Renaud Sirdey

Régression flexible

Éditions des Nik's News
www.niksnews.com/editions/

1998

Introduction

Le but de ce travail est de produire des estimations de fonctions de régression par les méthodes

1. de Nadaraya-Watson (estimateur de Parzen) ;
2. de régression spline ;
3. de régression par réseaux de neurones multicouches.

Les méthodes de régressions flexibles se veulent fournir une solution au problème d'estimation de la fonction $y = f(x)$ qui lie l'entrée et la sortie d'un système. L'intérêt de ces méthodes est qu'elles sont prédictives sans être accompagnées de la « lourdeur » d'une modélisation physique.

1 Méthode de Nadaraya-Watson

1.1 L'estimateur de Parzen

Moyennant l'utilisation de l'ensemble d'apprentissage $\{x_i, y_i\}_{i=1,\cdots,N}$, l'estimateur de Parzen de la fonction f s'écrit

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} y_i K\left(\frac{(x-x_i)^2}{b}\right)}{\sum_{i=1}^{N} K\left(\frac{(x-x_i)^2}{b}\right)}$$

Où K est une fonction noyau. On choisira par exemple

$$K(\xi) = \frac{1}{\sqrt{\pi b}} e^{-\xi}$$

Où $\xi = \frac{(x-xi)^2}{b}$.

Cette dernière expression correspond à une gaussienne centrée de variance $\frac{b}{2}$. En fonction de b l'estimateur aura donc le comportement suivant

- si b est grand, les voisins éloignés auront une contribution non négligeable et la courbe aura tendance à être lisse ;
- dans le cas contraire, la courbe sera irrégulière.

1.2 Choix de b

L'idée consiste à trouver la valeur de b qui minimise un certain critère de qualité pour $\hat{f}$, on choisit l'erreur quadratique moyenne

$$E\left[(y - \hat{f}(x))^2\right]$$

On utilise alors un ensemble d'apprentissage et un ensemble de test [1]. La méthode devient « choisir b sur l'ensemble d'apprentissage tel que l'erreur soit minimale sur l'ensemble de test ».
L'erreur s'estime de la manière suivante [2]

$$\epsilon(b) = \sum_{i=1}^{M} (\tilde{y}_i - \hat{f}(\tilde{x}_i))^2 \tag{1}$$

Où les $\{\tilde{x}_i, \tilde{y}_i\}_{i=1,\cdots,M}$ représentent les éléments de l'ensemble de test.

Dans le cas des données `motorcycle` on procède à une minimisation manuelle du critère de qualité, les valeurs essayées sont $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4, 10^5\}$ puis $\{1, \cdots, 10\}$ qui semble être une zone pertinente. La figure 1 présente les erreurs calculées pour les différentes valeurs de b testées ainsi que l'estimateur de Parzen, pour $b^* = 5$ qui réalise l'erreur estimée minimale.

2 Régression spline

2.1 Estimateur

Moyennant l'utilisation de l'ensemble d'apprentissage $\{x_i, y_i\}_{i=1,\cdots,N}$, l'estimateur par régression spline de la fonction f s'écrit

$$\hat{f}(x) = \sum_{i=1}^{N} \alpha_i(\lambda) G\left(\frac{(x - x_i)^2}{b}\right) + \sum_{j=1}^{d} \beta_j H_j(x)$$

Où

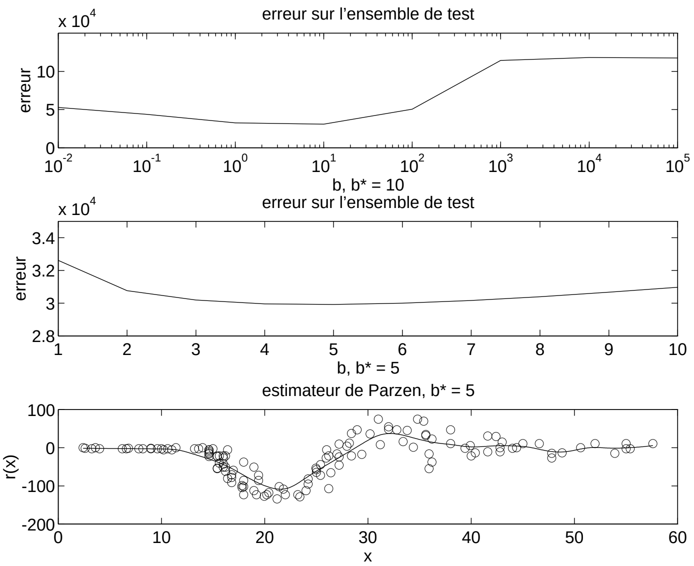
- λ [3] est un paramètre d'attache aux données ;

