Apprentissage d'un classifieur minimax pour données discrètes

Cyprien GILET[1],* Susana BARBOSA[2], Lionel FILLATRE[1]

[1]Université Côte d'Azur, CNRS, laboratoire I3S,
Sophia-Antipolis, France.

[2]Université de Côte d'Azur, CNRS, Institut de Pharmacologie Moléculaire et Cellulaire (IPMC),
Sophia-Antipolis, France.

gilet@i3s.unice.fr sudocarmo@gmail.com lionel.fillatre@i3s.unice.fr

Résumé – L'apprentissage d'un classifieur supervisé lorsque les proportions par classe de la base d'apprentissage diffèrent de celles de la base de test, ou lorsque que les données sont non-balancées, peut augmenter le risque d'erreurs de classification pour de nouvelles observations. Nous nous intéressons ici au classifieur de Bayes non-naïf lorsque les variables prédictives sont discrètes ou discrétisées. Nous montrons que, sous ces conditions et pour une fonction de perte positive quelconque, le risque de Bayes considéré comme une fonction des proportions des classes est concave non-différentiable et affine par morceaux. Nous proposons un algorithme de sous-gradient projeté permettant d'estimer les proportions qui maximise ce risque de Bayes. Le classifieur minimax obtenu minimise le risque conditionnel maximum.

Abstract – Learning a classifier when the class proportions in the training set differ from the state of nature, or when the training set is imbalanced, may increase the misclassification risks when classifying some test samples. This paper studies the non-naive minimax Bayes classifier for classifying discrete features between multiple classes. It shows that when considering any positive loss function, the optimal Bayes risk considered as a function of the class proportions is a concave non-differentiable multivariate piecewise affine function. The maximum value of the optimal Bayes risk corresponds to the class proportions of the minimax classifier. To compute these class proportions, we derive a projected subgradient algorithm whose convergence is established. The resulting minimax classifier minimized the maximum conditional risk.

1 Introduction

Contexte : Nous souhaitons utiliser la classification supervisée pour traiter des problèmes de diagnostics médicaux tels que celui décrit dans [6]. Les données médicales contiennent à la fois des variables catégorielles et des variables numériques. Afin de faciliter la manipulation de ces données, une approche efficace est de discrétiser l'ensemble des variables [9]. Par ailleurs, pour ce type de problèmes, des coûts d'erreurs de classification sont au préalable imposés afin de pénaliser différemment les erreurs de décision qui peuvent avoir des impacts importants sur le traitement des patients. Enfin, dans le domaine médical, les classes à prédire sont fréquemment déséquilibrées, ce qui conduit les classifieurs à délaisser les classes les moins représentées. Le fait de ré-équilibrer la base d'apprentissage peut détériorer le risque moyen d'erreurs pour de nouvelles prédictions. Ce risque moyen peut également se dégrader lorsque les proportions des classes de la base d'apprentissage sont incertaines ou qu'elles évoluent dans le temps. L'objectif de cet article est d'entraîner un classifieur statistique qui 1) exploite des données discrètes, 2) soit le moins sensible possible face aux changements de proportions des classes, et 3) prenne en compte les coûts d'erreurs de classifications.

*Les auteurs remercient la région Provence-Alpes-Côte d'Azur pour son soutien financier.

Position du problème : Le problème de classification supervisée consiste à minimiser le risque empirique moyen d'erreurs de classification à partir d'un ensemble fini d'observations étiquetées $\{(Y_i, X_i), i \in \mathcal{I}\}$. Définissons $K \geq 2$ le nombre de classes, $\mathcal{Y} := \{1, \ldots, K\}$ l'ensemble des classes observées, $\hat{\mathcal{Y}} = \mathcal{Y}$ l'ensemble des classes à prédire, $\mathcal{X}$ l'espace sur lequel l'ensemble des variables observées sont définies, et m le nombre d'observations composant la base d'apprentissage. On note Y_i la variable aléatoire caractérisant la classe de l'observation i, et $X_i = [X_{i1}, \ldots, X_{id}] \in \mathcal{X}$ le vecteur aléatoire regroupant l'ensemble des d variables descriptives associées à l'observation i. Définissons $\Delta = \{\delta : \mathcal{X} \to \hat{\mathcal{Y}}\}$ l'ensemble des classifieurs et considérons une règle de décision $\delta \in \Delta$. Enfin, considérons la matrice des coûts $C \in \mathbb{R}_+^{K \times K}$ telle que, pour tout $(k, l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$, C_{kl} représente le coût d'attribuer la classe $\hat{Y}_i = l$ au patient i alors que sa véritable classe Y_i est k. Le risque empirique $\hat{r}(\delta)$ de δ peut s'écrire [8] :

