PRML の 4 章のための数学

サイボウズ・ラボ 光成滋生

2011 年 5 月 17 日

1 概要

この文章は『パターン認識と機械学習』（以下 PRML）の 4 章を理解するために必要な数学の一部です. 間違い, 質問などございましたら herumi@nifty.com または twitterID:herumi までご連絡ください.

2 行列の微分の復習

$A = (a_{ij})$ とかいた.

$$(AB)_{ij} = \sum_k a_{ik} b_{kj}.$$

$$\mathrm{tr}(A) = \sum_i a_{ii}$$

$$A^{\boldsymbol{T}} = (a_{ji})$$

などを思い出しておく.

さて A, B を適当な行列として

$$\frac{\partial}{\partial A} \mathrm{tr}(AB) = B^{\boldsymbol{T}}$$

なぜなら,

$$(\frac{\partial}{\partial A} \mathrm{tr}(AB))_{ij} = \frac{\partial}{\partial a_{ij}} \sum_{s,t} a_{st} b_{ts} = b_{ji}.$$

ここで $\frac{\partial}{\partial a_{ij}} a_{st} = \delta_{is}\delta_{jt}$ を使った. つまり添え字 s, t が走るときに, $s = i$, $t = j$ のときのみが生き残るというわけである.

慣れるためにもう一つやっておこう.

$$\frac{\partial}{\partial A} \mathrm{tr}(ABA^{\boldsymbol{T}}) = A(B + B^{\boldsymbol{T}}).$$

なぜなら,

$$\begin{aligned}
\frac{\partial}{\partial a_{ij}} \mathrm{tr}(ABA^{\boldsymbol{T}}) &= \frac{\partial}{\partial a_{ij}} \sum_{s,t,u} a_{st} b_{tu} a_{su} \\
&= \sum_{s,t,u} b_{tu} \frac{\partial}{\partial a_{ij}} (a_{st} a_{su}) \\
&= \sum_{s,t,u} b_{tu} (\delta_{is}\delta_{jt} a_{su} + a_{st}\delta_{is}\delta_{ju}) \\
&= \sum_{u} b_{ju} a_{iu} + \sum_{t} b_{tj} a_{it} \\
&= \sum_{u} a_{iu} b_{ju} + \sum_{t} a_{it} b_{tj} \\
&= (AB^{\boldsymbol{T}})_{ij} + (AB)_{ij} \\
&= (A(B + B^{\boldsymbol{T}}))_{ij}.
\end{aligned}$$

3 多クラス

K 個の線形関数を使った K クラス識別を考える.

$$y_k(x) = w_k^{\boldsymbol{T}} x + w_{k0}.$$

ここで w_k は重みベクトル, w_{k0} はバイアスパラメータでスカラー, x が分類したい入力パラメータでベクトルである.

クラス分類を次の方法で定義する: x に対して, ある k が存在し, 全ての $j \neq k$ にたいして $y_k(x) > y_j(x)$ であるとき x はクラス C_k に割り当てるとする.

これは well-defined である. つまり

- (一意性) x が二つの異なるクラス C_k に C'_k に属することはない. なぜならそういう k, k' があったとすると $y_k(x) > y'_k(x) > y_k(x)$ となり矛盾するから.
- (存在性) x が与えられたとき $\{y_k(x)\}$ の最大値 m を与える k_0 がその候補である. もしも $m = y_k(x)$ となる k が複数個存在 (k_1, k_2) したとすると, クラス分類はできないが, そういう x の集合は $\{x|y_{k_1}(x) = y_{k_2}(x)\}$ の部分集合となり, 通常次元が落ちる. つまり無理できるぐらいしかない.

上記で分類されたクラス C_k に属する空間は凸領域となる. すなわち x, x' を C_K の点とすると, 任意の $\lambda \in [0, 1]$ に対して $x'' = \lambda x + (1 - \lambda)x'$ も C_k に属する.

なぜなら x, $x' \in C_k$ より任意の $j \neq k$ にたいして $y_k(x) > y_j(x)$, $y_k(x') > y_j(x')$. $y_k(x)$ は x について線形なので $\lambda \geq 0$, $1 - \lambda \geq 0$ より

$$y_k(x'') = \lambda y_k(x) + (1 - \lambda) y_k(x') > \lambda y_j(x) + (1 - \lambda) y_j(x') = y_j(x'')$$

が成り立つからである.

