Interpolation et Régression

Pierre Barbillon

21 avril 2008

Introduction

Le problème posé est de reconstituer une fonction à partir de la seule connaissance de ses valeurs en un nombre limité de points. L'interpolation et la regression permettent de répondre à ce problème.
L'interpolation consiste à trouver une fonction passant exactement par les points à notre disposition tandis que la régression ajuste au mieux une fonction simple à ces points.

Dans un premier temps, nous essayerons de trouver des méthodes d'interpolation en partant d'exemples simples que nous chercherons à généraliser. Ensuite, nous verrons quel sens donner à ajuster au mieux dans le cadre d'une régression linéaire et comment l'obtenir. Nous garderons à l'esprit que ces deux méthodes visent au même but mais dans des contextes différents.

1 Interpolation

Soit f une fonction inconnue

$$\begin{aligned} f : \mathbb{R} &\rightarrow \mathbb{R} \\ f(x) &= y \end{aligned} \tag{1.1}$$

Nous connaissons en certains points $(x_1, x_2 \cdots x_N)$, les valeurs de f :

$$y_1 = f(x_1),\ y_2 = f(x_2), \cdots y_N = f(x_N)$$

Nous souhaiterions trouver $g : \mathbb{R} \to \mathbb{R}$ telle que

$$\left\{ \begin{array}{rcccl} y_1 & = & f(x_1) & = & g(x_1) \\ y_2 & = & f(x_2) & = & g(x_2) \\ & & \cdots & & \\ y_N & = & f(x_N) & = & g(x_N) \end{array} \right.$$

1.1 Méthodes polynomiales

Commençons par un problème simple, comment interpoler deux points de coordonnées (x_1, y_1), (x_2, y_2) ?
Voici la forme d'une fonction affine, il nous reste à trouver les coefficients a et b pour deux points donnés.

$$g(x) = a \cdot x + b$$

En résolvant le système,

$$\left\{ \begin{array}{rcl} y_1 & = & a \cdot x_1 + b \\ y_2 & = & a \cdot x_2 + b \end{array} \right.$$

nous obtenons

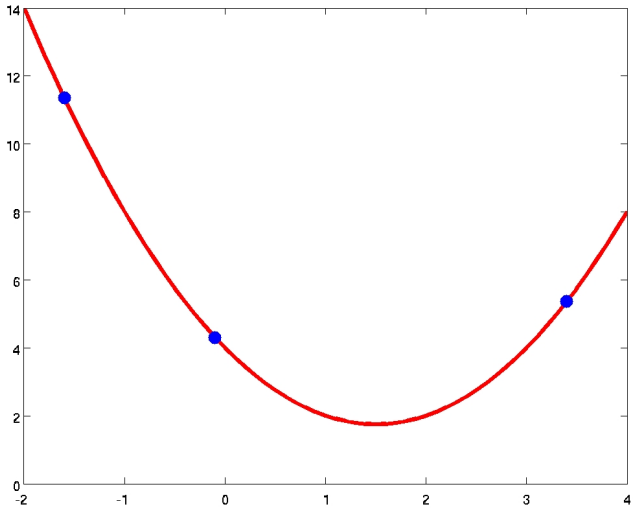
$$\begin{array}{rcl} a & = & \dfrac{y_2 - y_1}{x_2 - x_1} \\ b & = & y_1 - a \cdot x_1 \end{array}$$

Et trois ?
S'ils sont non-alignés, il n'est pas possible d'utiliser une droite. Il faut complexifier notre fonction par un degré polynomial plus élevé.

$$g(x) = a \cdot x^2 + b \cdot x + c$$

Maintenant il faut résoudre le système à 3 équations et 3 inconnues (a,b,c) suivant :

$$\left\{ \begin{array}{rcl} y_1 & = & a \cdot x_1^2 + b \cdot x_1 + c \\ y_2 & = & a \cdot x_2^2 + b \cdot x_2 + c \\ y_3 & = & a \cdot x_3^2 + b \cdot x_3 + c \end{array} \right.$$



Interpolation de 3 points par une parabole

Que penser si l'on souhaite interpoler N points ?