$$\hat{r}(\delta) = \sum_{k \in \mathcal{Y}} \hat{\pi}_k \hat{R}_k(\delta), \tag{1}$$

où, pour tout $k \in \mathcal{Y}$, $\hat{\pi}_k$ correspond à la proportion d'observations appartenant à la classe k,

$$\hat{R}_k(\delta) = \sum_{l \in \hat{\mathcal{Y}}} C_{kl} \hat{\mathbb{P}}(\delta(X_i) = l \mid Y_i = k), \tag{2}$$

avec $\hat{\mathbb{P}}(\delta(X_i) = l \mid Y_i = k)$ détaillée en (4).

Classifieur minimax : Notons $\delta_{\hat{\pi}}$ un classifieur $\delta \in \Delta$ calibré à partir des observations $(Y_i, X_i)_{i\in\mathcal{I}}$ dont les proportions par classe sont $\hat{\pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_K]$. Ce classifieur est ensuite utilisé pour prédire la classe Y_i' de nouvelles observations de test $\{(Y_i', X_i'), i \in \mathcal{I}'\}$ à partir des variables descriptives associées $X_i' \in \mathcal{X}$. Considérons l'ensemble $\mathcal{I}'$ composé de m' observations vérifiant les proportions $\pi' = [\pi_1', \ldots, \pi_K']$. Le risque d'erreur associé au classifieur $\delta_{\hat{\pi}}$ et aux proportions π' est alors noté et défini par $\hat{r}(\pi', \delta_{\hat{\pi}}) = \sum_{k\in\mathcal{Y}} \pi_k' \hat{R}_k(\delta_{\hat{\pi}})$. Comme illustré sur la figure 1, ce risque évolue et peut se dégrader linéairement lorsque les proportions π' diffèrent de $\hat{\pi}$. Comme décrit dans [8], une solution pour rendre une règle de décision robuste à de possibles différences de proportions entre $\mathcal{I}$ et $\mathcal{I}'$ consiste à minimiser $\max_{k\in\mathcal{Y}} \hat{R}_k(\delta_{\hat{\pi}})$, dans le but de minimiser le pire cas possible au vue des changements potentiels de proportions π'. En notant $\mathbb{S}$ le simplexe probabiliste de dimension K, ce problème minimax peut alors se réécrire

$$\delta_{\hat{\pi}}^B \;=\; \underset{\delta\in\Delta}{\operatorname{argmin}} \max_{\pi\in\mathbb{S}} \hat{r}(\pi, \delta_\pi) \;=\; \underset{\delta\in\Delta}{\operatorname{argmin}} \max_{\pi\in\mathbb{S}} \hat{r}(\delta_\pi). \quad (3)$$

Comme démontré dans [8], le classifieur minimax est associé au pire cas possible pour la distribution a priori des proportions des classes. De ce fait, le risque moyen du classifieur minimax est moins sensible à de potentiels changements de proportions.

État de l'art : Dans [4], les auteurs se sont intéressés au problème (3) en proposant un algorithme qui alterne une étape de ré-échantillonnage de la base d'apprentissage avec une étape d'estimation du risque conditionnel par classe. Cette approche ne fonctionne que pour un risque d'erreur différentiable, ce qui n'est pas le cas dans notre contexte (voir la Proposition 1). Dans [5], les auteurs ont proposé une approche minimax basée sur une approximation normale des probabilités d'erreur pour des problèmes de classification supervisée binaires. Plus récemment, [3] propose de construire un ensemble restreint de distributions de probabilités centrées autour de la distribution empirique de la base d'apprentissage, puis de chercher le classifieur minimax par rapport à cet ensemble restreint. Pour que le problème d'optimisation lié au calcul du classifieur minimax soit faisable, la définition de l'ensemble restreint est plutôt délicate et peu pertinente dans notre contexte. Comme évoqué dans [4], les méthodes basées sur le *cost-sensitive learning* cherchent à modifier la matrice de coûts C pour contrebalancer des proportions de classes déséquilibrées. La modification des coûts n'est pas souhaitable dans notre cas car cela modifie les préconisations médicales sur l'impact des traitements. De plus, cette modification est délicate en présence d'au moins trois classes distinctes.

