus probable, et ii) les modèles discriminatifs qui modélisent directement $\mathbb{P}(c|x)$. Cette modélisation plus directe explique que les modèles discriminatifs, en particulier la régression logistique, soient très utilisés avec les réseaux de neurones profonds [4].

Cet article s'intéresse en particulier à la comparaison entre la régression logistique (RL) et le Classifieur Naïf de Bayes (CNB). Lorsque le modèle probabiliste des données est complètement connu et que les variables sont indépendantes conditionnellement à la classe prédite, le CNB est le classifieur optimal. De plus, la structure du CNB est intrinsèquement explicable puisqu'il s'appuie sur des rapports de vraisemblances qui, pour une variable d'entrée, mesure la contribution de cette variable pour la prise de décision [7]. L'objectif de l'étude est donc d'établir dans quelle mesure la régression logistique peut approximer le CNB. De nombreux travaux se sont intéressés à cette question, notamment [8, 10, 3, 2, 1, 5, 12], mais le lien entre ces deux modèles n'est toujours pas parfaitement déterminé, sauf dans le cas de données Gaussiennes ou de données binaires. Autrement dit, lorsqu'une RL est estimée, on ne sait toujours pas si elle correspond (ou non) à un, voire plusieurs, CNB. Dans le cas de données discrètes, nous proposons dans cet article une étude approfondie de cette équivalence.

Plus précisément, nous explorons les deux principaux aspects de l'équivalence entre deux classifieurs statistiques : l'erreur d'approximation et l'erreur d'estimation [7]. Nous montrons que l'erreur d'approximation entre la RL et le CNB est nulle pourvu que les données soient préalablement encodées avec un code "one-hot" [7]. Par contre, nous montrons que l'erreur d'estimation, liée à l'entraînement de la RL, est significative : il n'est pas possible de retrouver le modèle probabiliste à l'origine des données. En fait, à partir des coefficients estimés par la RL, nous pouvons accéder à de nombreux modèles probabilistes reliés les uns aux autres. Cette relation entre les modèles est explicitement décrite dans cet article.

Nos contributions principales sont les suivantes. Premièrement, nous proposons un encodage spécifique mettant en avant l'équivalence structurelle entre le CNB et la RL tout en garantissant l'unicité des paramètres de la RL estimée. Ensuite, nous montrons que la RL peut être associée à de nombreux classifieurs optimaux, tous reliés entre eux par un paramétrage affine. Enfin, nous en déduisons qu'une même RL est associée à plusieurs modèles génératifs sans qu'il soit possible d'identifier le modèle génératif qui a généré les données traitées.

La structure de l'article est la suivante. La Section 2 démontre l'équivalence struturelle entre la RL et le CNB. La section 3 démontre que l'estimation d'une RL permet d'accéder à une classe de modèles génératifs reliés les uns aux autres par un paramétrage affine. La section 4 illustre sur des données simulées la pertinence des réultats théoriques établis dans l'article. La section 5 conclut l'article.

2 Erreur d'approximation

Nous considérons une base de données $\mathcal{D} = \{(\underline{x}_i, c_i)\}_{i=1}^N$ composée de N échantillons $\underline{x}_i$ avec leur étiquette $c_i \in \{0, 1\}$. Un échantillon, noté $\underline{x} = (x_1, \ldots, x_d)$, est la réalisation d'un vecteur aléatoire $\underline{X} = (X_1, \ldots, X_d)$ composé de d variables notées X_i. On notera $\mathcal{D}_{\underline{x}} = \{\underline{x}_i\}_{i=1}^N$ l'ensemble des vecteurs $\underline{x}_i$ de $\mathcal{D}$. Chaque variable X_i est à valeurs dans un ensemble $\Omega_i = \{\omega_{i,1}, ..., \omega_{i,m_i}\}$ fini de taille m_i. Les échantillons $\underline{x}_i$ prennent donc leurs valeurs dans l'ensemble fini $\Omega = \Omega_1 \times \Omega_2 \times ... \times \Omega_d$. Sous l'hypothèse d'indépendance entre les composantes du vecteur $\underline{X}$ conditionnellement à la classe prédite, la distribution discrète suivie par $\underline{X}$ lorsqu'il est issu

