Conditions de marge pour la quantification vectorielle

Clément Levrard [1]

[1] *clement.levrard@math.u-psud.fr*

Résumé. Le principe de la quantification vectorielle est de représenter une distribution de probabilité P sur $\mathbb{R}^d$ avec un nombre fini, noté ici k, de points. Les mises en applications de ce principe sont nombreuses et variées, allant de la compression de signal dont il est issu à la classification non supervisée, notamment le pré-processing de jeux de données. N'ayant accès qu'à un n-échantillon tiré suivant la loi originelle P, la stratégie adoptée est naturellement de minimiser un risque empirique. Nous nous restreindrons au cas où le risque empirique est la moyenne sur l'échantillon de la plus petite distance au k-vecteur considéré au carré, ce qui revient à considérer l'estimateur familier des k means. Il a récemment été prouvé que, sous certaines conditions, la vitesse de convergence de cet estimateur vers l'optimum était de l'odre de $1/n$. Nous expliquerons en quoi ces conditions peuvent être comparées aux conditions de marge utilisées dans le contexte de la classification supervisée, puis discuterons de l'influence des autres paramètres impliqués dans cette vitesse de convergence.

Mots-clés. classification non supervisée, conditions de marge ...

Abstract. Vector quantization aims to answer the issue of representing a probability distribution P over $\mathbb{R}^d$ with k significant points, where k denotes a positive integer. There is a wide bunch of application for vector quantization, from signal compression to classification, for instance when pre-processing large data sets. Since only a n-sample drawn from P is available, most strategies attempt to minimize an empirical risk designed from the sample. We focus here on the celebrated k-means strategy, where the risk is the average squared distance to the closest element to the k-vector we intend to evaluate. Recent results show that, under some restrictive conditions, fast rates of convergence can be derived for the k-means strategy, namely $1/n$. We will expose why these conditions may be thought of as margin type conditions, such as one Mammen and Tsybakov introduced for the statistical learning theory. Then the dependancy of this convergence rate on other natural parameters will be discussed.

Keywords. classification, margin condition ...

1 Introduction à la quantification vectorielle

Cette première partie introduit les notions et définitions utiles à la compréhension du reste du texte. Durant tout l'exposé k désignera un nombre entier fixe, et P une distribution de probabilité sur $\mathbb{R}^d$. Les candidats pour représenter P sont des k-vecteurs d'élément de P, qu'on nommera dictionnaires. De fait, un dictionnaire $\mathbf{c}$ est alors formé de k mots $(c_1, \ldots, c_k)$, chaque c_i étant un élément de $\mathbb{R}^d$. La capacité pour un dictionnaire $\mathbf{c}$ à représenter la distribution P est mesurée par la fonction de risque

$$R(\mathbf{c}) = P\left(\min_{j=1,\ldots,k} \|x - c_j\|^2\right),$$

où $Pf(.)$ signifie intégration par rapport à P. Une première remarque est que ce risque correspond à l'erreur de la stratégie de classication non-supervisée définie par $Q(x) = c_j$ si $x \in V_j(\mathbf{c})$, où $V_j(\mathbf{c})$ est la cellule de Voronoï associée à c_j dans le diagramme de Voronoï engendré par $c_1, \ldots, c_k$. Cela se traduit par la formule

$$R(\mathbf{c}) = P(\|x - Q(x)\|^2) = \sum_{j=1}^{k} \|x - c_j\|^2 \mathbb{1}_{V_j(\mathbf{c})}(x).$$

Une seconde remarque est que cette fonction de risque $R(.)$ se met sous la forme $P\gamma(\mathbf{c}, .)$, où la fonction $\gamma(\mathbf{c}, .)$ est alors définie par

$$\gamma : \begin{cases} \left(\mathbb{R}^d\right)^k \times \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ (\mathbf{c}, x) & \longmapsto & \min\limits_{j=1,\ldots,k} \|x - c_j\|^2 \end{cases}.$$

Le but étant de trouver un dictionnaire minimisant R, on notera $\mathbf{c}^*$ un tel dictionnaire optimal, et on supposera par commodité que ce dictionnaire optimal est unique à permutation près.

