PRMLの4章のための数学

サイボウズ・ラボ 光成滋生

2011年6月7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章の式変形を一部埋めたものです．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 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．

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

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．

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

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

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

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,

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

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.

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

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 $j \neq k$ に対して $y_k(x) > y_j(x)$ 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 $j \neq k$ に対して $y_k(x) > y_j(x)$, $y_k(x') > y_j(x')$．$y_k(x)$ 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．

任意のループを連続的に 1 点につぶすことが出来る様な領域を単連結（simply connected）という．凸領域は単連結である．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．

単連結であることを簡単に示しておこう：X を凸領域とする．

$$f : S^1 \to X$$

を S^1 から X への連続関数とする．任意のループは $f(S^1)$ で表される．$t \in [0,1]$, $\lambda \in [0,1]$ に対して

$$f_\lambda(t) = \lambda f(t) + (1-\lambda)f(0)$$

とすると X の凸性から $f_\lambda(t) \in X$．つまり $f_\lambda(S^1)$ は X 内のループ．$f_0(t) = f(0)$ は X のある 1 点．$f_1(t) = f(t)$ は元のループだから，これはループ $f(S^1)$ を X の中で連続的に一点 $f(0)$ につぶすことが出来ることを示している．つまり X は単連結．

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，$\tilde{w}_k = (w_{k0}, w_k^{\boldsymbol{T}})^{\boldsymbol{T}}$, $\tilde{x} = (1, x^{\boldsymbol{T}})^{\boldsymbol{T}}$ 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

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

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 $y = w^{\boldsymbol{T}}x$ 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．ラグランジュの未定乗数法を用いて

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

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.

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）．

$$\begin{aligned}
\{\} &= \frac{1}{N}(N_1m_1 + N_2m_2)(N_1m_1 + N_2m_2)^{\boldsymbol{T}} - N_1m_1m_1^{\boldsymbol{T}} - N_2m_2m_2^{\boldsymbol{T}} + \frac{N_1N_2}{N}(m_1m_2^{\boldsymbol{T}} + m_2m_2^{\boldsymbol{T}}) \\
&= (\frac{N_1^2}{N} - N_1 + \frac{N_1N_2}{N})m_1m_1^{\boldsymbol{T}} + (\frac{2}{N}N_1N_2 - \frac{2}{N}N_1N_2)m_1m_2^{\boldsymbol{T}} + (\frac{N_2^2}{N} - N_2 + \frac{N_1N_2}{N})m_2m_2^{\boldsymbol{T}}.
\end{aligned}$$

$$\begin{aligned}
\frac{N_1^2}{N} - N_1 + \frac{N_1N_2}{N} &= \frac{N_1}{N}(N_1 - N + N_2) = 0, \\
\frac{N_2^2}{N} - N_2 + \frac{N_1N_2}{N} &= \frac{N_2}{N}(N_2 - N + N_1) = 0.
\end{aligned}$$

よって

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

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

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

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

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

よって

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

とおくと $p(C_1|x) = \sigma(w^{\boldsymbol{T}}x + w_0)$. 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

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

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

（注）PRML（4.63）の a_k の定義だと x の 2 次の項が残るため, 式（4.68）を出すにはそれを除かなければならない.

よって

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

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 $\{x_n, t_n\}$, $n = 1, \ldots, N$．$t_n = 1$ 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

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

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．$\sum p_i = 1$．尤度関数は

$$p(t, X|p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, A) = \prod_{n=1}^N \prod_{i=1}^K p_i\mathcal{N}(x_n|\mu_i, A))^{t_{ni}}.$$

この対数に未定乗数法の $\lambda(\sum p_i - 1)$ の項を加え, p_i に関する部分を抜き出すと

$$\sum_n t_{ni} \log p_i + \lambda p_i.$$

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

$$-p_i \lambda = \sum_n t_{ni} = N_i$$

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 μ_i について最大化してみよう. μ_1 についての部分は

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

μ_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

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

μ_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. A に関する部分の対数は

$$\begin{aligned}
&-\frac{1}{2} \sum_{n=1}^{N} (t_n \log |A| + t_n (x_n - \mu_1)^{\boldsymbol{T}} A^{-1} (x_n - \mu_1) + (1 - t_n) \log |A| + (1 - t_n)(x_n - \mu_2)^{\boldsymbol{T}} A^{-1} (x_n - \mu_2)) \\
&= -\frac{N}{2} \log |A| - \frac{1}{2} \mathrm{tr}(\sum_{n=1}^{N} (t_n A^{-1} (x_n - \mu_1)(x_n - \mu_1)^{\boldsymbol{T}} + (1 - t_n) A^{-1} (x_n - \mu_2)(x_n - \mu_2)^{\boldsymbol{T}})) \\
&= -\frac{N}{2} \log |A| - \frac{1}{2} \mathrm{tr}(A^{-1}(\sum_{n \in C_1} (x_n - \mu_1)(x_n - \mu_1)^{\boldsymbol{T}} + \sum_{n \in C_2} (x_n - \mu_2)(x_n - \mu_2)^{\boldsymbol{T}})) \\
&= -\frac{N}{2} \log |A| - \frac{1}{2} \mathrm{tr}(A^{-1}(N_1 S_1 + N_2 S_2)) \\
&= -\frac{N}{2} \log |A| - \frac{N}{2} \mathrm{tr}(A^{-1} S)
\end{aligned}$$

ここで最後の式変形に

$$S_i = \frac{1}{N_i} \sum_{n \in C_i} (x_n - \mu_i)(x_n - \mu_i)^{\boldsymbol{T}}, S = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$$

