Classification binaire sur les régions extrêmes

Hamid Jalalzai[1], Stéphan Clémençon[1] & Anne Sabourin[1]

[1] *LTCI, Télécom ParisTech,*
75013, Paris, France
prenom.nom@telecom-paristech.fr

Résumé. Parmi les diverses applications relatives à la détection d'anomalies les observations extrêmes jouent un rôle essentiel car les anomalies correspondent souvent à de grandes observations. La question clé est alors de faire la distinction entre les grandes observations issue la classe normale et celles provenant de la classe des anomalies. C'est un problème de classification binaire dans les régions extrêmes. Cependant, les observations extrêmes contribuent de manière négligeable à l'erreur empirique, au vue de leur rareté. Par conséquent, les minimiseurs du risque empirique ne bénéficient pas de garanties appropriées dans les régions extrêmes. Nous proposons un cadre général pour la classification des valeurs extrêmes. Plus précisément, dans le cadre d'hypothèses de distributions à queues lourdes non paramétriques, nous introduisons une version asymptotique du risque d'erreur mesurant la performance prédictive dans les régions extrêmes. Nous montrons que les minimiseurs d'une version empirique de ce risque ont une bonne capacité de généralisation, au moyen d'inégalités de concentration dans les régions à faible probabilité. Des expériences numériques illustrent la pertinence de l'approche développée.

Mots-clés. Apprentissage supervisé, Classification, Théorie des Valeurs Extrêmes

Abstract.

In a wide variety of applications related to anomaly detection, extreme observations play a key role because anomalies often correspond to large observations. The key issue is then to distinguish between large observation from the normal class and large observations from the anomaly class. This task can thus be formulated as a binary classification problem in extreme regions. However, extreme observations generally contribute in a negligible manner to the (empirical) error, simply because of their rarity. As a consequence, empirical risk minimizers generally perform very poorly in extreme regions. This paper develops a general framework for classification of extreme values. Precisely, under non-parametric heavy-tail assumptions, we propose a natural and asymptotic notion of risk accounting for predictive performance in extreme regions. We prove that minimizers of an empirical version of this dedicated risk lead to classification rules with good generalization capacity, by means of maximal deviation inequalities in low probability regions. Numerical experiments illustrate the relevance of the approach developed.

Keywords. Supervised Learning, Classification, Extreme Value theory

1 Introduction

Nous présentons ici les principaux résultat de Jalalzai et al. (2018). Dans un contexte de classification supervisée, considérons une paire aléatoire $(X, Y) \in \mathbb{R}^d \times \{-1, 1\}$ de loi jointe P inconnue. Le but de l'apprentissage est alors de construire une fonction de prédiction $g : \mathbb{R}^d \to \{-1, 1\}$ minimisant le coût $L_P(g) = \mathbb{P}\{Y \neq g(X)\}$ à l'aide de données $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ composées de n copies i.i.d de (X, Y). L'approche par Minimisation du Risque Empirique (MRE sous forme abrégée, voir par exemple Devroye et al. (1996)) consiste à considérer l'estimateur g_n, une solution du problème de minimisation $\min_{g \in \mathcal{G}} \widehat{L}_n(g)$, où $\widehat{L}_n(g) = (1/n) \sum_{i=1}^n \mathbb{1}\{Y_i \neq g(X_i)\}$ est le risque empirique. Ici $\mathbb{1}\{\mathcal{E}\}$ désigne la fonction indicatrice associée à l'évènement $\mathcal{E}$. On notera par la suite g^* le classifieur de Bayes, $g^*(x) = 2\,\mathbb{1}\{\eta(x) \geq 1/2\} - 1$, où $\eta(X) = \mathbb{P}\{Y = 1 \mid X\}$ est la fonction de régression.

Une observation X est dite extrême si $\|X\|$ dépasse un seuil $t > 0$ relativement grand. Ces observations étant rares, elles sont de fait sous-représentées parmi les données d'apprentissage $\mathcal{D}_n$. Par conséquent rien ne garantit que la MRE produise un classifieur optimal sur les régions extrêmes $\{x : \ \|x\| > t\}$. Nous nous intéressons donc au risque de classification au dessus d'un seuil t et à sa limite supérieure en t, respectivement

$$L_t(g) = L_{P_t}(g) = \mathbb{P}\{Y \neq g(X) \mid \|X\| > t\}, \qquad L_\infty(g) = \limsup_{t \to \infty} L_t(g), \tag{1}$$

P_t étant la distribution conditionnelle de (X, Y) sachant $\|X\| > t$. Nous appellerons L_∞ le *risque asymptotique* et noterons $L_\infty^* = \inf_{g \text{ mesurable}} L_\infty(g)$. On montre facilement que g^* est un minimiseur de L_t. Ainsi, pour tout classifieur $g, L_t(g) \geq L_t(g^*)$, en passant à la limite, on obtient que g^* minimise L_∞. Donc $L_\infty^* = L_\infty(g^*)$, le classifieur de Bayes est donc aussi optimal pour le risque L_∞. On va voir que sous des hypothèses appropriées de variation régulière, précisées au paragraphe suivant, il existe un autre minimiseur de L_∞, celui-ci construit à partir de la loi limite de (X, Y) conditionnellement à $\|X\| > t$.

