目次

前回，機械学習のアルゴリズムを「教師あり」と「なし」に大別した．教師あり学習では，個々の学習データが「入力」と「それに対する出力の正解」のペアとして与えられることを学んだ．さらに，教師あり学習の問題は，出力として量的な値を扱う「回帰」と，カテゴリを扱う「識別」に分けられることも学んだ．今回は，回帰のための教師あり学習を取り上げる．

★11　パターン認識と機械学習 (2) — 回帰のための教師あり学習

教師あり学習の目的は，学習データを用いて入力 $\boldsymbol{x}$ に対するシステムの出力 $\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{w})$ が正解に近づくようにパラメータ $\boldsymbol{w}$ を調節し，未知の入力データに対しても望ましい出力が得られるようにすることである．今回は，出力 $\boldsymbol{y}$ として量的な値を扱う問題，すなわち**回帰問題**を対象とし，その最も簡単な解法として，**最小二乗法**を取り上げる．

★11.1　最小二乗法による直線の当てはめ

回帰問題の最も簡単な形は，入力も出力も 1 次元で，入出力の関係として直線

$$y = f(x; a, b) = ax + b \tag{1}$$

を考える場合である．パラメータは a, b である．学習データとなる入力と出力の正解のペアが $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ の N 個与えられるとすると，この問題は，これら学習データに当てはまる直線を見つける問題といえる (図 1 参照)．つまり，$n = 1, 2, \ldots, N$ のそれぞれについて

$$y_n \approx ax_n + b \tag{2}$$

となるようにパラメータ a,b を定める問題である (☆ 1)．このように考えると，この問題を解くには，学習データに対する直線当てはめの「良さ」を定義して，最も良いパラメータ (a, b) を選べばよいことがわかる．このような回帰問題を含めたデータ解析の最も基本的な手法の一つである**最小二乗法** (☆ 2) では，この「良さ」の規準として，入力 x_n に対する出力 $f(x_n)$ とその正解 y_n との**二乗誤差** (☆ 3) を考える (☆ 4)．いま考えている直線当てはめ (直線回帰) の問題では，二乗誤

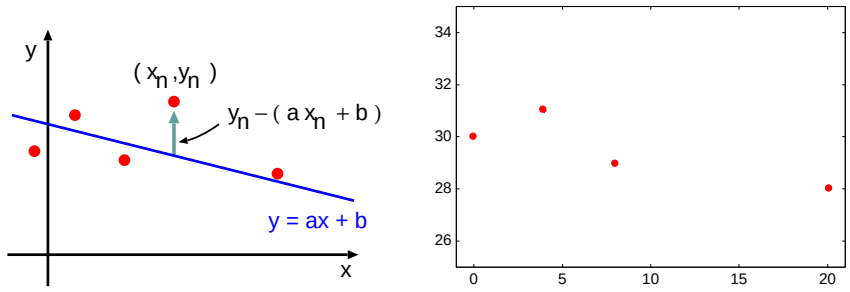

図 1: 左: 直線当てはめ，右: Q1 のデータを表した図

☆ 1) 記号 ≈ は,「近似的に等しい」という意味 .

☆ 2) 最小二乗法: least squares method. 最小自乗法とも.

☆ 3) 二乗誤差: squared error. その平均は平均二乗誤差 (mean squared error, MSE).

☆ 4) 当てはまりの「良さ」の規準に二乗誤差を採用することにはちゃんとした理由があるが，この授業では省略する.

差は

$$(y_n - f(x_n))^2 = (y_n - (ax_n + b))^2 \tag{3}$$

と表される．このとき，学習データに対する二乗誤差の和として誤差関数 $E(a,b)$ を次のように定義する．(☆ 5)．

$$E(a,b) = \frac{1}{2}\sum_{n=1}^{N}(y_n - (ax_n + b))^2 \tag{4}$$

$E(a,b)$ が最小となるような a,b を求めたいので，$\frac{\partial E}{\partial a} = 0, \frac{\partial E}{\partial b} = 0$ とおく．すると，次式が得られる (☆ 6)．

$$\frac{\partial E}{\partial a} = \sum_{n=1}^{N}(y_n - (ax_n + b))(-x_n) = a\sum_{n=1}^{N}x_n^2 + b\sum_{n=1}^{N}x_n - \sum_{n=1}^{N}x_n y_n = 0 \tag{5}$$

$$\frac{\partial E}{\partial b} = \sum_{n=1}^{N}(y_n - (ax_n + b))(-1) = a\sum_{n=1}^{N}x_n + b\sum_{n=1}^{N}1 - \sum_{n=1}^{N}y_n = 0 \tag{6}$$