を用いた. これを A で微分する. `https://github.com/herumi/prml/raw/master/prml2.pdf` で示した行列式の対数の微分の公式（2）:

$$\frac{\partial}{\partial A} \log |A| = (A^{-1})^{\boldsymbol{T}}$$

と 13 ページで示した式：

$$\frac{\partial}{\partial A}\operatorname{tr}(A^{-1}B) = -(A^{-1}BA^{-1})^{\boldsymbol{T}}$$

を使うと

$$-\frac{N}{2}((A^{-1})^{\boldsymbol{T}} - (A^{-1}SA^{-1})^{\boldsymbol{T}}) = 0.$$

よって $A = S$ となる．これは 2 クラスの各クラスの共分散行列の重みつき平均である．またフィッシャーの判別基準で求めた総クラス内共分散行列 S_W を N で割ったものに等しいことにも注意する．

11 ロジスティック回帰

2 クラス分類問題において, ある程度一般的な仮定の元で C_1 の事後確率を

$$p(C_1|\phi) = y(\phi) = \sigma(w^{\boldsymbol{T}}\phi)$$

とかけた．もちろん $p(C_2|\phi) = 1 - p(C_1|\phi)$ である．この式の導出に使った仮定を忘れ, これを出発点としこの形の関数を使うモデルをロジスティック回帰（logistic regression）という．このモデルにおけるパラメータを最尤法で求める．

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

としたとき

$$\sigma'(x) = -\frac{-e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}(1 - \frac{1}{1+e^{-x}}) = \sigma(x)(1-\sigma(x)).$$

データ集合 $\{\phi_n, t_n\}$, $t_n \in \{0, 1\}$, $\phi_n = \phi(x_n)$, $n = 1, \ldots, N$, $t = (t_1, \ldots, t_N)^{\boldsymbol{T}}$, $y_n = p(C_1|\phi_n) = \sigma(a_n)$, $a_n = w^{\boldsymbol{T}}\phi_n$ とする．尤度関数は

$$p(t|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}.$$

誤差関数は

$$E(w) = -\log p(t|w) = -\sum(t_n \log y_n + (1-t_n)\log(1-y_n)).$$

w で微分してみよう．まず

$$\frac{\partial}{\partial w} y_n = \sigma'(a_n)\phi_n = y_n(1-y_n)\phi_n.$$

よって

$$\begin{aligned}
\frac{\partial}{\partial w} E &= -\sum(t_n(1-y_n)\phi_n + (1-t_n)(-y_n)\phi_n) \\
&= \sum(-t_n + t_n y_n + y_n - y_n t_n)\phi \\
&= \sum(y_n - t_n)\phi_n.
\end{aligned}$$

$y_n - t_n$ は目的値とモデルの予測値の誤差なので, 線形回帰モデルのときと同じ形．

12 反復再重み付け最小二乗

いわゆるニュートン・ラフラソン法は関数 $f(x)$ の零点を求める方法：
零点の近似解 x_n が与えられたときにより近い値 x_{n+1} を見つける．x_n における接線の方程式

$$f'(x_n)(x-x_n)+f(x_n)=0$$

の解を x_{n+1} とすると

$$x_{n+1}=x_n-(f'(x_n))^{-1}f(x_n).$$

関数 $E(w)$ を最小化するためのベクトル w を与える更新式を考えてみると，最小化を与える w は $\frac{\partial}{\partial w}E(w)$ の零点．$\frac{\partial}{\partial w}E(w)$ の零点を求める問題にニュートン・ラフラソン法を適用する．w を古い値，w' を新しい値，$H(w)$ を $E(w)$ のヘッシアンとする．$f \longleftrightarrow \frac{\partial}{\partial w}E(w)$, $f' \longleftrightarrow H(w)$ という対応により

$$w'=w-H(w)^{-1}\frac{\partial}{\partial w}E(w).$$

この式を線形回帰モデルに適用してみる．

$$E(w)=\frac{1}{2}\sum(t_n-w^{\boldsymbol{T}}\phi(x_n))^2, \phi_n=\phi(x_n)$$

を w で微分して $\Phi=(\phi_1,\ldots)^{\boldsymbol{T}}$ とおくと

$$\frac{\partial}{\partial w}E(w)=\sum(t_n-w^{\boldsymbol{T}}\phi_n)\phi_n=\sum\phi_n\phi_n^{\boldsymbol{T}}w-\sum\phi_n t_n=\Phi^{\boldsymbol{T}}\Phi w-\Phi^{\boldsymbol{T}}t.$$

$$H=H(w)=\frac{\partial^2}{\partial w_i\partial w_j}E(w)=\Phi^{\boldsymbol{T}}\Phi.$$

更新式に代入すると

$$w'=w-(\Phi^{\boldsymbol{T}}\Phi)^{-1}(\Phi^{\boldsymbol{T}}\Phi w-\Phi^{\boldsymbol{T}}t)=(\Phi^{\boldsymbol{T}}\Phi)^{-1}\Phi^{\boldsymbol{T}}t.$$

これは最小二乗解である．つまり 1 回の更新で厳密解に到達した．線形回帰モデルは w に関して 2 次なので $\frac{\partial}{\partial w}E(w)$ に関しては 1 次だからである．
次にロジスティック回帰に適用してみる．

$$E(w)=-\sum(t_n\log y_n+(1-t_n)\log(1-y_n)), y_n=\sigma(a_n)=\sigma(w^{\boldsymbol{T}}\phi_n).$$

$$\frac{\partial}{\partial w}E(w)=\sum(y_n-t_n)\phi_n.$$

$y_n'=y_n(1-y_n)\phi_n$ だったので

$$H=\sum\phi_n y_n(1-y_n)\phi_n^{\boldsymbol{T}}=\Phi^{\boldsymbol{T}}R\Phi.$$

ここで $R=\mathrm{diag}(R_n)=\mathrm{diag}(y_n(1-y_n))$.
$H>0$ を確認する．任意の縦ベクトル u に対して $v=\Phi u$ とおくと $v\neq 0$ ならば

$$u^{\boldsymbol{T}}Hu=v^{\boldsymbol{T}}Rv=\sum y_n(1-y_n)v_n^2>0.$$

最後の不等号では $0<y_n<1$ を用いた．ヘッシアンが正定値であることが分かったので交差エントロピー誤差関数は唯一の最小解を持つ．

w の更新式を見てみよう.

