

第 2 章

「確率分布」のための数学

この章は『パターン認識と機械学習』（PRML）の 2 章、ガウス分布のところを理解するために必要な数学をまとめてみたものです。目標はガウス分布の最尤推定の式変形をきちんと追えるようになることです。いくつかの定理は証明せずに認めますが、可能な限り self-contained であることを目指してみました。概ね PRML に従っていますが、違う方法をとっているところもあります。

2.1 微積分の復習

2.1.1 微分の定義

微分の積の公式も忘れたなあという人のために、微分について軽く復習しておこう。関数 $y = f(x)$ が与えられたとき、点 $x = a$ における微分係数 $f'(a)$ とはその点でのグラフの接線の傾きのことであった。

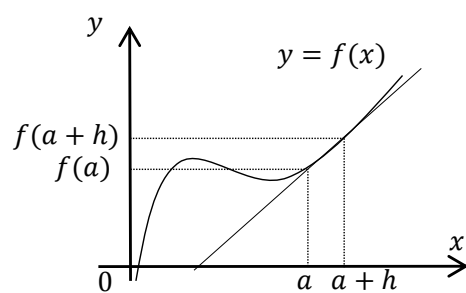


図 2.1 $y = f(x)$

h を十分小さい値ならば、 $x = a$ での接線の傾きは区間 $[a, a+h]$ での平均の傾きで近似

できるだろう：

$$f'(a) = a \text{ における傾き } \approx \frac{f(a+h) - f(a)}{(a+h) - a} = \frac{f(a+h) - f(a)}{h}.$$

両辺を h 倍して移行すると

$$f(a+h) \approx f(a) + f'(a)h.$$

この式は f の点 a における値 $f(a)$ と傾き $f'(a)$ で a の付近の値を直線で近似したということを表している (h について線形) . a を x で置き換えて

$$f(x+h) = f(x) + f'(x)h + \epsilon$$

とかくことにする. ϵ は h に比べて十分小さい h の関数である.

さて, 二つの関数 $f(x)$ と $g(x)$ があったとき, その積の関数 $s(x) = f(x)g(x)$ の微分はどうなるだろうか.

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \epsilon_1 \\ g(x+h) &= g(x) + g'(x)h + \epsilon_2 \end{aligned}$$

を $s(x+h)$ に代入して計算してみよう：

$$\begin{aligned} s(x+h) &= f(x+h)g(x+h) = (f(x) + f'(x)h + \epsilon_1)(g(x) + g'(x)h + \epsilon_2) \\ &= f(x)g(x) + (f'(x)g(x) + f(x)g'(x))h + (h \text{ より十分小さい}) \\ &= s(x) + (f'(x)g(x) + f(x)g'(x))h + \epsilon_3 \end{aligned}$$

となる. つまり

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

が成り立つ. これが積の微分である.

もう一つ試してみよう. 今度は $y = f(x)$ と $z = g(y)$ という関数があったときにその合成関数 $z = g(f(x))$ の微分を考えてみる.

$$\begin{aligned} g(f(x+h)) &= g(f(x) + f'(x)h + \epsilon_1) = g(f(x)) + g'(f(x))(f'(x)h + \epsilon_1) + \epsilon_2 \\ &= g(f(x)) + g'(f(x))f'(x)h + \epsilon_3. \end{aligned}$$

つまり

$$(g(f(x)))' = g'(f(x))f'(x)$$

が成り立つ. これは合成関数の微分である. $dy/dx = f'(x)$ という微分の記号を使うと

$$\frac{d(g(f(x)))}{dx} = \frac{dg}{dy} \Big|_{y=f(x)} \frac{dy}{dx}.$$

dy/dx という記号があたかも約分できるように見えるのが面白い. 変数変換ではこの記法が活躍する.

2.1.2 変数変換

$$\int f(x) dx$$

で $x = g(y)$ とすると $dx = g'(y)dy$ より

$$\int f(g(y))g'(y) dy.$$

多変数関数の場合は $g'(y)$ の部分がヤコビ行列の行列式（ヤコビアン）になる.

$x_i = g_i(y_1, \dots, y_n)$ for $i = 1, \dots, n$ とすると

$$\det \left(\frac{\partial(x_1, \dots, x_n)}{\partial(y_1, \dots, y_n)} \right) = \det \left(\frac{\partial x_i}{\partial y_j} \right).$$

ヤコビアンは変数変換したときのある点における微小区間の拡大率を意味する.

適当な条件の下で

$$\begin{aligned} & \int \cdots \int f(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &= \int \cdots \int f(g_1(y_1, \dots, y_n), \dots, g_n(y_1, \dots, y_n)) \left| \det \left(\frac{\partial x_i}{\partial y_j} \right) \right| dy_1 \cdots dy_n. \end{aligned}$$

2.1.3 奇関数の積分

全ての x について $f(-x) = f(x)$ が成り立つとき f を偶関数, $f(-x) = -f(x)$ が成り立つとき f を奇関数という. 奇関数 f について

$$I = \int_{-\infty}^{\infty} f(x) dx = 0.$$

なぜなら $I = \int_{-\infty}^0 f(x) dx + \int_0^{\infty} f(x) dx$ と積分区間を半分に分けてみよう. 第 1 項で $x = -y$ と変数変換すると $f(x)dx = -f(-y)dy = f(y)dy$ となる. 積分範囲は ∞ から 0 になり, 向きが逆転するので入れ換えると符号がひっくり返る. よって第 1 項 $= -\int_0^{\infty} f(y) dy$. 第 2 項と打ち消しあって $I = 0$ となるからである.