de la classe c est $\mathbb{P}_c(\underline{X} = \underline{x}) = \prod_{i=1}^{d} \mathbb{P}_c(X_i = x_i)$ avec la notation $\mathbb{P}_c(\underline{X} = \underline{x}) = \mathbb{P}(\underline{X} = \underline{x}|C = c)$. La distribution de $\underline{x}$ pour la classe c est donc défini par les d vecteurs des d distributions marginales

$$\underline{p}_{c,i} = [\mathbb{P}_c(X_i = \omega_{i,j})]_{j=1}^{m_i} \in \mathcal{S}_{m_i}, \quad i = 1, \ldots, d, \quad (1)$$

où $\mathcal{S}_M$ désigne le simplex probabiliste de dimension M. Le problème de classification binaire est alors défini par d couples de distributions $(\underline{p}_{0,i}, \underline{p}_{1,i})$.

2.1 Classifieur Naïf de Bayes (CNB)

Le Classifieur Naïf de Bayes (CNB) est le classifieur optimal lorsque les variables X_i sont indépendantes conditionnellement à la classe c [7]. Il fait parti de la famille des modèles génératifs puisque son calcul nécessite de connaître les couples $(\underline{p}_{0,i}, \underline{p}_{1,i})$. La décision $f^*(\underline{x}) \in \{0, 1\}$ du CNB s'écrit

$$f^*(\underline{x}) = \arg\max_{c \in \{0,1\}} \mathbb{P}(C = c) \prod_{i=1}^{d} \mathbb{P}_c(X_i = x_i) \quad (2)$$

Il est immédiat de vérifier que le CNB peut se ré-écrire sous la forme $f^*(\underline{x}) = \mathbb{1}\{h^*(\underline{x}) > 0\}$ où

$$h^*_{\underline{\alpha}}(\underline{x}) = \alpha_0 + \sum_{i=1}^{d} \sum_{j=1}^{m_i} \mathbb{1}\{x_i = \omega_{i,j}\}\alpha_{i,j}, \quad (3)$$

$\underline{\alpha}$ est le vecteur contenant les coefficients α_0 et $\alpha_{i,j}$,

$$\underline{\alpha} = \left[\alpha_0, \alpha_{1,1}, ..., \alpha_{1,m_1}, ..., \alpha_{d,1}, ..., \alpha_{d,m_d}\right]^\top, \quad (4)$$

et $\mathbb{1}\{A\}$ est la fonction indicatrice qui vaut 1 si l'évènement A est vrai et 0 sinon. Les coefficients α_0 et $\alpha_{i,j}$ sont donnés par

$$\alpha_0 = \ln \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)}, \quad \alpha_{i,j} = \ln \frac{\mathbb{P}_1(X_i = \omega_{i,j})}{\mathbb{P}_0(X_i = \omega_{i,j})}. \quad (5)$$

Apprendre un CNB revient à estimer $\underline{\alpha}$ à partir de $\mathcal{D}$.

2.2 Régression logistique équivalente au CNB

L'encodage "one-hot" est une technique couramment utilisée pour représenter des variables catégorielles [7]. L'encodage "one-hot", modélisé par la fonction $\tilde{e}_H(\cdot)$, est défini par

$$\underline{\tilde{x}} = \tilde{e}_H(\underline{x}) = (1, \tilde{e}_{H,1}(x_1), ..., \tilde{e}_{H,d}(x_d))^\top \in \mathbb{R}^m \quad (6)$$

où $m = 1 + \sum_{i=1}^{d} m_i$ et chaque composante x_j est encodée avec un code "one-hot" $\tilde{e}_{H,j}(x_j)$:

$$\tilde{e}_{H,j}(x_j) = \left[\mathbb{1}\{x_j = \omega_{j,1}\}, \ldots, \mathbb{1}\{x_j = \omega_{j,m_j}\}\right]. \quad (7)$$

La fonction de score (3) du CNB s'écrit alors :

$$h^*_{\underline{\alpha}}(\underline{x}) = \underline{\alpha}^\top \tilde{e}_H(\underline{x}) = \underline{\alpha}^\top \underline{\tilde{x}} = \alpha_0 + \sum_{j=1}^{d} \underline{\alpha}_j^\top \tilde{x}_j \quad (8)$$

où $a^\top$ est la transposée de a et $\underline{\alpha}_j = \left[\alpha_{j,1}, ..., \alpha_{j,m_j}\right]^\top$.