$$\begin{aligned} w' &= w - (\Phi^{\boldsymbol{T}} R\Phi)^{-1}\Phi^{\boldsymbol{T}}(y-t) \\ &= (\Phi^{\boldsymbol{T}} R\Phi)^{-1}(\Phi^{\boldsymbol{T}} R\Phi w - \Phi^{\boldsymbol{T}}(y-t)) \\ &= (\Phi^{\boldsymbol{T}} R\Phi)^{-1}\Phi^{\boldsymbol{T}} R(\Phi w - R^{-1}(y-t)) \\ &= (\Phi^{\boldsymbol{T}} R\Phi)^{-1}\Phi^{\boldsymbol{T}} Rz. \end{aligned}$$

ここで $z = \Phi w - R^{-1}(y-t)$ である. R は y_n つまり w に依存しているので正規方程式は更新式ごとに計算し直す必要がある. 反復最重み付き最小二乗法（IRLS:iterative reweighted least squares method）という.

$t=1$ をクラス C_1, $t=0$ をクラス C_2 に割り当てて, それぞれの確率は y, $1-y$ だから

$$E[t] = y = \sigma(x).$$

$t^2 = t$ だから

$$\mathrm{var}[t] = E[t^2] - E[t]^2 = E[t] - E[t]^2 = y - y^2 = y(1-y).$$

つまり重み付け対角行列 R の対角成分は分散である.

IRLS を線形近似の解として解釈することも出来る. すなわち $a = w^{\boldsymbol{T}}\phi$, $y = \sigma(a)$ という関係を通じて a を y の関数とみなし, a_n を目標値 t_n の変数とみなして近次解 $y_n = \sigma(w_{\mathrm{old}}^{\boldsymbol{T}}\phi)$ のまわりで一次近似を行うと

$$\begin{aligned} a_n &\approx a_n(y_n) + \frac{\partial}{\partial y_n} a_n\Big|_{t_n=y_n}(t_n - y_n) \\ &= w_{\mathrm{old}}^{\boldsymbol{T}}\phi + \frac{1}{y_n(1-y_n)}(t_n - y_n) \\ &= w_{\mathrm{old}}^{\boldsymbol{T}}\phi - \frac{y_n - t_n}{y_n(1-y_n)} \\ &= z_n. \end{aligned}$$

つまり z_n は線形近似したときの目標変数値と解釈できる.

13 Jensen の不等式

実数上の実数値関数 $f(x)$ が凸関数であるとする. すなわち任意の x, y, $0 \le t \le 1$ に対して

$$tf(x) + (1-t)f(y) \ge f(tx + (1-t)y)$$

である.

$p_1, \ldots, p_n$ を足して 1 になる非負の数, すなわち $\sum_{i=1}^n p_i = 1$, $p_i \ge 0$ とする.

このとき n 個の任意の実数 $x_1, \ldots, x_n$ に対して

$$\sum_{i=1}^n p_i f(x_i) \ge f(\sum_{i=1}^n p_i x_i).$$

これを Jensen の不等式という.

証明は数学的帰納法を使う. $n=1$ のときは自明. n のとき成り立つとし,

$$\sum_{i=1}^{n+1} p_i = 1$$

とする．$q=\sum_{i=1}^n p_i$ とおくと $q+p_{n+1}=1$．$q=0$ のときは $p_{n+1}=1$ となり上記不等式は自明になりたつ．よって $q\neq 0$ とすると

$$\sum_{i=1}^n (p_i/q)=1$$

$$\begin{aligned}\sum_{i=1}^{n+1} p_i f(x_i) &= q\sum_{i=1}^n (p_i/q)f(x_i)+p_{n+1}f(x_{n+1})\\ &\quad \text{帰納法の仮定を用いて}\\ &\geq qf(\sum_{i=1}^n (p_i/q)x_i)+p_{n+1}f(x_{n+1})\\ &\quad f\text{ が凸関数であることを用いて}\\ &\geq f(q(\sum_{i=1}^n (p_i/q)x_i)+p_{n+1}x_{n+1})\\ &= f(\sum_{i=1}^{n+1} p_i x_i).\end{aligned}$$

14 多クラスロジスティック回帰

多クラス分類の事後確率を

$$p(C_k|\phi)=y_k(\phi)=\frac{\exp(a_k)}{\sum_j \exp(a_j)}, a_k=w_k^{\boldsymbol{T}}\phi$$

で与えたときに最尤法を用いて直接 w_k を求めよう．

$$\frac{\partial}{\partial a_k}y_k=\frac{\exp(a_k)(\sum\exp(a_j))-\exp(a_k)\exp(a_k)}{(\sum\exp(a_j))^2}=y_k-y_k^2.$$

$k\neq j$ として

$$\frac{\partial}{\partial a_j}y_k=-\frac{\exp(a_k)\exp(a_j)}{(\sum\exp(a_j))^2}=-y_k y_j.$$

よってこの二つをまとめて

$$\frac{\partial}{\partial a_j}y_k=y_k(\delta_{kj}-y_j).$$

目的変数ベクトル t_n を k 番目の要素だけが 1 であるものとする．つまり $t_n=(t_{nk})$．$y_{nk}=y_k(\phi_n)$，$T=(t_nk)$ とすると

$$p(T|w_1,\ldots,w_k)=\prod_{n=1}^N\prod_{k=1}^K p(C_k|\phi_n)^{t_{nk}}=\prod_{n,k} y_{nk}^{t_{nk}}.$$

