Séminaire Mathématique de Béjaïa (LaMOS)
Volume 2, 2004, pp. 71 - 74

12

Choix du paramètre de lissage dans la méthode du noyau

A. BARECHE[1]

Laboratoire de Modélisation et d'Optimisation des Systèmes LAMOS
Université de Béjaïa 06000, Algerie.
email : aicha_barecheyahoo.fr

Résumé la méthode du noyau est la méthode non paramétrique la plus utilisée pour l'estimation d'une densité de probabilité. Plusieurs auteurs se sont penchés sur le choix de paramètre de lissage lié à cette méthode. Ce travail fournit une discussion sur les différentes méthodes existant dans la littérature concernant la selection de ce paramètre.

Mots clés : Méthode du noyau, Paramètre de lissage, Cross-validation, Plus proche voisin, Bootstrap.

12.1 Introduction

Dans la pratique, on est généralement confronté à des situations où la fonction densité de l'une des lois régissant un système de files d'attente donné est inconnue et doit être estimée. Il existe plusieurs méthodes d'estimation de la densité de probabilité. Nous pouvons citer : les histogrammes, la méthode du noyau et les estimateurs par fonctions orthogonales. Nous discutons dans cet article le choix du paramètre de lissage dans la méthode du noyau.

12.2 Méthode du noyau

Soit $X_1, X_2, ..., X_n$ un n-échantillon issu d'une v.a X de loi de probabilité f. L'estimateur de Parzen-Rosenblatt ([4], [5]) s'écrit sous la forme suivante :

$$\hat{f}_h(x) = f_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{12.1}$$

La fonction K est appelée noyau et elle vérifie :

$$H : \int_{-\infty}^{\infty} K(y)dy = 1, \int_{-\infty}^{\infty} |K(y)|dy < \infty, \sup_{-\infty<y<\infty} |K(y)| < \infty \text{ et } \lim_{|y|\to\infty} |y|K(y) = 0$$

Le paramètre h_n est appelé paramètre de lissage (fenêtre), c'est une suite de réels positifs tendant vers zéro.

12.2.1 Choix du paramètre de lissage

Le choix du noyau ne pose pas problème, par contre, le choix du paramètre de lissage est crucial dans la méthode du noyau. Nous pouvons citer quelques approches pour le choix de ce dernier.

1. Le moyen le plus simple est de choisir une loi pour la densité f. Si f est la loi normale et K est le noyau gaussien alors :

$$h_{\text{opt}} = 1.059 \cdot S_n \cdot n^{-\frac{1}{5}}, \tag{12.2}$$

où S_n est l'écart-type de l'échantillon $X_1, ..., X_n$.

2. **Heuristique adaptée :** En se basant sur la formule :

$$h_n^* = \alpha(K) * \beta(f) * n^{(-1/5)} \tag{12.3}$$

trouvée en minimisant $\text{MISE}(f_n(x)) = \mathbb{E}\int_{-\infty}^{\infty}(f_n(x) - f(x))^2 dx$, où

$$\alpha(K) = \left\{ \frac{f(x)\int K^2(y)dy}{(\int y^2 K(y)dy)^2} \right\}^{(1/5)} \tag{12.4}$$

$$\beta(f) = \left\{(f''(x))^2\right\}^{(-1/5)}; \tag{12.5}$$

Scott, Tapia et Thampson (1977) ont élaboré un algorithme, appelé algorithme (S.T.T), basé sur une méthode itérative pour trouver la solution de h_n en remplaçant $f^{(2)}$ par son estimateur $f_n^{(2)}$ et $\beta(f)$ par son estimateur $\beta(f_n)$ (ou $\beta(f_n(., h_n))$). Les étapes de cet algorithme, appelé algorithme (S.T.T), sont :

i) $h_{(0)}$ = étendue de l'échantillon.
ii) $h_{(i)} = \alpha(K)\beta(f_n(., h_{(i-1)}))n^{-1/5}$.
L'étape (ii) est répétée jusqu'à ce que la suite $h_{(i)}$ devient presque constante, c'est à dire :

$$\left|\frac{h_{(i)} - h_{(i-1)}}{h_{(i)}}\right| < \varepsilon \tag{12.6}$$

où ε est une précision donnée.

