Quantification et Clustering avec des Divergences de Bregman

Aurélie Fischer

Quantification et Clustering avec des Divergences de Bregman

Aurélie Fischer

Université Pierre et Marie Curie - Paris VI
Laboratoire de Statistique Théorique et Appliquée (LSTA)
175, rue du Chevaleret, boîte 158
75013 PARIS

Résumé

Nous nous intéressons au problème de quantification pour une variable aléatoire X à valeurs dans un espace de Banach séparable et réflexif, et à la question liée du clustering de n variables aléatoires indépendantes distribuées selon la loi de X. Nous utilisons une méthode de quantification avec une classe de mesures de distorsion appelées divergences de Bregman. Nous donnons des conditions assurant l'existence d'un quantificateur optimal et d'un quantificateur empirique optimal. Nous discutons aussi des vitesses de convergence.

Abstract

We are interested in the quantization problem for a random variable X taking its values in a separable reflexive Banach space, and in the related question of clustering n independent random variables with the same distribution as X. We use a quantization scheme with a class of distortion measures called Bregman divergences. We give conditions ensuring existence of an optimal quantizer and an empirically optimal quantizer. We also discuss rates of consistency.

Mots-clés : divergences de Bregman, quantification, classification non supervisée, centres mobiles

Keywords : Bregman divergences, quantization, clustering, k-means

Soit $(E, \|\cdot\|)$ un espace de Banach séparable et réflexif, et soit $k \geq 1$ un entier. La quantification, présentée notamment par Gersho et Gray (1992), Graf et Luschgy (2000) et Linder (2002), consiste à représenter une variable aléatoire X de loi μ, à valeurs dans E, par $q(X)$ où q est une application borélienne $q : E \to \mathbf{c} = \{c_1, \ldots, c_k\} \subset E$. Chaque élément de E est affecté à l'un des c_j, ce qui définit une partition de E en k cellules $\{S_1, \ldots, S_k\}$. Un quantificateur q est déterminé par sa table de codage $\mathbf{c}$ et sa partition. L'erreur que l'on fait en remplaçant X par $q(X)$ est mesurée par la distorsion de q quantifiant X :

$$W(\mu, q) = \mathbb{E}d(X, q(X)) = \int_E d(x, q(x))d\mu(x) \tag{1}$$

où $d(\cdot, \cdot)$ est une fonction mesurable à valeurs dans $\mathbb{R}^+$ appelée mesure de distorsion. Notons $W_k^*(\mu)$ l'erreur minimale. Pour une table de codage $\mathbf{c}$ donnée, la plus petite erreur est obtenue pour le quantificateur associé à la partition de Voronoi, qui est définie par :

$$S_1 = \{x \in E, d_\phi(x, c_1) \leq d_\phi(x, c_j), j = 1 \ldots, k\}$$
$$S_i = \{x \in E, d_\phi(x, y_i) \leq d_\phi(x, c_j), j = 1, \ldots, k\} \setminus \bigcup_{\ell=1}^{i-1} S_\ell \quad \text{pour } i = 2, \ldots, k$$

Ainsi, trouver un quantificateur optimal revient à trouver une table de codage optimale. A la question probabiliste de la quantification est reliée celle du clustering : dans un contexte statistique, on ne connaît pas la loi de X, mais on dispose de données $X_1, \ldots, X_n$ supposées distribuées selon cette loi, que l'on désire répartir en k groupes de manière à ce que les données à l'intérieur d'un même groupe soient très semblables et que les différents groupes soient aussi séparés que possible. Pour cela, on cherche à minimiser le critère empirique correspondant à (1) :

$$W(\mu, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} d(X_i, c_j) \tag{2}$$

Dans $\mathbb{R}^d$, la mesure de distorsion la plus fréquemment utilisée est la distance euclidienne au carré. Dans de nombreux domaines, les données à classer sont des courbes, donc des données de dimension infinie. Biau, Devroye et Lugosi (2008) considèrent ainsi une distance hilbertienne. D'autre part, Teboulle, Berkhin, Dhillon, Guan et Kogan (2007) remarquent que pour classer certains types de données, une mesure de distorsion non symétrique est plus appropriée. Les travaux de Banerjee, Merugu, Dhillon et Ghosh (2005) ainsi que Wu, Xiong, Chen et Zhou (2007) nous invitent à considérer une classe de mesures de distorsion introduite par Bregman (1967). Si $\mathcal{C} \subset E$ est un convexe et $\phi : \mathcal{C} \to \mathbb{R}$ une fonction strictement convexe, différentiable sur l'intérieur relatif de $\mathcal{C}$, la divergence de Bregman associée est définie par