交差エントロピー誤差関数は

$$E=-\log p(T|w_1,\ldots,w_k)=-\sum_{n,k} t_{nk}\log y_{nk}.$$

よって

$$\begin{aligned}\frac{\partial}{\partial w_j}E &= -\sum_{n,k} t_{nk}\frac{y_k(\phi_n)(\delta_{kj}-y_j(\phi_n))}{y_k(\phi_n)}\phi_n \\ &= -\sum_{n,k} t_{nk}(\delta_{kj}-y_{nj})\phi_n = -\sum_n((\sum_k t_{nk}\delta_{kj}) - (\sum_k t_{nk})y_{nj})\phi_n \\ &= -\sum_n (t_{nj}-y_{nj})\phi_n \\ &= \sum (y_{nj}-t_{nj})\phi_n.\end{aligned}$$

やはり誤差 $y_{nj}-t_{nj}$ と基底関数 ϕ_n の積となる.

ヘッシアンをみる.

$$\frac{\partial}{\partial w_k}y_{nj} = \frac{\partial}{\partial w_k}y_j(\phi_n) = y_k(\phi_n)(\delta_{kj}-y_j(\phi_n))\phi_n = y_{nk}(\delta_{kj}-y_{nj})\phi_n$$

より

$$H = \frac{\partial^2}{\partial w_k \partial w_j}E = \sum_n y_{nk}(\delta_{kj}-y_{nj})\phi_n\phi_n^{\boldsymbol{T}}.$$

H の正定値であることを示そう.

任意の $M\times K$ 次元ベクトルを $u=(u_1^{\boldsymbol{T}},\ldots,u_K^{\boldsymbol{T}})^{\boldsymbol{T}}$, u_k は M 次元ベクトルとする. $v_{nk}=u_k^{\boldsymbol{T}}\phi_n$, $f(x)=x^2$ とおく. $f(x)$ は下に凸.

$$\begin{aligned}u^{\boldsymbol{T}}Hu &= \sum_{n,k,j} y_{nk}(\delta_{kj}-y_{nj})(u_k^{\boldsymbol{T}}\phi_n)(\phi_n^{\boldsymbol{T}}u_j) \\ &= \sum_n(\sum_{k,j} y_{nk}\delta_{kj}v_{nk}v_{nj} - \sum_{k,j} y_{nk}y_{nj}v_{nk}v_{nj}) \\ &\quad \text{一般に } (\sum_k x_k)^2 = (\sum_k x_k)(\sum_j x_j) = \sum_{k,j} x_k x_j \text{ より} \\ &= \sum_n\{\sum_k y_{nk}v_{nk}^2 - (\sum_k y_{nk}v_{nk})^2\}.\end{aligned}$$

ここで $\sum_k y_{nk}=1$, $0<y_{nk}<1$ より Jensen の不等式を適用すると

$$u^{\boldsymbol{T}}Hu \geq 0.$$

15 プロビット回帰

指数型分布族で表される条件付き確率分布に対して, クラスの事後確率はある線形関数とロジスティック（またはソフトマックス）関数の合成で表された.

$a=w^{\boldsymbol{T}}\phi$, $f(a)$ を活性化関数として

$$p(t=1|a)=f(a).$$

とかける範囲でもう少し考察する. $f(a)$ がある確率密度 $p(\theta)$ の累積分布関数で表されるとする. とくに $p(\theta)=\mathcal{N}(\theta|0,1)$ のとき累積分布関数は

$$\Phi(a) = \int_{-\infty}^{a}\mathcal{N}(\theta|0,1)\,d\theta.$$

この逆関数をプロビット関数（probit）という．誤差関数を

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp(-\theta^2)\, d\theta$$

で定義する．

$x = \theta/\sqrt{2}$ とおくと $dx = d\theta/\sqrt{2}$.

$$\begin{aligned}\Phi(a) &= \int_{-\infty}^a = \int_{-\infty}^0 + \int_0^a = \frac{1}{2} + \int_0^a \frac{1}{\sqrt{2\pi}} \exp(-\frac{\theta^2}{2})\, d\theta \\ &= \frac{1}{2} + \int_0^{a/\sqrt{2}} \frac{1}{\sqrt{2\pi}} \exp(-x^2)\sqrt{2}\, dx \\ &= \frac{1}{2}(1 + \frac{2}{\sqrt{\pi}} \int_0^{a/\sqrt{2}} \exp(-x^2)\, dx) \\ &= \frac{1}{2}\{1 + \mathrm{erf}(\frac{a}{\sqrt{2}})\}.\end{aligned}$$

プロビット活性化関数を用いた一般化線形モデルをプロビット回帰という．

$x \to \infty$ でロジスティック回帰の微分は $\sigma'(x) = \exp(-x)/(1+\exp(-x))^2 \sim \exp(-x)$. プロビット回帰の微分は $\Phi'(x) \sim \exp(-x^2)$. つまりプロビット関数の逆関数はロジスティック関数よりも急速に 1 に近づき平らになる．プロビット回帰の方が外れ値に敏感.