凸領域は単連結 (simply connected) である. つまりその領域の中に空洞は無い. 任意の凸領域の 2 点を結ぶ線分が凸領域に入ることから直感的には明らかであろう.

4　分類における最小二乗

前節では重みベクトル w_{k0} を別扱いしたが, $\tilde{w}_k = (w_{k0}, w_k^{\boldsymbol{T}})^{\boldsymbol{T}}$, $\tilde{x} = (1, x^{\boldsymbol{T}})^{\boldsymbol{T}}$ と 1 次元増やすと $y_k(x) = \tilde{w}^{\boldsymbol{T}}\tilde{x}$ とかける. 面倒なので $\tilde{x}$ を x と置き換えてしまおう.

さらにまとめて $y(x) = W^{\boldsymbol{T}}x$ としよう. x, y はベクトル, W は行列である.

二乗誤差関数

$$E_D(W) = \frac{1}{2}\,\mathrm{tr}((XW-T)^{\boldsymbol{T}}(XW-T))$$

を最小化する W を求めよう.

$$\begin{aligned}
\frac{\partial}{\partial w_{ij}}E_D(W) &= \frac{1}{2}\frac{\partial}{\partial w_{ij}}\sum_{s,t}((XW-T)_{st})^2 \\
&= \sum_{s,t}(XW-T)_{st}\frac{\partial}{\partial w_{ij}}(XW-T)_{st} \\
&= \sum_{s,t}(XW-T)_{st}\frac{\partial}{\partial w_{ij}}(\sum_u x_{su}w_{ut}) \\
&= \sum_{s,t,u}(XW-T)_{st}x_{su}\delta_{iu}\delta_{jt} \\
&= \sum_s(XW-T)_{sj}x_{si} \\
&= \sum_s(X^{\boldsymbol{T}})_{is}(XW-T)_{sj} \\
&= (X^{\boldsymbol{T}}(XW-T))_{ij}.
\end{aligned}$$

よって

$$\frac{\partial}{\partial W}E_D(W) = X^{\boldsymbol{T}}(XW-T).$$

$= 0$ とおいて $X^{\boldsymbol{T}}XW = X^{\boldsymbol{T}}T$ より

$$W = (X^{\boldsymbol{T}}X)^{-1}X^{\boldsymbol{T}}T.$$

5　フィッシャーの線形判別

まず D 次元のベクトル x の入力に対して $y = w^{\boldsymbol{T}}x$ で 1 次元に射影する. $y \geq w_0$ なら C_1, そうでないなら C_2 に分類する. C_1 の点が N_1 個, C_2 の点が N_2 個とする. C_i の点の平均は

$$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{n\in C_i}x_n.$$

$m_i = w^{\boldsymbol{T}}\boldsymbol{m_i}$ として, $|w|^2 = \sum_i w_i^2 = 1$ の制約下で

$$m_2 - m_1 = w^{\boldsymbol{T}}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

を最大化してみよう.

$$f(w) = w^{\boldsymbol{T}}(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \lambda(1-|w|^2)$$

とおくと

$$\frac{\partial}{\partial w} f = \boldsymbol{m}_2 - \boldsymbol{m}_1 - 2\lambda w = 0.$$

よって

$$w = \frac{1}{2\lambda}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1).$$

$$\frac{\partial}{\partial \lambda} f = 1 - |w|^2 = 0$$

より $|w| = 1$. ただしこの手法ではそれぞれのクラスの重心 $\boldsymbol{m}_1$ と $\boldsymbol{m}_2$ とだけで w の向きが決まってしまい, 場合によっては二つのクラスの射影が大きく重なってうまく分離できないことがある. そこでクラス間の重なりを最小にするように分散も加味してみる.

クラス C_k から射影されたデータのクラス内の分散を

$$y_n = w^{\boldsymbol{T}} x_n, s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

で定義し, 全データに対する分散を $s_1^2 + s_2^2$ とする.