\mathbf{x} が n 次元ベクトルのときも同様に全ての \mathbf{x} について $f(-\mathbf{x}) = -f(\mathbf{x})$ となるとき f を奇関数という. やはり

$$I = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) dx_1 \cdots dx_n = 0.$$

なぜなら

$$I = \int_{-\infty}^0 \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} + \int_0^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}$$

と二つの領域に分けて $\mathbf{x} = -\mathbf{y}$ と変数変換すると, $d\mathbf{x} = (-1)^n d\mathbf{y}$. 第1項の積分範囲は $(0, -\infty) \times (\infty, -\infty) \times \cdots \times (\infty, -\infty)$ になり, 第2項の積分範囲に合わせると $(-1)^n$ がでる. よって $f(-\mathbf{y}) = -f(\mathbf{y})$ を使うと

$$I = \int_0^\infty \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty f(-\mathbf{y}) d\mathbf{y} + \int_0^\infty \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty f(\mathbf{x}) d\mathbf{x} = 0.$$

2.1.4 $\exp(-x^2)$ の積分

$$I = \int_0^\infty \exp(-x^2) dx$$

とおくと

$$I^2 = \int_0^\infty \int_0^\infty \exp(-(x^2 + y^2)) dx dy.$$

ここで $x = r \cos(\theta)$, $y = r \sin(\theta)$ と置くと $x^2 + y^2 = r^2$. ヤコビアンは

$$\det \begin{pmatrix} \frac{\partial(x, y)}{\partial(r, \theta)} \end{pmatrix} = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} = r(\cos^2 \theta + \sin^2 \theta) = r.$$

積分範囲は x, y が (x, y) 平面の第一象限全体なので r は 0 から ∞ , θ は 0 から $\pi/2$ を渡る. よって

$$I^2 = \int_0^{\pi/2} \int_0^\infty \exp(-r^2) r dr d\theta = \pi/2 \left[-\frac{1}{2} \exp(-r^2) \right]_0^\infty = \pi/4.$$

よって $I = \sqrt{\pi}/2$. x^2 は偶関数なので積分範囲を $-\infty$ から ∞ にすると 2 倍になって

$$\int_{-\infty}^\infty \exp(-x^2) dx = \sqrt{\pi}.$$

本当は積分の順序を交換したりしているところを気にしないといけないが, ここでは自由に交換できるものと思っておく.

2.1.5 ガウス分布の積分

前章の積分で $a > 0$ をとり $x = \sqrt{a}y$ とすると $dx = \sqrt{a}dy$.

$$\int_{-\infty}^\infty \exp(-x^2) dx = \int_{-\infty}^\infty \exp(-ay^2) \sqrt{a} dy = \sqrt{\pi}.$$

よって

$$\int_{-\infty}^\infty \exp(-ax^2) dx = \sqrt{\pi/a}.$$

ここで両辺を a に関して微分する. 積分の中身は $\frac{\partial}{\partial a} \exp(-ax^2) = -x^2 \exp(-ax^2)$. 気にせず積分と微分を交換することで

$$-\int_{-\infty}^{\infty} x^2 \exp(-ax^2) dx = -1/2\sqrt{\pi}a^{-3/2}.$$

$a = 1/(2\sigma^2)$ と置き換えることで

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx = \sqrt{2\pi}\sigma. \quad (2.1)$$

$$\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx = \sqrt{2\pi}\sigma^3. \quad (2.2)$$

式 (2.1) は正規化項が $\sqrt{2\pi}\sigma$ であることを示している. つまりガウス分布を

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

とすると

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1.$$

平均は

$$x\mathcal{N}(x|\mu, \sigma^2) = (x-\mu)\mathcal{N}(x|\mu, \sigma^2) + \mu\mathcal{N}(x|\mu, \sigma^2)$$

とわけると, 第 1 項は $(x-\mu)$ に関して奇関数なので積分すると消えて

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x\mathcal{N}(x|\mu, \sigma^2) dx = \mu.$$

分散は $x^2 = (x-\mu)^2 + 2\mu(x-\mu) + \mu^2$ なので積分すると第 1 項は式 (2.2) より σ^2 . 第 2 項は 0. 第 3 項は μ^2 . よって

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2\mathcal{N}(x|\mu, \sigma^2) dx = \sigma^2 + \mu^2.$$

2.2 線形代数の復習

2.2.1 行列の積

以下, 特に断らない限り行列の数値は複素数とする.