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y)$$

avec $D_y\phi$ la différentielle de ϕ en y. Lorsque E est un espace de Hilbert, pour le produit scalaire $\langle\cdot,\cdot\rangle$, on peut écrire aussi

$$d_\phi(x,y) = \phi(x) - \phi(y) - \langle x-y, \nabla\phi(y)\rangle$$

La notion a d'abord été définie pour $E = \mathbb{R}^d$; on trouve chez Alber et Butnariu (1997) et Frigyik, Srivastava et Gupta (2008) l'extension à la dimension infinie. On a $\forall(x,y) \in E$, $d_\phi(x,y) \in \mathbb{R}^+$ et $x = y \Leftrightarrow d_\phi(x,y) = 0$, mais les divergences de Bregman ne sont pas toutes des distances, car elles ne vérifient pas nécessairement la propriété de symétrie et l'inégalité triangulaire. De plus, elles sont convexes en la première variable, mais elles ne le sont pas forcément en la seconde. Selon la fonction ϕ choisie, on obtient différentes mesures de distorsion, en dimension finie ou infinie. On retrouve ainsi le carré de la distance euclidienne, la distance de Mahalanobis, le carré de la distance L^2 ou la distance de Kullback-Leibler.

Sous certaines hypothèses, on peut montrer l'existence d'un quantificateur optimal pour les critères (1) et (2), $d(\cdot,\cdot)$ étant une divergence de Bregman $d_\phi(\cdot,\cdot)$, ainsi que les convergences

$$\lim_{n\to\infty} W(\mu, \mathbf{c}_n^*) = W_k^*(\mu) \quad p.s.$$

et

$$\lim_{n\to\infty} \mathbb{E}W(\mu, \mathbf{c}_n^*,) = W_k^*(\mu)$$

où $\mathbf{c}_n^*$ désigne un minimiseur du risque empirique (2). Ainsi, lorsque la taille n de l'échantillon augmente, la performance du quantificateur empirique optimal approche celle du quantificateur optimal. Afin d'évaluer la valeur de n à partir de laquelle $\mathbf{c}_n^*$ est suffisamment proche du quantificateur optimal, on s'intéresse également aux vitesses de convergence.

Bibliographie

[1] Alber, Y. et Butnariu, D. (1997) Convergence of Bregman Projection Methods for Solving Consistent Convex Feasibility Problems in Reflexive Banach Spaces, *Journal of Optimization Theory and Applications*, 92, 1, 33–61.

[2] Banerjee, A., Merugu, S., Dhillon, I. S. et Ghosh, J. (2005) Clustering with Bregman Divergences, *Journal of Machine Learning Research*, 6, 1705–1749.

[3] Biau, G., Devroye, L. et Lugosi, G. (2008) On the performance of clustering in Hilbert spaces, *IEEE Transactions on Information Theory* 54, 2, 781-790.

[4] Bregman, L. M. (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

[5] Frigyik, B. A., Srivastava, S. et Gupta, M. R. (2008) Functional Bregman Divergence and Bayesian Estimation of Distributions, *IEEE Transactions on Information Theory*, 54, 11, 1681–1685.
[6] Gersho, A. et Gray, R. M. (1992) *Vector Quantization and Signal Compression*, Kluwer Academic Press, Boston.
[7] Graf, S. et Luschgy, H. (2000) *Foundations of Quantization for Probability Distributions*, Springer, Lectures Notes in Mathematics, 1730, Berlin.
[8] Linder, T. (2002) Learning-Theoretic Methods in Vector Quantization, in Györfi, L. (éditeur), *Principles of Nonparametric Learning*, Springer, Wien, New York.
[9] Teboulle, M., Berkhin, P., Dhillon, I., Guan, Y. et Kogan, J. (2006) Clustering with Entropy-Like k-Means Algorithms, *in Grouping Multidimensional Data : Recent Advances in Clustering*, chapitre 5, 127-160, Springer.
[10] Wu, J., Xiong, H., Chen, J. et Zhou, W. (2007) A Generalization of Proximity Functions for K-means, *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, IEEE Computer Society, 361–370.