Couplage d'un problème de classification et d'estimation de densité par des noyaux gaussiens

Aaron Catherine

SAMOS-MATISSE
Université Paris 1
90 rue de Tolbiac
Paris, France, 75013

RÉSUMÉ. La classification et l'estimation de densité d'un nuage de point issu du tirage d'un mélange de lois sont deux problèmes intimement liés. En effet la connaissance de la densité induit une classification naturelle dans laquelle le nombre de classe est connu (et correspond au nombre de modes), d'autre part la connaissance de la classification permet de localiser dans l'espace les points correspondant a chacune des composantes du mélange et simplifie le problème de l'estimation de densité. Dans la pratique aucune de ses deux données n'est disponible. Dans ce papier on propose une méthode permettant de résoudre conjointement ces deux problèmes.

MOTS-CLÉS : Classification, estimation de densité, noyaux gaussien, taille de fenêtre, cross-validation

1 Introduction

1.1 Problématique

On dispose de N observations dans $\mathbb{R}^p$ qui correspondent, par hypothèse, aux réalisation d'un mélange de k lois uni-modales et on se propose de résoudre le double problème : estimation de la densité du nuage de point (par une méthode à noyau) et segmentation des données. Dans un premier temps on va montrer en quoi ces deux problèmes sont implicitement liés.

1.2 Classification sous hypothèse de densité connue

Si on suppose que la densité totale du nuage est connue, Wishart's a proposé, en 1969, une méthode de classification des données autour des domaines d'attraction des modes. Il y a autant de groupes que de modes et chaque point est affecté à la classe du mode qui « l'attire » (on peut lier le point au mode par un chemin croissant de densité). Visuellement cette classification est relativement naturelle (voir figure 1)

1.3 Densité sous hypothèse de classification connue

La principale difficulté pour l'estimation de densité par une méthode à noyau consiste a trouver une « bonne » taille pour la fenêtre des noyaux. Dans le cas d'une densité uni-modale on peut utiliser la cross-validation qui donne des résultats pertinents. Dans le cas de densités multi-modales avec hétérogénéité des dispersion autour des modes la recherche d'une taille unique (sur tout le nuage) de fenêtre est vouée à l'échec (voir figure 2). Dans ce cas on cherchera des tailles de fenêtres dépendant des points. Si la

classification des données est connue, une méthode naturelle serait de chercher une taille de fenêtre par classe.

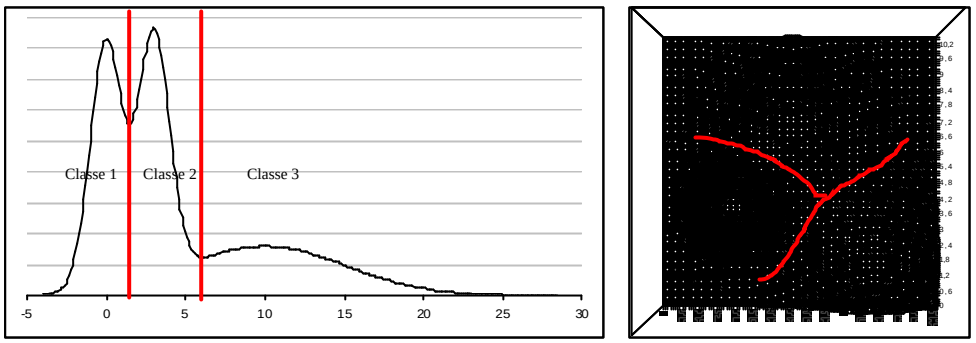

Figure 1 : classification sous hypothèse de densité connue

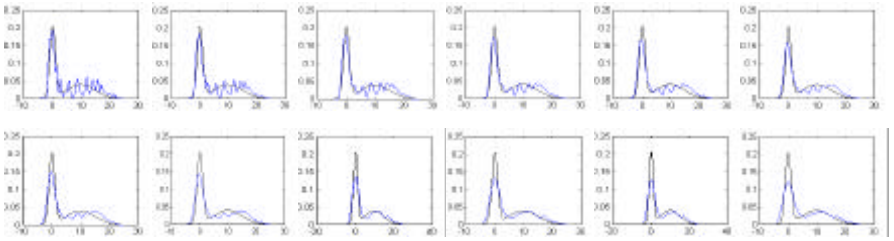

Figure 2 : Densité multi-modale avec hétérogénéité de dispersion et taille de fenêtre unique on n'estime jamais les deux composantes a la fois

2 Estimation de densité par noyaux et cross-validation

2.1 Estimation de densité par noyaux gaussiens

Dans toute la suite nous ne traiterons que le cas des variables uni-dimensionnelles pour simplifier l'écriture des équations mais la généralisation au cas multidimensionnel est aisé.