[1] Dans le cas particulier des données `motorcycle` les ensembles d'apprentissage et de test correspondent respectivement à 66% et 33% du jeu de données initial.

[2] On n'a pas besoin du terme $\frac{1}{N}$ pour procéder à une minimisation.

[3] Résultat de l'introduction, par la méthode des multiplicateurs de Lagrange, d'une contrainte sur la dérivée des fonctions utilisées.

FIG. 1 – Courbes d'erreur et estimateur de Parzen pour les données `motorcycle`.



- b est un paramètre de lissage (qui se comporte de la même façon que pour la méthode de Nadaraya-Watson) ;
- $G(\xi)$ est une fonction noyau (par exemple $e^{-\xi}$) ;
- $H_{j,j=1,\cdots,d}$ données (par exemple, $d = 1$ et $H_1(x) = k$).

Les $\alpha_{i,i=1,\cdots,N}$ et les $\beta_{j,j=1,\cdots,d}$ sont déterminés par résolution d'un système linéaire.

Finalement, on utilise le modèle simplifié suivant

$$\hat{r}(x) = \sum_{i=1}^{N} \alpha_i G\left(\frac{(x - x_i)^2}{b}\right) + \beta$$

Et l'on a deux paramètres à choisir, à savoir λ et b.

2.2 Choix des paramètres λ et b

Comme dans le cas de la méthode de Nadaraya-Watson, on utilise un ensemble d'apprentissage pour procéder à l'estimation et un ensemble de test pour mesurer sa qualité.

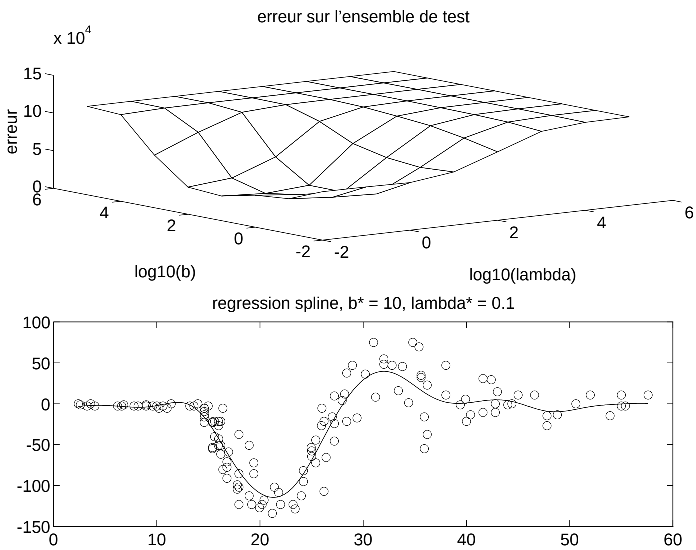
Là encore, on procède à une minimisation manuelle. On essaye les différents couples possibles pour

$$b = \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$$
$$\lambda = \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$$

La figure 2 présente la surface d'erreur en fonction de b et de λ ainsi que l'estimation par régression spline pour le couple $(b* = 10, \lambda* = 10^{-1})$ qui réalise le minimum de l'erreur estimée.

On note que dans le cadre d'utilisation des splines incomplets on a la possibilité de réduire le nombre de fonctions utilisées à $m << M$. Ceci est intéressant en terme de temps de calcul et au niveau théorique, puisqu'utiliser moins de fonctions augmente la régularité de la solution. Dans ce cas, le paramètre m doit aussi être choisi afin de fournir la meilleure estimation au sens du critère (1).