2 Variation régulière et meilleur classifieur asymptotique

Un vecteur aléatoire $X = (X^{(1)}, \ldots, X^{(d)}) \in \mathbb{R}_+^d$est dit à variation régulière d'indice $\alpha > 0$ si il existe une mesure de Radon μ définie sur $E = [0, \infty]^d \backslash \{0\}$, appelée *mesure exponentielle*, telle que pour tout borélien $A \subset E$ pour lequel $0 \notin \partial A$ et $\mu(\partial A) = 0$,

$$t^\alpha \mathbb{P}\{X \in tA\} \xrightarrow[t \to \infty]{} \mu(A),$$

voir par exempleResnick (1987, 2007) pour une introduction. Pour des exemples récents de travaux utilisant la théorie des valeurs extrêmes dans un cadre d'apprentissage statistique,

citons par exemple Goix et al. (2016); Carpentier and Valko (2014); Roos et al. (2006); Ohannessian and Dahleh (2012); Brownlees et al. (2015) ou Mendelson (2018).

La mesure exponentielle vérifie alors $\mu(tC) = t^{-\alpha}\mu(C)$ pour tout $t > 0$ quelque soit le borélien $C \subset E$. Cette propriété d'homogénéité se traduit par une forme produit en coordonnées polaires. Pour tout $x = (x_1, \ldots, x_d) \in \mathbb{R}^d_+$, soit $R(x) = \|x\|$ et $\Theta(x) = \frac{1}{R(x)}x \in S$, où S est la restriction de la sphère (associée à $\|\cdot\|$) au quadrant positif de $\mathbb{R}^d$. Soit Φ la *mesure angulaire* définie sur S comme $\Phi(B) = \mu\{r\theta \ : \ \theta \in B, r \geq 1\}, \quad B \subset S$, mesurable. Cette mesure est finie et l'on a la convergence en loi suivante :

$$\begin{aligned}\lim_{t\to\infty} \mathbb{P}\left\{R(X)/t > r, \Theta(X) \in B \mid R(X) > t\right\} &= c\,\mu\{x \ : \ R(x) > r, \Theta(x) \in B\} \\ &= c\,r^{-\alpha}\,\Phi(B),\end{aligned}$$

où $c = \mu\{x : R(x) > 1\}^{-1} = \Phi(S)^{-1}$ est une constante de normalisation.

Nous désignerons par F_+ et F_- les distributions conditionnelles de X sachant que $Y = +/-1$, nous travaillerons sous l'hypothèse de variation régulière suivante :

Hypothèse 1 *Pour tout $\sigma \in \{-, +\}$, $\forall A \subset [0,\infty]^d \setminus \{0\}$ un ensemble mesurable tel que $0 \notin \partial A$ et $\mu(\partial A) \neq 0$,*

$$t\mathbb{P}\left\{t^{-1}X \in A \,\middle|\, Y = \sigma\, 1\right\} \xrightarrow[t\to\infty]{} \mu_\sigma(A), \qquad \sigma \in \{-,+\},$$

La distribution marginale de X, $F = pF_+ + (1-p)F_-$, avec $p = \mathbb{P}\{Y = +1\} > 0$, est alors également à variation régulière d'indice 1. En effet : $t\mathbb{P}\{t^{-1}X \in A\} \xrightarrow[t\to\infty]{} \mu(A) := p\mu_+(A) + (1-p)\mu_-(A)$. De plus, en notant $\Omega = \{x \in \mathbb{R}^d_+ \ : \ \|x\| \leq 1\}$, on a

$$p_t = \mathbb{P}\{Y = +1 \mid \|X\| > t\} \xrightarrow[t\to\infty]{} p\frac{\mu_+(\Omega^c)}{\mu(\Omega^c)} = p\frac{\Phi_+(S)}{\Phi(S)} \stackrel{\text{def}}{=} p_\infty. \tag{2}$$

Remarque 1 *L'hypothèse 1 revient à supposer que chaque classe est à variation régulière de même indice $\alpha^+ = \alpha^- = 1$. Lorsque les deux classes sont à variation régulière d'indices $\alpha^+ \neq \alpha^-$, la classe d'indice le plus faible devient largement majoritaire au delà de seuils élevés. Asymptotiquement le meilleur classifieur est alors le classifieur déterministe prédisant systématiquement la classe en question et le problème est facile à traiter. En revanche il arrive fréquemment que les indices de queues des différentes composantes $X^{(j)}$ des prédicteurs soient différents. Une solution classique consiste à travailler avec des variables standardisées $V^{(j)} = 1/(1 - F_j(X^{(j)})) \in [1,\infty]$ et $V = (V^{(1)}, \ldots, V^{(d)})$. Remplacer X par V permet de travailler légitimement avec $\alpha = 1$.*

