Une approche PAC-bayésienne d'un problème de ranking binaire en grande dimension

Benjamin Guedj [1] & Sylvain Robbiano [2]

[1] *Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France.*
benjamin.guedj@upmc.fr
[2] *CIMFAV, Pedro Montt 2421, Valparaíso, Chili.*
sylvain.robbiano@uv.cl

Résumé. Le *ranking* binaire est un problème d'apprentissage supervisé qui consiste à apprendre, d'un échantillon initial $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, à ranger des observations $\mathbf{X}_i$ dans le même ordre que leurs labels $Y_i \in \{\pm 1\}$. Nous nous situons dans le cadre de la grande dimension, en considérant des observations $\mathbf{X}_i \in \mathbb{R}^d$, où $d \gg n$. Il est pertinent de chercher à résoudre cette tâche en définissant la notion de fonction de *scoring*. Nous proposons d'approcher la fonction de *scoring* optimale par une procédure basée sur la distribution *a posteriori* de Gibbs, favorisant les estimateurs parcimonieux et s'écrivant de façon additive en les covariables. Ce schéma présente l'avantage de faciliter l'interprétation de l'effet de chaque covariable, tout en préservant une formulation non paramétrique. Nous proposons pour cette procédure une étude théorique à l'aide des outils PAC-bayésiens, ainsi qu'une mise en œuvre faisant appel à des techniques de simulation par MCMC.

Mots-clés. Théorie PAC-bayésienne, ranking, agrégation, apprentissage supervisé, grande dimension et parcimonie.

Abstract. The bipartite ranking problem consists in learning from a sample $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ to *rank* observations $\mathbf{X}_i$, while preserving the order of their associated labels $Y_i \in \{\pm 1\}$. We consider this problem in the high dimensional situation, where the observations $\mathbf{X}_i$s lie in a space of dimension d, possibly much larger than the sample size n. A standard approach in this context involves the introduction of a *scoring function*. We propose to estimate the optimal scoring function using the so-called Gibbs posterior distribution, which favors sparse additive estimators. This procedure appears valuable to assess the effect of each covariate on the score of an observation. Using elements from the PAC-Bayesian theory, we provide theoretical guarantees about our method, along with an implementation through MCMC.

Keywords. PAC-Bayesian theory, ranking, aggregation, supervised learning, high dimension and sparsity.

1 Notations

Soit $(\mathbf{X}, Y)$ une variable aléatoire à valeurs dans $\mathbb{R}^d \times \{\pm 1\}$ [1]. Notons $\mathbb{P} = (\mu, \eta)$ la loi de $(\mathbf{X}, Y)$, où μ désigne la loi marginale de $\mathbf{X}$, et $\eta = \mathbb{P}\{Y = 1|X\}$. L'objectif du *ranking* bipartite est d'ordonner l'espace $\mathcal{X} = \mathbb{R}^d$, de façon à préserver l'ordre sur les labels. En d'autres termes, il s'agit de définir une relation d'ordre sur $\mathbb{R}^d$ qui soit cohérente avec l'ordre sur $\{\pm 1\}$.

Une façon naturelle de construire un tel ordre sur cet espace $\mathbb{R}^d$ est d'introduire la notion de *fonction de scoring*, notée $s : \mathcal{X} \to \mathbb{R}$. Le problème que nous étudions peut alors se reformuler de la façon suivante : il s'agit de produire une fonction s de *scoring*, telle que pour toute paire $(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2$, $s(\mathbf{x}) \leq s(\mathbf{x}') \Leftrightarrow \eta(\mathbf{x}) \leq \eta(\mathbf{x}')$. Du point de vue de l'apprentissage statistique, il s'agit de construire une fonction de *scoring* sur la base d'un échantillon $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ composé de réplications i.i.d. de la variable $(\mathbf{X}, Y)$.

2 *Ranking* et perte par paires

Pour attester de la qualité d'une fonction de *scoring*, il est naturel d'introduire le risque dit de *ranking*, basé sur la fonction de perte de classification par paires, que nous définissons comme suit : soient $(\mathbf{X}, Y)$ et $(\mathbf{X}', Y')$ deux variables indépendantes et toutes deux de loi $\mathbb{P}$. Le risque de *ranking* d'une procédure de *scoring* s est défini par

$$L(s) \stackrel{def}{=} \mathbb{P}\left\{(s(\mathbf{X}) - s(\mathbf{X}')) \cdot (Y' - Y) < 0\right\}. \tag{1}$$

Dans [1], il a été démontré que la fonction de *scoring* optimale pour la perte de *ranking* est la loi *a posteriori* η, qui n'est bien sûr pas accessible au statisticien.

Une quantité d'intérêt théorique majeur est l'excès de risque de *ranking*, noté $\mathcal{E}(\cdot) = L(\cdot) - L(\eta)$. Il est également démontré dans [1] que cet excès de risque peut se reformuler comme

$$\mathcal{E}(s) = \mathbb{E}\left[|\eta(\mathbf{X}) - \eta(\mathbf{X}')| \, \mathbb{1}_{\{(s(\mathbf{X})-s(\mathbf{X}'))(\eta(\mathbf{X}')-\eta(\mathbf{X}))<0\}}\right], \quad \forall s.$$

L'approche que nous adoptons dans ce travail consistera à borner cet excès de risque, sous la forme d'inégalités oracles, dépendant donc de la classe de fonctions de *scoring* envisagée.

