最小二乗法

大枝 真一

2011 年 4 月 8 日

1 はじめに

複雑なデータや関数を簡単な関数の和で近似する代表的な手法が「最小二乗法」である．これはコンピュータによるデータ解析の最も重要な基礎である．本授業ではこれを学ぶと共に，ベクトルや行列による線形計算に慣れることを目的とする．

2 最小二乗法とは

N 個のデータ $(x_1, y_1), \ldots, (x_N, y_N)$ に直線を当てはめたいとする．当てはめたい直線を $y = ax + b$ と置く．a, b はこれから定める未知の定数である．

理想的には $y_\alpha = ax_\alpha + b, \alpha = 1, \ldots, N$，となることが望ましいが，データ点 (x_α, y_α) が厳密に同一直線上にあるとは限らないので，a, b をどう選んでも多くの α に対して $y_\alpha \neq ax_\alpha + b$ となる．そこで

$$y_\alpha \approx ax_\alpha + b, \qquad \alpha = 1, \ldots, N \tag{1}$$

となるように a, b を定める (図 1)．記号 $\approx$ は「ほぼ等しい」という意味である．これを次のように解釈する．ただし $\to$ はその左側の式を最小にすることを表す．

$$J = \frac{1}{2} \sum_{\alpha=1}^{N} (y_\alpha - (ax_\alpha + b))^2 \to \min \tag{2}$$

これは食い違いの二乗の和を最小にする方法であることから，**最小二乗法**と呼ばれている．全体を 1/2 倍するのは後の計算を見やすくするためである．

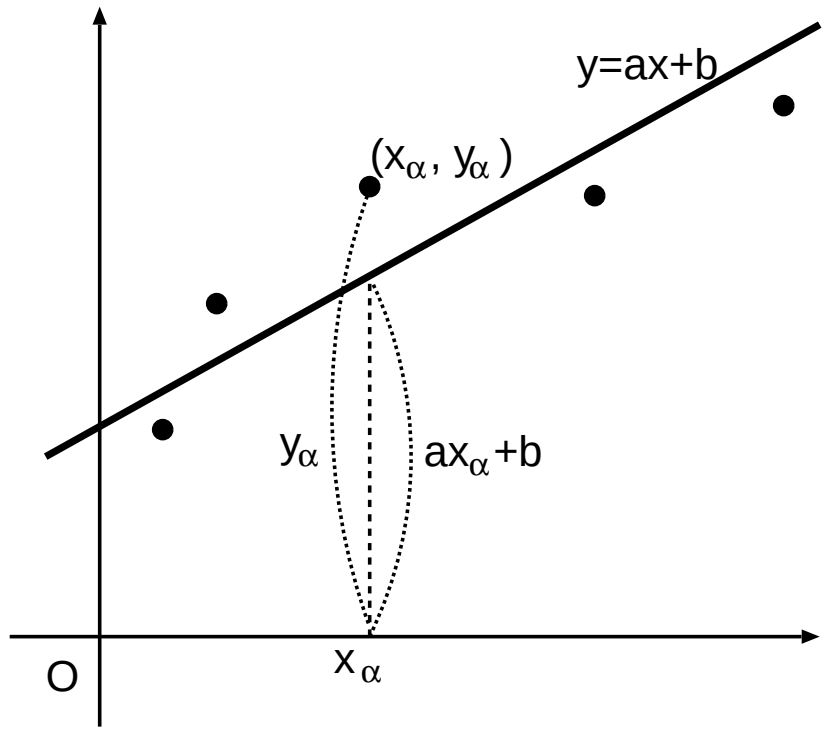

図 1　直線の当てはめ

2.1　1 次の最小二乗法の正規方程式

N 個のデータ $(x_1, y_1), \ldots, (x_N, y_N)$ に直線 $y = ax + b$ を当てはめよ.

(解)

式 (2) は a, b の関数である. 解析学で知られるように多変数の関数が最大値や最小値をとる点では，各変数に関する偏導関数が 0 でなければならない．したがって,

$$\frac{\partial J}{\partial a} = 0, \qquad \frac{\partial J}{\partial b} = 0 \tag{3}$$

を解いて a, b を定めればよい．式 (2) を a, b でそれぞれ偏微分すると次式を得る.

$$\frac{\partial J}{\partial a} = \sum_{\alpha=1}^{N} (y_\alpha - a x_\alpha - b)(-x_\alpha) = a \sum_{\alpha=1}^{N} x_\alpha^2 + b \sum_{\alpha=1}^{N} x_\alpha - \sum_{\alpha=1}^{N} x_\alpha y_\alpha = 0$$

$$\frac{\partial J}{\partial b} = \sum_{\alpha=1}^{N}(y_\alpha - ax_\alpha - b)(-1) = a\sum_{\alpha=1}^{N} x_\alpha + b\sum_{\alpha=1}^{N} 1 - \sum_{\alpha=1}^{N} y_\alpha = 0 \tag{4}$$

これから次の連立 1 次方程式を得る.

$$\begin{pmatrix} \sum_{\alpha=1}^{N} x_\alpha^2 & \sum_{\alpha=1}^{N} x_\alpha \\ \sum_{\alpha=1}^{N} x_\alpha & \sum_{\alpha=1}^{N} 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{\alpha=1}^{N} x_\alpha y_\alpha \\ \sum_{\alpha=1}^{N} y_\alpha \end{pmatrix} \tag{5}$$

これを**正規方程式**と呼ぶ. これを解いて a, b が定まる.