3. **Cross-validation :** Habbema et al [2] et Duin [1] proposent de choisir h qui maximise la "pseudo-vraisemblance" :

$$L(h) = \prod_{i=1}^{n} \hat{f}_{h,i}(x_i), \tag{12.7}$$

où :

$$\hat{f}_{h,i}(x_i) = \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x_i - X_j}{h}\right) \tag{12.8}$$

4. **Estimateur par les plus proches voisins :** Pour chaque observation x_i, on module le paramètre h en fonction de la concentration des observations autour de x_i. On écrit alors :

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - x_i}{h_i}\right), \tag{12.9}$$

où $h_i = hd(i)$, h étant l'estimateur de pseudo-vraisemblance. Pour calculer $d(i)$, on fixe à priori un nombre entier $k(n)$ tel que $0 < k(n) < n$, on détermine les $k(n)$ plus proches voisins de x_i parmi $x_1, ..., x_n$.
Loftsgaarden-Quesenberry [3], en se basant sur plusieurs travaux empiriques, ont donné une approximation pour $k(n)$ par $\sqrt{n}$.

5. **La méthode Bootstrap :** Le bootstrap a été utilisé dans l'estimation fonctionnelle de la densité de probabilité par la méthode du noyau dans le but de déterminer le paramètre de lissage h_n [6].

$$h_{\text{opt}} = n^{-1/5} \cdot S_{\text{opt}}, \tag{12.10}$$

avec :

$$S_{\text{opt}} \equiv S_{\text{opt}}(x) = \left(\frac{f(x) \int K^2(y)dy}{[f''(x) \int y^2 K(y)dy]^2}\right) \tag{12.11}$$

En remplaçant $f(x)$ et $f''(x)$ par des estimateurs consistants dans la définition de S_{opt}, on aboutit à :

$$h_{opt} = n^{-1/5} \cdot S_n, \tag{12.12}$$

avec $S_n \equiv S_n(x)$ étant consistant pour S_{opt}.

La plus récente procédure de sélection du paramètre de lissage h_n consiste à remplacer le MSE ($MSE(f_n(x)) = \mathbb{E}(f_n(x) - f(x))^2$) par la version bootstrap MSE^*. Si l'échantillon initial est $X_1, X_2, ..., X_n$, les variables bootstrap $X_{n1}^*, X_{n2}^*, ..., X_{nn}^*$ sont choisies i.i.d avec la densité :

$$\tilde{f}_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} L\left(\frac{x - X_i}{b_n}\right), \quad x \in \mathbb{R}, \tag{12.13}$$

où L est un autre noyau et b_n un paramètre de lissage que nous autorisons à dépendre des données $X_1, X_2, ..., X_n$.

En observant que par (12.10), le choix du paramètre de lissage se réduit au problème de la sélection du "paramètre échelle" s dans $h = n^{-1/5} \cdot s$, on obtient :

$$f^*_{n,s}(x) = \frac{1}{n^{4/5} \cdot s} \sum_{i=1}^{n} K\left(\frac{x - X^*_{ni}}{n^{-1/5} \cdot s}\right) \tag{12.14}$$

comme la version bootstrap de l'estimateur de Parzen-Rosenblatt donné dans la formule (12.1), et :

$$MSE^*_{n,s}(x) = \mathbb{E}^*((f^*_{n,s}(x) - \tilde{f}_n(x))^2) \tag{12.15}$$

comme le bootstrap correspondant du MSE.

Maintenant, un paramètre de lissage empirique peut être sélectionné par :

$$h_n = n^{-1/5} \cdot \arg\min_s MSE^*_{n,s} \tag{12.16}$$

Sous de fortes conditions de régularité sur f et le noyau, h_n dans (12.16) est de la forme (12.12) avec $S_n \to S_{opt}$ a.s.

12.3 Conclusion

Dans cet exposé, nous avons porté accent sur le choix du paramètre de lissage dans l'estimation fonctionnelle de la densité de probabilité par la méthode du noyau. En effet, beaucoup d'auteurs ce sont intéressé à cet aspect, et par conséquent plusieurs travaux récents sont apparus dans le domaine. C'est pourquoi, on parle du retour à la mode de la méthode du noyau.

Références

1. Duin R.P.W., 1976, On the choice of smoothing para meters for Parzen estimators of probability density functions. IEEE Trans. on Computers **C-25** 1175-1177.
2. Habbema J.D.F., Hermans J. and Van Den Broek K., 1974, A stepwise discriminant analysis program using density estimation. In Compstat 1974, ed. G. Bruckman, 101-110, physica Verlag, Vienna.
3. Loftsgaarden D.O. and Quesenberry C.P., 1965, A non parametric estimate of a multivariate density function, Ann. Math. Stat., 36.
4. Parzen E., 1962, on estimation of probability density function and mode. Ann. Math. Stat., 33 :1065.
5. Rosenblatt M., 1962, Curves estimates. Ann. Math. Stat., 33 :1815.
6. Ziegler K., 2002, On local bootstrap bandwidth choice in kernel density estimation, University of Munich. Germany.