Apprentissage d'un classifieur minimax pour données discrètes

Lionel Fillatre [1] & Cyprien Gilet [1] & Susana Barbosa [2] *

[1] *Université Côte d'Azur, CNRS, I3S, France*
Email: fillatre@i3s.unice.fr, gilet@i3s.unice.fr
[2] *Université Côte d'Azur, CNRS, IPMC, France*
Email: sudocarmo@gmail.com

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, 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 maximisent ce risque de Bayes. Le classifieur minimax obtenu minimise le risque conditionnel maximum.

Mots-clés. Apprentissage supervisé, variables discrètes, classifieur minimax.

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 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.

Keywords. Supervised learning, discrete features, minimax classifier.

1 Introduction

Contexte : Cet article s'intéresse au problème de classification supervisée qui consiste à minimiser le risque empirique moyen d'erreurs de classification à partir d'un ensemble fini d'observations étiquetées. Définissons $K \geq 2$ le nombre de classes, $\mathcal{Y} := \{1, \ldots, K\}$

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

l'ensemble des classes observées, $\mathcal{X}$ l'espace sur lequel l'ensemble des variables observées sont définies, et n le nombre d'observations dans la base d'apprentissage. On note Y_i la variable aléatoire caractérisant la classe de l'observation i, et $X_{i:} = [X_{i1}, \ldots, X_{id}]$ le vecteur aléatoire regroupant l'ensemble des d variables descriptives associées à l'observation i. Définissons $\Delta = \{\delta : \mathcal{X} \to \mathcal{Y}\}$ l'ensemble des classifieurs et considérons une règle de décision $\delta \in \Delta$. Comme décrit dans Poor (1994), le risque empirique $\hat{r}(\delta)$ de δ, associé à la fonction de perte $L_{0\text{-}1}$, s'écrit

$$\hat{r}(\delta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\delta(X_{i:}) \neq Y_i\}} = \sum_{k \in \mathcal{Y}} \hat{\pi}_k \hat{R}_k(\delta), \text{ avec } \hat{R}_k(\delta) = \frac{1}{n} \sum_{i:\, Y_i = k} \mathbb{1}_{\{\delta(X_{i:}) \neq k\}}, \qquad (1)$$

$\hat{\pi}_k = \hat{n}_k/n$ correspond à la proportion d'observations appartenant à la classe k, $\hat{n}_k = \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = k\}}$ est le nombre d'observations de la classe k et $\mathbb{1}_{\{\cdot\}}$ est la fonction indicatrice.

Introduction du classifieur minimax : Notons $\delta_{\hat{\pi}}$ le classifieur $\delta \in \Delta$ calibré à partir des observations $\mathcal{D}_n = (Y_i, X_{i:})_{i=1,\ldots,n}$ de la base d'apprentissage dont les proportions par classe sont $\hat{\pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_K]$. Ce classifieur est ensuite utilisé pour prédire la classe de nouvelles observations de test $\mathcal{D}'_m = (Y_i, X_{i:})_{i=n+1,\ldots,n+m}$. Supposons que l'ensemble des données de test vérifie les proportions $\hat{\pi}' = [\hat{\pi}'_1, \ldots, \hat{\pi}'_K]$, le risque d'erreurs associé au classifieur $\delta_{\hat{\pi}}$ et aux proportions $\hat{\pi}'$ est alors noté et défini par $\hat{r}(\hat{\pi}', \delta_{\hat{\pi}}) = \sum_{k \in \mathcal{Y}} \hat{\pi}'_k \hat{R}_k(\delta_{\hat{\pi}})$. Comme illustré sur la figure 1, ce risque évolue linéairement lorsque les proportions $\hat{\pi}'$ diffèrent de $\hat{\pi}$. Ceci peut donc poser problème lorsque les proportions de la base d'apprentissage sont incertaines, ou que les données sont non-balancées et que l'on cherche à les rééquilibrer. Comme décrit dans Poor (1994), une solution pour rendre une règle de décision robuste à de possibles différences de proportions entre $\mathcal{D}_n$ et $\mathcal{D}'_m$ consiste à minimiser $\max_{k \in \mathcal{Y}} \hat{R}_k(\delta_{\hat{\pi}})$. Ce problème minimax se réécrit

$$\bar{\delta} \;=\; \underset{\delta \in \Delta}{\operatorname{argmin}} \max_{\hat{\pi} \in \mathbb{S}} \hat{r}(\hat{\pi}, \delta_{\hat{\pi}}) \;=\; \underset{\delta \in \Delta}{\operatorname{argmin}} \max_{\hat{\pi} \in \mathbb{S}} \hat{r}(\delta_{\hat{\pi}}), \qquad (2)$$

où $\mathbb{S}$ correspond au simplexe probabiliste de dimension K.