Une Régresssion Logistique (RL) usuelle [7] se modélise sous la forme $f^\dagger(\underline{x}) = \mathbb{1}\{h^\dagger_{\underline{\beta}}(\underline{x}) > 0\}$ où $h^\dagger_{\underline{\beta}}(\underline{x})$ est une fonction linéaire paramétrée par $\underline{\beta}$. Au lieu d'appliquer la RL directement au vecteur $\underline{x}$, nous proposons de l'appliquer au vecteur "one-hot" $\underline{\tilde{x}}$ pour obtenir le score :

$$h^\dagger_{\underline{\beta}}(\underline{x}) = \underline{\beta}^\top \tilde{e}_H(\underline{x}) = \underline{\beta}^\top \underline{\tilde{x}} = h^*_{\underline{\beta}}(\underline{x}) \quad (9)$$

où $\beta \in \mathbb{R}^m$. L'encodage "one-hot" permet ainsi d'établir qu'un modèle CNB coincide parfaitement avec un modèle LR avec un encodage "one-hot" des variables lorsque $\underline{\beta} = \underline{\alpha}$.

3 Erreur d'estimation

3.1 Non-unicité de l'estimation

Le modèle RL avec encodage "one-hot", appelé RLO, est estimé avec le critère habituel, l'entropie croisée binaire $\mathcal{L}(\underline{\beta})$:

$$\mathcal{L}(\underline{\beta}) = -\sum_{i=1}^{N} c_i \ln\left(\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\right) + (1 - c_i) \ln\left(1 - \sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\right) \quad (10)$$

où $\underline{\tilde{x}}_i$ désigne l'encodage (6) de $\underline{x}_i$. Soit $\underline{\hat{\beta}}$ un estimateur qui minimise $\mathcal{L}(\underline{\beta})$. La RLO (9) est entièrement caractérisée par l'ensemble des scores $\sigma(\underline{\hat{\beta}}^\top \underline{\tilde{x}}_i)$. Sous l'hypothèse que $\Omega \subset \mathcal{D}_{\underline{x}}$, c'est-à-dire que la base d'apprentissage contient tous les vecteurs discrets qui peuvent se réaliser, une RLO est donc définie de façon unique par le vecteur des scores $\tilde{X}\underline{\hat{\beta}}$ où $\tilde{X}$ est la matrice des données d'entrées (un échantillon par ligne) :

$$\tilde{X} = \left[\underline{\tilde{x}}_1^\top, \ldots, \underline{\tilde{x}}_N^\top\right]^\top \in \mathbb{R}^{N \times m}. \quad (11)$$

L'estimateur $\underline{\hat{\beta}}$, s'il existe, est unique si et seulement si $\mathcal{L}(\underline{\beta})$ est strictement convexe. Le hessien $\nabla^2\mathcal{L}(\underline{\beta})$ de $\mathcal{L}(\underline{\beta})$ s'écrit

$$\nabla^2\mathcal{L}(\underline{\beta}) = \tilde{X}^\top D(\underline{\beta})\tilde{X}, \quad (12)$$

$$D(\underline{\beta}) = \text{diag}\left(\left[\sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\left(1 - \sigma(\underline{\beta}^\top \underline{\tilde{x}}_i)\right)\right]\right), \quad (13)$$

où $\text{diag}(\underline{u})$ désigne la matrice diagonale dont la diagonale est le vecteur $\underline{u}$. On en déduit le lemme suivant.

Lemme 1 *Le hessien $\nabla^2\mathcal{L}(\underline{\beta})$ est semi-défini positif 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.*

Par conséquent, nous n'avons aucune garantie sur l'erreur d'estimation et sur la fiabilité de l'estimation $\underline{\hat{\beta}}$.