A を m 行 n 列の行列とする. 横に n 個, 縦に m 個数字が並んでいる. A の i 行 j 列の値が a_{ij} であるとき, $A = (a_{ij})$ とかく. $m = n$ のとき n 次正方行列という. 並んでいる数字が実数値のみからなる行列を実行列という.

A を l 行 m 列の行列, B を m 行 n 列の行列とすると, 積 AB を $(AB)_{ij} = \sum_{k=1}^m a_{ik}b_{kj}$ で定義する. AB は l 行 n 列の行列になる.

1. A, B が正方行列だったとしても $AB = BA$ とは限らない.
2. A, B, C がその順序で掛け算できるとき $(AB)C = A(BC)$ が成り立つ.
 なぜなら $((AB)C)_{ij} = \sum_k (AB)_{ik} c_{kj} = \sum_k (\sum_l a_{il} b_{lk}) c_{kj} = \sum_{k,l} a_{il} b_{lk} c_{kj}$.
 $A(BC)_{ij} = \sum_l a_{il} (BC)_{lj} = \sum_l a_{il} (\sum_k b_{lk} c_{kj}) = \sum_{k,l} a_{il} b_{lk} c_{kj}$ だから.

2.2.2 トレース

A が n 次正方行列のとき $\text{tr}(A) = \sum_{i=1}^n a_{ii}$ と A のトレースと呼ぶ.

$$\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B).$$

$$\text{tr}(AB) = \text{tr}(BA).$$

なぜなら $\text{tr}(AB) = \sum_i (AB)_{ii} = \sum_i \left(\sum_j a_{ij} b_{ji} \right) = \sum_j \left(\sum_i b_{ji} a_{ij} \right) = \sum_j (BA)_{jj} = \text{tr}(BA)$. 3 個の行列の積については $\text{tr}(ABC) = \sum_i (ABC)_{ii} = \sum_i \left(\sum_{j,k} a_{ij} b_{jk} c_{ki} \right) = \sum_{i,j,k} a_{ij} b_{jk} c_{ki}$ より

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB).$$

$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ のときは $\text{tr}(A) = a + d$.

2.2.3 行列式

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

のとき, $|A| = ad - bc$ を A の行列式という. 一般には次のように定義する:

S_n を $1, \dots, n$ の順序を並び替える操作全体の集合とする. たとえば S_2 は何も動かさない操作と 1 を 2 に, 2 を 1 に並び替える操作の二つの操作からなる. n 個の要素を並び替える組み合わせは $n \times (n-1) \times \dots \times 1 = n!$ 通りある.

$D = \prod_{i < j} (x_i - x_j)$ とし, $S_n \ni \sigma$ に対して $\sigma D = \prod_{i < j} (x_{\sigma(i)} - x_{\sigma(j)})$ とすると, D と σD は符号しか変わらない. $\sigma D = \text{sgn}(\sigma) D$ で $\text{sgn}(\sigma) \in \{1, -1\}$ を定義する.

$\{\sigma(1), \dots, \sigma(n)\}$ を 2 個ずつ順序を入れ換えて $\{1, \dots, n\}$ に並び替えられたとき, 偶数回でできたら $\text{sgn}(\sigma) = 1$, 奇数回でできたら $\text{sgn}(\sigma) = -1$ である. これを使って行列式を定義する.

A を n 次正方行列 (n 行 n 列) とするとき,

$$\det(A) = |A| = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)}.$$

A が 2 次正方行列のときを見直してみる. S_2 は 2 個の要素しかもたなかった. 一つは何も動かさない操作でそれに対して sgn は 1. もう一つは 1 と 2 を入れ換える操作で sgn は -1 となる. よって

$$|A| = a_{11}a_{22} - a_{12}a_{21}.$$

ここで二つの n 次正方行列 A, B に対して $|AB| = |A||B|$ が成り立つ. 2 次のときのみ確認しておこう.

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, B = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$$

とすると,

$$\begin{aligned} |AB| &= \begin{vmatrix} ax+bz & ay+bw \\ cx+dz & cy+dw \end{vmatrix} = (ax+bz)(cy+dw) - (ay+bw)(cx+dz) \\ &= (ad-bc)(xw-yz) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} \begin{vmatrix} x & y \\ z & w \end{vmatrix} = |A||B|. \end{aligned}$$

一般のときの証明は省略する. $|A|$ は負の数にもなることに注意する (絶対値の記号と間違えないように).

2.2.4 行列の種類

A を m 行 n 列の行列とする. A に対して $A^T = (a_{ji})$ を A の転置行列という. これは n 行 m 列の行列である. $|A^T| = |A|$, $(AB)^T = B^T A^T$ である.

$\overline{A} = (\overline{a_{ij}})$ を A の複素共役行列, $A^* = \overline{A^T}$ を随伴行列という. $|A^*| = \overline{|A|}$, $(AB)^* = B^* A^*$ である.

a_{ii} を対角成分という. 対角成分以外の項が 0 である行列を対角行列といい $\text{diag}(a_0, \dots, a_n)$ と書く.

$$\delta_{ij} = \begin{cases} 1 & (i=j) \\ 0 & (i \neq j) \end{cases}$$

をクロネッカーの δ といい, $I_n = (\delta_{ij})$ を n 次単位行列という. I と略すこともあるし E と書くこともある. A を n 次正方行列とすると $AI_n = I_n A = A$.