これを整理すると，次の連立一次方程式を得る．これを正規方程式という．

$$\begin{pmatrix} \sum_{n=1}^{N}x_n^2 & \sum_{n=1}^{N}x_n \\ \sum_{n=1}^{N}x_n & \sum_{n=1}^{N}1 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N}x_n y_n \\ \sum_{n=1}^{N}y_n \end{pmatrix} \tag{7}$$

これを解けば，学習データに対する二乗誤差を最小とするようなパラメータ a,b を得ることができる．

まとめると，この場合の最小二乗法の手続きは次のようになる．

1. 学習データ $\{(x_n,y_n)|n = 1,2,\ldots,N\}$ を用いて正規方程式 (7) を求める．
2. その解 (a,b) を求める．

☆ 5) $\frac{1}{2}$ 倍してるのは，微分した後が楽になるように．

☆ 6) $E(a,b)$ は「下に凸な」関数なので，その偏導関数がいずれも 0 のときに最小となる．

Q1. ほげお君は，こつこつためたお金で自動車を購入したいと思い，たまに中古車情報を調べている．ほげお君が気になっている自動車は，価格を調べ始めた日には 30 万円だったが，それ以降右の表のような価格推移を示している．調べ始めた日から経過した週数を x_n，価格差を y_n として，$n = 1,2,3,4$ に対する (x_n,y_n) の値（例えば $(x_4,y_4) = (20,-2)$ である）に最小二乗法を適用してみよう．

経過週数	0	4	8	20
価格 [万円]	30	31	29	28
価格差 [万円]	0	1	−1	−2

(1) (x_n,y_n) $(n = 1,2,3,4)$ の値から，直線のパラメータ a,b に対する正規方程式を求めなさい．

(2) (1) で求めた正規方程式を解いて a,b を求めなさい．

(3) (2) の結果を用いると，調べ始めから 36 週後には価格はいくらになっていると予測されるか答えなさい．

(5) (2) の結果を用いると，この自動車の価格が 25 万円になるのはいつ頃と予測されるか答えなさい．

★ 11.2　最小二乗法による平面・多項式の当てはめ

先の例では，入力も出力も 1 次元で与えられ，これに直線を当てはめる場合を考えた．最小二乗法は，このような場合に限らず，任意の次元数のデータに対する回帰問題に適用できる．また，直線のみならず，様々な曲線／平面／曲面等の当てはめにも用いることができる．

★ 11.2.1　平面の当てはめ

入力が D 次元，出力が 1 次元の回帰問題を考える．入力 $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$ と出力 y の関係が平面 (D 次元空間中の $(D-1)$ 次元超平面) の式で表せると仮定すると，$(D+1)$ 個のパラメータ $w_0, w_1, \ldots, w_D$ を用いて次のように書ける．

$$y = f(\boldsymbol{x}; w_0, w_1, \ldots, w_D) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D \tag{8}$$

ここで，見通しを良くするために，1 と $\boldsymbol{x}$ の要素をならべて作った $(D+1) \times 1$ 行列を $\widetilde{\boldsymbol{x}} = (1 \ x_1 \ x_2 \ \ldots \ x_D)^\top$ とおき，パラメータの $(D+1) \times 1$ 行列を $\widetilde{\boldsymbol{w}} = (w_0 \ w_1 \ \ldots \ w_D)^\top$ とおく．すると，式 (8) は次のように表せる．

$$y = f(\boldsymbol{x}; \widetilde{\boldsymbol{w}}) = \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}} \tag{9}$$

このとき，学習データ $\{(\boldsymbol{x}_n, y_n) | n = 1, 2, \ldots, N\}$ に対する二乗誤差を最小にするパラメータ $\widetilde{\boldsymbol{w}}$ を求めるための正規方程式を，直線当てはめの場合と同様に導出することができる．