3.2 Encodage suffisant

Pour garantir l'unicité de l'estimation, la proposition suivante introduit un nouvel encodage, appelé l'encodage suffisant, qui garantit la stricte convexité de $\mathcal{L}(\underline{\beta})$.

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

$$\tilde{X}Q = \left[\tilde{S} \;\middle|\; \tilde{R}\right], \quad Q^\top\beta = \begin{bmatrix}\underline{\theta}^S \\ \underline{\theta}^R\end{bmatrix}, \quad (14)$$

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 colone $m - d$,*
- *il existe une matrice L de taille $(m - d) \times d$, de rang plein colonne d, telle que $\tilde{R} = \tilde{S}L$.*
- *le calcul des scores sur $\mathcal{D}_{\underline{x}}$ est préservé, c'est-à-dire*

$$\tilde{X}\underline{\beta} = \tilde{S}(\underline{\theta}^S + L\underline{\theta}^R) = \tilde{S}\underline{\theta}. \quad (15)$$

La proposition introduit une nouvelle matrice d'encodage $\tilde{S}$ déduite de l'encodage initial $\tilde{X}$. Désormais, chaque vecteur $\underline{x}_i$ est encodé sous une forme plus compacte $\tilde{s}_i = \tilde{e}_S(\underline{x}_i)$ qui ne conserve que $m-d$ bits de l'encodage "one-hot" initial $\tilde{\underline{x}}_i$. Ce nouvel encodage définit une nouvelle RL, appelée la RLOS :

$$h^\star_{\underline{\theta}}(\underline{x}) = \underline{\theta}^\top \tilde{e}_S(\underline{x}) = \underline{\theta}^\top \underline{\tilde{s}}, \tag{16}$$

dont les scores, d'après (15), sont égaux à ceux de $h^\dagger_{\underline{\beta}}(\underline{x})$. Les deux RL sont donc parfaitement équivalentes. Cependant, l'estimation du modèle suffisant $h^\star_{\underline{\theta}}$ est nettement plus fiable, ce qui est démontré par la proposition suivante.

Lemme 2 *Soit $\mathcal{L}(\underline{\theta})$ la fonction à minimiser pour estimer $h^\star_{\underline{\theta}}$. 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.*

Il est à noter qu'une régularisation strictement convexe de $\mathcal{L}(\underline{\theta})$ permettrait également de garantir l'unicité de l'estimation. Cependant, la solution dépendrait alors de l'hyperparamètre de régularisation. Bien que l'estimation de $\underline{\theta}$ soit désormais aisée et fiable, il est important de reconstruire $\underline{\beta}$ pour retrouver une interprétation de la RLOS en terme de rapports de vraisemblance comme pour le CNB dans (3)-(4). La proposition suivante montre qu'il existe une infinité de vecteurs $\underline{\beta}$.

Proposition 2 *Soit la matrice de permutation Q et la matrice L définies dans la Proposition 1. Soit $\hat{\underline{\theta}}$ l'estimateur obtenu à partir du Lemme 2. Tout vecteur $\hat{\underline{\beta}}$ de la forme*

$$\hat{\underline{\beta}} = Q\left(\begin{bmatrix}\hat{\underline{\theta}}\\ 0\end{bmatrix} + \begin{bmatrix}-L\\ I_d\end{bmatrix}\underline{\theta}^R\right) = Q\left(b(\hat{\underline{\theta}}) + A\underline{\theta}^R\right), \tag{17}$$

où $\underline{\theta}^R \in \mathbb{R}^d$ est un vecteur choisi arbitrairement, permet d'obtenir une RL $h^\dagger_{\hat{\underline{\beta}}}$ telle que $h^\dagger_{\hat{\underline{\beta}}} = h^\star_{\hat{\underline{\theta}}}$.