Considérons la paire $(X_\infty, Y_\infty) \in \Omega^c \times \{-1, +1\}$, de loi P_∞ définie par $\mathbb{P}\{Y_\infty = +1\} = p_\infty$ (voir (2)) et telle que la loi de X_∞ sachant $Y_\infty = \sigma 1$, $\sigma \in \{-,+\}$ est $\mu_\sigma(\Omega^c)^{-1}\mu_\sigma(\,\cdot\,)$. Ainsi pour tout $A \subset \Omega^c$,

$$\begin{aligned}\mathbb{P}\{X_\infty \in A, Y_\infty = +1\} &= \frac{p_\infty\mu_+(A)}{\mu_+(\Omega^c)} = \frac{p\mu_+(A)}{\mu(\Omega^c)} = \frac{p\lim_t t\mathbb{P}\{X \in tA \mid Y = +1\}}{\lim_t t\mathbb{P}\{X \in t\Omega^c\}} \\ &= \lim_{t\to\infty} \mathbb{P}\{X \in tA, Y = +1 \mid \|X\| > t\}.\end{aligned}$$

Soient φ_+, φ_- les densités respectives de Φ_-, Φ_+ sur S par rapport à une mesure de référence ρ quelconque, (*ex.* $\rho = \Phi_+ + \Phi_-$). Par homogénénité de μ_+, μ_-, la distribution conditionnelle de Y_∞ sachant $X_\infty = x$ est

$$\begin{aligned}\eta_\infty(x) &\stackrel{\text{def}}{=} \mathbb{P}\left\{Y_\infty = 1 \mid X_\infty = x\right\} = \frac{p_\infty \varphi_+(\Theta(x))/\Phi_+(S)}{p_\infty \varphi_+(\Theta(x))/\Phi_+(S) + (1-p_\infty)\varphi_-(\Theta(x))/\Phi_-(S)} \\ &= \frac{p\varphi_+(\Theta(x))}{p\varphi_+(\Theta(x)) + (1-p)\varphi_-(\Theta(x))}.\end{aligned}$$

Notons que η_∞ ne dépend pas de la composante radiale de X_∞. Le classifieur optimal pour la paire aléatoire (X_∞, Y_∞) par rapport à la perte classique L_{P_∞} est $g^*_\infty(x) = 2\mathbb{1}\{\eta_\infty(x) \geq 1/2\} - 1$.

Théorème 1 (Classifieurs optimaux propres aux extrêmes) *Sous l'hypothèses 1 et la condition de convergence uniforme*

$$\sup_{\theta \in S} |\eta(\Theta(t\theta)) - \eta_\infty(\theta)| \xrightarrow[t\to\infty]{} 0,$$

le risque optimal au -delà d'un seuil t converge, $L^*_t \xrightarrow[t\to\infty]{} L^*_{P_\infty}$*, de sorte que* $L^*_\infty = L^*_{P_\infty}$*. De plus,* g^*_∞ *minimise le risque asymptotique dans les régions extrêmes :*

$$\inf_{g \text{ measurable}} L_\infty(g) = L_\infty(g^*_\infty) = \mathbb{E}\left\{\min(\eta_\infty(\Theta_\infty), 1 - \eta_\infty(\Theta_\infty))\right\}.$$

Le théorème 1 nous assure que le risque asymptotique minimal est atteint par g^*_∞ qui n'est fonction que de la composante angulaire du point considéré, et pas de son rayon.

3 Minimisation du risque empirique dans les extrêmes

Soit $\mathcal{G}_S$ une classe de classifieurs $g : \theta \in S \mapsto g(\theta) \in \{-1, +1\}$ sur S ne dépendant que de la composante angulaire du point à classifier. Considérons la statistique d'ordre $\|X_{(1)}\| > \ldots > \|X_{(n)}\|$ et $\{Y_{(i)}\}_{i\in\{1,\ldots,n\}}$ les étiquettes associées. Soit $\tau > 0$ une faible portion des données, soit t_τ le quantile d'ordre $(1-\tau)$ de la variable $\|X\|$. Soit $k = \lfloor n\tau \rfloor$, on définit l'estimateur du risque extrême

$$\widehat{L}_k(g) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\{Y_{(i)} \neq g(\Theta(X_{(i)}))\} \tag{3}$$