A が n 次正方行列で, $|A| \neq 0$ のとき A を正則といい, $AB = BA = I$ となる行列 B が存在する. B を逆行列といい, A^{-1} と書く.

1. 逆行列は存在すればただ一つである. なぜなら B, B' を逆行列とすると $B = BI = B(AB') = (BA)B' = IB' = B'$.
2. 有限次元では $AB = I$ ならば $BA = I$ であることを示すことができる. つまり $AB = I$ だが $BA \neq I$ なものは存在しない (無限次元ではそのような行列を構成できる).

n 次正方行列 A について

1. $|A| \neq 0$ なものの全体を $GL_n(\mathbb{C})$ と書く. 実正則行列全体は $GL_n(\mathbb{R})$ と書く.
2. $|A| = 1$ なものの全体を $SL_n(\mathbb{C})$ とかく. 実行列のときは $SL_n(\mathbb{R})$.
3. $AA^* = I$ となるときユニタリー行列といい, その全体を $U(n)$ と書く. このとき $|AA^*| = ||A||^2 = 1$. ここで $||A||$ の内側の $||$ は行列式, 外側の $||$ は数値の絶対値である. ユニタリー行列であって, 更に $|A| = 1$ なものの全体を $SU(n)$ と書く.
4. 実行列 A が $AA^T = I$ となるとき, 直交行列といい, その全体を $O(n)$ と書く. このとき $||A||^2 = 1$. 更に $|A| = 1$ なものの全体を $SO(n)$ と書く. $|A| \in \mathbb{R}$ なので $|A| = \pm 1$.
5. $A = A^T$ となるとき対称行列という.

2.2.5 ブロック行列の逆行列

A, D を正方行列として (B, C は正方行列とは限らない)

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

の逆行列を求めてみよう. 逆行列を

$$X^{-1} = \begin{pmatrix} M & N \\ L & P \end{pmatrix}$$

とおくと

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} M & N \\ L & P \end{pmatrix} = \begin{pmatrix} AM + BL & AN + BP \\ CM + DL & CN + DP \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

2-1 ブロックに左から D^{-1} を掛けて $L = -D^{-1}CM$. これを 1-1 ブロックに代入して

$$AM + BL = AM - BD^{-1}CM = (A - BD^{-1}C)M = I.$$

よって $M = (A - BD^{-1}C)^{-1}$. 今度は

$$\begin{pmatrix} M & N \\ L & P \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} MA + NC & MB + ND \\ LA + PC & LB + PD \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

の 1-2 ブロックに右から D^{-1} を掛けて $N = -MBD^{-1}$.

2-2 ブロックに右から D^{-1} を掛けて $P = D^{-1} - LBD^{-1} = D^{-1} + D^{-1}CMBD^{-1}$.

よって $M = (A - BD^{-1}C)^{-1}$ として

$$\begin{pmatrix} M & N \\ L & P \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}.$$

これが X の逆行列となることは容易に確認できる.

(以下余談) $R = MB, S = D^{-1}C$ とおくと

$$X = \begin{pmatrix} M^{-1} & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I + MBD^{-1}C & MB \\ D^{-1}C & I \end{pmatrix} = \begin{pmatrix} M^{-1} & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I + RS & R \\ S & I \end{pmatrix}$$

と変形できることはすぐ分かる.

$$\begin{pmatrix} M^{-1} & 0 \\ 0 & D \end{pmatrix}^{-1} = \begin{pmatrix} M & 0 \\ 0 & D^{-1} \end{pmatrix},$$

$$\begin{pmatrix} I + RS & R \\ S & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -R \\ -S & I + SR \end{pmatrix}$$

なので X^{-1} もすぐ求められる. 更にこの行列の行列式は 1 なので

$$|X| = \begin{vmatrix} M^{-1} & 0 \\ 0 & D \end{vmatrix} = |M|^{-1}|D| = |A - BD^{-1}C||D|.$$

2.2.6 三角化

n 次正方行列 A に対して $a_{ij} = 0 (i > j)$ のとき (上半) 三角行列という.

$$A = \begin{pmatrix} a_{11} & \dots & * \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{nn} \end{pmatrix}.$$

この $*$ は任意の値が入っていることを示す.

このとき $|A| = \prod_i a_{ii}$ である. なぜなら行列式の定義で 1 行ごとに異なる列のものをとっていったものの積を考えるわけだが, 最初に a_{11} 以外の $a_{1j} (j > 1)$ を選択すると, 残り $n-1$ 個をとる中で 0 でないものは $n-2$ 個しかない. したがって必ず 0 になる. 以下同様にして対角成分を拾ったものしか残らないからである.

さて次の定理を証明無しで認める:

任意の n 次正方行列 A に対して,
あるユニタリー行列 P があって $P^{-1}AP$ を三角化できる. (2.3)

(注意) 一般の行列が常に対角化できるとは限らないが三角化は常にできる.