D'après la proposition 1, il faut d'abord apprendre la RLOS $h^\star_{\underline{\theta}}(\underline{x})$ dans (16) car le paramètre $\underline{\theta}$ peut être estimé de façon unique en résolvant une optimisation strictement convexe. Il existe alors une infinité de RLO d'après la Proposition 2 qui sont égales à la RLOS en revenant au paramétrage "one-hot" initial. Toutes ces RLO sont reliées par un paramétrage affine défini en (17) qui se résume à l'espace affine noté $\mathcal{C}_{\hat{\underline{\theta}}}$:

$$\mathcal{C}_{\hat{\underline{\theta}}} = \left\{\hat{\underline{\beta}} \in \mathbb{R}^m : \hat{\underline{\beta}} = Q\left(b(\hat{\underline{\theta}}) + A\underline{\theta}^R\right),\ \ \underline{\theta}^R \in \mathbb{R}^d\right\}. \tag{18}$$

3.3 Estimation des modèles génératifs

Pour d couples de distributions génératives donné $(\underline{p}_{0,j}, \underline{p}_{1,j})$, il existe un vecteur de coefficients $\underline{\alpha}$ qui définit un CNB et donc une RLOS. Réciproquement, la proposition suivante montre qu'à une RLOS estimée correspond en fait une infinité de couples génératifs $(\hat{\underline{p}}_{0,j}, \hat{\underline{p}}_{1,j})$ pour lesquels toutes les RLO ont exactement les mêmes performances.

Proposition 3 *Soit $\hat{\underline{\beta}} = \left[\hat{\beta}_0, \hat{\underline{\beta}}_1, \ldots, \hat{\underline{\beta}}_d\right]^\top \in \mathcal{C}_{\hat{\underline{\theta}}}$ defini dans (18). Soit $\mathcal{P}_{\hat{\underline{\beta}}}$ l'ensemble des modèles génératifs défini par*

$$\mathcal{P}_{\hat{\underline{\beta}}} = \Big\{\{(\underline{p}_{0,i}, \underline{p}_{1,i})\}_{i=1}^d : \underline{p}_{0,i}, \underline{p}_{1,i} \in \mathcal{S}_{m_i}, \tag{19}$$

$$\left(\exp\left(\hat{\underline{\beta}}_j\right) - 1\right)^\top \underline{p}_{0,j} = 0, \tag{20}$$

$$\underline{p}_{1,j} = \operatorname{diag}\left(\exp\left(\hat{\underline{\beta}}_j\right)\right)\underline{p}_{0,j}\Big\}, \tag{21}$$

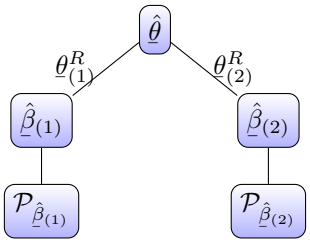

FIGURE 1 : Une estimation de la RLOS permet de retrouver de nombreux modèles génératifs.

où $\exp(\underline{x}) = \left[\exp(x_1), \ldots, \exp(x_d)\right]^\top$. Alors, tous les CNB (3) associés aux couples $\{(\underline{p}_{0,i}, \underline{p}_{1,i})\}_{i=1}^d \in \mathcal{P}_{\hat{\underline{\beta}}}$ sont égaux. De plus, ils coïncident tous avec l'unique RLOS $h^\star_{\hat{\underline{\theta}}}$.

La Figure 1 résume la proposition 3. La RLOS permet d'estimer un paramètre $\hat{\underline{\theta}}$ qui génère une infinité $\mathcal{C}_{\hat{\underline{\theta}}}$ de vecteur de rapport de vraisemblance $\hat{\underline{\beta}}_{(k)}$. Chaque vecteur $\hat{\underline{\beta}}_{(k)}$ dépend d'un paramètre $\underline{\theta}^R_{(k)}$ spécifique. Chacun de ces vecteurs $\hat{\underline{\beta}}_{(k)}$ génère une CNB avec des coefficients différents mais qui, par l'intermédiaire de compensations linéaires internes à la fonction de score, donnent exactement la même fonction de score et donc le même CNB. Pour un vecteur $\hat{\underline{\beta}}_{(k)}$ donné, les distributions discrètes qui partagent les mêmes valeurs du rapport de vraisemblance peuvent être retrouvées en résolvant (19)-(21) avec, par exemple, les résultats décrits dans [9].