フィッシャーの判別基準は

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

で定義される. この定義を書き直してみよう.

$$S_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^{\boldsymbol{T}},$$

$$S_W = \sum_{n \in C_1} (x_n - \boldsymbol{m}_1)(x_n - \boldsymbol{m}_1)^{\boldsymbol{T}} + \sum_{n \in C_2} (x_n - \boldsymbol{m}_2)(x_n - \boldsymbol{m}_2)^{\boldsymbol{T}}$$

とする. S_B をクラス間共分散行列, S_W を総クラス内共分散行列という.

$w^{\boldsymbol{T}} S_B w = w^{\boldsymbol{T}} (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^{\boldsymbol{T}} w = (m_2 - m_1)^2$.

$$w^{\boldsymbol{T}} S_W w = w^{\boldsymbol{T}} \sum_{n \in C_1} (x_n - \boldsymbol{m}_1)(x_n - \boldsymbol{m}_1)^{\boldsymbol{T}} w + w^{\boldsymbol{T}} \sum_{n \in C_2} (x_n - \boldsymbol{m}_2)(x_n - \boldsymbol{m}_2)^{\boldsymbol{T}} w = \sum_{n \in C_1} (y_n - m_1)^2 + \sum_{n \in C_2} (y_n - m_2)^2$$

より

$$J(w) = \frac{w^{\boldsymbol{T}} S_B w}{w^{\boldsymbol{T}} S_W w}.$$

これが最大となる w の値を求めてみよう. 大きさはどうでもよくて向きが重要である.

$$\frac{\partial}{\partial w} J(w) = (2(S_B w)(w^{\boldsymbol{T}} S_W w) - 2(w^{\boldsymbol{T}} S_B w)(S_W w))/(w^{\boldsymbol{T}} S_W w)^2 = 0.$$

よって

$$(w^{\boldsymbol{T}} S_B w) S_w w = (w^{\boldsymbol{T}} S_W w) S_B w.$$

$S_B w = (\boldsymbol{m}_2 - \boldsymbol{m}_1)((\boldsymbol{m}_2 - \boldsymbol{m}_1)^{\boldsymbol{T}} w) \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1)$ だから

$$w \propto S_W^{-1} S_B w \propto S_W^{-1} (\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

のときに $J(w)$ が最大となる. これをフィッシャーの線形判別（linear discriminant）という.

6 最小二乗との関連

- 最小二乗法：目的変数の値の集合にできるだけ近いように
- フィッシャーの判別基準：クラスの分離を最大化するように

2 クラスの分類のときは最小二乗の特別な場合がフィッシャーの判別基準であることをみる．フィッシャーの判別基準が, 最小二乗と関係があることが分かるとそちらの議論が使えていろいろ便利なことがある．

クラス C_i に属するパターンの個数を N_i として全体を $N = N_1 + N_2$ とする．クラス C_1 に対する目的変数値を N/N_1, クラス C_2 に対する目的変数値を $-N/N_2$ とする．

この条件下で二乗和誤差

$$E = \frac{1}{2}\sum_{n=1}^{N}(w^{\boldsymbol{T}}x_n + w_0 - t_n)^2$$

を最大化してみよう．

$$\frac{\partial}{\partial w_0}E = \sum(w^{\boldsymbol{T}}x_n + w_0 - t_0) = 0$$

より $\boldsymbol{m} = (1/N)\sum x_n$ とおくと $Nw^{\boldsymbol{T}}m + Nw_0 - \sum t_n = 0.$

$$\sum t_n = N_1(N/N_1) + N_2(-N/N_2) = 0$$

より $w_0 = -w^{\boldsymbol{T}}\boldsymbol{m}$. また

$$\sum(w^{\boldsymbol{T}}x_n)x_n = \sum(x_n^{\boldsymbol{T}}w)x_n = \sum(x_n x_n^{\boldsymbol{T}})w.$$

$$\sum w_0 x_n = Nw_0\boldsymbol{m} = -N(w^{\boldsymbol{T}}\boldsymbol{m})\boldsymbol{m} = -N(\boldsymbol{m}\boldsymbol{m}^{\boldsymbol{T}})w.$$

$$\sum t_n x_n = \sum_{n\in C_1} t_n x_n + \sum_{n\in C_2} t_n x_n = N/N_1(N_1\boldsymbol{m}_1) + (-N/N_2)(N_2\boldsymbol{m}_2) = N(\boldsymbol{m}_1 - \boldsymbol{m}_2).$$