État de l'art : Guerrero-Curieses et al. (2004) se sont intéressés au problème (2) pour des fonctions de perte SSB (Strict Sense Bayesian) et ont proposé un algorithme qui consiste à alterner une étape de ré-échantillonnage de la base d'apprentissage avec une étape d'estimation du risque conditionnel par classe. Kaizhu et al. (2004) ont eux également proposé une approche minimax intéressante (Minimum Error Minimax Probability Machine) adaptée aux problèmes de classification supervisée binaires ($K = 2$). Plus récemment, Farnia and Tse (2016) ont proposé une approche qui cherche à calibrer une règle de décision en estimant la distribution des données qui minimise le pire cas d'erreurs de classifications sur un ensemble de distributions centrées sur la distribution empirique. Bien que leur approche ne soit pas directement liée aux problèmes de proportions, elle reste assez proche du problème (2).

Contributions : Nous nous intéressons dans cet article au cas où les variables observées sont discrètes ($\mathcal{X} \subset \mathbb{N}^d$), et où nous avons $K \geq 2$ classes à prédire. Nous montrons que, sous ces conditions, nous pouvons calculer la règle de décision $\delta^B_{\hat{\pi}}$ qui minimise le risque d'erreurs associé à la base d'apprentissage $\mathcal{D}_n$. Nous proposons ensuite un algorithme permettant d'estimer le classifieur minimax solution de (2) sans avoir besoin de ré-échantillonner la base d'apprentissage à chaque itération. 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 estimé par notre algorithme sur une base de données simulées puis sur une base de données réelles.

2 Classifieur minimax pour variables discrètes

Supposons que l'ensemble des d variables descriptives soient discrètes ou discrétisées avec un nombre fini de valeurs. Il existe un nombre fini T de "profiles" possibles permettant de caractériser chaque combinaison des d variables. Nous avons donc $\mathcal{X} = \{x_1, \ldots, x_T\}$ où $x_t \in \mathbb{N}^d$. Pour tout classifieur $\delta \in \Delta$, le risque (1) se réécrit alors

$$\hat{r}\left(\delta\right) = 1 - \sum_{t \in \mathcal{T}} \sum_{k \in \mathcal{Y}} \mathbb{1}_{\{\delta(x_t)=k\}}\, \hat{\pi}_k\, \hat{p}_{kt}, \quad \text{avec} \quad \hat{p}_{kt} := \frac{1}{\hat{n}_k} \sum_{i:\, Y_i = k} \mathbb{1}_{\{X_{i:}=x_t\}}, \tag{3}$$

où $\mathcal{T} := \{1, \ldots, T\}$. La règle de décision $\delta^B_{\hat{\pi}}$ définie par

$$\delta^B_{\hat{\pi}} : X_{i:} \mapsto \underset{k \in \{1,\ldots,K\}}{\arg\max} \sum_{t \in \mathcal{T}} \hat{\pi}_k\, \hat{p}_{kt}\, \mathbb{1}_{\{X_{i:}=x_t\}} \tag{4}$$

minimise (3). Le classifieur $\delta^B_{\hat{\pi}}$ est appelé le classifieur de Bayes empirique. Soit $\hat{V} : \hat{\pi} \mapsto \hat{r}\left(\hat{\pi}, \delta^B_{\hat{\pi}}\right)$ le risque de Bayes empirique minimum défini sur le simplexe $\mathbb{S}$. Le calcul du classifieur minimax (2) revient alors à résoudre le problème d'optimisation :

$$\max_{\hat{\pi}} \hat{r}\left(\hat{\pi}, \delta^B_{\hat{\pi}}\right) = \max_{\hat{\pi}} \hat{V}(\hat{\pi}) \quad \text{s.t.} \quad \hat{\pi} \in \mathbb{S}. \tag{5}$$