Polynôme de Lagrange Pour nos N points, $(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)$

Définition : Le polynôme de Lagrange associé au point (x_i, y_i) est

$$l_i(x) = \frac{x - x_1}{x_i - x_1} \cdots \frac{x - x_{i-1}}{x_i - x_{i-1}} \cdot \frac{x - x_{i+1}}{x_i - x_{i+1}} \cdots \frac{x - x_N}{x_i - x_N}$$

Les l_i sont de degré $N - 1$.
On peut vérifier qu'on a $l_i(x_i) = 1$ et $l_i(x_j) = 0$.

Grâce à ces polynômes, nous pouvons interpoler tout ensemble de N points par un polynôme de degré $N - 1$. Et nous avons même plus...

Théorème : Le polynôme $L(x) = y_1 \cdot l_1(x) + y_2 \cdot l_2(x) + \cdots y_N \cdot l_N(x)$ est l'**unique** polynôme de degré au plus $N - 1$ vérifiant $L(x_i) = y_i$.

On peut vérifier qu'on retrouve bien les interpolateurs calculés dans le cadre

de nos exemples.

1.2 Autres méthodes

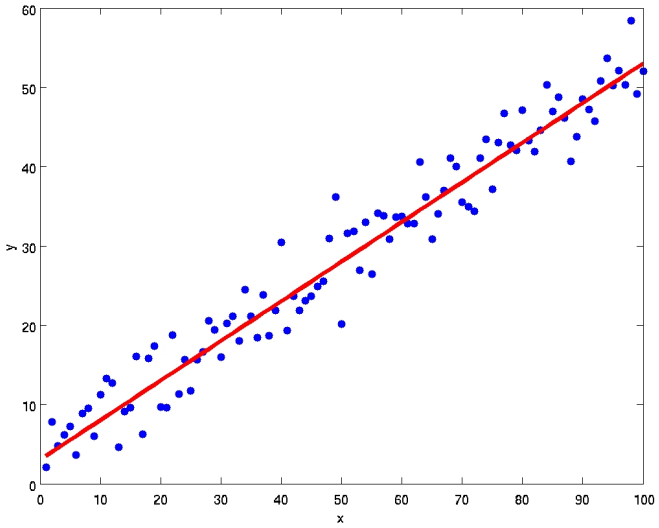
Il existe des méthodes dites des plus proches voisins qui consistent à approximer un point par la moyenne d'un nombre choisi de points voisins.
Il est possible de pondérer cette moyenne par la distance aux points. Ce qui en pratique, si nous choisissons de considérer les deux plus proches voisins, nous ramène à une interpolation linéaire entre les deux points les plus proches. La solution sera une fonction affine par morceaux.

2 Régression

Nous disposons toujours des points

$$(x_1, y_1 = f(x_1)), (x_2, y_2 = f(x_2)) \cdots (x_N, y_N = f(x_N))$$

Nous ne souhaitons plus ici passer exactement par ces points, mais en être "proche" et avoir une fonction g simple (une fonction affine par exemple).
Cette problématique est sensée notamment parce qu'il arrive souvent, en pratique, que les observations des évaluations soient bruitées. C'est à dire que nous observons l'évaluation $f(x_i)$ + une petite erreur. Celle-ci peut provenir d'une imprécision de mesure par exemple.



Un exemple de régression linéaire

S'il on trace deux droites passant à peu près par les points, il faut se donner un critère permettant de décider laquelle est la mieux ajustée. Ce critère doit mesurer **l'écart entre la droite et les points**. Ainsi la droite rendant minimum un tel critère est la mieux ajustée.

Quel critère proposer ?
Pour une fonction g,

$$\begin{aligned} R_1(g) &= |y_1 - g(x_1)| + |y_2 - g(x_2)| + \cdots |y_N - g(x_N)| \\ R_2(g) &= (y_1 - g(x_1))^2 + (y_2 - g(x_2))^2 + \cdots (y_N - g(x_N))^2 \end{aligned}$$

Le deuxième critère est souvent celui qu'on retient. La fonction affine g le minimisant est appelée droite des moindres carrés.
Il nous reste à l'expliciter.
g s'écrit

$$g(x) = a \cdot x + b$$

Il faut trouver a et b rendant minimum le critère R_2.

$$R_2(a,b) = (y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + \cdots (y_N - ax_N - b)^2$$