2.2.7 対称行列

A を n 次実対称行列とする.

n 次実対称行列 A に対して, ある行列 P が存在して $P^{-1}AP$ を実対角化できる. (2.4)

定理 (2.3) を用いて証明しよう.

A に対してあるユニタリー行列 P があって $P^{-1}AP$ を三角化できる:

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}.$$

この両辺の随伴をとる. P はユニタリー行列なので $PP^* = I$. つまり $P^{-1} = P^*$. さらに A は実対称行列なので $A^* = A$ に注意すると

$$P^*A^*(P^{-1})^* = P^{-1}AP = \begin{pmatrix} \overline{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ * & \cdots & \overline{\lambda_n} \end{pmatrix}.$$

この二つの式が同一なので $\overline{\lambda_i} = \lambda_i$ かつ $*$ の部分が 0. これは $\lambda_i \in \mathbb{R}$ で, $P^{-1}AP$ はもともと対角行列であったことを意味する.

実は P が実行列であるようにもできる. そのときは P は直交行列になるので $|P| = \pm 1$.

もし $|P| = -1$ だったとすると, I' を単位行列の 1 行目と 2 行目を入れ換えたものとして $P' = PI'$ とおいて

$$P'^{-1}AP' = I'(P^{-1}AP)I' = I' \text{diag}(\lambda_1, \dots, \lambda_n)I' = \text{diag}(\lambda_2, \lambda_1, \lambda_3, \dots, \lambda_n).$$

これはもとの対角成分の 1 番目と 2 番目を入れ換えたものである. $|I'| = -1$, $|P'| = |P||I'| = 1$ なのでもともと $|P| = 1$ だったとしてもよい. 従ってより強く

n 次実対称行列 A に対して, ある $P \in SO(n)$ が存在して $P^{-1}AP$ を実対角化できる. (2.5)

がいえる. PRML では直交行列の行列式が 1 であることを暗に仮定しているときがあるが不正確 (cf. (C.37) 付近). たとえば PRML 式 (2.54) で $|J|^2 = 1$ から $|J| = 1$ を出しているが, $|J| = -1$ の可能性もある. 予め U を $SO(n)$ の元としてとっておけば

$$|J| = |U| = 1$$

ですむ. ただし多重積分を考えるときはヤコビアン絶対値のみが関係するのでここでは $\|J\| = 1$ が言えれば十分である.

2.2.8 2 次形式

A を一般に n 次正方行列とし, \mathbf{x} を n 次元縦ベクトルとする.

$$\mathbf{x}^T A \mathbf{x} = \sum_i x_i (A\mathbf{x})_i = \sum_i x_i \left(\sum_j a_{ij} x_j \right) = \sum_{i,j} a_{ij} x_i x_j \quad (2.6)$$

を x の 2 次形式という.

A が与えられたときに $S = (A + A^T)/2$, $T = (A - A^T)/2$ とすると, $A = S + T$, $S^T = S$, $T^T = -T$ となる. $T^T = -T$ ということは $t_{ij} = -t_{ji}$ なので (標数 2 ではないから) $t_{ii} = 0$. 式 (2.6) の和を $i = j$ と $i \neq j$ の二つに分けて $A = T$ として適用すると

$$\mathbf{x}^T T \mathbf{x} = \sum_i t_{ii} x_i x_i + \sum_{i < j} (t_{ij} + t_{ji}) x_i x_j.$$

第 1 項は $t_{ii} = 0$ より 0. 第 2 項も $t_{ij} = -t_{ji}$ より 0. つまり $T^T = -T$ のとき 2 次形式の値は 0 となる. よって $\mathbf{x}^T A \mathbf{x} = \mathbf{x}^T S \mathbf{x} + \mathbf{x}^T T \mathbf{x} = \mathbf{x}^T S \mathbf{x}$. つまり 2 次形式を考えるときは一般性を失うことなく A を対称行列としてよい.

2 変数のときを見てみる. 行列の計算は分かりにくければとりあえず 2 次で書いてみる.

$$\begin{pmatrix} x & y \end{pmatrix}^T \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix}^T \begin{pmatrix} ax + by \\ bx + cy \end{pmatrix} = ax^2 + 2bxy + cy^2.$$

ブロック行列なら A, C を対称行列として

$$\begin{pmatrix} x & y \end{pmatrix}^T \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix}^T \begin{pmatrix} Ax + By \\ B^T x + Cy \end{pmatrix} = \mathbf{x}^T A \mathbf{x} + 2\mathbf{x}^T B \mathbf{y} + \mathbf{y}^T C \mathbf{y}.$$

ここで $\mathbf{x}^T B \mathbf{y}$ はスカラー値なので転置しても変わらない, つまり

$$\mathbf{x}^T B \mathbf{y} = (\mathbf{x}^T B \mathbf{y})^T = \mathbf{y}^T B^T (\mathbf{x}^T)^T = \mathbf{y}^T B^T \mathbf{x}$$

を用いた. 対称行列は $SO(n)$ の元 P を用いて対角化できた ($PP^T = I$). $\mathbf{y} = P^{-1}\mathbf{x}$ とおくと

$$\mathbf{x}^T A \mathbf{x} = \mathbf{y}^T P^T A P \mathbf{y} = \mathbf{y}^T \text{diag}(\lambda_1, \dots, \lambda_n) \mathbf{y} = \sum_{i=1}^n \lambda_i y_i^2.$$

つまり 2 次形式は対角化すれば単なる成分ごとの直和になる.