4 Expériences numériques

Pour confirmer l'équivalence entre la RLOS et le CNB, nous testons ces méthodes sur des jeux de données simulés discrets. Le premier jeu de données utilisé est appelé GM pour Gaussian Mixture. Il est composé de deux variables, X_1 et X_2, et d'une classe binaire équiprobable. Pour $C=0$, X_1 sera échantillonné avec un mélange de deux Gaussiennes centrées respectivement en -2 et 10 et d'écart type 1, X_2 sera échantillonné avec une Gaussienne centrée en 4 et d'écart type 2. Pour $C=1$, l'échantillonnage de X_1 et X_2 est inversé. Les probabilités marginales pour cette base de données sont présentées en colonne C de la Figure 3. Le deuxième jeu de données utilisé, Circles, provient de scikit-learn : X_1 et X_2 représentent les coordonnées de deux cercles imbriqués, voir ligne 2 de la Figure 2 ; chaque cercle correspond à une classe. Contrairement au jeu de données GM, les variables X_1 et X_2 dans Circles ne seront pas indépendantes ; cette base de données sert à tester la robustesse des différentes méthodes. Pour faire nos expériences, nous normalisons les variables X_1 et X_2 puis les discrétiserons sur 10 segments uniformes pour former le quadrillage présent dans la Figure 2.

Les méthodes CNB, RL classique (sans encodage "one-hot") et RLOS sont comparées sur les jeux de données discrétisés. Les frontières de décision sont donc également discrètes. Sur la Figure 2, les colonnes correspondent aux méthodes utilisées sauf pour la première colonne qui montre les jeux de données ; chaque ligne correspond à une base de données, GM ou Circles. Pour mettre en avant le fait que la RL classique n'est pas équivalente à un CNB, nous avons sélectionné deux jeux de données qui ne sont pas linéairement séparables. De

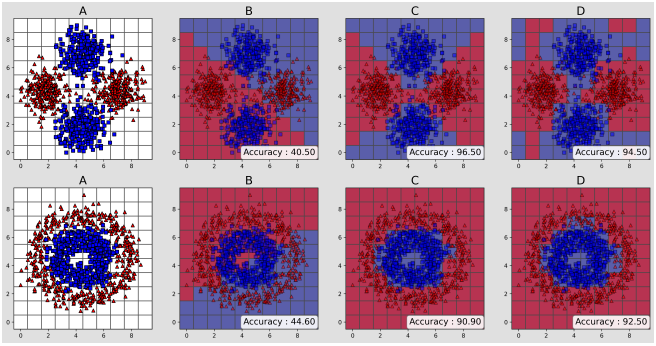

FIGURE 2 : Frontière de décision : **B**) - Du CNB, **C**) - De la RL et **D**) - De la RLOS sur les jeux de données GM et Circles en **A**). Avec en bleu et rouge les différentes classes et X_1,X_2 respectivement en abscisse et ordonnée.

ce fait, la RL classique a une accuracy proche de 50% puisque sa frontière de décision est une droite. Comme cela est démontré dans la section 2, le CNB et la RLOS séparent de façon non-linéaire les deux classes, ce qui confirme bien l'importance de l'encodage "one-hot" (6). La RLOS obtient de très bon résultats (> 90%) sur la base de données GM malgré un bruit observable dans les coins de la Figure 2 (colonne D). Ce bruit, dû a l'absence d'observations dans ces zones, est cependant négligeable comme le montre l'accuracy. Les résultats du CNB et de RLOS sur la base de données Circles confirment la robustesse de ces méthodes.