Comment trouver a et b ? Voici une méthode
Développons les polynômes et ordonnons les en b :

$$\begin{aligned} R_2(a,b) &= ((y_1 - ax_1) - b)^2 + ((y_2 - ax_2) - b)^2 + \cdots ((y_N - ax_N) - b)^2 \\ &= Nb^2 - 2((y_1 - ax_1) + \cdots (y_N - ax_N))b + (y_1 - ax_1)^2 + \cdots (y_N - ax_N)^2 \end{aligned}$$

Nous considèrons la fonction $b \mapsto R(a,b)$. Afin d'obtenir son minimum, il nous faut trouver où sa dérivée s'annule.
Il suffit de faire le calcul. On cherche b telle que la dérivée est nulle

$$2Nb - 2((y_1 - ax_1) + \cdots (y_N - ax_N)) = 0$$

d'où

$$b = \frac{1}{N}((y_1 - ax_1) + \cdots (y_N - ax_N)) = \overline{y} - a\overline{x}$$

on a posé $\overline{y} = \frac{1}{N}(y_1 + \cdots y_N)$ et $\overline{x} = \frac{1}{N}(x_1 + \cdots x_N)$.
En réinjectant cela dans $R_2(a,b)$,

$$\begin{aligned} R_2(a,b) &= ((y_1 - ax_1) - b)^2 + \cdots ((y_N - ax_N) - b)^2 \\ &= ((y_1 - ax_1) - \overline{y} - a\overline{x})^2 + \cdots ((y_N - ax_N) - \overline{y} - a\overline{x})^2 \\ &= ((y_1 - \overline{y}) - a(x_1 - \overline{x}))^2 + \cdots ((y_N - \overline{y}) - a(x_N - \overline{x}))^2 \\ R_2(a,b) &= S(a) \end{aligned}$$

Nous avons un polynôme qui ne dépend plus que de a, nous développons et nous ordonnons en a comme nous avons fait précédemment avec b

$$\begin{aligned} S(a) = {} & ((x_1 - \overline{x})^2 + \cdots (x_N - \overline{x})^2)a^2 - 2((x_1 - \overline{x})(y_1 - \overline{y}) + \cdots (x_N - \overline{x})(y_N - \overline{y}))a \\ & + ((y_1 - \overline{y})^2 + \cdots (y_N - \overline{y})^2) \end{aligned}$$

En dérivant par rapport à a, et en cherchant où la dérivée est nulle

$$S'(a) = 2((x_1 - \overline{x})^2 + \cdots (x_N - \overline{x})^2)a - 2((x_1 - \overline{x})(y_1 - \overline{y}) + \cdots (x_N - \overline{x})(y_N - \overline{y})) = 0$$