16　正準連結関数

活性化関数として正準連結関数（canonical link function）と呼ばれるものを使い, 条件付き確率分布に指数型分布族を選んだときに誤差関数の微分が「誤差」×「特徴ベクトル」という形でかけることを示そう．

$$p(t|\eta, s) = \frac{1}{s} h(t/s) g(\eta) \exp(\frac{\eta t}{s})$$

とする．確率なので $\int p(t|\eta, s)\, dt = 1$. つまり

$$g(\eta) \int h(t/s) \exp(\frac{\eta t}{s})\, dt = s.$$

η で微分して

$$\begin{aligned}&(\frac{\partial}{\partial\eta} g(\eta)) \int h(t/s) \exp(\frac{\eta t}{s})\, dt + g(\eta) \int (t/s) h(t/s) \exp(\frac{\eta t}{s})\, dt \\ &= (\frac{\partial}{\partial\eta} g(\eta))(\frac{s}{g(\eta)}) + \int t p(t|\eta, s)\, dt \\ &= s\frac{\partial}{\partial\eta} \log g(\eta) + E[t] = 0.\end{aligned}$$

よって

$$y = E[t] = -s\frac{\partial}{\partial\eta} \log g(\eta).$$

y が η の関数として表せた．この逆関数が存在するとしてそれを $\eta = \psi(y)$ と書くことにする．

y を連結関数 $f(a)$ と w の線形関数の合成,

$$y = f(w^{\boldsymbol{T}} \phi)$$

とかけるモデルを考える．対数尤度関数は

$$\log p(t|\eta, s) = \sum_{n=1}^{N} \log p(t_n|\eta, s) = \sum_{n=1}^{N} (\log g(\eta_n) + \frac{\eta_n t_n}{s}) + \mathrm{const}$$

を考える．ここで s と η は独立, $\eta_n = \phi(y_n)$, $y_n = f(a_n)$, $a_n = w^{\boldsymbol{T}} \phi_n$.

パラメータが多いので依存関係に注意して微分する．

$$\frac{\partial}{\partial w} \eta_n = \psi'(y_n) f'(a_n) \phi_n.$$

$$\frac{\partial}{\partial w} \log g(\eta_n) = \frac{g'(\eta_n)}{g(\eta_n)} \psi'(y_n) f'(a_n) \phi_n = -\frac{y_n}{s} \phi'(y_n) f'(a_n) \phi_n.$$

よって

$$\frac{\partial}{\partial w} \log p(t|\eta, s) = \sum_n \frac{1}{s} (t_n - y_n) \psi'(y_n) f'(a_n) \phi_n.$$

連結関数として $f^{-1}(y) = \psi(y)$ となるものを使ってみよう．$f(\psi(y)) = y$ を y で微分して

$$f'(\psi(y)) \psi'(y) = 1.$$

同じことだが $a = f^{-1}(y) = \psi(y)$ を使って

$$f'(a) \psi'(y) = 1.$$

よって

$$\frac{\partial}{\partial w} E(w) = -\frac{\partial}{\partial w} \log p(t|\eta, s) = \frac{1}{s} \sum_n (y_n - t_n) \phi_n.$$

「誤差」×「特徴ベクトル」という形でかけることが分かった．

17　ラプラス近似

ロジスティック回帰のベイズ的な扱いは解析的に難しい．ここではラプラス近似というものを紹介する．これは連続変数上の確率密度分布をあるガウス分布で近似することである (当然複数の山があると辛い).

$$p(z) = \frac{1}{Z} f(z)$$

$Z = \int f(z)\, dz$ は正規化係数で未知とする．$p(z)$ のモード, つまり最大値を与える z_0 を探そう．暗に山形を仮定しているので $f(z_0) > 0$, $f''(z) < 0$ とする．$f'(z_0) = 0$ だから $\log f(z)$ を $z = z_0$ の付近でテイラー展開すると

$$\begin{aligned}\log f(z) &\approx \log f(z_0) + \frac{f'(z_0)}{f(z_0)}(z - z_0) + \frac{1}{2} \frac{d^2}{dz^2} \log f(z_0)(z - z_0)^2 \\ &= \log f(z_0) + \frac{1}{2} \frac{d^2}{dz^2} \log f(z_0)(z - z_0)^2.\end{aligned}$$

$$A = -\frac{d^2}{dz^2} \log f(z) \Big|_{z=z_0}$$

とおくと

$$\log f(z) \approx \log f(z) - \frac{1}{2} A (z - z_0)^2.$$

つまり

$$f(z) \approx f(z_0)\exp(-\frac{1}{2}A(z-z_0)^2).$$

よって正規分布で近似すると

$$q(z) = (\frac{A}{2\pi})^{1/2}\exp(-\frac{1}{2}A(z-z_0)^2).$$

M 次元の場合を考える．$p(z) = (1/Z)f(z)$．$z = z_0$ が最大値を与えるなら $p(z_0) > 0$, $\frac{\partial}{\partial z}f(z_0) = 0$．1 次元と同様に

$$A = -\frac{\partial^2}{\partial z_i \partial z_j}\log f(z)\Big|_{z=z_0}$$

とすると $A > 0$（正定値）で

$$f(z) \approx f(z_0)\exp(-\frac{1}{2}(z-z_0)^{\boldsymbol{T}}A(z-z_0)).$$

よって

$$q(z) = \mathcal{N}(z|z_0, A^{-1}) = \sqrt{\frac{|A|}{(2\pi)^M}}\exp(-\frac{1}{2}(z-z_0)^{\boldsymbol{T}}A(z-z_0)).$$

山が複数ある多峰的なときはどのモードを選ぶかでラプラス近似は異なる．総的的にデータ数が多くなるとガウス分布に近づくので近似はよくなるが, ある点での近傍の情報しか利用していないため大域的な特徴がとらえられるとは限らない.