La Figure 3 illustre la proposition 3 sur la base de données GM uniquement. Les colonnes correspondent aux différents couples de distributions et les deux lignes aux deux classes. Comme décrit dans la Figure 1, nous estimons $\hat{\underline{\theta}}$ en appprenant la RLOS puis, en choisissant $\underline{\theta}_1^R \neq \underline{\theta}_2^R$, nous estimons deux vecteurs $\hat{\underline{\beta}}_{(1)}$ et $\hat{\underline{\beta}}_{(2)}$ distincts avec lesquels nous calculons deux couples de distributions génératifs. Ces distributions sont présentées dans la Figure 3. Les $\underline{\theta}^R$ sont choisis de sorte à avoir un écart $||\hat{\underline{\beta}}_1 - \hat{\underline{\beta}}_2|| = 4.5$ significatif mais un écart nul entre les scores de classification $||\tilde{X}\hat{\underline{\beta}}_1 - \tilde{X}\hat{\underline{\beta}}_2|| = 0$, comme établit dans la proposition 2. La distribution des couples générées (**A**

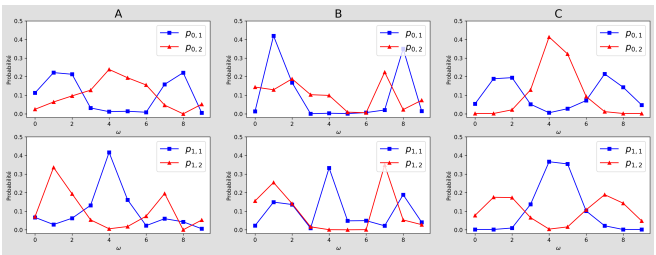

FIGURE 3 : Histogramme des probabilités : **A**) générées à partir de $\hat{\beta}_1$, **B**) générées à partir de $\hat{\beta}_2$, **C**) simulées dans la Figure 2, ligne du haut. L'histogramme de X_1 est en bleu et celui de X_2 est en rouge.

et **B**) est similaire à la distribution initiale (**C**) mais présente des différences assez notables. Malgré cela un phénomène de compensation linéaire fait que les différentes RLO associées au ratio de probabilités de ces nouvelles distributions ont toutes le même score et la même accuracy sur la base de données GM. Toutes ces RLO correspondent à une unique RLOS.

5 Conclusion

Cet article étudie l'équivalence entre le classifieur naïf de Bayes et la regression logistique dans le cas de données discrètes. Avec un encodage "one-hot" approprié, une unique RL permet d'approximer parfaitement un grand nombre de CNB qui sont dérivés des distributions génératives décrites dans la proposition 3. Par conséquent, la régression logistique est un modèle très versatile. Toutefois, il n'est pas possible de retrouver la distribution qui a généré les données à partir des seules informations contenues dans la régression logistique.

Références

[1] Daniel BEREND et Aryeh KONTOROVICH : A finite sample analysis of the naive bayes classifier. *J. Mach. Learn. Res.*, 16(1):1519–1545, jan 2015.

[2] Jose BERNARDO *et al.* : Generative or discriminative ? getting the best of both worlds. *Bayesian Statistics*, 8:3–24, 01 2007.

[3] Concha BIELZA et Pedro LARRANAGA : Discrete bayesian network classifiers : A survey. *ACM Computing Surveys*, 47:1–43, 07 2014.

[4] Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE : *Deep Learning*. MIT Press, 2016.

[5] Tapan KUMAR BHOWMIK : Naive bayes vs logistic regression : Theory, implementation and experimental validation. *Inteligencia Artificial*, 18(56):14–30, Dec. 2015.

[6] Thomas LÉVI-STRAUSS *et al.* : Radiomics, a promising new discipline : Example of hepatocellular carcinoma. *Diagnostics*, 13(7), 2023.

[7] Kevin P. MURPHY : *Machine learning : a probabilistic perspective*. MIT Press, 2013.

[8] Andrew NG et Michael JORDAN : On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. *NIPS*, 14, 2001.

[9] Steven ROMAN : *Positive Solutions to Linear Systems : Convexity and Separation*, pages 395–408. Springer New York, 2005.

[10] Gherardo VARANDO *et al.* : Decision boundary for discrete bayesian network classifiers. *Journal of Machine Learning Research*, 16(1):2725–2749, 2015.

[11] Kun-Hsing YU, Andrew L. BEAM et Isaac S. KOHANE : Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731, octobre 2018.

[12] Harry ZHANG : The optimality of naive bayes. *In Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. AAAI Press, 2004.