Contributions : Comme introduit précédemment, nous nous intéressons dans cet article au cas où les variables observées sont discrètes ou discrétisées ($\mathcal{X} \subset \mathbb{N}^d$), où nous avons $K \geq 2$ classes à prédire, et où la matrice de coûts $C \in \mathbb{R}_+^{K\times K}$ est quelconque et fixée. Nous montrons que sous ces conditions nous pouvons calculer la règle de décision $\delta_{\hat{\pi}}^B$ qui minimise le risque d'erreurs associé à la base d'apprentissage. Nous proposons ensuite un algorithme permettant d'estimer le classifieur minimax solution de (3) qui est applicable à toute famille de lois discrètes et qui ne nécessite ni de manipuler les distributions de la base d'apprentissage, ni de modifier la matrice de coûts C. La convergence de cet algorithme est établie et nous montrons que nous pouvons borner la vitesse de convergence. Enfin, nous illustrons la robustesse du classifieur minimax obtenu sur une base de données simulées puis sur une base de données réelles.

2 Classifieur minimax

Supposons que chacune des d variables prend un nombre fini de valeurs. Il existe alors un nombre fini T de "profils" possibles permettant de caractériser chaque observation composée de ces d variables : chaque profil correspond à une combinaison des valeurs discrètes prises par ces d variables. L'ensemble $\mathcal{X}$ s'écrit alors $\mathcal{X} = \{x_1, \ldots, x_T\}$ où, pour tout $t \in \{1, \ldots, T\}$, $x_t \in \mathbb{N}^d$. Notons $\mathcal{T} := \{1, \ldots, T\}$, $\mathcal{I}_k := \{i \in \mathcal{I} : Y_i = k\}$ l'ensemble des données d'apprentissage de la classe k et $m_k := \sum_{i\in\mathcal{I}} \mathbb{1}_{\{Y_i=k\}}$ le nombre d'échantillons de la classe k. Soit

$$\hat{\mathbb{P}}\left(\delta_{\hat{\pi}}(X_i) = l \mid Y_i = k\right) \;=\; \frac{1}{m_k} \sum_{i\in\mathcal{I}_k} \mathbb{1}_{\{\delta_{\hat{\pi}}(X_i)=l\}} \quad (4)$$

la probabilité conditionnelle empirique de choisir la classe l lorsque la vraie classe est k. Le risque (1) pour $\delta_{\hat{\pi}} \in \Delta$ s'écrit

$$\hat{r}\left(\delta_{\hat{\pi}}\right) = \sum_{t\in\mathcal{T}} \sum_{k\in\mathcal{Y}} \sum_{l\in\hat{\mathcal{Y}}} \mathbb{1}_{\{\delta_{\hat{\pi}}(x_t)=l\}}\, C_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}, \quad (5)$$

où
$$\hat{p}_{kt} := \frac{1}{m_k} \sum_{i\in\mathcal{I}_L} \mathbb{1}_{\{X_i=x_t\}} \quad (6)$$

est la probabilité empirique d'observer le profil x_t dans la classe k et où $\mathbb{1}_{\{.\}}$ est la fonction indicatrice. On peut ensuite montrer que la règle de décision $\delta_{\hat{\pi}}^B$ définie par

$$\delta_{\hat{\pi}}^B : X_i \mapsto \underset{l\in\hat{\mathcal{Y}}}{\arg\min} \sum_{t\in\mathcal{T}} \sum_{k\in\mathcal{Y}} C_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{X_i=x_t\}} \quad (7)$$

minimise (5). Le classifieur $\delta_{\hat{\pi}}^B$ est de ce fait appelé le classifieur de Bayes empirique. De plus, pour tout $k \in \mathcal{Y}$ et à partir de (1), (5) et (7), $\hat{R}_k\left(\delta_{\hat{\pi}}^B\right)$ peut se réécrire

$$\hat{R}_k\left(\delta_{\hat{\pi}}^B\right) = \sum_{t\in\mathcal{T}} \sum_{l\in\hat{\mathcal{Y}}} C_{kl}\, \hat{p}_{kt}\, \mathbb{1}_{\{\lambda_{lt}=\min_{q\in\hat{\mathcal{Y}}} \lambda_{qt}\}}, \quad (8)$$

où
$$\lambda_{lt} := \sum_{k\in\mathcal{Y}} C_{kl}\, \hat{\pi}_k\, \hat{p}_{kt}, \quad \forall (l,t) \in \hat{\mathcal{Y}} \times \mathcal{T}. \quad (9)$$