よって

$$\frac{\partial}{\partial w}E = \sum(w^{\boldsymbol{T}}x_n + w_0 - t_n)x_n = 0$$

を使うと

$$\sum(x_n x_n^{\boldsymbol{T}})w = N(\boldsymbol{m}\boldsymbol{m}^{\boldsymbol{T}})w + N(\boldsymbol{m}_1 - \boldsymbol{m}_2).$$

これらの式を使って S_w を計算する．

$$\begin{aligned} S_W &= \sum_{n\in C_1} x_n x_n^{\boldsymbol{T}} - 2\sum_{C_1} x_n\boldsymbol{m}_1^{\boldsymbol{T}} + \sum_{C_1}\boldsymbol{m}_1\boldsymbol{m}_1^{\boldsymbol{T}} + \sum_{C_2} x_n x_n^{\boldsymbol{T}} - 2\sum_{C_2} x_n\boldsymbol{m}_2^{\boldsymbol{T}} + \sum_{C_2}\boldsymbol{m}_2\boldsymbol{m}_2^{\boldsymbol{T}} \\ &= \sum x_n x_n^{\boldsymbol{T}} - N_1\boldsymbol{m}_1\boldsymbol{m}_1^{\boldsymbol{T}} - N_2\boldsymbol{m}_2\boldsymbol{m}_2^{\boldsymbol{T}} \\ &= N(\boldsymbol{m}\boldsymbol{m}^{\boldsymbol{T}})w + N(\boldsymbol{m}_1 - \boldsymbol{m}_2) - N_1\boldsymbol{m}_1\boldsymbol{m}_1^{\boldsymbol{T}} - N_2\boldsymbol{m}_2\boldsymbol{m}_2^{\boldsymbol{T}}. \end{aligned}$$

よって

$$(S_W + \frac{N_1N_2}{N}S_B)w = N(\boldsymbol{m}_1 - \boldsymbol{m}_2) + \{N\boldsymbol{m}\boldsymbol{m}^{\boldsymbol{T}} - N_1\boldsymbol{m}_1\boldsymbol{m}_1^{\boldsymbol{T}} - N_2\boldsymbol{m}_2\boldsymbol{m}_2^{\boldsymbol{T}} + \frac{N_1N_2}{N}(\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^{\boldsymbol{T}}\}w.$$

{} 内が 0 であることを示す（めんどうなので $\boldsymbol{m}_i$ を m_i と略する）.

$$\begin{aligned}\{\} &= \frac{1}{N}(N_1 m_1 + N_2 m_2)(N_1 m_1 + N_2 m_2)^{\boldsymbol{T}} - N_1 m_1 m_1^{\boldsymbol{T}} - N_2 m_2 m_2^{\boldsymbol{T}} + \frac{N_1 N_2}{N}(m_1 m_2^{\boldsymbol{T}} + m_2 m_2^{\boldsymbol{T}}) \\ &= (\frac{N_1^2}{N} - N_1 + \frac{N_1 N_2}{N}) m_1 m_1^{\boldsymbol{T}} + (\frac{2}{N} N_1 N_2 - \frac{2}{N} N_1 N_2) m_1 m_2^{\boldsymbol{T}} + (\frac{N_2^2}{N} - N_2 + \frac{N_1 N_2}{N}) m_2 m_2^{\boldsymbol{T}}.\end{aligned}$$

$$\frac{N_1^2}{N} - N_1 + \frac{N_1 N_2}{N} = \frac{N_1}{N}(N_1 - N + N_2) = 0,$$

$$\frac{N_2^2}{N} - N_2 + \frac{N_1 N_2}{N} = \frac{N_2}{N}(N_2 - N + N_1) = 0.$$

よって

$$(S_W + \frac{N_1 N_2}{N} S_B) w = N(\boldsymbol{m}_1 - \boldsymbol{m}_2).$$

$S_B w \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1)$ なので $w \propto S_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$.