1. 学習データに対する二乗誤差の和として誤差関数 $E(\widetilde{\boldsymbol{w}})$ を定義する．

$$E(\widetilde{\boldsymbol{w}}) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}}_n)^2 \tag{10}$$

2. $E(\widetilde{\boldsymbol{w}})$ が最小となるようなパラメータを求めるため，これを $\widetilde{\boldsymbol{w}}$ の各要素について偏微分して 0 とおく．

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} (y_n - \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}}_n) \frac{\partial}{\partial w_i} (y_n - \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}}_n) \tag{11}$$

$$= \sum_{n=1}^{N} (y_n - \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}}_n) \left(-\frac{\partial}{\partial w_i} (w_0 + w_1 x_{n,1} + \cdots + w_D x_{n,D}) \right) \tag{12}$$

$$= \sum_{n=1}^{N} (y_n - \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}}_n)(-x_{n,i}) = 0 \qquad (i = 0, 1, 2, \ldots, D) \tag{13}$$

ただし，$x_{n,0} \equiv 1$ とおいた．これを整理すると (☆ 7)，正規方程式は

$$\left(\sum_{n=1}^{N} \widetilde{\boldsymbol{x}}_n \widetilde{\boldsymbol{x}}_n^\top \right) \widetilde{\boldsymbol{w}} = \sum_{n=1}^{N} y_n \widetilde{\boldsymbol{x}}_n \tag{14}$$

となる．

☆7) 途中かなり省略している．

3. 正規方程式を解いて，誤差 $E(\widetilde{\boldsymbol{w}})$ を最小にするパラメータ $\widetilde{\boldsymbol{w}}$ を求める．

ここで，学習データをならべた $(D+1) \times N$ 行列 $\boldsymbol{X}$ と $1 \times N$ 行列 $\boldsymbol{Y}$ を

$$\boldsymbol{X} = (\widetilde{\boldsymbol{x}}_1 \ \widetilde{\boldsymbol{x}}_2 \ \ldots \ \widetilde{\boldsymbol{x}}_N) \tag{15}$$

$$\boldsymbol{Y} = (y_1 \ y_2 \ \ldots \ y_N) \tag{16}$$

とおくと，式 (14) の正規方程式は次式のように簡単な形になる．

$$\boldsymbol{X}\boldsymbol{X}^\top\tilde{\boldsymbol{w}} = \boldsymbol{X}\boldsymbol{Y}^\top \tag{17}$$

★ 11.2.2　多項式のあてはめ

平面当てはめの問題の D 次元入力の各要素を $x, x^2, x^3, \ldots, x^D$ に置き換えると，式 (8) は

$$y = f(x; w_0, w_1, \ldots, w_D) = w_0 + w_1 x + w_2 x^2 + \cdots + w_D x^D \tag{18}$$

という D 次多項式となる．したがって，この場合に最小二乗法を適用すると，1 次元入力 1 次元出力のデータを近似する D 次多項式を求めることができる．

$D = 2$ の場合について具体的に正規方程式を書き表してみると，

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_N \\ x_1^2 & x_2^2 & \ldots & x_N^2 \end{pmatrix} \qquad \boldsymbol{Y} = (y_1\ y_2\ \ldots\ y_N) \tag{19}$$

より

$$\begin{pmatrix} \sum 1 & \sum x_n & \sum x_n^2 \\ \sum x_n & \sum x_n^2 & \sum x_n^3 \\ \sum x_n^2 & \sum x_n^3 & \sum x_n^4 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \sum y_n \\ \sum x_n y_n \\ \sum x_n^2 y_n \end{pmatrix} \tag{20}$$

となる（和の添字は省略した）．

最小二乗法による多項式当てはめの例を図 2 に示す．3 次の結果はそれほど悪くないが，17 次ともなると学習データにはよく当てはまっているものの汎化能力が低くなっていることがわかる (☆ 8)．