Soit $V : \pi \mapsto \hat{r}\left(\pi, \delta_\pi^B\right)$ le risque de Bayes empirique minimum défini sur le simplexe $\mathbb{S}$. La valeur $V(\pi)$ est donc le risque minimum possible pour un classifieur appliqué à la base d'apprentissage lorsque les proportions par classes sont données par π et que les probabilités $\hat{p}_{kt}$ demeurent inchangées. Ainsi, le calcul du classifieur minimax défini dans (3) revient à résoudre le problème d'optimisation

$$\max_\pi \hat{r}\left(\pi, \delta_\pi^B\right) = \max_\pi V(\pi) \quad \text{s.c.} \quad \pi \in \mathbb{S}. \quad (10)$$

Remarque 1. *Les probabilités $\hat{p}_{kt}$ calculées en (5) jouent un rôle majeur ici puisque, avec les proportions $\hat{\pi}$, celles-ci sont suffisantes pour modéliser le risque de Bayes empirique V. Étant donné que, en pratique, nous ne disposons que des observations de la base d'apprentissage pour modéliser V, nous allons pour la suite considérer ces proportions $\hat{p}_{kt}$ fixes. En effet, au vue de la base d'apprentissage, les $\hat{p}_{kt}$ sont les meilleures estimations possibles des probabilités d'occurrences des profils.*

La proposition suivante caractérise la surface $\pi \mapsto V(\pi)$.

Proposition 1. *Considérons les probabilités $\hat{p}_{kt}$ calculées en (5) fixes. La fonction $V(\pi)$ est concave sur le simplexe $\mathbb{S}$ et affine par morceaux avec un nombre fini de faces. De plus, s'il existe $\pi, \pi' \in \mathbb{S}$ et $k \in \mathcal{Y}$ tels que $\hat{R}_k\left(\delta^B_\pi\right) \neq \hat{R}_k\left(\delta^B_{\pi'}\right)$, alors V est non-différentiable sur $\mathbb{S}$.*

Démonstration. La preuve est omise par manque de place. □

Pour estimer les proportions qui maximisent la surface de Bayes V dans le cas général où V est non différentiable (notre algorithme restant tout de même applicable pour le cas particulier où l'hypothèse "pour tout $(\pi, \pi', k) \in \mathbb{S} \times \mathbb{S} \times \mathcal{Y}, \hat{R}_k\left(\delta^B_\pi\right) = \hat{R}_k\left(\delta^B_{\pi'}\right)$" est vérifiée), nous proposons d'utiliser une méthode de sous-gradient projeté (voir détails dans [1]) basée le schéma itératif suivant

$$\pi^{(n+1)} = \mathrm{P}_{\mathbb{S}}\left(\pi^{(n)} + \frac{\gamma_n}{\eta_n} g^{(n)}\right), \tag{11}$$

où pour chaque itération $n \geq 1$, $g^{(n)}$ est un sous-gradient de V au point $\hat{\pi}^{(n)}$, γ_n correspond au pas du sous-gradient, $\eta_n = \max\{1, \|g^{(n)}\|_2\}$, et $\mathrm{P}_{\mathbb{S}}$ dénote la projection sur $\mathbb{S}$.

Lemme 1. *Soit $\pi \in \mathbb{S}$, $\hat{R}\left(\delta^B_\pi\right) := [\hat{R}_1\left(\delta^B_\pi\right), \ldots, \hat{R}_K\left(\delta^B_\pi\right)]$ est un sous-gradient du risque de Bayes empirique V au point π. De plus, $\hat{R}\left(\delta^B_\pi\right)$ ne s'annule jamais.*