Dans [1], les auteurs étudient le cas où η est supposée appartenir à un ensemble fonctionnel de VC-dimension finie. La procédure d'estimation proposée consiste à minimiser la contrepartie empirique du risque de *ranking*, définie par

$$L_n \colon s \mapsto \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{1}_{\{(Y_i - Y_j)(s(\mathbf{X}_i) - s(\mathbf{X}_j)) < 0\}},$$

1. avec la notation usuelle $\mathbf{X} = (X_1, \ldots, X_d)$.

Dans [2] et [6], les auteurs étendent l'étude précédente au cas où η appartient à une certaine classe de Hölder, et satisfait une hypothèse de marge. Nous proposons dans ce travail d'étudier le cas où η admet une représentation parcimonieuse, faisant intervenir un petit nombre d_0 de covariables, tel que $d_0 \ll n$. Notons que dans un travail parallèle, cette situation a été étudiée par [5], pour des fonctions de *scoring* linéaires. Nous considérons un cas plus général, en introduisant des fonctions de *scoring* non paramétriques et se décomposant de façon additive en les covariables.

3 Approche PAC-bayésienne parcimonieuse

Dans la lignée des travaux de [3] et [4] en régression dans les modèles additifs généralisés, nous proposons le schéma d'estimation suivant. La collection de fonctions de *scoring* que nous envisageons est

$$\mathcal{S}_\Theta = \left\{ s_\theta \colon \mathbf{x} \mapsto \sum_{j=1}^{d} \sum_{k=1}^{M} \theta_{jk} \phi_k(x_j), \quad \theta \in \mathbb{R}^{dM} \right\},$$

où $\mathbb{D} = \{\phi_1, \ldots, \phi_M\}$ désigne une collection—un *dictionnaire*—de fonctions déterministes connues du statisticien. Ainsi, l'enjeu est de produire un certain vecteur $\hat{\theta}$, que l'on souhaite parcimonieux (*i.e.*, avec un grand nombre de composantes nulles), pour ensuite construire l'estimateur *plug-in* $s_{\hat{\theta}}$. Notre procédure repose sur la théorie PAC-bayésienne, et démarre par la définition d'une mesure *a priori* π sur l'espace Θ muni de sa tribu borélienne. En notant $\mathbf{m} = (m_1, \ldots, m_d) \in \{0,1\}^d$ le modèle codant la présence des covariables (si $m_i = 0$, la covariable i est absente du modèle et $\theta_{i\cdot} = (\theta_{i1}, \ldots, \theta_{iM}) = 0$), nous proposons :

$$\pi(\mathrm{d}\theta) \propto \sum_{\mathbf{m}} \binom{d}{|\mathbf{m}|_0}^{-1} \alpha^{|\mathbf{m}|_0} \mathrm{Unif}_{\mathcal{B}_\mathbf{m}}(\theta), \tag{2}$$

où $|\mathbf{m}|_0 = \sum_{j=1}^{d} m_j$, et $\mathcal{B}_\mathbf{m}$ désigne la boule unité en norme ℓ_2 dans $\mathbb{R}^{|\mathbf{m}|_0}$.

Nous nous intéressons aux estimateurs θ randomisés, échantillonnés suivant une certaine distribution ρ absolument continue par rapport à π. Il est pertinent dans ce cadre de chercher à résoudre le problème suivant d'optimisation sous contrainte :

$$\arg\min_{\rho} \left\{ \int_\Theta L_n(s_\theta) \rho(\mathrm{d}\theta) + \frac{\lambda}{n} \mathcal{KL}(\rho, \pi) \right\}.$$

En utilisant les conditions de Karush-Kuhn-Tucker, il apparaît que l'unique solution est la distribution *a posteriori* de Gibbs, notée $\hat{\rho}_\lambda$ et définie par

$$\hat{\rho}_\lambda(\mathrm{d}\theta) \propto \exp[-\lambda L_n(s_\theta)] \pi(\mathrm{d}\theta),$$

où $\lambda > 0$ joue le rôle d'un paramètre de température inverse. En d'autre termes, cette distribution va "tordre" la distribution *a priori* π en fonction de l'adéquation aux données, mesurée à l'aune du risque empirique de *ranking*.

L'estimateur final que nous proposons est noté $s_{\hat{\theta}}$, où $\hat{\theta} = \int_\Theta \theta \hat{\rho}_\lambda(\mathrm{d}\theta) = \mathbb{E}_{\hat{\rho}_\lambda}\theta$. Nous présenterons pour cet estimateur une inégalité oracle attestant de ses mérites théoriques, et évoquerons également la mise en œuvre pratique de cette procédure par MCMC.

Références

[1] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36 :844–874, 2008.

[2] S. Clémençon and S. Robbiano. Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of ICML '11*, 2011.

[3] B. Guedj. *Agrégation d'estimateurs et de classificateurs : théorie et méthodes*. PhD thesis, Université Pierre & Marie Curie – Paris VI, 2013.

[4] B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7 :264–291, 2013.

[5] C. Li, W. Jiang, and M. A. Tanner. General Oracle Inequalities for Gibbs Posterior with Application to Ranking. *JMLR : Workshop and Conference Proceedings*, 30 :1–10, 2013.

[6] S. Robbiano. Upper bounds and aggregation in bipartite ranking. *Electronic Journal of Statistics*, 7 :1249–1271, 2013.