1 French version

1.1 Méthode des moindres carrés pour l'ajustement de données

La méthode des moindres carrés est utilisée lorsque l'on souhaite ajuster une fonction à un ensemble de données. Plus précisément, nous considérons une famille de fonctions F dont les éléments f_a dépendent d'un paramètre a appartenant à R^d. Par exemple,

$$F = \{f_a : R \to R; x \mapsto ax\}$$

est l'ensemble des fonctions linéaires paramétrées par un paramètre réel $a \in R$.

Nous disposons d'un ensemble de points de données $(\{(x_i, y_i)\}_{1 \leq i \leq n})$ tels que

$$\forall i \in \{1, \ldots, n\}, \quad y_i = f_{a_0}(x_i) + \epsilon_i$$

où $x_i \in R^p$ et $y_i \in R$ pour un certain paramètre a_0. La présence de ϵ_i peut être due au bruit d'acquisition (arrondi numérique, erreur ou imprécision) et/ou à l'approximation dans la modélisation.

Dans de nombreuses applications, nous cherchons à trouver la meilleure fonction de F qui modélise la génération de données. Autrement dit, étant donné l'ensemble de données $u = \{(x_i, y_i)\}_{1 \leq i \leq n}$, quelle est la fonction $f_a \in F$ pour laquelle le nuage de points $v_{f_a} = \{(x_i, f_a(x_i))\}_{1 \leq i \leq n}$ correspond le mieux aux données initiales u?

Ces deux ensembles de points appartiennent à un espace euclidien, donc nous pouvons mesurer la distance entre eux en utilisant la distance de Frobenius, écrite comme

$$\|v_{f_a} - u\|_F = \sum_{i=1}^{n} \|f_a(x_i) - y_i\|_2^2$$

Le problème d'ajustement peut ainsi être écrit sous la forme de problème d'optimisation suivant :

$$\min_{f_a \in F} \sum_{i=1}^{n} \|f_a(x_i) - y_i\|_2^2$$

ce qui revient à minimiser l'erreur quadratique moyenne entre les données y_i et le modèle théorique $f_a(x_i)$; d'où le nom "méthode des moindres carrés". Puisque la famille F est paramétrée par le vecteur $a \in R^d$, la forme finale du problème d'ajustement devient

$[\min_{a \in R^d} \sum_{i=1}^{n} \|f_a(x_i) - y_i\|_2^2$.

1.2 Loi d'Ohm et ajustement linéaire

Prenons l'exemple de la loi d'Ohm, qui stipule qu'il existe une relation linéaire entre l'intensité I (Ampère) du courant électrique qui traverse un conducteur et la tension U (Volt) à ses bornes :

$$U = RI$$

Le rapport R (constant) entre la tension et l'intensité est appelé résistance (Ohm). Afin d'estimer la valeur de la résistance d'un conducteur donné, on peut faire passer un courant d'une intensité connue I_0 à travers le conducteur, puis mesurer la tension U_0 associée ; enfin, faire le rapport et en déduire une valeur de résistance. Si l'on réitère l'expérience avec une autre valeur d'intensité I_1 connue, on devrait trouver la même valeur de résistance. En pratique, ce n'est pas le cas (et ce, même si $I_0 = I_1$). De nombreux facteurs expliquent une telle différence : le voltmètre utilisé n'est pas suffisamment précis, l'affichage tronque la valeur de la tension, la loi d'Ohm n'est qu'une loi théorique qui approche le comportement réel du conducteur... Ainsi, si on fait une série de n mesures de tension U_i associées à n valeurs d'intensité connues I_i, on a seulement

$$\forall i = 1, \dots, n, \quad U_i = RI_i$$

Pour estimer la valeur de la résistance, nous allons procéder à un ajustement linéaire sur les données $\{(I_i, U_i)\}_{1 \leq i \leq n}$, pour lequel on peut appliquer la méthode des moindres carrés.

2 English version

2.1 Least Squares Method for Data Fitting

The least squares method is used when one wants to fit a function to a set of data. More precisely, we consider a family of functions F whose elements f_a depend on a parameter a belonging to R^d. For example,

$$F = \{f_a : R \to R; x \mapsto ax\}$$

is the set of linear functions parameterized by a real parameter $a \in R$.
We have a set of data points $(\{(x_i, y_i)\}_{1 \leq i \leq n})$ such that

$$\forall i \in \{1, \dots, n\}, \quad y_i = f_{a_0}(x_i) + \epsilon_i$$

where $x_i \in R^p$ and $y_i \in R$ for some parameter a_0. The presence of ϵ_i may be due to acquisition noise (round-off, error, or inaccuracy) and/or approximation in modeling.

In many applications, we seek to find the best function from F that models the data generation. In other words, given the dataset $u = \{(x_i, y_i)\}_{1 \leq i \leq n}$, what is the function $f_a \in F$ for which the point cloud $v_{f_a} = \{(x_i, f_a(x_i))\}_{1 \leq i \leq n}$ best matches the initial data u?

These two sets of points belong to a Euclidean space, so we can measure the distance between them using the Frobenius distance, written as

$$\|v_{f_a} - u\|_F = \sum_{i=1}^{n} \|f_a(x_i) - y_i\|_2^2$$

The fitting problem can thus be written in the form of the following optimization problem:

$$\min_{f_a \in F} \sum_{i=1}^{n} \|f_a(x_i) - y_i\|_2^2$$

which amounts to minimizing the mean square error between the data y_i and the theoretical model $f_a(x_i)$; hence the name "least squares method". Since the family F is parameterized by the vector $a \in R^d$, the final form of the fitting problem becomes

$[\min_{a \in R^d} \sum_{i=1}^{n} \|f_a(x_i) - y_i\|_2^2$.

2.2 Law of Ohm and Linear Fitting

Let's take the example of Ohm's law, which states that there exists a linear relationship between the current intensity I (Amperes) flowing through a conductor and the voltage U (Volts) across it:

$$U = RI$$

The ratio R (constant) between voltage and current is called resistance (Ohm). To estimate the value of the resistance of a given conductor, one can pass a current of known intensity I_0 through the conductor, then measure the associated voltage U_0; finally, take the ratio and deduce a resistance value. If we repeat the experiment with another known intensity value I_1, we should find the same resistance value. In practice, this is not the case (even if $I_0 = I_1$). Many factors account for such a difference: the voltmeter used is not precise enough, the display truncates the voltage value, Ohm's law is only a theoretical law that approximates the real behavior of the conductor... Thus, if we make a series of n measurements of voltage U_i associated with n known current values I_i, we have only

$$\forall i = 1, \ldots, n, \quad U_i = RI_i$$

To estimate the value of the resistance, we will proceed with a linear fitting on the data $\{(I_i, U_i)\}_{1 \leq i \leq n}$, for which we can apply the least squares method.