アルゴリズム第 2 期末レポート

61908697 佐々木良輔

問 3 最小二乗法の拡張

最小二乗法は測定で得られた離散的なデータを関数で近似するために用いられる．ここでは特に近似を行う関数形が m 次多項式の場合を考える．すなわち関数は

$$g(x) = \sum_{k=0}^{m} a_k x^k = a_0 + a_1 x + \cdots + a_m x^m \tag{1}$$

という形で表される. ここで n 組の離散データの組が (x_i, y_i) で与えられるとき最小二乗法ではこれらのデータと $g(x)$ の残差の二乗和 E が最小となるように a_k を定める．すなわち目的関数は以下で与えられる. ただし $a_0, \ldots, a_m$ を a_k と略記している.

$$\text{minimize} : E(a_k) = \sum_{i=0}^{n-1} (y_i - g(x_i))^2 \tag{2}$$

ここで a_k について $E(a_k)$ が唯一極小値を持つことを仮定すると $E(a_k)$ が最小となる条件は

$$\begin{aligned}\frac{\partial E}{\partial a_k} &= \frac{\partial}{\partial a_k} \sum_{i=0}^{n-1} (y_i - (a_0 + \cdots + a_k x_i^k + \cdots + a_m x_i^m))^2 \\ &= 2 \sum_{i=0}^{n-1} (y_i - g(x_i))(-x_i^k) = 0\end{aligned} \tag{3}$$

すなわち

$$\sum_{i=0}^{n-1} (a_0 x_i^k + a_1 x_i^{k+1} + \cdots + a_m x_i^{k+m}) = \sum_{i=0}^{n-1} y_i x_i^k \tag{4}$$

となる．$k = 0, \ldots, m$ までを並べて行列とすると

$$\begin{pmatrix} n & \overline{x} & \cdots & \overline{x^m} \\ \overline{x} & \overline{x^2} & \cdots & \overline{x^{m+1}} \\ \vdots & \vdots & \vdots & \vdots \\ \overline{x^m} & \overline{x^{m+1}} & \cdots & \overline{x^{2m}} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{xy} \\ \vdots \\ \overline{x^m y} \end{pmatrix} \tag{5}$$

ここで $\overline{x^k y}$ などは $\sum_{i=0}^{n-1} x_i^k y_i$ を意味する. 以上から (5) 式の連立方程式を解くことで $E(a_k)$ を最小化する a_k が得られる．また m 次多項式での最小二乗法は係数行列の生成に $O(m^2 \times n)$ の総和計算, 連立方程式を解くのには $O(m^2)$ の計算量が必要と見積もられる.

以上の最小二乗法を数値計算するプログラム (lsm_polynomial.c) を c 言語で実装した. ソースコードをに示す．入力データは 1 列目に時刻, 2 列目に値を持ったテキストファイルである．また入力ファイルの 1 行 1 列にはデータ数を, 1 行 2 列には多項式の次数を与える．連立方程式のソルバーとしてガウスの消去法を用いた.

まず動作確認として, 第 10 回授業で配布された xy9.txt と同一のデータを直線回帰し, その結果を同じく第 10 回授業で配布された lsm.c の出力結果と比較した．それぞれのプログラムで得られた回帰直線は以下の通りである．この結果から 2 つの lsm_polynomial.c は lsm.c と全く同一の出力をしており, 正常に動作していると考えられる.

表 1　動作確認の結果

プログラム	回帰直線
lsm.c	$y = 12.068x - 5.011$
lsm_polynomial.c	$y = 12.068x - 5.011$