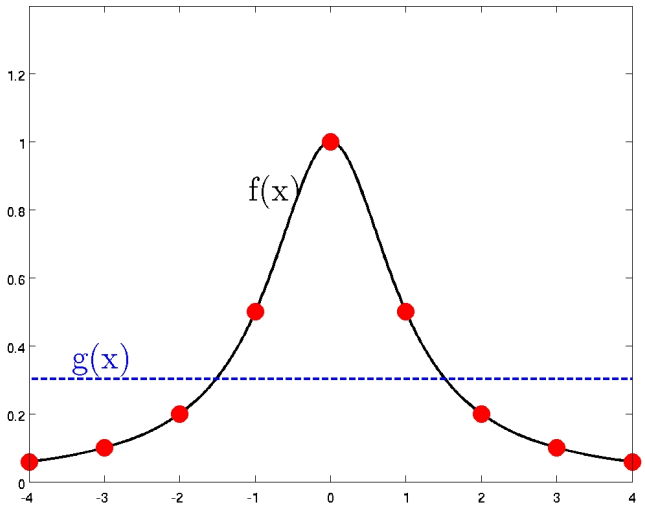
Ainsi

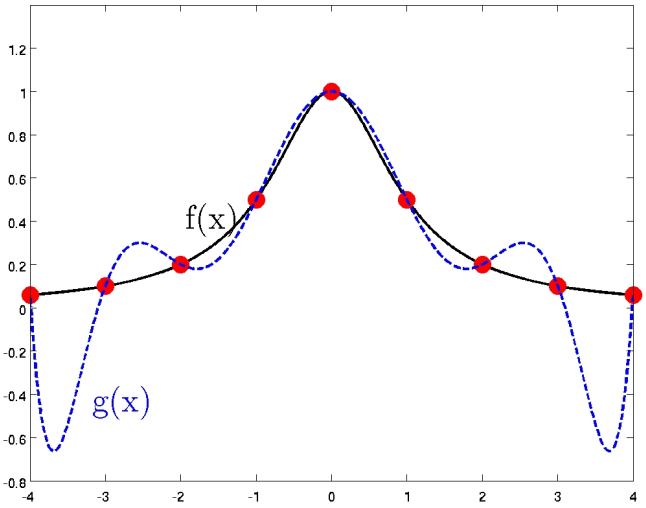
$$a = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots (x_N - \overline{x})(y_N - \overline{y})}{(x_1 - \overline{x})^2 + \cdots (x_N - \overline{x})^2}$$

Nous avons donc une écriture explicite des coefficients a et b de la droite des moindres carrés.

Conclusion

Maintenant que nous nous sommes familiarisés avec les notions d'interpolation et de régression, il est intéressant de se poser la question de quelle méthode utiliser dans quelles circonstances.
Illustrons notre propos par un graphique,

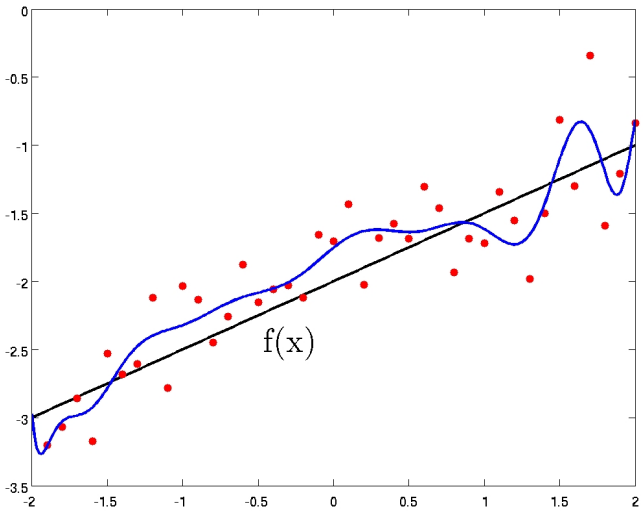




Nous retrouvons notre fonction à reconstituer f en trait plein dont nous connaissons quelques évaluations (ces points sont représentés par des cercles), et notre fonction g en pointillés. Sur le dessin du haut g est la droite des moindres carrés pour une régression linéaire et sur le dessin du bas g est un interpolateur polynomial de degré 8 (puisqu'il y a 9 points).
L'on se rend compte que les deux sont mauvais car l'un est trop simple, la vraie fonction ne peut s'approcher par une droite et l'autre est trop complexe. Le fait d'interpoler les points induits de trop grande variations du polynôme entre ces points. Ceci s'appelle le phénomène de Runge et on peut montrer

qu'il est préfèrable de prendre un polynôme de degré intermédiaire.

A présent considérons l'exemple où nous disposons de données bruitées. Nous observons des observations d'une fonction affine f (représentée en noir ci-dessous) bruitées par une petite perturbation.



Nous avons représenté en bleu, une approximation par un polynôme de degré 15 que nous avons essayé d'ajuster à nos données. Cette approximation est mauvaise car elle dépend trop de nos observations. C'est à dire que si nous effectuons d'autres observations le polynôme de degré 15 sera très différent de celui obtenu. Nous obtenons quasiment la droite f en ajustant la droite des moindres carrés et ce peu importe le jeu de données employé.

Par nos exemples, nous avons appréhendé le dilemme fondamental de la statistique : *le compromis biais-variance*. En effet, dans les deux exemples précédents, prendre un modèle trop complexe (ici un degré polynomiale trop grand) induit une trop grande dépendance en les observations ce qui donnera une variance trop grande tandis qu'un modèle trop simple induira un biais important puisque notre modèle ne nous permettra pas de rendre compte des observations (la droite dans le premier exemple).