18 モデルの比較と BIC

前節のラプラス近似を行うと正規化係数 Z の近似も分かる．ガウス分布の特性から

$$Z = \int f(z)\,dx \approx \int f(z_0)\exp(-\frac{1}{2}(z-z_0)^{\boldsymbol{T}}A(z-z_0))\,dx = f(z_0)\sqrt{\frac{(2\pi)^M}{|A|}}.$$

データ集合 D, パラメータ $\{\theta_i\}$ の集合 $\{M_i\}$ を考えて各モデルに対して $p(D|\theta_i, M_i)$ を定義する．事前確率 $p(\theta_i|M_i)$ を決めてモデルエビデンス $p(D|M_i)$ を計算してみよう．以下 M_i を略す．また行列作用素 $(\frac{\partial^2}{\partial\theta_i\partial\theta_j})$ を ∇^2 と書くことにする.

$$p(D) = \int p(D|\theta)p(\theta)\,d\theta.$$

$f(\theta) = p(D|\theta)p(\theta)$, $Z = p(D)$ とすると $\theta = \theta_{\mathrm{MAP}}$ のときのラプラス近似を用いて

$$A = -\nabla^2 \log p(D|\theta_{\mathrm{MAP}})p(\theta_{\mathrm{MAP}}).$$

$$\log p(D) = \log Z \approx \log f(\theta_{\mathrm{MAP}}) + \log\sqrt{\frac{(2\pi)^M}{|A|}} = \log p(D|\theta_{\mathrm{MAP}}) + (\log p(\theta_{\mathrm{MAP}}) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|A|).$$

括弧内の後ろ三項を Occam 係数という.

この値をごくごく荒く近似してみる.

$$p(\theta) = \mathcal{N}(\theta|m, B^{-1}),$$

$$H = -\nabla^2 \log p(D|\theta_{\mathrm{MAP}})$$

とすると

$$A = H - \nabla^2 \log p(\theta_{\mathrm{MAP}}) = H + B.$$

$$\log p(\theta) = \log |B|^{1/2} - (M/2)\log(2\pi) - (1/2)(\theta - m)^{\boldsymbol{T}} B(\theta - m)$$

を使って

$$\log p(D) \approx \log p(D|\theta_{\mathrm{MAP}}) - \frac{1}{2}(\theta_{\mathrm{MAP}} - m)^{\boldsymbol{T}} B(\theta_{\mathrm{MAP}} - m) - \frac{1}{2}\log|H+B| + \frac{1}{2}\log|B|.$$

データ集合 D の点の個数を N とし, それぞれが独立であると考えると H の各要素は $O(N)$. B は N や次元の数 M には依存しないので M, N を大きくしたとき無視できるとすると C を定数として

$$\log|H| \approx \log \prod^{M}(NC) = M\log N + M\log C \approx M\log N.$$

よって

$$\log p(D) \approx \log p(D|\theta_{\mathrm{MAP}}) - \frac{1}{2}M\log N.$$

これをベイズ情報量基準（Bayesian Information Criterion, BIC）と呼ばれるモデルの良さを評価するための指標に一致する. かなり無理筋な近似ではあるが, ラプラス近似が（別の方法で導出される）BIC と関連があることを暗示している.

19 ディラックのデルタ関数

ヘヴィサイド関数 $H(x)$ を $\mathbb{R}$ から $\mathbb{R}$ への関数で

$$H(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x \le 0) \end{cases}$$

と定義する. 気持ち的にはデルタ関数はヘヴィサイド関数の微分である.

$$H'(x) = \delta(x).$$

$x = 0$ 以外では微分すると 0 で, $x = 0$ では無限大になるので

$$\delta(x) = \begin{cases} \infty & (x = 0) \\ 0 & (x \ne 0) \end{cases}$$

という感じである. ただ, 実際にはデルタ関数は積分を通してしか扱われない. 厳密には測度論や関数解析の理論を用いて正当化されなければならないが次のような関係式が成り立つ.

$$\int_{-\infty}^{\infty} \delta(x)\,dx = \Big[H(x)\Big]_{-\infty}^{\infty} = H(\infty) - H(-\infty) = 1.$$

実数値関数 $f(x)$ に対して部分積分を適用すると

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\delta(x)\,dx &= \Big[H(x)f(x)\Big]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} H(x)f'(x)\,dx \\
&= f(\infty) - \int_{0}^{\infty} f'(x)\,dx \\
&= f(\infty) - (f(\infty) - f(0)) \\
&= f(0).
\end{aligned}$$

20 ロジスティックシグモイド関数とプロビット関数の逆関数

ロジスティックシグモイド

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

プロビット関数の逆関数

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,d\theta.$$

$\Phi(\lambda a)$ と $\sigma(a)$ を近似するように λ を調節する．原点で同じ向きになるようにしよう．

$$\sigma'(0) = \sigma(a)(1-\sigma(a))\Big|_{a=0} = \frac{1}{2}(1-\frac{1}{2}) = \frac{1}{4}.$$

$$\Phi'(\lambda a) = \frac{1}{\sqrt{2\pi}}\lambda \exp(-\frac{1}{2}(\lambda a)^2)$$

より $\Phi'(0) = \lambda/\sqrt{2\pi}$. この二つが等しいので

$$\frac{\lambda}{\sqrt{2\pi}} = \frac{1}{4}.$$

よって $\lambda = \sqrt{\pi/8}$.

Φ と $\mathcal{N}$ に関する畳み込み計算の関係式：$\lambda > 0$ として

$$\int_{-\infty}^{\infty} \Phi(\lambda a)\mathcal{N}(a|\mu,\sigma^2)\,da = \Phi(\frac{\lambda\mu}{\sqrt{1+\lambda^2\sigma^2}}). \tag{1}$$

これを示そう．左辺を L, 右辺を R とおく．まずガウス分布に関する積分を簡単にする．

$$\mathcal{N}(a|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2).$$

$$f(x) = \mathcal{N}(x|0,1) = \frac{1}{2\pi}\exp(-\frac{1}{2}x^2)$$

とおく．

$$\Phi(x) = \int_{-\infty}^{x} f(y)\,dy.$$

$a = \mu + \sigma x$ とおくと $da = \sigma dx$ で

$$\mathcal{N}(a|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(\mu+\sigma x-\mu)^2) = \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2) = \frac{f(x)}{\sigma}.$$