Notons $X_1, \ldots, X_n$ l'échantillon de taille n, et P_n la distribution empirique qui y est associée. Le risque empirique est alors naturellement défini par

$$\hat{R}_n(\mathbf{c}) = P_n\gamma(\mathbf{c}, .) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|X_i - c_j\|^2.$$

La stratégie adoptée est celle du minimiseur du risque empirique, c'est à dire que de manière pratique on trouve un dictionnaire empirique $\hat{\mathbf{c}}_n$ minimisant $\hat{R}_n$, et on cherche à évaluer sa performance par rapport au dictionnaire optimal. En d'autres termes on cherche à évaluer la perte

$$\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) = R(\hat{\mathbf{c}}_n) - R(\mathbf{c}^*),$$

en espérance ou avec grande probabilité sur l'échantillon.

Au vu de la forme des fonctions de risque, il semble naturel que les bonnes propriétés de ce dictionnaire empirique vont essentiellement découler des propriétés de déviations entre la distribution P et la distribution empirique P_n, mesurées sur l'ensemble des fonctions de contraste $\gamma(\mathbf{c}, .)$, $\mathbf{c}$ variant dans l'espace des dictionnaires. Cette approche a notamment été développée pour la quantification vectorielle par Linder dans [4] pour obtenir des vitesses de convergence de l'ordre de $1/\sqrt{n}$.

2 Conditions de marge

Pour obtenir une vitesse de convergence de l'ordre de $1\sqrt{n}$, l'hypothèse requise est généralement que P a un support borné (dans [4, Theorem 4] par exemple). Cette hypothèse restera présente dans les deux sections suivantes.

2.1 Cas des distributions régulières

Cependant, d'autres conditions on été introduites par Pollard dans les années 80 (voir [5]) pour établir des résultats de convergence asymptotique de $\hat{\mathbf{c}}_n$, dans le cas où P admet une densité continue f. Plus précisément, on introduit, pour un dictionnaire $\mathbf{c}$, la matrice formée de blocs de dimension $d \times d$

$$H(\mathbf{c})_{i,j} = \begin{cases} 2P(V_i(\mathbf{c})) - 2\sum_{\ell \neq i} r_{i\ell}^{-1} \sigma\left[f(x)(x-c_i)(x-c_i)^t \mathbf{1}_{\partial(V_i(\mathbf{c}) \cap V_\ell(\mathbf{c}))}\right] & \text{pour} \quad i = j \\ -2{r_{ij}}^{-1} \sigma\left[f(x)(x-c_i)(x-c_j)^t \mathbf{1}_{\partial(V_i(\mathbf{c}) \cap V_j(\mathbf{c}))}\right] & \text{pour} \quad i \neq j \end{cases},$$

où $r_{ij} = \|c_i - c_j\|^{-1}$ et σ désigne l'intégrale de Lebesgue en dimension $d-1$.

Cette matrice est la matrice Hessienne de la fonction de risque $R(\mathbf{c})$. La condition de régularité de Pollard est alors la suivante.

Definition 1 (Condition de régularité de Pollard) *La distribution P satisfait la condition de régularité de Pollard si*

1. *P est à support borné,*
2. *P admet une densité continue f,*
3. *$H(\mathbf{c}^*)$ est définie positive.*

D'un point de vue technique, si la condition de régularité de Pollard est satisfaite, alors la fonction de risque R se trouve être localement de l'odre de $\|\mathbf{c} - \mathbf{c}^*\|^2$ aux alentours de $\mathbf{c}^*$. Plus précisément, il est prouvé dans [2] que si il est acquis que P est à support borné et admet une densité continue f, alors la condition de régularité de Pollard est équivalente à

$$\ell(\mathbf{c}, \mathbf{c}^*) \approx \|\mathbf{c} - \mathbf{c}^*\|^2 \approx \mathrm{Var}(\gamma(\mathbf{c}, .) - \gamma(\mathbf{c}^*, .)).$$