Proposition 1. *Supposons que les probabilités estimées $\hat{p}_{kt}$ dans (3) sont fixes. La fonction $\hat{V}(\hat{\pi})$ est concave sur le simplexe $\mathbb{S}$ et affine par morceaux avec un nombre fini de faces. De plus, s'il existe $\hat{\pi} \in \mathbb{S}$ tel que $\hat{V}(\hat{\pi}) > 0$, alors $\hat{V}$ est non-différentiable sur $\mathbb{S}$.*

Nous supposons dans la suite que $\hat{V}(\hat{\pi})$ est non-uniformément nulle sur $\mathbb{S}$. D'après la proposition 1, une méthode du gradient projeté n'est pas adapté ici pour traiter le problème d'optimisation non-différentiable (5). Nous proposons d'utiliser une méthode du sous-gradient projeté (voir détails dans Alber et al. (1998)) suivant le schéma itératif

$$\hat{\pi}^{(n+1)} = \mathrm{P}_{\mathbb{S}}\left(\hat{\pi}^{(n)} + \frac{\gamma_n}{\eta_n}\, g^{(n)}\right), \tag{6}$$

où pour chaque itération $n \geq 1$, $g^{(n)}$ est un sous-gradient de $\hat{V}$ 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 $\hat{\pi} \in \mathbb{S}$. Le vecteur $\hat{R}_{\hat{\pi}}$ défini par*

$$\hat{R}_{\hat{\pi}} := \left[\hat{R}_1\left(\delta^B_{\hat{\pi}}\right), \ldots, \hat{R}_K\left(\delta^B_{\hat{\pi}}\right)\right] \tag{7}$$

est un sous-gradient de $\hat{V}(\hat{\pi})$ au point $\hat{\pi}$. De plus, $\hat{R}_{\hat{\pi}}$ ne s'annule jamais.

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

$$\inf_{n\geq 1} \gamma_n > 0, \quad \sum_{n=1}^{+\infty} \gamma_n^2 < +\infty, \quad \sum_{n=1}^{+\infty} \gamma_n = +\infty, \tag{8}$$

alors la séquence des itérés (6) converge vers une solution $\bar{\pi}$ de (5), quelque soit l'initialisation $\hat{\pi}^{(1)} \in \mathbb{S}$. De plus, l'erreur de convergence maximum jusqu'à l'itération N vérifie

$$\left|\max_{n\leq N}\left\{\hat{r}\left(\hat{\pi}^{(n)}, \delta^B_{\hat{\pi}^{(n)}}\right)\right\} - \bar{r}\right| \;\leq\; \frac{\rho^2 + \sum_{n=1}^N \gamma_n^2}{2\sum_{n=1}^N \gamma_n}, \tag{9}$$

où ρ est une constante satisfaisant $\|\hat{\pi}^{(1)} - \bar{\pi}\|_2 \leq \rho$ et $\bar{r} = \max_{\hat{\pi}\in\mathbb{S}} \hat{V}(\hat{\pi})$.

Il est à noter que la séquence $(\hat{\pi}^{(n)})_{n\geq 1}$ générée par (6) est infinie. Puisque la borne de droite dans (9) converge vers 0 lorsque $N \to \infty$, nous pouvons choisir $\varepsilon > 0$ tel que (9) soit bornée par ε à partir d'un certain $N = N_\varepsilon$ dépendant de ε. L'algorithme peut alors être arrêté à l'itération N_ε. En pratique, nous utilisons l'algorithme de Condat (2016) pour réaliser la projection sur $\mathbb{S}$. La valeur finale de l'algorithme est notée $\hat{\pi}^\star$.

