

目次

- [回帰] 最小二乗法による平面・多項式の当てはめ
- [回帰] 応用例: 時系列の線形予測, 顔画像からの年齢推定
- [識別] 素朴なアイデア — 最短距離法
- [識別] 最近傍法と k 最近傍法

★ 12 パターン認識と機械学習 (3) — 回帰 (承前)

★ 12.2 最小二乗法による平面・多項式の当てはめ

最小二乗法は, 前回説明したような直線の場合に限らず, 様々な曲線/平面/曲面等を用いた回帰問題に適用できる. D 次元空間中の $(D - 1)$ 次元超平面の当てはめ, 放物線等の多項式の当てはめについて簡単に述べる.

★ 12.2.1 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 $\mathbf{x} = (x_1, x_2, \dots, x_D)$ 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, \dots, w_D を用いて次のように書ける.

$$y = f(\mathbf{x}; w_0, w_1, \dots, w_D) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_D x_D \quad (1)$$

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

$$y = f(\mathbf{x}; \tilde{\mathbf{w}}) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}} \quad (2)$$

このとき, 学習データ $\{(\mathbf{x}_n, y_n) | n = 1, 2, \dots, N\}$ に対する二乗誤差を最小にするパラメータ $\tilde{\mathbf{w}}$ 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 $E(\tilde{\mathbf{w}})$ を定義する.

$$E(\tilde{\mathbf{w}}) = \frac{1}{2} \sum_{n=1}^N (y_n - \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_n)^2 \quad (3)$$

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

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^N (y_n - \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_n) \frac{\partial}{\partial w_i} (y_n - \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_n) \quad (4)$$

$$= \sum_{n=1}^N (y_n - \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_n) \left(-\frac{\partial}{\partial w_i} (w_0 + w_1 x_{n,1} + \dots + w_D x_{n,D}) \right) \quad (5)$$

$$= \sum_{n=1}^N (y_n - \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}_n) (-x_{n,i}) = 0 \quad (i = 0, 1, 2, \dots, D) \quad (6)$$

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

$$\left(\sum_{n=1}^N \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n^\top \right) \tilde{\mathbf{w}} = \sum_{n=1}^N y_n \tilde{\mathbf{x}}_n \quad (7)$$

となる.

☆ 1) 途中かなり省略している.

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

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

$$\mathbf{X} = (\tilde{x}_1 \tilde{x}_2 \dots \tilde{x}_N) \tag{8}$$

$$\mathbf{Y} = (y_1 y_2 \dots y_N) \tag{9}$$

とおくと、式 (7) の正規方程式は次式のように簡単な形になる.

$$\mathbf{X}\mathbf{X}^\top \tilde{w} = \mathbf{X}\mathbf{Y}^\top \tag{10}$$

★ 12.2.2 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, \dots, x^D に置き換えると、式 (1) は

$$y = f(x; w_0, w_1, \dots, w_D) = w_0 + w_1x + w_2x^2 + \dots + w_Dx^D \tag{11}$$

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、

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_N \\ x_1^2 & x_2^2 & \dots & x_N^2 \end{pmatrix} \quad \mathbf{Y} = (y_1 y_2 \dots y_N) \tag{12}$$

より

$$\begin{pmatrix} \sum 1 & \sum x_n & \sum x_n^2 \\ \sum x_n & \sum x_n^2 & \sum x_n^3 \\ \sum x_n^2 & \sum x_n^3 & \sum x_n^4 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \sum y_n \\ \sum x_n y_n \\ \sum x_n^2 y_n \end{pmatrix} \tag{13}$$

となり、前回の宿題で登場した式と等価な結果が得られる (☆ 2).

最小二乗法による放物線当てはめの例を図 1 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 (☆ 3).

☆ 2) パラメータの並び順を前回の宿題の式に合わせれば、全く同じ式であることがわかる.

☆ 3) このような現象を過学習という. 詳細はいずれまた.