よって

$$L = \int_{-\infty}^{\infty} \Phi(\lambda\mu+\lambda\sigma x)\frac{f(x)}{\sigma}\sigma\,dx = \int_{-\infty}^{\infty}\Phi(\lambda\mu+\lambda\sigma x)f(x)\,dx.$$

まず L と R の μ に関する微分が等しいことを示す．$\Phi'(x) = f(x)$ なので

$$\frac{\partial}{\partial\mu}R = f(\frac{\lambda\mu}{\sqrt{1+\lambda^2\sigma^2}})\frac{\lambda}{\sqrt{1+\lambda^2\sigma^2}}.$$

$$\frac{\partial}{\partial\mu}L = \int_{-\infty}^{\infty}\lambda f(\lambda\mu+\lambda\sigma x)f(x)\,dx = \lambda\int_{-\infty}^{\infty}\Big(\frac{1}{\sqrt{2\pi}}\Big)^2\exp(-\frac{1}{2}(\lambda\mu+\lambda\sigma x)^2 - \frac{1}{2}x^2)\,dx.$$

exp() の内側を x について平方完成しよう：

$$-\frac{1}{2}((\lambda\sigma)^2+1)x^2+2\lambda^2\mu\sigma x+\lambda^2\mu^2)=-\frac{1}{2}\Big((\sqrt{1+\lambda^2\sigma^2}x+\frac{\lambda^2\mu\sigma}{\sqrt{1+\lambda^2\sigma^2}})^2+\lambda^2\mu^2-\frac{\lambda^4\mu^2\sigma^2}{1+\lambda^2\sigma^2}\Big).$$

定数項は

$$\lambda^2\mu^2-\frac{\lambda^4\mu^2\sigma^2}{1+\lambda^2\sigma^2}=\frac{\lambda^2\mu^2+\lambda^4\mu^2\sigma^2-\lambda^4\mu^2\sigma^2}{1+\lambda^2\sigma^2}=\frac{\lambda^2\mu^2}{1+\lambda^2\sigma^2}.$$

$$\int_{-\infty}^{\infty}\exp(-\frac{1}{2\sigma^2}x^2)\,dx=\sqrt{2\pi}\sigma$$

より

$$I:=\int_{-\infty}^{\infty}\exp(-\frac{1}{2}(\sqrt{1+\lambda^2\sigma^2}x+\frac{\lambda^2\mu\sigma}{\sqrt{1+\lambda^2\sigma^2}})^2)\,dx=\frac{\sqrt{2\pi}}{\sqrt{1+\lambda^2\sigma^2}}.$$

よって

$$\begin{aligned}\frac{\partial}{\partial\mu}L&=\frac{\lambda}{2\pi}\exp(-\frac{1}{2}\Big(\frac{\lambda^2\mu^2}{1+\lambda^2\sigma^2}\Big))I\\&=\frac{\lambda}{2\pi}\exp(-\frac{1}{2}\Big(\frac{\lambda^2\mu^2}{1+\lambda^2\sigma^2}\Big))\frac{\sqrt{2\pi}}{\sqrt{1+\lambda^2\sigma^2}}\\&=\frac{\lambda}{\sqrt{1+\lambda^2\sigma^2}}f(\frac{\lambda\mu}{\sqrt{1+\lambda^2\sigma^2}})\\&=\frac{\partial}{\partial\mu}R.\end{aligned}$$

つまり L と R は定数の差を除いて等しいことが分かった．定数項が 0 であることを示す．$\mu\to\infty$ で

$$L\to\int_{-\infty}^{\infty}\Phi(\infty)f(x)\,dx=\int_{-\infty}^{\infty}f(x)\,dx=1.$$

$$R\to\int_{-\infty}^{\infty}f(y)\,dy=1.$$

よって $L=R$ が示された．

21 ベイズロジスティック回帰

ベイズロジスティック回帰にラプラス近似を行ってみよう．

事前確率分布：$y_n=\sigma(w^{\boldsymbol{T}}\phi_n)$ とおいて

$$p(t|w)=\prod_{n=1}^{N}y_n^{t_n}(1-y_n)^{1-t_n}.$$

事前ガウス分布：ハイパーパラメータ m_0, S_0 を用いて

$$p(w)=\mathcal{N}(w|m_0,S_0).$$

事後分布は N 次元縦ベクトル $t=(t_0,\dots,t_N)^{\boldsymbol{T}}$ を用いて

$$p(w|t)\propto p(w)p(t|w).$$

よって

$$\log p(w|t) = -\frac{1}{2}(w - m_0)^{\boldsymbol{T}} S_0^{-1}(w - m_0) + \sum_{n=1}^{N}(t_n \log y_n + (1 - t_n)\log(1 - y_n)) + \mathrm{const}.$$

これを最大化する最大事後確率（MAP）の解 w_{MAP} を何らかの方法で求める．ラプラス近似をするために w_{MAP} を平均とするガウス分布で近似しよう．以後 w_{MAP} を m_N と書く．共分散は今まで何度もやった通り

$$S_N^{-1} = -\nabla^2 \log p(w|t) = S_0^{-1} + \sum_{n=1}^{N} y_n(1 - y_n)\phi_n \phi_n^{\boldsymbol{T}}.$$

よって事後確率分布をガウス分布で近似すると

$$q(w) = \mathcal{N}(w|m_N, S_N).$$

次に C_1 についての予測分布を求める．

$$p(C_1|\phi, t) = \int p(C_1|\phi, w)p(w|t)\,dw \approx \int \sigma(w^{\boldsymbol{T}}\phi)q(w)\,dw.$$

デルタ関数の性質から

$$\sigma(w^{\boldsymbol{T}}\phi) = \int \delta(a - w^{\boldsymbol{T}}\phi)\sigma(a)\,da.$$