3 Expériences numériques

Simulations : Nous avons généré des données pour laquelle $K = 2$ classes et $d = 3$ variables. Pour chaque observation $i = 1, \ldots, n+m$, $Y_i \sim \mathrm{Cat}(K, \pi)$ avec $\pi = [0.2, 0.8]$, où $\mathrm{Cat}(K, \pi)$ désigne la loi catégorielle à valeurs dans $\{1, \ldots, K\}$ telle que la probabilité d'obtenir $k \in \{1, \ldots, K\}$ est π_k. Pour tout $j \in \{1, \ldots, d\}$, nous générons les variables aléatoires X_{ij} 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 :

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

Les proportions de la base d'apprentissage sont $\hat{\pi} = [0.201, 0.799]$, ce qui est très proche de π. Tous les classifieurs sont appris, une unique fois, sur cette base d'apprentissage. La Figure 1 compare les risques d'erreurs associés au test de Bayes $\delta^{\mathrm{B}}_{\hat{\pi}}$ (4), à la régression logistique $\delta^{\mathrm{LR}}_{\hat{\pi}}$, à la forêt aléatoire $\delta^{\mathrm{Forest}}_{\hat{\pi}}$, au SVM $\delta^{\mathrm{SVM}}_{\hat{\pi}}$, et au classifieur minimax $\delta^B_{\hat{\pi}^\star}$. Le

risque associé à chaque classifieur évolue linéairement lorsque $\hat{\pi}'$ diffère des proportions d'apprentissage $\hat{\pi}$. Les résultats de test sont une moyenne des risques calculés sur 200 segments indépendants. Chaque segment de test contient 1000 observations qui ont été générées aléatoirement de sorte que tous les segments ont les mêmes proportions $\hat{\pi}'$. Les proportions $\hat{\pi}'$ sont choisies de sorte à échantillonner de façon uniforme le simplexe $\mathbb{S}$. Tous les classifieurs comparés sont évalués sur chaque segment de test. Les classifieurs $\delta^{\mathrm{LR}}_{\hat{\pi}}$ et $\delta^{\mathrm{SVM}}_{\hat{\pi}}$ utilisent les données brutes tandis que $\delta^{\mathrm{B}}_{\hat{\pi}}$, $\delta^{\mathrm{Forest}}_{\hat{\pi}}$ et $\delta^{B}_{\hat{\pi}^\star}$ utilisent des données discrétisées. Après $N = 100$ itérations, l'algorithme (6) converge vers $\hat{\pi}^\star = [0.42, 0.58]$. La Figure 1 illustre la robustesse de $\delta^{B}_{\hat{\pi}^\star}$ lorsque les proportions $\hat{\pi}'$ different de $\hat{\pi}$.

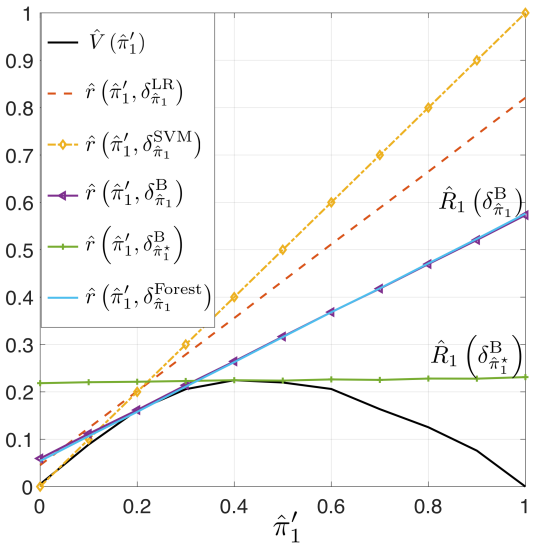

Figure 1: Évolution des risques empiriques associés aux règles de décisions régression logistique $\delta^{\mathrm{LR}}_{\hat{\pi}}$, SVM $\delta^{\mathrm{SVM}}_{\hat{\pi}}$, forêt aléatoire $\delta^{\mathrm{Forest}}_{\hat{\pi}}$, test de Bayes $\delta^{\mathrm{B}}_{\hat{\pi}}$ et le classifieur minimax $\delta^{B}_{\hat{\pi}^\star}$ lorsque $\hat{\pi}'$ diffère des proportions d'apprentissage $\hat{\pi}$. Puisque $K = 2$ et $\hat{\pi}' \in \mathbb{S}$, $\hat{r}$ peut se réécrire comme fonction de $\hat{\pi}'_1$.