2.3 多変量ガウス分布

A を n 次実対称行列, \mathbf{z} を n 次元縦ベクトルとしてする. まず

$$f(\mathbf{z}) = \exp\left(-\frac{1}{2}\mathbf{z}^T A^{-1}\mathbf{z}\right)$$

を考える. これは \mathbf{z} について偶関数である. A を直交行列 P で対角化する. $P^{-1}AP = \text{diag}(\lambda_1, \dots, \lambda_n)$ より, $P^{-1}A^{-1}P = \text{diag}(\lambda_1^{-1}, \dots, \lambda_n^{-1})$. $\mathbf{y} = P^{-1}\mathbf{z}$ と置いて前節の変形を行うと

$$f(\mathbf{z}) = \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{z_i^2}{\lambda_i}\right) = \prod_{i=1}^n \exp\left(-\frac{z_i^2}{2\lambda_i}\right).$$

ここで区間 $(-\infty, \infty)$ の積分を考えるが, そうすると積分値が発散しないためには全ての $\lambda_i > 0$ である必要がある. 以下この条件を仮定する. このとき $|A| = \prod_i \lambda_i > 0$.

積分値は式 (2.1) より

$$\int f(\mathbf{z}) d\mathbf{z} = \prod_{i=1}^n \sqrt{2\pi\lambda_i} = \sqrt{2\pi}^n \sqrt{|A|}.$$

よって

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, A) = \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{|A|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T A^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

とすると正規化されている. これが多変量版のガウス分布である.

平均を求めよう:

$$\mathbf{x}f(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})f(\mathbf{x} - \boldsymbol{\mu}) + \boldsymbol{\mu}f(\mathbf{x} - \boldsymbol{\mu})$$

とすると第1項は $(\mathbf{x} - \boldsymbol{\mu})$ に関して奇関数なので積分すると消える. 後者は $\boldsymbol{\mu}$ が定数で外に出るので

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}.$$

次に分散を考える. $\mathbf{x} - \boldsymbol{\mu} = P\mathbf{y}$ とおくと

$$\mathbf{x}\mathbf{x}^T = (P\mathbf{y} + \boldsymbol{\mu})(P\mathbf{y} + \boldsymbol{\mu})^T = P\mathbf{y}\mathbf{y}^T P^T + P\mathbf{y}\boldsymbol{\mu}^T + \boldsymbol{\mu}\mathbf{y}^T P^T + \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

これに $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, A)$ をかけた値を積分するわけだが, 第2項, 第3項は奇関数になるので0. 第4項は $\boldsymbol{\mu}\boldsymbol{\mu}^T$ が定数に出る.

第1項を考えよう: $P = (\mathbf{p}_1, \dots, \mathbf{p}_n)$ とすると $(P\mathbf{y})_i = \sum_{j=1}^n p_{ij}y_j$ だから

$$P\mathbf{y} = \sum_{j=1}^n y_j \mathbf{p}_j.$$

よって第1項 $\times f(\mathbf{x} - \boldsymbol{\mu})$ は

$$\sum_{i,j} \mathbf{p}_i \mathbf{p}_j^T y_i y_j \prod_{k=1}^n \exp\left(-\frac{y_k^2}{2\lambda_k}\right).$$

積分すると $i \neq j$ のところでは $y_i \exp(-y_i^2/(2\lambda_i))$ が奇関数になるので0.

$i = j$ のところでは $y_i^2 \exp(-\frac{y_i^2}{2\lambda_i})$ から λ_i がで, それ以外では1. よって

$$\sum_i \mathbf{p}_i \mathbf{p}_i^T \lambda_i = (\mathbf{p}_1, \dots, \mathbf{p}_n) \text{diag}(\lambda_1, \dots, \lambda_n) \begin{pmatrix} \mathbf{p}_1^T \\ \vdots \\ \mathbf{p}_n^T \end{pmatrix} = P \text{diag}(\lambda_1, \dots, \lambda_n) P^T = A.$$

第4項と合わせて, 結局

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + A.$$

2.4 行列の微分

ここではガウス分布の最尤推定で使ういくつかの公式を列挙する。PRML の付録 C ではたとえば (C.22) を (C.33) や (C.47) を使って示せとあるが、それだと対称行列や対角化ができる行列に対してしか示せていない中途半端なものである。これらはもっと一般的な行列で成り立つ。