☆ 8) このような現象を**過学習**という．詳細はいずれまた．

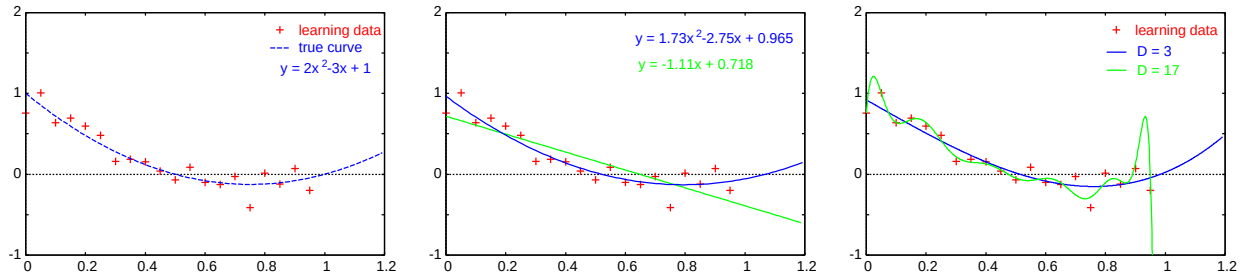

図 2: ノイズを含むデータに対する最小二乗法．左: データの真の関数は青い破線であったが，赤い点が表すように y にノイズの入った値が観測された．中: 左の赤い点を学習データとして最小二乗法による放物線 (青) と直線 (緑) の当てはめを行った結果．右: 3 次 ($D = 3$) と 17 次 ($D = 17$) の多項式の当てはめ結果．

★11.3 応用例: 時系列の線形予測

最小二乗法の時系列の予測問題への応用例を紹介する．時系列データが $\{x_1, x_2, x_3, \dots\}$ と与えられるときに，時刻 t の値 x_t を，それより過去の値 $x_{t-1}, x_{t-2}, \dots x_{t-D}$ の線形和で近似したい．すなわち，

$$x_t \approx f(x_{t-1}, x_{t-2}, \dots, x_{t-D}) = \sum_{j=1}^{D} w_j x_{t-j} \tag{21}$$

$$= w_0 + w_1 x_{t-1} + w_2 x_{t-2} + \cdots + w_D x_{t-D} \tag{22}$$

という関係が成り立つようにしたい．そのためには，

$$E = \frac{1}{2} \sum_t (x_t - f(x_{t-1}, x_{t-2}, \dots, x_{t-D}))^2 \tag{23}$$

を最小にするパラメータ $w_0, w_1, \dots, w_D$ を最小二乗法によって定めればよい．このように線形の式で時系列の未来の値を予測する仕組みを線形予測器という．

図 3 は，ノイズの加わった正弦波を学習データにして線形予測器 ($D = 10$) を学習させ，未知の時系列データの一時刻先の値を予測させた結果を示している．$t = 100$ までは学習データと同周期・同振幅の時系列なので予測の誤差は小さい．$t = 100$ で時系列の振幅と位相を突然変化させると誤差が大きくなるが，しばらくすると元に戻っている．しかし，$t = 150$ 以降ではうまく予測できなくなっている．これは，$t = 150$ 以降の時系列の周期が学習データのものと異なっているいるためである．

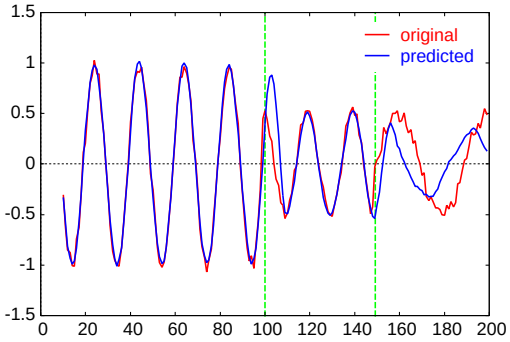

図 3: 線形予測器による時系列予測．赤が未知の時系列データ (真の値)，青が予測器の出力．

★11.4 応用例: 顔画像からの年齢推定

顔画像の画素値を入力，年齢を出力として教師あり学習すれば，顔画像から年齢を推定する仕組みを作れます．最小二乗法による線形回帰（平面当てはめ）でどれくらいできるもんでしょう？ 時間があればデモします…(^_^;