

# 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 18 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](mailto:herumi@nifty.com) または [twitterID:herumi](https://twitter.com/herumi) までご連絡ください。

## 2 クラス分類問題

いくつに分類したいかを決めて入力空間を相異なる空間に分割し、それぞれの空間をクラス  $C_k$  とすること。訓練データ  $x$  が与えられたときに、推論段階と決定段階を経てクラスに割り当てる。

訓練データ  $x \rightarrow$  モデル  $p(C_k|x)$  を作る  $\rightarrow$  事後確率を使ってクラスに割り当てる

訓練データからどの情報を使って分類するかによって三つの方法がある：

- 生成モデル (generative model)

$p(x|C_k)$  を  $C_k$  ごとに決める。  $p(C_k)$  も決める。 そうすると同時分布  $p(x, C_k)$  が分かり、

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

で事後確率を求める。

$p(x)$  は  $p(x) = \sum p(x|C_k)p(C_k)$  で求められる。  $p(x|C_k)$  があると自分でさいころを振って各  $C_k$  に対して  $x$  を作ることができるという点で、生成モデルという。

- 識別モデル (discriminative model)

$p(x|C_k)$  を求めずにいきなり事後確率  $p(C_k|x)$  を決める推論問題を解く。 決定理論を使って  $x$  をあるクラスに割り当てる。

- 識別関数 (discriminant function)

確率モデルを考えずに入力関数によって定まる識別関数  $f : x \mapsto k$  を作る。

## 3 行列の微分の復習

$A = (a_{ij})$  とかいた。

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

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

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

などを思い出しておく.

さて  $A, B$  を適当な行列として

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

なぜなら,

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

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.

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

なぜなら,

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

## 4 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.

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

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。つまり無視できるくらいしかない。

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。

## 5 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、 $\tilde{w}_k = (w_{k0}, w_k^T)^T$ ,  $\tilde{x} = (1, x^T)^T$  と 1 次元増やすと  $y_k(x) = \tilde{w}^T \tilde{x}$  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  $y(x) = W^T x$  としよう。  $x, y$  はベクトル、  $W$  は行列である。

二乗誤差関数

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

を最小化する  $W$  を求めよう。

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

よって

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

$= 0$  において  $X^T X W = X^T T$  より

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

## 6 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  $y = w^T x$  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

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

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

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

を最大化してみよう.

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

とおくと

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

よって

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

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

より  $|w| = 1$ . ただしこの手法ではそれぞれのクラスの重心  $\mathbf{m}_1$  と  $\mathbf{m}_2$  とだけで  $w$  の向きが決まってしまう, 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

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

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.

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

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

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.

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

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

より

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

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.

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

よって

$$(w^T S_B w) S_W w = (w^T S_W w) S_B w.$$

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

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

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.

## 7 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おことがある.

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

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

を最大化してみよう.

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

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

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

より  $w_0 = -w^T \mathbf{m}$ . また

$$\sum (w^T x_n) x_n = \sum (x_n^T w) x_n = \sum (x_n x_n^T) w.$$

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

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

よって

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

を使うと

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

これらの式を使って  $S_w$  を計算する.

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

よって

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

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

$$\begin{aligned} \{ \} &= \frac{1}{N} (N_1 m_1 + N_2 m_2)(N_1 m_1 + N_2 m_2)^T - N_1 m_1 m_1^T - N_2 m_2 m_2^T + \frac{N_1 N_2}{N} (m_1 m_2^T + m_2 m_1^T) \\ &= (\frac{N_1^2}{N} - N_1 + \frac{N_1 N_2}{N}) m_1 m_1^T + (\frac{2}{N} N_1 N_2 - \frac{2}{N} N_1 N_2) m_1 m_2^T + (\frac{N_2^2}{N} - N_2 + \frac{N_1 N_2}{N}) m_2 m_2^T. \\ &\quad \frac{N_1^2}{N} - N_1 + \frac{N_1 N_2}{N} = \frac{N_1}{N} (N_1 - N + N_2) = 0, \\ &\quad \frac{N_2^2}{N} - N_2 + \frac{N_1 N_2}{N} = \frac{N_2}{N} (N_2 - N + N_1) = 0. \end{aligned}$$

よって

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

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

## 8 確率的生成モデル

分類を確率的な視点から見る. 生成的アプローチ

- $p(x|C_k)$ : モデル化されたクラスの条件付き確率密度
- $p(C_k)$ : クラスの事前確率

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)}$$

とする. ロジスティックシグモイド関数を

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

と定義し,

$$a = \log \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} \text{ とすると}$$

$$\sigma(a) = \frac{1}{1 + \frac{p(x|C_2)p(C_2)}{p(x|C_1)p(C_1)}} = p(C_1|x).$$

ロジスティックシグモイド関数の関数の性質:

$$\sigma(-a) = \frac{1}{1 + e^a} = 1 - \frac{e^a}{1 + e^a} = 1 - \frac{1}{1 + e^{-a}} = 1 - \sigma(a).$$

$$\sigma(a) = \frac{e^a}{e^a + 1}$$

より  $e^a(\sigma(a) - 1) = -\sigma(a)$ . よって

$$a = \log \frac{\sigma(a)}{1 - \sigma(a)}.$$

この関数をロジット関数という.

$K > 2$  クラスの場合,  $a_k = \log(p(x|C_k)p(C_k))$  より

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

となる. この関数は正規化指数関数, あるいはソフトマックス関数という.