Démonstration. La preuve est omise par manque de place. □

Théorème 1. *Considérons $g^{(n)} = \hat{R}\left(\delta^B_{\pi^{(n)}}\right)$ et une suite de pas $(\gamma_n)_{n\geq 1}$ satisfaisant*

$$\inf_{n\geq 1} \gamma_n > 0, \quad \sum_{n=1}^{+\infty} \gamma_n^2 < +\infty, \quad \sum_{n=1}^{+\infty} \gamma_n = +\infty, \tag{12}$$

alors la séquence des itérés suivant le schéma (11) converge vers une solution $\bar{\pi}$ de (10), quelque soit l'initialisation $\pi^{(1)} \in \mathbb{S}$. De plus, l'erreur de convergence maximum jusqu'à l'itération N vérifie

$$\left|\max_{n\leq N}\left\{V\left(\pi^{(n)}\right)\right\} - V\left(\bar{\pi}\right)\right| \leq \frac{\rho^2 + \sum_{n=1}^{N}\gamma_n^2}{2\sum_{n=1}^{N}\gamma_n}, \tag{13}$$

où ρ est une constante satisfaisant $\|\pi^{(1)} - \bar{\pi}\|_2 \leq \rho$.

Démonstration. La preuve est omise par manque de place. □

En considérant $g^{(n)} = \hat{R}\left(\delta^B_{\pi^{(n)}}\right)$, il est à noter, d'après le lemme 1, que la séquence $(\pi^{(n)})_{n\geq 1}$ générée par (11) est infinie. Puisque la borne de droite dans (13) converge vers 0 lorsque $N \to \infty$, nous pouvons choisir $\varepsilon > 0$ assez petit comme critère d'arrêt. En effet, pour tout ε fixé, on calcule $N = N_\varepsilon$ tel que (13) soit borné par ε. L'algorithme peut alors être arrêté à l'itération N_ε. En pratique nous utilisons l'algorithme de Condat [2] pour réaliser la projection sur $\mathbb{S}$. Par la suite, la valeur finale de l'algorithme (11) sera notée $\pi^\star$.

3 Expériences numériques

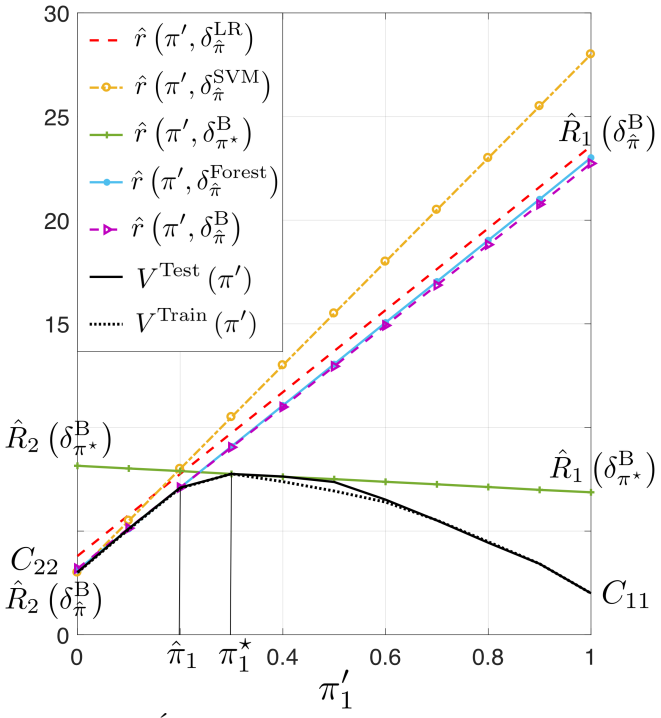

FIGURE 1 – Évolution des risques empiriques associés à différentes règles de décisions lorsque π' diffère des proportions d'apprentissage $\hat{\pi} = [0.2, 0.8]$. Puisque $K = 2$ et $\pi' \in \mathbb{S}$, chaque risque peut se réécrire comme une fonction de π'_1.

