Apprentissage statistique : le point de vue PAC-bayésien

Benjamin Guedj [1]

[1] *Inria, benjamin.guedj@inria.fr*
http://researchers.lille.inria.fr/bguedj/

Résumé. L'analyse PAC (*Probably Approximately Correct*) d'estimateurs quasi-bayésiens, ou théorie PAC-bayésienne, s'est affirmée ces dernières années comme une approche performante d'apprentissage statistique, notamment dans le cadre de données massives. Cet exposé a pour ambition de présenter les concepts qui président à cette théorie, les bornes usuelles sur le risque d'estimateurs PAC-bayésiens, ainsi que quelques algorithmes liés.

Mots-clés. Apprentissage statistique, théorie PAC-bayésienne.

Abstract. The PAC-Bayesian theory consists in the PAC analysis of quasi-Bayesian estimators, and has attracted some attention connected to the big data landslide. I will present the fundamentals of PAC-Bayesian learning and comment on recent advances, ranging from risk bounds to algorithmic considerations.

Keywords. PAC-Bayesian learning, Statistical learning theory.

La théorie statistique de l'apprentissage s'est affirmée en quelques décennies comme une discipline dynamique, empruntant tant à la statistique mathématique qu'à l'optimisation et au machine learning, et s'enrichit de nombreuses méthodes pour lesquelles existent de solides garanties mathématiques et algorithmiques. En particulier, la confrontation de la statistique et des données massives et complexes (le phénomène dit du *big data*) a conduit à de profonds bouleversements des approches classiques.

Ce document présente de manière succincte l'une de ces méthodes d'apprentissage. La théorie PAC-bayésienne consiste en une analyse PAC (produisant des bornes en déviations sur le risque) d'estimateurs quasi-bayésiens : je détaille dans ce qui suit cette terminologie. Cette approche, initiée par [14] et [13], a été notamment formalisée par [5], [4], [1] et [8]. Ces dernières années ont vu l'extension de l'arsenal PAC-bayésien à de nombreux modèles (tant en analyse *batch* qu'*online*), voir par exemple [2], [9], [10], et [8], parmi d'autres. Une attention particulière sera donnée dans l'exposé aux récents travaux [11], [3] et [12], proposant une version PAC-bayésienne des problèmes de *ranking* binaire, de factorisation de grandes matrices aléatoires et d'*online clustering*, respectivement.

Considérons un échantillon $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ de copies i.i.d. d'une variable aléatoire $(\mathbf{X}, Y) \in \mathcal{X} \times \mathcal{Y}$. Nous nous proposons de construire, à partir de $\mathcal{D}_n$, une procédure $\widehat{\phi}\colon \mathcal{X} \to \mathcal{Y}$ fournissant une approximation de Y pour tout nouveau point $\mathbf{X}$ collecté, *i.e.*, pour laquelle Y et $\widehat{\phi}(\mathbf{X})$ sont proches en un certain sens. Nous nous munissons pour cela d'une fonction de perte $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ à laquelle nous associons la notion de risque d'une procédure $\widehat{\phi}$:

$$R\left(\widehat{\phi}\right) = \mathbb{E}\ell\left(Y, \widehat{\phi}(\mathbf{X})\right).$$

Ce risque dépendant de la distribution de $(\mathbf{X}, Y)$, nous lui substituons son alter ego empirique :

$$R_n\left(\widehat{\phi}\right) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i, \widehat{\phi}(\mathbf{X}_i)\right).$$

Le cœur de l'approche PAC-bayésienne consiste à fabriquer des estimateurs $\widehat{\phi}$ sur un certain espace fonctionnel probabilisé $\mathcal{F}$, échantillonnés selon la distribution $\widehat{\pi}_\lambda$ définie par

$$\mathrm{d}\widehat{\pi}_\lambda\left(\widehat{\phi}\right) \propto \exp\left(-\lambda R_n\left(\widehat{\phi}\right)\right) \mathrm{d}\pi\left(\widehat{\phi}\right), \quad \forall\, \widehat{\phi} \in \mathcal{F},$$

où $\lambda > 0$ et π est une distribution de référence. Par analogie avec la terminologie bayésienne, cette distribution π est généralement désignée comme *prior*, le terme exponentiel est appelé quasi-vraisemblance, avec comme conséquence le terme de quasi-*posterior* pour $\widehat{\pi}_\lambda$. Le recours à cette mesure $\widehat{\pi}_\lambda$ (également connue sous le nom de *posterior* de Gibbs, ou poids exponentiels, voir [7]) est motivé par le lemme suivant, dû à [6] et utilisé par [5].