On note $j(x,y,h) = \frac{1}{\sqrt{2p\,h^2}} \exp\left(-\frac{(x-y)^2}{2h^2}\right)$

L'estimation de la densité d'un nuage de N points x_i par noyaux gaussien de taille de fenêtre h est donné par : $\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} j(x, x_i, h)$ tout le problème se résumant à déterminer un h « correct ». Pour cela nous avons choisi la méthode dite de « cross-validation » exposée ci-dessous

2.2 Dans le cas d'une seule taille de fenêtre

On note $\hat{f}_{-i}^{h}(x) = \frac{1}{N-1}\sum_{j\neq i} j(x, x_j, h)$ la densité estimée si on avait toute la base sauf le point i et on cherche a maximiser en h la pseudo-vraisemblance $L(h) = \prod_i \hat{f}_{-i}^{h}(x_i)$. L'annulation de la dérivée en h mène à : $h^2 = \frac{1}{N}\sum_i \frac{\sum_{j\neq i} j(x_i, x_j, h)(x_i - x_j)^2}{\sum_{j\neq i} j(x_i, x_j, h)}$

2.3 Dans le cas de plusieurs tailles de fenêtres

Dans le cas où il y a K classes avec $s(j)$ la classe de j la densité estimée est :

$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} j(x, x_i, h_{s(i)})$ et la maximisation de la pseudo-vraisemblance donne :

$$h_k^2 = \sum_i \frac{\sum_{j\neq i, s(j)=k} j(x_i, x_j, h_k)(x_i - x_j)^2}{\sum_{j\neq i} j(x_i, x_j, h_{s(j)})} \Bigg/ \sum_i \frac{\sum_{j\neq i, s(j)=k} j(x_i, x_j, h_k)}{\sum_{j\neq i} j(x_i, x_j, h_{s(j)})}$$

3 Algorithme de classification

L'algorithme proposé est un algorithme stochastique en effet, ici, l'aspect stochastique a un double avantage d'une part on évite de converger vers le maximum de la vraisemblance en N classes avec une taille de fenêtre tendant vers 0 et, d'un point de vue pratique on diminue notoirement le temps de calcul.

A l'état initial on a une seule classe et une taille de fenêtre h_0 puis, on itère Nit_1 :

- On tire $N_1 < N$ points y_j sans remise qui serviront de points sur lesquels on « posera » les noyaux
- on transforme les tailles de fenêtres pour maximiser la pseudo-vraisemblance de l'ensemble de la base en en effectuant Nit_2 fois :

$$h_k^1(it+1) := \sqrt{\sum_i \frac{\sum_{y_j\neq x_i, s_{it}(j)=k} j(x_i, y_j, h_k(it))(x_i - y_j)^2}{\sum_{y_j\neq x_i} j(x_i, y_j, h_{s_{it}(j)}(it))}} \Bigg/ \sum_i \frac{\sum_{y_j\neq x_i, s_{it}(j)=k} j(x_i, y_j, h_k(it))}{\sum_{y_j\neq x_i} j(x_i, y_j, h_{s_{it}(j)}(it))}$$

- on classe les données autour des modes observés pour obtenir la fonction s_{it+1}
- enfin obtient les nouvelles tailles de fenêtres par moyenne des valeurs de $h_k^1(it+1)$ sur la nouvelle classification : $h_{k'}(it+1) = \frac{\sum_{s_{it+1}(i)=k'} h^1_{s_{it}(i)}(it+1)}{\sum_{s_{it+1}(i)=k'} 1}$

Pour finir on effectue un dernier tour dopération sur lensemble de la base et non uniquement sous un sous ensemble tiré aléatoirement