Simulations : Nous avons généré une base de données contenant 95000 observations pour laquelle $K = 2$ classes et $d = 3$ variables descriptives. Cette base de données a ensuite été décomposée en une base d'apprentissage $\{(Y_i, X_i), i \in \mathcal{I}\}$ contenant 55000 observations et une base de test $\{(Y'_i, X'_i), i \in \mathcal{I}'\}$ regroupant les observations restantantes. Pour chaque observation $i \in \mathcal{I} \cup \mathcal{I}'$, $Y_i \sim \mathrm{Cat}(K, \hat{\pi})$ avec $\hat{\pi} = [0.2, 0.8]$, où $\mathrm{Cat}(K, \hat{\pi})$ désigne la loi catégorielle à valeurs dans $\{1, \ldots, K\}$ telle que la probabilité d'obtenir $k \in \{1, \ldots, K\}$ est $\hat{\pi}_k$. Pour tout $j \in \{1, \ldots, d\}$, les variables aléatoire X_{ij} ont été générées de la façon suivante : $X_{ij} = \mathbb{1}_{\{Y_i=1\}} U_i + \mathbb{1}_{\{Y_i=2\}} V_i$, où $U_i \sim \mathcal{N}(\mu_{1j}, \sigma^2_{1j})$, $V_i \sim \mathcal{N}(\mu_{2j}, \sigma^2_{2j})$, et où μ et σ sont définies par :

$$\mu = \begin{bmatrix} 37.5 & 6.5 & 19 \\ 39 & 7 & 20 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 1 & 1.5 & 1.2 \\ 2 & 0.8 & 2 \end{bmatrix}.$$

Nous avons ensuite discrétiser chaque variable descriptive en six catégories. Enfin, la matrice de coûts C est

$$C = \begin{bmatrix} 2 & 28 \\ 20 & 3 \end{bmatrix}.$$

La Figure 1 illustre la comparaison des risques d'erreurs associés au test de Bayes $\delta^{\mathrm{B}}_{\hat{\pi}}$ (7), à la régression logistique $\delta^{\mathrm{LR}}_{\hat{\pi}}$, aux forêts aléatoires $\delta^{\mathrm{Forest}}_{\hat{\pi}}$, aux SVM $\delta^{\mathrm{SVM}}_{\hat{\pi}}$, et au classifieur minimax $\delta^{B}_{\pi^\star}$. Ces résultats sont une moyenne des risques calculés sur 200 folds pour lesquels 4000 observations ont été aléatoirement sélectionnées dans l'ensemble $\mathcal{I}'$ de sorte que notre sous échantillon respecte les proportions π'. Le risque associé à chaque classifieur évolue linéairement lorsque π' diffère des proportions d'apprentissage $\hat{\pi} = [0.2, 0.8]$. Cela implique que les classifieurs non-minimax voient leurs risques très fortement augmentés lorsque π'_1 augmente. Après $N = 100$ itérations, l'algorithme (11) a convergé vers $\pi^\star = [0.29, 0.71]$. Cette figure illustre la robustesse de $\delta^{B}_{\pi^\star}$ lorsque les proportions π' diffèrent de $\hat{\pi}$. Le risque d'erreurs reste assez stable sur le simplexe $\mathbb{S}$ et on remarque que $\max_{k\in\mathcal{Y}} \hat{R}_k(\delta^{B}_{\hat{\pi}})$ est minimisé.

Remarque 2. *Pour les différentes proportions π' considérées sur la Figure 1, nous avons tracé la surface de Bayes $V^{\mathrm{Train}}(\pi')$ estimée à partir de la base d'apprentissage $\mathcal{I}$ (surface sur laquelle l'algorithme du sous-gradient (11) est appliqué), puis la surface de Bayes $V^{\mathrm{Test}}(\pi')$ estimée à partir des observations du sous-échantillon test (vérifiant les proportion π') à partir de la règle de décision $\delta^{\mathrm{B}}_{\pi'}$ (7), tout en gardant les probabilités $\hat{p}_{kt}$ fixes. Sur la figure 1, ces deux surfaces ne sont pas tout à fait identiques. Ceci témoigne de l'erreur de généralisation (plutôt faible) entre la phase d'apprentissage et la phase de test.*

TABLE 1 – Abalone dataset : le risque de chaque méthode est résumée par [moyenne $\pm$ écart-type].