よって

$$\begin{aligned}
p(C_1|\phi, t) &\approx \int \Big(\int \delta(a - w^{\boldsymbol{T}}\phi)\sigma(a)q(w)\,da\Big)dw \\
&= \int \Big(\int \delta(a - w^{\boldsymbol{T}}\phi)q(w)\,dw\Big)\sigma(a)\,da \\
&= \int p(a)\sigma(a)\,da.
\end{aligned}$$

ただし

$$p(a) = \int \delta(a - w^{\boldsymbol{T}}\phi)q(w)\,dw$$

とおいた．

$p(a)$ の平均を求める．$q(w)$ の平均が m_N であることに注意すると，

$$\begin{aligned}
\mu_a = E[a] &= \int p(a)a\,da = \int\int \delta(a - w^{\boldsymbol{T}}\phi)q(w)a\,dwda \\
&= \int \Big(\int \delta(a - w^{\boldsymbol{T}}\phi)a\,da\Big)q(w)\,dw = \int q(w)(w^{\boldsymbol{T}}\phi)\,dw \\
&= (\int q(w)w\,dw)^{\boldsymbol{T}}\phi = E[w]^{\boldsymbol{T}}\phi = m_N^{\boldsymbol{T}}\phi.
\end{aligned}$$

分散は

$$\begin{aligned}
\sigma(a)^2 &= \int p(a)(a^2 - E[a]^2)\,da = \int (\delta(a - w^{\boldsymbol{T}}\phi)a^2\,da)q(w)\,dw - \int (\delta(a - w^{\boldsymbol{T}}\phi)E[a]^2\,da)q(w)\,dw \\
&= \int q(w)(w^{\boldsymbol{T}}\phi)^2\,dw - \int q(w)(m_N^{\boldsymbol{T}}\phi)^2\,dw \\
&= \phi^{\boldsymbol{T}}\Big(\int q(w)(ww^{\boldsymbol{T}} - m_N m_N^{\boldsymbol{T}})\,dw\Big)\phi.
\end{aligned}$$

括弧内は

$$E[ww^{\boldsymbol{T}}] - m_N m_N^{\boldsymbol{T}} \int q(w)\,dw = (m_N m_N^{\boldsymbol{T}} + S_N) - m_N m_N^{\boldsymbol{T}} = S_N$$

より

$$\sigma_a^2 = \phi^{\boldsymbol{T}} S_N \phi.$$

これは線形回帰モデルの予測分布の分散（PRML 式 3.59）

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^{\boldsymbol{T}} S_N \phi(x)$$

でノイズを消したもの（$(1/\beta) \to 0$）は σ_a^2 に一致する．

よって予測分布の近似は

$$p(C_1|t) = \int \sigma(a)p(a)\,da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)\,da.$$

別の方法でも求めてみる：
w の座標を座標変換して第一成分の単位基底ベクトル e が ϕ と同じ向きになるようにとり，残りは ϕ と直交するようにとる．つまり

$$e = \phi/||\phi||.$$

それに関する w の係数を $w = (w_1, w_2)^{\boldsymbol{T}}$ と書く．w_1 は 1 次元ベクトルの係数で非負，w_2 は $M-1$ 次元である．すると w と ϕ の内積は e の方向のみが残るので

$$w^{\boldsymbol{T}}\phi = w_1||\phi||.$$

$q(w) = q(w_1, w_2) = q(w_2|w_1)q(w_1)$ より

$$\begin{aligned}
p &= \int \sigma(w^{\boldsymbol{T}}\phi)q(w)\,dw = \int \sigma(w_1||\phi||) \int q(w_2|w_1)\,dw_2 dw_1 \\
&= \int \sigma(w_1||\phi||)\Big(\int q(w_2|w_1)\,dw_2\Big)q(w_1)dw_1 \\
&\quad \text{括弧内は 1 なので} \\
&= \int \sigma(w_1||\phi||)q(w_1)\,dw_1.
\end{aligned}$$

$q(w)\mathcal{N}(w|m_N, S_N)$ で $w = (w_1, w_2)^{\boldsymbol{T}}$ と書いたときの w_1 に関する周辺分布は e 方向への射影化になる：

$$q(w_1) = \mathcal{N}(w_1|e^{\boldsymbol{T}} m_N, e^{\boldsymbol{T}} S_N e).$$

$a = w_1||\phi||$ と書くと a に関しては平均は $||\phi||$ 倍，分散は $||\phi||^2$ 倍されるので

$$q(a) = \mathcal{N}(a|(||\phi||e)^{\boldsymbol{T}} m_N, ||\phi||e^{\boldsymbol{T}} S_N ||\phi||e) = \mathcal{N}(a|\phi^{\boldsymbol{T}} m_N, \phi^{\boldsymbol{T}} S_N \phi).$$

ロジスティックシグモイド関数でのガウス分布のたたみこみ積分は解析的には難しい．ここではロジスティックシグモイド関数をプロビット関数の逆関数で近似することでたたみ込み積分を解析的に扱おう．

前節の結果から $\lambda = \sqrt{\pi/8}$ として $\sigma(a) \approx \Phi(\lambda a)$ と近似すると

$$\begin{aligned}
\int \sigma(a)\mathcal{N}(a|\mu, \sigma^2)\,da &\approx \sigma\Big(\frac{\mu}{\sqrt{1+\lambda^2\sigma^2}}\Big) \\
&\quad \kappa(\sigma^2) = 1/\sqrt{1+(\pi/8)\sigma^2} \text{とおくと} \\
&= \sigma(\kappa(\sigma^2)\mu).
\end{aligned}$$

よって

$$p(C_1|\phi,t) \approx \sigma(\kappa(\sigma_a^2)\mu_a)$$

という近似予測分布が得られた.