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13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,

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

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

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

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

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