APPRENTISSAGE	$\hat{\pi} = [0.02, 0.64, 0.28, 0.05, 0.01]$
$\hat{r}\left(\hat{\pi}, \delta^{\mathrm{LR}}_{\hat{\pi}}\right)$	0.4178 ± 0.0057
$\hat{r}\left(\hat{\pi}, \delta^{\mathrm{Forest}}_{\hat{\pi}}\right)$	0.4164 ± 0.0059
$\hat{r}\left(\hat{\pi}, \delta^{B}_{\hat{\pi}}\right)$	0.3914 ± 0.0054
ESTIMATION	$\pi^\star = [0.03, 0.48, 0.02, 0.10, 0.36]]$
$\hat{r}\left(\pi^\star, \delta^{B}_{\pi^\star}\right)$	0.9051 ± 0.0408
$\hat{r}\left(\hat{\pi}, \delta^{B}_{\pi^\star}\right)$	0.9158 ± 0.1025
TEST	$\pi' = [0.24, 0.16, 0.19, 0.29, 0.11]$
$\hat{r}\left(\pi', \delta^{\mathrm{LR}}_{\hat{\pi}}\right)$	1.6218 ± 0.1374
$\hat{r}\left(\pi', \delta^{\mathrm{Forest}}_{\hat{\pi}}\right)$	1.6323 ± 0.1394
$\hat{r}\left(\pi', \delta^{B}_{\hat{\pi}}\right)$	1.6823 ± 0.1497
$\hat{r}\left(\pi', \delta^{B}_{\pi^\star}\right)$	1.2717 ± 0.1374

Base de données Abalone : La base de données Abalone [7] contient les mesures physiques (8 variables : 1 catégorielle et 7 numériques) de 4177 abalones de Tasmanie. Les 7 variables numériques ont été discrétisées en 3 catégories. À partir de ces 8 variables, l'objectif est de prédire l'âge de chaque abalone s'étendant de 1 an à 29 ans. Nous avons décidé de considérer $K = 5$ classes $\{A_1, A_2, A_3, A_4, A_5\}$ représentant le tranches d'âges $\{[1, 4], [5, 10], [11, 15], [16, 20], [\geq 21]\}$ dont les proportions par classe sont $\hat{\pi} = [0.02, 0.64, 0.28, 0.05, 0.01]$. Ces tranches d'âge sont non-balancées. Pour cette expérience nous avons considéré une matrice de coûts quadratiques : pour tout $(k, l) \in \mathcal{Y} \times \hat{\mathcal{Y}}$, $C_{kl} = (k-l)^2$, de sorte que plus la tranche d'âge prédite est éloignée de la tranche d'âge réelle, plus le coût de cette erreur sera important. Comme illustré dans le tableau 1, le test minimax présente le risque d'erreur le plus élevé sur la base d'apprentissage. Ce risque reste stable que les proportions soient $\pi^\star$ ou $\hat{\pi}$. Dans le cas où π' diffère considérablement de $\hat{\pi}$ dans la base de test, le risque empirique des classifieurs non-minimax change significativement alors que celui du classifieur minimax $\delta^{B}_{\pi^\star}$ varie moins.

4 Conclusion

Cet article propose un algorithme de sous-gradient projeté permettant d'estimer le classifieur minimax à partir de données d'apprentissage discrètes ou discrétisées. La robustesse de ce classifieur est illustrée sur deux expériences numériques. Nos prochains travaux s'intéresseront à l'erreur de généralisation.

Références

[1] Ya. I. Alber, A. N. Iusem, and M. V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a hilbert space. *Mathematical Programming*, 81(1) :23–35, Mar 1998.

[2] Laurent Condat. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1) :575–585, 2016.

[3] F. Farnia and D. Tse. A minimax approach to supervised learning. In *Advances in NIPS 29*, pages 4240–4248. 2016.

[4] A. Guerrero-Curieses, R. Alaiz-Rodriguez, and J. Cid-Sueiro. A fixed-point algorithm to minimax learning with neural networks. *IEEE Transactions on Systems, Man and Cybernetics, Part C, Applications and Reviews*, 34(4) :383–392, Nov 2004.

[5] Huang Kaizhu, Yang Haiqin, King Irwin, R. Lyu Michael, and Laiwan Chan. The minimum error minimax probability machine. *Journal of Machine Learning Research*, page 1253–1286, 2004.

[6] Emanuela Martinuzzi and Susana Barbosa et al. Stratification and prediction of remission in first-episode psychosis patients : the optimise cohort study. *Translational Psychiatry*, 9, 01 2019.

[7] Warwick J. Nash et al. The population biology of abalone (haliotis species) in tasmania. 1, blacklip abalone (h. rubra) from the north coast and the islands of bass strait. *Sea Fisheries Division, Technical Report*, (48), 1994.

[8] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag New York, 2nd edition, 1994.

[9] Ying Yang and Geoffrey I. Webb. Discretization for naive-bayes learning : managing discretization bias and variance. *Machine Learning*, 74(1) :39–74, Jan 2009.