FIG. **2** – Surface d'erreur et régression spline pour les données `motorcycle`.



3 Perceptron multicouche

3.1 Estimateur

Le cadre d'utilisation d'un réseau de neurone multicouche, fournit une généralisation du modèle des splines incomplets. En effet

1. on ne se limite plus à des fonctions centrées sur les points de l'ensemble d'apprentissage ;
2. chaque fonction posséde son propre paramètre de régularité $b_{i,i=1,\cdots,m<<M}$.

Le modèle devient donc le suivant

$$\hat{r}(x) = \sum_{i=1}^{m<<M} \alpha_i G\left(\frac{x-c_i}{b_i}\right) + \beta$$

Le principal problème est que l'on a rompu avec la linéarité et que l'optimisation des paramètres doit se faire en utilisant une méthode de gradient.

Dans le cadre de ce travail, nous nous sommes contentés de choisir le nombre de neurones de la couche cachée du réseau.

3.2 Choix de m

Comme dans les cas précédents on procède à une minimisation manuelle du critère. L'expérience montre que l'erreur estimée est minimale pour $m^* = 4$ neurones[4]. La figure 3 présente la courbe d'erreur en fonction de m ainsi que l'estimation réalisée pour $m^* = 4$.

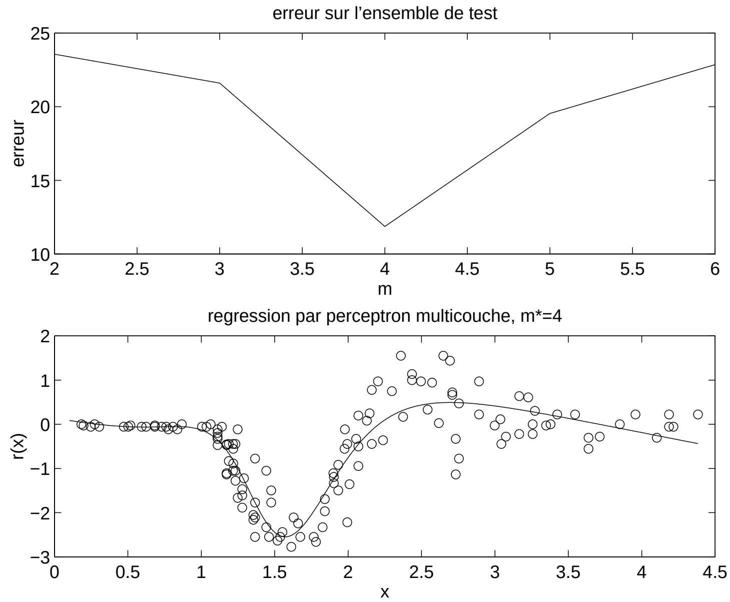
Conclusion

Les exemples d'estimation présentés précédemment montrent la puissance des méthodes de régression flexible. Cependant, l'optimisation du critère de qualité est restée un peu primaire, puisque manuelle. Il serait intéressant de développer des méthodes d'optimisation systématique des paramètres[5] basées sur un algorithme de type gradient et permettant d'obtenir un minimum (local?).

[4] L'ensemble des valeurs testées est $m = \{2, 3, 4, 5, 6\}$.

[5] Ce qui est au moins possible dans le cas de l'estimateur de Parzen.

FIG. 3 – Courbe d'erreur et régression par perceptron multicouche pour les données `motorcycle`.



En terme de qualité, on notera que l'erreur minimale est, pour les trois méthodes, du même ordre de grandeur [6]. Cependant, le fait d'introduire plus de paramètres (splines, perceptron) permet d'avoir plus de degrés de liberté mais il est nécessaire de contrôler la complexité du modèle.

[6]. Attention, les valeurs obtenues dans le cas de l'estimation par perceptron multicouche, sont dues au fait que les données ont été normalisées.