A を n 次正方行列とする。

2.4.1 2 次形式の別の表現

\mathbf{x} を n 次縦ベクトルとすると、

$$\mathbf{x}^T A \mathbf{x} = \sum_{i,j} a_{ij} x_i x_j = \sum_i \left(\sum_j a_{ij} (\mathbf{x} \mathbf{x}^T)_{ji} \right) = \sum_i (A \mathbf{x} \mathbf{x}^T)_{ii} = \text{tr}(\mathbf{x} \mathbf{x}^T). \quad (2.7)$$

この式は A が対称行列でなくても成り立つことに注意する。

2.4.2 内積の微分

\mathbf{x}, \mathbf{y} を縦ベクトルとして

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^T \mathbf{y}) &= \mathbf{y}. \\ \frac{\partial}{\partial \mathbf{y}} (\mathbf{x}^T \mathbf{y}) &= \mathbf{x}. \end{aligned}$$

ここで $\frac{\partial}{\partial \mathbf{x}}$ は $\frac{\partial}{\partial x_i}$ を縦に並べた縦ベクトルとする。 $\frac{\partial}{\partial \mathbf{x}}$ を ∇ と書くこともあるが PRML では場所によって縦ベクトル (2.228) だったり、横ベクトル (3.13) だったりする。常に縦ベクトルとしたほうが混乱は少ない。

証明は $\mathbf{x}^T \mathbf{y} = \sum_j x_j y_j$ なので

$$\begin{aligned} \frac{\partial}{\partial x_i} (\mathbf{x}^T \mathbf{y}) &= \sum_j \delta_{ij} y_j = y_j. \\ \frac{\partial}{\partial y_i} (\mathbf{x}^T \mathbf{y}) &= \sum_j x_j \delta_{ij} = x_j. \end{aligned}$$

2.4.3 2 次形式の微分

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^T A \mathbf{x}) = (A + A^T) \mathbf{x}. \quad (2.8)$$

証明は

$$\begin{aligned}\frac{\partial}{\partial x_i}(\mathbf{x}^T A \mathbf{x}) &= \sum_{s,t} a_{st} \frac{\partial}{\partial x_i}(x_s x_t) = \sum_{s,t} a_{st}(\delta_{is} x_t + x_s \delta_{it}) \\ &= \left(\sum_t a_{it} x_t \right) + \left(\sum_s a_{si} x_s \right) = (A \mathbf{x})_i + (A^T \mathbf{x})_i = ((A + A^T) \mathbf{x})_i.\end{aligned}$$

特に A が対称行列のときは

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T A \mathbf{x}) = 2A \mathbf{x}.$$

2.4.4 逆行列の微分

$AA^{-1} = I$ の両辺を x で微分すると

$$\left(\frac{\partial}{\partial x} A \right) A^{-1} + A \frac{\partial}{\partial x} (A^{-1}) = 0.$$

左から A^{-1} をかけることによって

$$\frac{\partial}{\partial x} (A^{-1}) = -A^{-1} \left(\frac{\partial}{\partial x} A \right) A^{-1}. \quad (2.9)$$

2.4.5 行列式の対数の微分の公式 (1)

$|A| > 0$ となる行列に対して

$$\frac{\partial}{\partial x} \log |A| = \text{tr} \left(A^{-1} \frac{\partial}{\partial x} A \right).$$

(証明) A を P で三角化する.

$$A = P^{-1} \begin{pmatrix} \lambda_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} P.$$

ここで計算を見やすくするために

$$\begin{pmatrix} \lambda_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} = \text{tri}(\lambda_i)$$

と略記する. すると上の式は

$$A = P^{-1} \text{tri}(\lambda_i) P$$

と表記できる. 逆行列は

$$A^{-1} = P^{-1} \text{tri}(\lambda_i)^{-1} P$$

となる. さて $|A| = \prod \lambda_i$ なので証明すべき式の左辺は

$$\frac{\partial}{\partial x} \left(\sum \log(\lambda_i) \right) = \sum \frac{\lambda'_i}{\lambda_i}.$$

ここで $\frac{\partial \lambda_i}{\partial x} = \lambda'_i$ と略記した. 証明すべき右辺を考えよう.

$$\frac{\partial A}{\partial x} = A' = (P^{-1} \operatorname{tri}(\lambda_i) P)' = (P^{-1})' \operatorname{tri}(\lambda_i) P + P^{-1} \operatorname{tri}(\lambda'_i) P + P^{-1} \operatorname{tri}(\lambda_i) P'.$$

第1項に式(2.9)を使うと

$$(P^{-1})' \operatorname{tri}(\lambda_i) P = -P^{-1} P' P^{-1} \operatorname{tri}(\lambda_i) P$$

更に $\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$ を使うと

$$\begin{aligned} \operatorname{tr}(A^{-1} A') &= -\operatorname{tr}((P^{-1} \operatorname{tri}(\lambda_i)^{-1} P) P^{-1} P' P^{-1} \operatorname{tri}(\lambda_i) P) \\ &\quad + \operatorname{tr}((P^{-1} \operatorname{tri}(\lambda_i)^{-1} P) P^{-1} \operatorname{tri}(\lambda'_i) P) + \operatorname{tr}((P^{-1} \operatorname{tri}(\lambda_i)^{-1} P) P^{-1} \operatorname{tri}(\lambda_i) P') \\ &= -\operatorname{tr}(P^{-1} \operatorname{tri}(\lambda_i)^{-1} P' P^{-1} \operatorname{tri}(\lambda_i) P) + \operatorname{tr}(P^{-1} \operatorname{tri}(\lambda_i)^{-1} \operatorname{tri}(\lambda'_i) P) + \operatorname{tr}(P^{-1} P') \end{aligned}$$