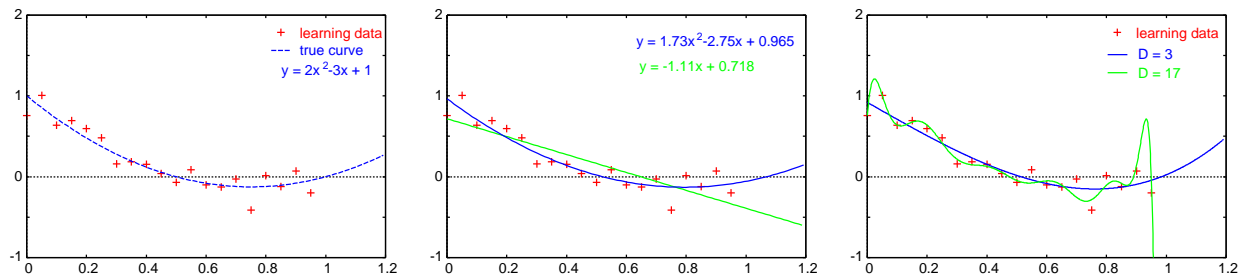


図 1: ノイズを含むデータに対する最小二乗法. 左: データの真の関数は青い破線であったが、赤い点が表すように y にノイズの入った値が観測された. 中: 左の赤い点を学習データとして最小二乗法による放物線 (青) と直線 (緑) の当てはめを行った結果. 右: 3 次 ($D = 3$) と 17 次 ($D = 17$) の多項式の当てはめ結果.

★ 12.3 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、

$$x_t \approx f(x_{t-1}, x_{t-2}, \dots, x_{t-D}) = \sum_{j=1}^D w_j x_{t-j} \quad (14)$$

$$= w_0 + w_1 x_{t-1} + w_2 x_{t-2} + \dots + w_D x_{t-D} \quad (15)$$

という関係が成り立つようにしたい。そのためには、

$$E = \frac{1}{2} \sum_t (x_t - f(x_{t-1}, x_{t-2}, \dots, x_{t-D}))^2 \quad (16)$$

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 2 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ためである。

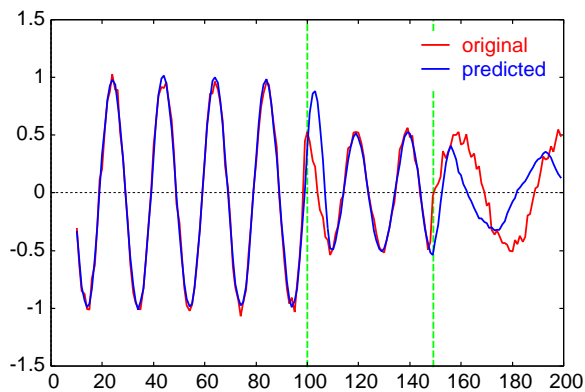


図 2: 線形予測器による時系列予測。赤が未知の時系列データ (真の値), 青が予測器の出力。

★ 12.4 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のでしょう? 準備が間に合えばデモします…(^_^);

★ 13 パターン認識と機械学習 (4) — 識別のための教師あり学習の例

「識別」のための教師あり学習の手法を取り上げる。

★ 13.1 素朴なアイデア — 最短距離法

最も単純な識別の手法は、どのクラスに所属するかがわかっている見本データ (これを**プロトタイプ** (☆4) という) をクラス毎に 1 つずつ用意しておき、未知のデータが入力されたら、そのデータと最も「近い」プロトタイプを探して、そのプロトタイプと同じクラスに所属すると判断する、というものである。これは、次節で解説する最近傍法の特殊な場合といえる。

☆4) プロトタイプ: prototype.

Q1. (身長 [cm], 体重 [kg]) の二次元のデータが与えられたときに、上記の手法で「人間」と「ほげ星人」を識別してみよう。「人間」のプロトタイプは (170, 65), 「ほげ星人」のプロトタイプは (100, 100) とする。また、2次元ベクトル間のユークリッド距離で「近さ」を測ることにする。次の2人はどちらに識別されるか。A さん: (135, 45), B さん: (135, 82).