Ce type de condition est au coeur des techniques de localisation (voir [1] par exemple), et mène à des inégalités oracles de type

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq \frac{C}{n},$$

où C dépend des constantes d'équivalence entre ces différentes quantités, et peut être considéré comme le pendant technique des conditions de marge en classification supervisée. Pour achever le parallèle, on introduit la condition suivante, qui implique la condition de régularité de Pollard.

Definition 2 (Condition de marge régulière) *Appelons N^* la zone correspondant aux arêtes du diagramme de Voronoï engendré par $\mathbf{c}^*$. Supposons que P admet une densité continue f et est à support borné. Alors P satisfait une condition de marge régulière ssi*

$$\forall x \in N^* \quad |f(x)| \leq C(d, k, P),$$

où $C(d, k, P)$ est une fonction (déterminée) de k, d et P.

Cette condition établit un lien entre conditions de marge en classification non supervisée et supervisée, où le rôle de la valeur $1/2$ pour la fonction de régression est joué par la zone critique N^*.

2.2 Cas quelconque

Notons $N^*(t)$ le t-voisinage des arêtes du diagramme de Voronoi optimal, et $p(t) = P(N^*(t))$ sa masse. On définit la condition suivante, introduite dans [3].

Definition 3 (Condition de marge générale) *Supposons que P soit à support borné. Alors P satisfait une condition de marge générale ssi*

$$\exists r_0 > 0 \quad \forall t \leq r_0 \quad p(t) \leq \kappa(P, k)t,$$

où κ est une fonction (déterminée) de k et P.

Cette condition de marge générale est satisfaite quand P est constituée de k pôles bien séparés. Il est intéressant de préciser que l'on peut fournir une condition suffisante dans le cas des mélanges gaussiens impliquant des paramètres naturels du mélange (voir [3][Proposition 3.2].

Notons ε la quantité $\inf \ell(\tilde{\mathbf{c}}, \mathbf{c}^*)$, où $\tilde{\mathbf{c}}$ parcourt l'ensemble des minimiseurs **locaux** du risque. Soient B la quantité définie par $\inf_{i \neq j} \|c_i^* - c_j^*\|$ et $p_{\min}$ le poids minimal d'une cellule optimale, plus précisément $p_{\min} = \inf_i P(V_i(\mathbf{c}^*))$. Supposons que P satisfait une condition de marge avec support borné par 1 (pour simplifier), on peut alors prouver (voir [3]) que

$$\mathbb{E}\ell(\hat{\mathbf{c}}_n, \mathbf{c}^*) \leq \left(\frac{1}{\varepsilon} \vee \frac{64}{p_{min} B^2 r_0^2}\right) \frac{C(k, d)}{n},$$

où $C(k,d)$ est connue. Cette inégalité oracle non asymptotique fait apparaître, outre des termes classiques de taille du dictionnaire k et de dimension d, des termes plus spécifiques au problème de la quantification, à savoir la plus petite distance inter mots optimaux B où le degré de séparation ε. Concernant ce dernier degré de séparation, une borne inférieure peut être obtenue sous certains régimes dépendant de n que l'on détaillera selon le temps disponible.

References

[1] Koltchinskii, V., Local Rademacher complexities and oracle inequalities in risk minimization, *The Annals of Statistics*, 2006

[2] Levrard, C., Fast rates for empirical vector quantization, *Electronic Journal of Statistics*, 2013

[3] Levrard, C., Margin conditions for vector quantization, *submitted*, 2013

[4] Linder, T., Learning-theoretic methods in vector quantization, *Principles of nonparametric learning (Udine, 2001)*, 2002

[5] Pollard, D., A central limit theorem for k-means clustering, *The Annals of Probability*, 1982