Introduisons les notations suivantes : pour deux mesures de probabilités μ_1 et μ_2, la divergence de Kullback-Leibler de μ_1 par rapport à μ_2 est notée $\mathcal{K}(\mu_1, \mu_2)$ et définie comme

$$\mathcal{K}(\mu_1, \mu_2) = \begin{cases} \int \log\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right) \mathrm{d}\mu_1 & \text{si } \mu_1 \ll \mu_2, \\ \infty & \text{sinon.} \end{cases}$$

Pour tout ensemble probabilisé $(A, \mathcal{A}, \mu)$, $\mathcal{M}_\mu(A)$ désigne l'ensemble des mesures de probabilité absolument continues par rapport à μ.

Lemme. *Soit $(A, \mathcal{A})$ un ensemble mesurable. Pour toute mesure de probabilité μ sur $(A, \mathcal{A})$ et toute fonction mesurable $h : A \to \mathbb{R}$ telle que $\int(\exp \circ h)\mathrm{d}\mu < \infty$,*

$$\log \int (\exp \circ h)\mathrm{d}\mu = \sup_{m \in \mathcal{M}_\mu(A)} \left\{ \int h \mathrm{d}m - \mathcal{K}(m, \mu) \right\},$$

avec la convention $\infty - \infty = -\infty$. De plus, si h est bornée sur le support de μ, le supremum dans le terme de droite est atteint pour la distribution de Gibbs g définie par

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp(h(a))}{\int(\exp \circ h)\mathrm{d}\mu}, \quad \forall a \in A.$$

Ce résultat est central dans l'obtention d'inégalités PAC, pour le choix $h(\cdot) = -\lambda R_n(\cdot)$. L'exposé s'attachera en particulier à présenter le schéma de preuve de telles inégalités. Le quasi-posterior $\widehat{\pi}_\lambda$ est ensuite utilisé pour construire des estimateurs, en particulier

$$\widehat{\phi} \sim \widehat{\pi}_\lambda \quad \text{ou} \quad \widehat{\phi} = \int \phi \, \mathrm{d}\widehat{\pi}_\lambda(\phi) \, .$$

Sous des hypothèses minimales et pour le choix $\lambda \propto n$, nous obtenons des inégalités PAC du type suivant : pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}\right) - R^\star \leq \inf_{\rho \in \mathcal{M}_\pi(\mathcal{F})} \left\{ \int R(\phi) \mathrm{d}\rho(\phi) - R^\star + \frac{\mathcal{K}(\rho, \pi) + \log \varepsilon^{-1}}{n} \right\} \right) \geq 1 - \varepsilon,$$

où

$$R^\star = \inf_{\phi \, \in \, \mathcal{Y}^{\mathcal{X}}} R(\phi)$$

désigne le risque de Bayes. Ces résultats (dits "oracles") portent de nombreux enseignements, et permettent en particulier l'adaptation à la grande dimension à travers le *prior*. En effet, pour un choix convenable de π, la divergence $\mathcal{K}(\cdot, \pi)$ fait apparaître des termes d'ordre $\mathcal{O}\left(n^{-1} \log \dim(\mathcal{X})\right)$.

L'exposé consistera en un panorama des résultats théoriques et algorithmiques, et se concentrera en particulier sur les arguments mathématiques cruciaux en théorie PAC-bayésienne.

Bibliographie

[1] P. Alquier (2006). *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. PhD thesis, Université Paris 6.

[2] P. Alquier and G. Biau (2013). Sparse Single-Index. *Journal of Machine Learning Research*.

[3] P. Alquier and B. Guedj (2016). Bayesian Non-Negative Matrix Factorization, arXiv preprint http://arxiv.org/abs/1601.01345

[4] J.-Y. Audibert (2004). *Une approche PAC-bayésienne de la théorie statistique de l'apprentissage*. PhD thesis, Université Paris 6.

[5] O. Catoni (2004). *Statistical Learning Theory and Stochastic Optimization*. Ecole d'été de probabilités de Saint-Flour, Springer.

[6] I. Csiszár (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*.

[7] A. S. Dalalyan and A. B. Tsybakov (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*.

[8] I. Giulini (2015). PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces, arXiv preprint, http://arxiv.org/abs/1511.06263

[9] B. Guedj (2013). Agrégation d'estimateurs et de classificateurs : théorie et méthodes. PhD thesis, Université Paris 6.
[10] B. Guedj and P. Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models. *Electronic Journal of Statistics*.
[11] B. Guedj and S. Robbiano (2015). PAC-Bayesian High Dimensional Bipartite Ranking, arXov preprint http://arxiv.org/abs/1511.02729
[12] L. Li, B. Guedj and S. Loustau (2016). PAC-Bayesian Online Clustering, arXiv preprint http://arxiv.org/abs/1602.00522
[13] D. A. McAllester (1999). Some PAC-Bayesian theorems. *Machine Learning*.
[14] J. Shawe-Taylor and R. C. Williamson (1997). A PAC analysis of a Bayes estimator. *Proceedings of the 10th COLT*.