Q2. Q1 の場合、(身長, 体重)-平面は「人間」に所属する点の集合と「ほげ星人」に所属する点の集合に二分される。その境界はどのような図形になるか。ヒント: 2点からの距離が等しい点の集合は何?

Q2 の結果が示すように、識別とは、入力データの空間を与えられたクラスのそれぞれに対応する領域に分割することであるといえる (☆5)。異なる2つのクラスに対応する領域の境界を**決定境界 (識別境界)** (☆6) という。上記の例では決定境界は直線であるが、識別手法によっては曲線 (より高次元のデータに対しては平面、曲面) となることもある。

☆5) この授業では説明しないが、入力各点を確定的に1つのクラスに所属させるのではなく、各クラスへの所属確率を算出するような方法もある。したがってこの記述はちょっと説明不足。

☆6) 決定境界: decision boundary.

★ 13.2 最近傍法と k 最近傍法

上述の手法を一般化すると、1クラスあたり複数のプロトタイプを用意する手法が考えられる。このような手法は、**最近傍法** (☆7) と呼ばれる。最近傍法の手順は、次のようになる。

☆7) 最近傍法: nearest neighbor method. NN法とも。画素値の補間の話で同じ名前が出て来たが、別もの。

- 学習データ $\{(\mathbf{x}_n, y_n) | n = 1, 2, \dots, N\}$ を用意する。ここで y_n は、プロトタイプ \mathbf{x}_n の所属するクラスを表す (クラスラベルという)。例えば、‘ほげ’, ‘ふが’, ‘へな’ という3つのクラスを識別する場合、 $y_n \in \{\text{‘ほげ’}, \text{‘ふが’}, \text{‘へな’}\}$ とすればよい。
- 所属クラスが未知のデータ \mathbf{x} が与えられたら、最も「近い」プロトタイプの番号 n^* を求める (☆8)。ここで、 $d(\mathbf{x}, \mathbf{y})$ は \mathbf{x} と \mathbf{y} の距離を表す (☆9)。

☆8) argmin の意味は第12回資料参照。

$$n^* = \operatorname{argmin}_{n=1,2,\dots,N} d(\mathbf{x}, \mathbf{x}_n) \quad (17)$$

- データ \mathbf{x} をクラス y_{n^*} (n^* 番目の学習データの所属クラス) に識別する。

☆9) ユークリッド距離以外の距離を用いる場合もあるので、このように一般化して書いている。データが数値で表せないような問題でも、それらの間の距離さえ定義できれば最近傍法は用いることができる。

最近傍法はさらに、未知データに最も近いプロトタイプを 1 つ選ぶかわりに、最も近い k 個を選んで多数決をとる、というように一般化することができる。このような手法は、 k **最近傍法**と呼ばれる (☆ 10)。**最近傍法**は、 $k = 1$ の場合に相当する。図 3 に、2 次元のデータを k 最近傍法で識別した例を示す。

最近傍法や k 最近傍法は、与えられた学習データを全て記憶しておき、未知データと全ての学習データとの距離を計算する手法である。そのため、識別に必要な計算コスト・記憶コストともに高くなる。しかし、コンピュータ性能の向上により、近年では大規模データに対しても実用されることもある。また、学習データが大量に得られる場合には、その全てをプロトタイプとはせず、教師なし学習手法であるクラスタリング等の手法によってプロトタイプを選別することもある。