Apprentissage	$\hat{\pi} = [0.02, 0.64, 0.28, 0.05, 0.01]$
$\hat{r}\left(\hat{\pi}, \delta^{\mathrm{LR}}_{\hat{\pi}}\right)$	28.32 ± 0.44 %
$\hat{r}\left(\hat{\pi}, \delta^{\mathrm{Forest}}_{\hat{\pi}}\right)$	30.37 ± 0.30 %
$\hat{r}\left(\hat{\pi}, \delta^{B}_{\hat{\pi}}\right)$	30.16 ± 0.31 %
Estimation	$\hat{\pi}^\star = [0.02, 0.26, 0.27, 0.27, 0.18]$
$\hat{r}\left(\hat{\pi}^\star, \delta^{B}_{\hat{\pi}^\star}\right)$	48.43 ± 1.12 %
$\hat{r}\left(\hat{\pi}, \delta^{B}_{\hat{\pi}^\star}\right)$	48.59 ± 2.80 %
Test	$\hat{\pi}' = [0.18, 0.05, 0.57, 0.11, 0.09]$
$\hat{r}\left(\hat{\pi}', \delta^{\mathrm{LR}}_{\hat{\pi}}\right)$	57.12 ± 4.08 %
$\hat{r}\left(\hat{\pi}', \delta^{\mathrm{Forest}}_{\hat{\pi}}\right)$	95.08 ± 0.01 %
$\hat{r}\left(\hat{\pi}', \delta^{B}_{\hat{\pi}}\right)$	71.64 ± 3.92 %
$\hat{r}\left(\hat{\pi}', \delta^{B}_{\hat{\pi}^\star}\right)$	55.83 ± 4.08 %

Table 1: Abalone dataset : Comparaison des risques obtenus par la régression logistique $\delta^{\mathrm{LR}}_{\hat{\pi}}$, la forêt aléatoire $\delta^{\mathrm{Forest}}_{\hat{\pi}}$ et le test de Bayes $\delta^{\mathrm{B}}_{\hat{\pi}}$ (4) qui ont été calibrés à partir des proportions $\hat{\pi}$. $\delta^{\mathrm{B}}_{\hat{\pi}^\star}$ correspond au classifieur minimax estimé après $N = 100$ itérations. Les résultats sont présentés comme suit : [moyenne $\pm$ écart-type].

Base de données Abalone : La base de données Abalone (Nash et al. (1994)) contient les mesures physiques (8 variables : 1 catégorielle et 7 numériques) de 4177 abalones. 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 $\{C_1, C_2, C_3, C_4, C_5\}$ 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. Comme illustré dans le tableau 1, dans le cas où $\hat{\pi}'$ diffère considérablement de $\hat{\pi}$, le risque empirique est plus robuste pour le classifieur minimax $\delta^B_{\hat{\pi}^\star}$. Les risques du tableau 1 sont une moyenne sur 80 essais. Pour chaque essai, nous avons aléatoirement sélectionné 75% de l'ensemble complet des abalones pour construire la base d'apprentissage en respectant les proportions $\hat{\pi}$. Ensuite, nous avons construit un sous-échantillon test en sélectionnant aléatoirement des abalones parmi les 25% restants de sorte que les proportions $\hat{\pi}'$ soient respectées où $\hat{\pi}'$ a été choisi de façon arbitraire. Les 7 variables numériques ont été discrétisées en 3 catégories.

4 Conclusion

Cet article propose un algorithme de sous-gradient projeté permettant d'estimer le classifieur minimax à partir de données d'apprentissage lorsque les variables sont discrètes. La convergence de cet algorithme est démontrée et illustrée sur des données réelles.

References

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

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

Farnia, F. and Tse, D. (2016). A minimax approach to supervised learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4240–4248. Curran Associates, Inc.

Guerrero-Curieses, A., Alaiz-Rodriguez, R., and Cid-Sueiro, J. (2004). 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.

Kaizhu, H., Haiqin, Y., Irwin, K., Michael, R. L., and Chan, L. (2004). The minimum error minimax probability machine. *Journal of Machine Learning Research*, page 1253–1286.

Nash et al., W. J. (1994). 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).

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