Théorème 2 *Supposons que la classe* $\mathcal{G}_S$ *soit de* VC *dimension* $V_{\mathcal{G}_S} < +\infty$*. Soit* $\widehat{g}_k$ *un minimiseur de* (3) *avec* $k = \lfloor n\tau \rfloor$*. alors, pour tout* $\delta \in (0,1)$*,* $\forall n \geq 1$*, avec probabilité*

supérieur à $1-\delta$ *:*

$$L_{t_\tau}(\widehat{g}_k) - L^*_{t_\tau} \leq \frac{1}{\sqrt{k}}\left(\sqrt{2(1-\tau)\log(2/\delta)} + C\sqrt{V_{\mathcal{G}_S}\log(1/\delta)}\right) + \frac{1}{k}\left(5 + 2\log(1/\delta) + \sqrt{\log(1/\delta)}(C\sqrt{V_{\mathcal{G}_S}} + \sqrt{2})\right) + \left\{\inf_{g\in\mathcal{G}_S} L_{t_\tau}(g) - L^*_{t_\tau}\right\},$$

avec C *une constante indépendente de* n*,* τ *et* δ*.*

Nous renvoyons le lecteur à Jalalzai et al. (2018) pour les preuves des théorème 1 et 2.

4 Applications numériques

Nous travaillons avec deux classes $\mathcal{G}_S$ de classifieurs : les forêts aléatoires (RF) et les k plus proches voisins (k-NN). Pour chaque classe $\mathcal{G}_S$, les performances du classifieur angulaire $\widehat{g}_k$ sont comparées à celle des algorithmes (RF ou k-NN) usuels *i.e.* s'appuyant sur les données sans standardisation ni troncature ou projection des points sur le simplexe. Nous renvoyons à la section équivalente de Jalalzai et al. (2018) pour une description détaillée des expériences.

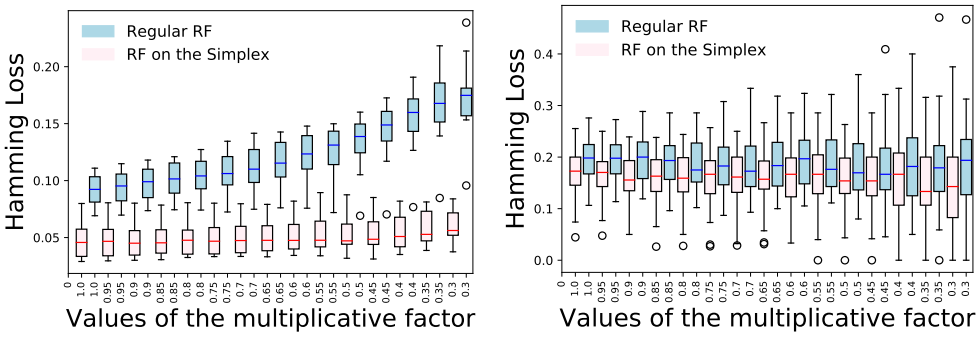

FIGURE 1 – risque comparé du classifieurs RF classique et du classifieur RF angulaire issu de l'approche proposée, en fonction du facteur multiplicatif κ, sur 10 jeux de données simulées (à gauche) et sur les données Ecoli (à droite)

La figure 1 présente l'évolution du risque d'erreur au-delà de seuils t_{test} croissants, correspondant à des valeurs décroissantes d'un facteur multiplicatif κ déterminant la proportion des données test considérées comme extrêmes. Les boites à moustache rendent compte respectivement des résultats sur 10 jeux de données simulés dans un modèle max-stable de valeurs extrêmes et sur différentes partition (entraînement/test) du jeu de données réelles Ecoli disponible sur l'archive « UCI ML repository ». Dans les deux cas, le risque du classifieur classique est supérieur à celui issu de l'approche proposée. L'écart de performance tend à se creuser avec le caractère extrême de la région test.

Références

Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6) :2507–2536.

Carpentier, A. and Valko, M. (2014). Extreme bandits. In *Advances in Neural Information Processing Systems 27*, pages 1089–1097. Curran Associates, Inc.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Applications of mathematics : stochastic modelling and applied probability. U.S. Government Printing Office.

Goix, N., Sabourin, A., and Clémençon, S. (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pages 75–83.

Jalalzai, H., Clémenç ton, S., and Sabourin, A. (2018). On binary classification in extreme regions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3096–3104.

Mendelson, S. (2018). Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1) :459–502.

Ohannessian, M. I. and Dahleh, M. A. (2012). Rare probability estimation under regularly varying heavy tails. In *Conference on Learning Theory*, pages 21–1.

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering.

Resnick, S. (2007). *Heavy-tail phenomena : probabilistic and statistical modeling*. Springer Science & Business Media.

Roos, T., Grünwald, P., Myllymäki, P., and Tirri, H. (2006). Generalization to unseen cases. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 1129–1136. MIT Press.