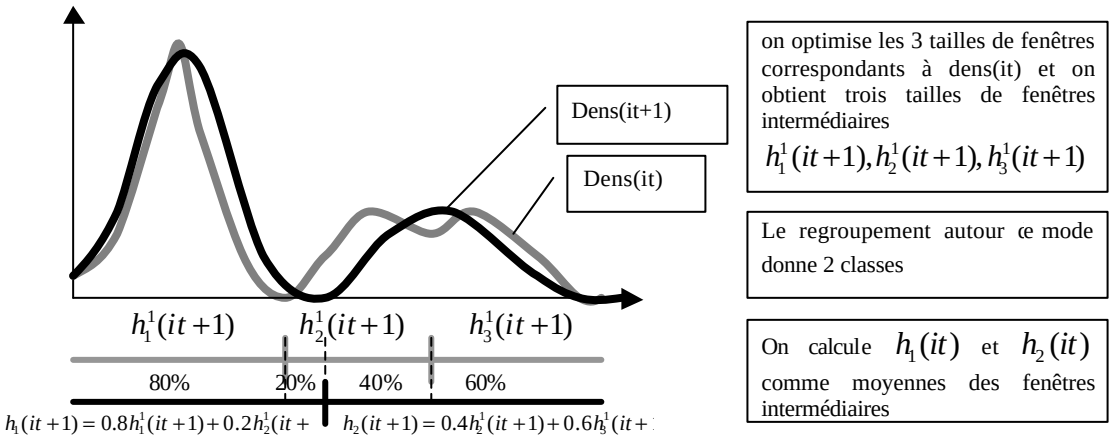

Figure 3 : Etapes de l'algorithme

4 Quelques résultats

Les résultats suivants sont le résultat d'estimation de densité et de classification sur des bases simulées en dimension 1 (ce qui permet de visualiser la différence entre les densités estimées et les « vraies » densités du tirage). Les lois simulées sont toutes des mélanges de gaussiennes. Les paramètres de l'algorithme sont, dans tous les cas : $N_1 = N/2$, $N_{it_1} = 10$ et $N_{it_2} = 3$

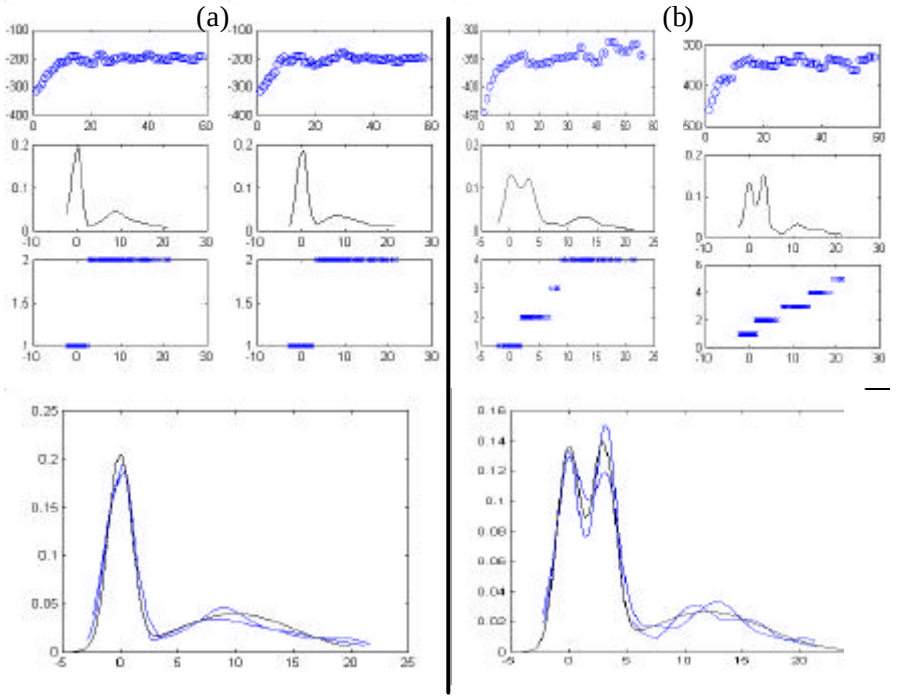

Figure 4 : Quelques résultats (a) : 200 points tirés pour moitié sur $\aleph(0,1)$ et pour moitié sur $\aleph(10,5)$ 2 exemples (pseudo-vraisemblance, estimation de densité et segmentation) et comparaison de la densité estimée à la vraie densité. (b) 300 points tirés pour tiers sur $\aleph(0,1)$, pour tiers sur $\aleph(3,1)$ et pour tiers sur $\aleph(12,5)$

5 Conclusion-perspectives

La méthode semble prometteuse mais nécessite encore des amélioration. En particulier on aimerait construire un indicateur nous permettant de déterminer quand un mode est « significatif » afin de ne pas scinder en un trop grand nombre de classe (cf figure4 exemple (b) ou des modes annexes apparaissent qui, visuellement ne semble pas « importants » mais qui numériquement induisent des erreurs de classification).

6 Bibliographie

.[BIC 0.3] BICEGO M, CRISTANI M., FUSIELLO A., MURINO V., *Watershed-based unsupervised clustering*, document de travail. http://profs.sci.univr.it/~bicego/bicego_murino_emmcvpr03.pdf.

[MIC 01] MICHALIS K., TITSIAS., ARISTIDIS C LIKAS., (2001), *Shared Kernel Models for Class Conditional Density Estimation*., IEE Transaction on Neural Network Vol 12 N°5

[SAI 96] SAIN S., SCOTT W., *On Locally Adaptative Density EStimation*, Journal of the American Statistical Association Vol 91 N°436

.[STU 03] STUETZLE W., *Estimation of the cluster tree of a density by analysing the minimal spanning tree of a sample*, http://www.stat.washington.edu/wxs/Learning-papers/mst.pdf

[WIS 69] WISHART D., *Mode Analysis : A Generalization of Nearest Neighbor which Reduce Chaining Effect*, Numerical Taxinomy, Ed A.J. Cole, Academic Press,282-311