2.2 2 次の最小二乗法の正規方程式

N 個のデータ $(x_1, y_1), \ldots, (x_N, y_N)$ に 2 次式 $y = ax^2 + bx + c$ を当てはめよ.

(解) 当てはめる 2 次式を $y = ax^2 + bx + c$ とし,

$$y_\alpha \approx ax_\alpha^2 + bx + c, \qquad \alpha = 1, \ldots, N \tag{6}$$

となる a, b, c を最小二乗法

$$J = \frac{1}{2}\sum_{\alpha=1}^{N}(y_\alpha - (ax_\alpha^2 + bx + c))^2 \to \min \tag{7}$$

によって定める. それには

$$\frac{\partial J}{\partial a} = 0, \qquad \frac{\partial J}{\partial b} = 0, \qquad \frac{\partial J}{\partial c} = 0 \tag{8}$$

を解いて a, b, c を定めればよい. 式 (7) を a, b, c でそれぞれ偏微分すると次式を得る.

$$\frac{\partial J}{\partial a} = \sum_{\alpha=1}^{N}(y_\alpha - ax_\alpha^2 - bx - c)(-x_\alpha^2) = a\sum_{\alpha=1}^{N} x_\alpha^4 + b\sum_{\alpha=1}^{N} x_\alpha^3 + c\sum_{\alpha=1}^{N} x_\alpha^2 - \sum_{\alpha=1}^{N} x_\alpha^2 y_\alpha = 0$$

$$\frac{\partial J}{\partial b} = \sum_{\alpha=1}^{N}(y_\alpha - ax_\alpha^2 - bx_\alpha - c)(-x\alpha) = a\sum_{\alpha=1}^{N} x_\alpha^3 + b\sum_{\alpha=1}^{N} x_\alpha^2 + c\sum_{\alpha=1}^{N} x_\alpha - \sum_{\alpha=1}^{N} x_\alpha y_\alpha = 0$$

$$\frac{\partial J}{\partial c} = \sum_{\alpha=1}^{N}(y_\alpha - ax_\alpha^2 - bx_\alpha - c)(-1) = a\sum_{\alpha=1}^{N} x_\alpha^2 + b\sum_{\alpha=1}^{N} x_\alpha + c\sum_{\alpha=1}^{N} 1 - \sum_{\alpha=1}^{N} y_\alpha = 0 \tag{9}$$

これから次の連立 1 次方程式を得る.

$$\begin{pmatrix} \sum_{\alpha=1}^N x_\alpha^4 & \sum_{\alpha=1}^N x_\alpha^3 & \sum_{\alpha=1}^N x_\alpha^2 \\ \sum_{\alpha=1}^N x_\alpha^3 & \sum_{\alpha=1}^N x_\alpha^2 & \sum_{\alpha=1}^N x_\alpha \\ \sum_{\alpha=1}^N x_\alpha^2 & \sum_{\alpha=1}^N x_\alpha & \sum_{\alpha=1}^N 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{\alpha=1}^N x_\alpha^2 y_\alpha \\ \sum_{\alpha=1}^N x_\alpha y_\alpha \\ \sum_{\alpha=1}^N y_\alpha \end{pmatrix} \tag{10}$$

これを解いて，a, b, c が定まる．

2.3 n 次の最小二乗法の正規方程式

N 個のデータ $(x_1, y_1), \ldots, (x_N, y_N)$ に n 次式
$y = c_0 x^n + c_1 x^{n-1} + \ldots + c_n$ を当てはめよ．

(解) 当てはめる n 次式を $y = c_0 x^n + c_1 x^{n-1} + \ldots + c_n$ とし，

$$y_\alpha \approx c_0 x^n + c_1 x^{n-1} + \ldots + c_n, \qquad \alpha = 1, \ldots, N \tag{11}$$

となる $c_1, \ldots, c_n$ を最小二乗法

$$J = \frac{1}{2} \sum_{\alpha=1}^N (y_\alpha - (c_0 x^n + c_1 x^{n-1} + \ldots + c_n))^2 \to \min \tag{12}$$

によって定める．それには

$$\frac{\partial J}{\partial c_0} = 0, \qquad \frac{\partial J}{\partial c_1} = 0, \qquad \ldots, \qquad \frac{\partial J}{\partial c_n} = 0 \tag{13}$$