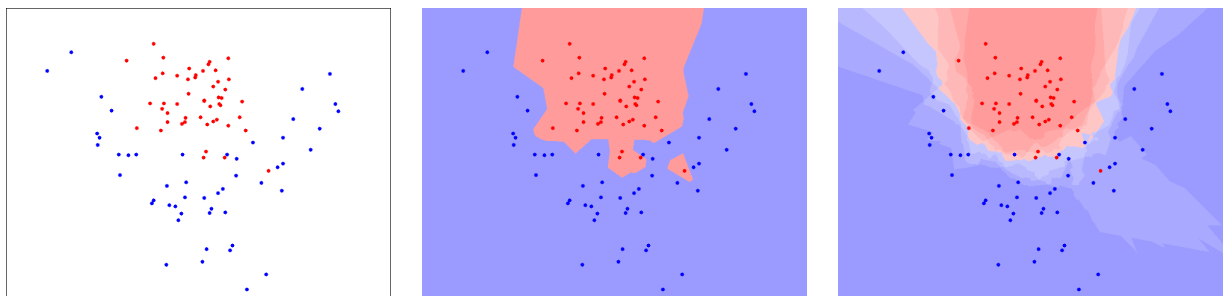


図 3: k 最近傍法による識別。左: 学習データ。赤と青の 2 クラス。中: $k = 1$ の識別結果に応じて領域を塗り分けたもの。右: $k = 7$ での結果。色の濃さは、 k 票中の多数票の多さに対応している。

★ 13.3 余談だよん

識別に関する手法は、画像や音声などのパターン情報を対象としてよく研究されてきたため、このような研究分野は**パターン認識** (☆ 11) と呼ばれている。この授業で取り上げることができたのは、この分野で知られている代表的な手法の中のほんの一部である。パターン認識の分野では、様々な意味で確率を取り入れたデータの扱い／アルゴリズムの構成をすることが一般的であるが、この授業では全く取り上げなかった。興味のある人は、次のような参考書を参照してほしい。

- 「わかりやすいパターン認識」 オーム社, ISBN4-274-13149-1
- 「パターン認識と機械学習 上下」 シュプリンガー・ジャパン, ISBN978-4-431-10013-3, 10031-7

☆ 10) k 最近傍法: k -nearest neighbor method. k -NN 法とも。

☆ 11) パターン認識: pattern recognition.

★ 宿題

Q3. 3次元ベクトルで特徴づけられるパターンを、「きゅぼ」、「ぼにゅ」、「にゅき」の3クラスに分類する識別問題を考える。識別手法には、ユークリッド距離を用いた最近傍法を用いることにする。学習データは $N = 4$ 個で、それぞれ

$((0, 0, 0), \text{「きゅぼ」}), ((1, 1, 0), \text{「きゅぼ」}), ((2, 2, 0), \text{「ぼにゅ」}), ((-1, 0, 1), \text{「にゅき」})$

であった。ただし、それぞれのデータは(入力, 出力の正解)を表している。このとき、特徴ベクトルが $(1, 1, 1)$ であるようなパターンはどのクラスに識別されるか。

Q4. 機械学習について述べた次の文章の箱の中に入る言葉を答えなさい。同じ番号の箱には同じ言葉が入る。箱の中に言葉が入っている場合、適切なものを選ぶこと。

機械学習を形式的に説明すると次のようになる。入力 x が与えられるとそれに応じた値を出力する仕組み f を考える。 f には(1)と呼ばれる調節可能な変数が複数あり、それらを変化させると同じ入力に対しても出力を変化させられるものとする。(1)を w で表して、 x に対する出力を $y = f(x; w)$ と表す。機械学習とは、このような仕組み f が様々な入力に対して望ましい出力を得られるように(1)を調節する手法である。

機械学習の手法は、(2)と(3)の二つに大きく分けられる。(2)は、出力の正解なしでデータの統計的性質などから(1)を自己組織的に調節するやり方であり、(3)は、与えられた正解に近い出力が得られるように(1)を調節するやり方である。

(3)では、入力とそれに対する出力の正解をペアとした学習データが複数与えられ、実際の(4)入力/出力と(5)出力/正解との間の誤差がなるべく(6)大きく/小さくなるようにパラメータを決めることが多い。多層パーセプトロンなどのように f が複雑で最適なパラメータが簡単には求まらないもの場合には、適当な初期値からはじめてパラメータを逐次的に変化させていく、勾配法などの最適化手法が用いられることが多い。