## 9 連続値入力

仮定: 条件付き確率密度がガウス分布, そのガウス分布の共分散行列 ( $\Sigma = A$ ) がすべてのクラスで共通

$$p(x|C_k) = \frac{1}{(2\pi)^{(D/2)}} \frac{1}{|A|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T A^{-1}(x - \mu_k)\right)$$

$$\begin{aligned} a &= \log \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} \\ &= \log \frac{p(C_1)}{p(C_2)} - \frac{1}{2}(x - \mu_1)^T A^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_2)^T A^{-1}(x - \mu_2) \\ &= \log \frac{p(C_1)}{p(C_2)} - \frac{1}{2}\mu_1^T A^{-1}\mu_1 + \frac{1}{2}\mu_2^T A^{-1}\mu_2 + (\mu_1 - \mu_2)^T A x. \end{aligned}$$

よって

$$w_0 = -\frac{1}{2}\mu_1^T A^{-1}\mu_1 + \frac{1}{2}\mu_2^T A^{-1}\mu_2 + \log \frac{p(C_1)}{p(C_2)},$$

$$w = A^{-1}(\mu_1 - \mu_2)$$

とおくと  $p(C_1|x) = \sigma(w^T x + w_0)$ . つまりロジスティックシグモイド関数の中は  $x$  について線形.

$K$  クラスの場合, 上で定義した  $a_k$  を用いると

$$\begin{aligned} a_k &= \log p(C_k) - \frac{1}{2}\mu_k^T A^{-1}\mu_k + \mu_k^T A^{-1}x - \frac{1}{2}x^T A^{-1}x + \text{const} \\ &= a'_k - \frac{1}{2}x^T A^{-1}x + \text{const}. \end{aligned}$$

ここで  $a'_k = w_k^T x + w_{k0}$ ,  $w_k = A^{-1}\mu_k$ ,  $w_{k0} = -(1/2)\mu_k^T A^{-1}\mu_k + \log p(C_k)$ .

(注) PRML (4.63) の  $a_k$  の定義だと  $x$  の 2 次の項が残ると思う.

よって

$$p(C_k|x) = \frac{\exp(a'_k) \exp(-(1/2)x^T A^{-1}x + \text{const})}{\sum_j \exp(a'_j) \exp(-(1/2)x^T A^{-1}x + \text{const})} = \frac{\exp(a'_k)}{\sum_j \exp(a'_j)}.$$

## 10 最尤解

仮定：条件付き確率分布がガウス分布, それらが共通の共分散行列を持つ.

2 クラスの場合を考える. データ集合  $\{x_n, t_n\}$ ,  $n = 1, \dots, N$ .  $t_n = 1$  はクラス  $C_1$ ,  $t_n = 0$  はクラス  $C_2$  とする. さらに  $p(C_1) = p$ ,  $p(C_2) = 1 - p$  という事前確率を割り当てる.  $N_i$  をクラス  $C_i$  のデータの個数,  $N = N_1 + N_2$  を総数とする.

$$p(x_n, C_1) = p(C_1)p(x_n|C_1) = p\mathcal{N}(x_n|\mu_1, A).$$

$$p(x_n, C_2) = p(C_2)p(x_n|C_2) = p(1 - p)\mathcal{N}(x_n|\mu_2, A).$$

尤度関数は

$$p(t, X|p, \mu_1, \mu_2, A) = \prod_{n=1}^N (p\mathcal{N}(x_n|\mu_1, A))^{t_n} ((1 - p)\mathcal{N}(x_n|\mu_2, A))^{1-t_n}.$$

このうち  $p$  に関する部分の対数は

$$\sum (t_n \log p + (1 - t_n) \log(1 - p)).$$

$p$  で微分して 0 とおく.

$$\frac{1}{p} \sum t_n - \frac{1}{1-p} \sum (1 - t_n) = \frac{1}{p} N_1 - \frac{1}{1-p} N_2 = \frac{(1-p)N_1 - pN_2}{p(1-p)} = 0.$$

よって  $p = N_1/(N_1 + N_2) = N_1/N$ . つまり  $p$  に関する最尤推定は  $C_1$  内の個数になる.

$K$  クラスのときを考えてみよう.  $\sum p_k = 1$ . 尤度関数は

$$p(t, X|p_k, \mu_k, A) = \prod_{n=1}^N (p_k \mathcal{N}(x_n|\mu_k, A))^{t_k}.$$

対数尤度関数の  $p_i$  に関する部分の未定乗数法を使うことで

$$\sum t_{ij} \log p_j + \lambda (\sum p_i - 1).$$

$p_i$  で微分して 0 とおくと  $\sum_j (t_{ij}/p_i) + \lambda = 0$ . よって

$$-p_i \lambda = \sum_j t_{ij} = N_i$$

また  $-\sum_i p_i \lambda = -\lambda = \sum_i N_i = N$  より  $p_i = -N_i/\lambda = N_i/N$ .

さて 2 クラスの問題に戻って  $\mu_i$  について最大化してみよう.  $\mu_1$  についての部分は

$$\sum t_n \log \mathcal{N}(x_n|\mu_1, A) = -\frac{1}{2} \sum t_n (x_n - \mu_1)^T A^{-1} (x_n - \mu_1) + \text{const.}$$

$\mu_1$  で微分して 0 とおくと

$$\sum t_n A^{-1} (x_n - \mu_1) = A^{-1} (\sum t_n x_n - \mu_1 \sum t_n) = A^{-1} (\sum t_n x_n - \mu_1 N_1) = 0.$$

よって

$$\mu_1 = \frac{1}{N_1} \sum t_n x_n.$$



$\mu_2$  については  $\sum (1 - t_n) \log \mathcal{N}(x_n | \mu_2, A)$  を考えて

$$\mu_2 = \frac{1}{N_2} \sum (1 - t_n) x_n.$$

最後に  $A$  に関する最尤解を求める.