を解いて $c_1, \ldots, c_n$ を定めればよい．式 (12) を c_k で偏微分すると次式を得る．

$$\frac{\partial J}{\partial c_k} = \sum_{\alpha=1}^N (y_\alpha - c_0 x_\alpha^n - c_1 x_\alpha^{n-1} - \cdots - c_n)(-x_\alpha^{n-k})$$

$$= c_0 \sum_{\alpha=1}^N x_\alpha^{2n-k} + c_1 \sum_{\alpha=1}^N x_\alpha^{2n-k-1} \cdots + c_n \sum_{\alpha=1}^N x_\alpha^{n-k} - \sum_{\alpha=1}^N x_\alpha^{n-k} y_\alpha \tag{14}$$

これを 0 と置いて $k = 0, 1, \ldots, n$ に対する式を並べると次の正規方程式を得る．

$$\begin{pmatrix} \sum_{\alpha=1}^N x_\alpha^{2n} & \sum_{\alpha=1}^N x_\alpha^{2n-1} & \cdots & \sum_{\alpha=1}^N x_\alpha^n \\ \sum_{\alpha=1}^N x_\alpha^{2n-1} & \sum_{\alpha=1}^N x_\alpha^{2n-2} & \cdots & \sum_{\alpha=1}^N x_\alpha^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{\alpha=1}^N x_\alpha^n & \sum_{\alpha=1}^N x_\alpha^{n-1} & \cdots & \sum_{\alpha=1}^N 1 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} \sum_{\alpha=1}^N x_\alpha^n y_\alpha \\ \sum_{\alpha=1}^N x_\alpha^{n-1} y_\alpha \\ \vdots \\ \sum_{\alpha=1}^N y_\alpha \end{pmatrix} \tag{15}$$

これを解いて，$c_1, \ldots, c_n$ が定まる．

3 演習課題

サンプルプログラムとサンプルデータは，/home/jugyou/j5/IntelligentSystem/2010/LeastSquaresMethod/LeastSquaresMethod.tar.gz にある．

問題 1

4 点 (4,-17), (15,-4), (30,-7), (100,50) に直線を当てはめよ．

(サンプルデータは，example1.dat になる．)

問題 2

ある実験によると，果実 A の直径 x(cm) と水分の含有率 $y(\%)$ に次の関係があった．これから直径が 5.7cm，6.5cm の果実の水分の含有率がどのように推定されるか求めよ．

x	5.6	5.8	6.0	6.2	6.4	6.4	6.4	6.6	6.8
y	30	26	33	31	33	35	37	36	33

(サンプルデータは，example2.dat になる．)

問題 3

makeData.c によって，サンプルデータを生成させ，このデータに直線を当てはめよ．

(サンプルデータは，example3.dat とする．)

問題 4

サンプルデータ example4.dat に 1 次方程式を当てはめよ．

発展問題 1

example4.dat に 2 次方程式を当てはめよ．

発展問題 2

サンプルデータ example4.dat に 1 次方程式を当てはめた場合と，2 次方程式を当てはめた場合で，どちらが誤差が少ないか検討しなさい．

発展問題 3

サンプルデータ example5.dat に 1 次方程式を当てはめた場合と，2 次方程式を当てはめた場合で，どちらが誤差が少ないか検討しなさい．

あるいは，もっと誤差の少ない方程式があるなら，それを求めなさい．

ヒント

時間内にすべてのプログラムを作成する自信のない人は，サンプルプログラムを用意したので，これを利用しても構わない．

1．サンプルプログラムを持って来る．

```
$ cp /home/jugyou/j5/IntelligentSystem/2010/LeastSquaresMethod/lsm.c .
$ cp /home/jugyou/j5/IntelligentSystem/2010/LeastSquaresMethod/example1.dat .
```

2．サンプルプログラムを動作させる．

```
$ gcc lsm.c -lm
% ./a.out 4 example1.dat
```

3．実際に `lsm.c` のソースコードを読む．

41 行目からの `lsm` 関数に手を加えれば良い．

参考

LeastSquaresMethod.tar.gz に各種ファイルを用意した．展開して使用して構わない．makeData.c では，targetFunction 関数で生成したデータにノイズとして，正規乱数を加えている．その正規乱数を発生するサンプルプログラムは，Box-MullerSample にある．このプログラムは，ボックスミュラー法と呼ばれる手法であり，一様乱数と三角関数があれば，正規乱数を生成できる．

参考文献

[1] 金谷健一, 「これなら分かる応用数学教室」, 共立出版, pp.1–pp.12, (2003).
(本資料はこの抜粋になります．多少加筆修正してます．)

[2] 河西朝雄, 「C 言語によるはじめてのアルゴリズム入門」, 技術評論社, pp.54–pp.55, (1992).

課題提出締切 2010 年 4 月 26 日

学籍番号 _____　　氏名 _____

作成したプログラムを印刷して提出しなさい．また，作成したプログラムをメールで oeda@j.kisarazu.ac.jp に送信しなさい．

(全ての問題は，ひとつのプログラムで解けるはずなので，問題 1 のプログラムだけで良いです．レポートにグラフがあると，評価は高くなる．)

問題 1

直線の式を書きなさい．

問題 2

直線の式を書きなさい．また，直径が 6.5cm のときの果実の水分の含有率の推定値を書きなさい．

問題 3

直線の式を書きなさい．

発展課題

自分なりにまとめて提出してください．