次に $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ を使ってトレースの中の積の順序を入れ換えて, 行列と逆行列の積を消していくと

$$\begin{aligned} \operatorname{tr}(A^{-1} A') &= -\operatorname{tr}(P' P^{-1} \operatorname{tri}(\lambda_i) P P^{-1} \operatorname{tri}(\lambda_i^{-1})) + \operatorname{tr}(\operatorname{tri}(\lambda_i)^{-1} \operatorname{tri}(\lambda'_i) P P^{-1}) + \operatorname{tr}(P^{-1} P') \\ &= -\operatorname{tr}(P' P^{-1}) + \operatorname{tr}(\operatorname{tri}(\lambda_i)^{-1} \operatorname{tri}(\lambda'_i)) + \operatorname{tr}(P^{-1} P') \\ &= \operatorname{tr}(\operatorname{tri}(\lambda_i)^{-1} \operatorname{tri}(\lambda'_i)). \end{aligned}$$

三角行列の逆行列はやはり三角行列であり, * の部分はもとの行列の部分とは異なる何かわからない値になる. しかし対角成分はもとの対角成分の逆数が並ぶ. つまり

$$\operatorname{tri}(\lambda_i)^{-1} = \operatorname{tri}(\lambda_i^{-1}).$$

よって

$$\operatorname{tr}(A^{-1} A') = \operatorname{tr}(\operatorname{tri}(\lambda_i^{-1} \lambda'_i)) = \sum \frac{\lambda'_i}{\lambda_i}.$$

これで左辺 = 右辺が示された.

2.4.6 行列式の対数の微分の公式 (2)

$|A| > 0$ となる行列に対して

$$\frac{\partial}{\partial A} \log |A| = (A^{-1})^T. \quad (2.10)$$

ここで行列 A で微分するというのは各要素 a_{ij} で微分したものを, 行列に並べたものを意味する. 今示した対数の微分の公式(1)より

$$\frac{\partial}{\partial a_{ij}} \log |A| = \operatorname{tr} \left(A^{-1} \frac{\partial}{\partial a_{ij}} A \right).$$

$\partial A / \partial a_{ij}$ は ij 成分のみが 1 でそれ以外は 0 の行列になる. その行列を I_{ij} と書くと,

$$\begin{aligned} \text{tr}(A^{-1}I_{ij}) &= \sum_s (A^{-1}I_{ij})_{ss} = \sum_s \left(\sum_t (A^{-1})_{st} (I_{ij})_{ts} \right) = \sum_s \left(\sum_t (A^{-1})_{st} \delta_{it} \delta_{js} \right) \\ &= (A^{-1})_{ji}. \end{aligned}$$

つまり $\log |A|$ を a_{ij} 成分で微分すると A^{-1} の ji 成分になることが分かったので証明完了.

実はこの式は三角化を使わなくても行列式の定義から直接示すことができる. 2 次正方行列で示してみよう. $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ とすると $|A| = ad - bc$. よって左辺は $\log |A|$ を a, b, c, d でそれぞれ微分して

$$\text{左辺} = \frac{1}{|A|} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix} = \text{右辺}.$$

一般のときは $|A| = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)}$ なので

$$|A|(\text{左辺})_{ij} = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \frac{\partial}{\partial a_{ij}} (a_{1\sigma(1)} \cdots a_{n\sigma(n)}).$$

a_{ij} による微分を考えると, 掛け算の中に a_{ij} があれば (微分が 1 なので) それを取り除き, なければ 0 になってしまう. a_{ij} が現れるのは $j = \sigma(i)$ を固定する σ についてのみである. つまり行列 A から i 行 j 列を取り除いたものになる.

実はこの式は A の余因子行列 \tilde{A} の余因子 \tilde{A}_{ji} と呼ばれるもので,

$$A\tilde{A} = |A|I$$

となることが示される (というか順序が逆で, 普通は逆行列をこれで構成する). つまり $\text{左辺} = (A^{-1})^T$.

2.5 ガウス分布の最尤推定

多変量ガウス分布から, N 個の観測値 $\mathbf{X} = \{\mathbf{x}_i\}$ が独立に得られたときに, 対数尤度関数

$$\log p(\mathbf{X}|\boldsymbol{\mu}, A) = -\frac{Nn}{2} \log(2\pi) - \frac{N}{2} \log |A| - \frac{1}{2} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^T A^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

を A についての最大化を求めてみる. ここでは PRML (演習 2.34) では対称性を仮定せずに解いた結果が対称であったという方針で解けとあるのに, 答えの式変形の途中で対称性を利用しているのはおかしい. せっくなので A の対称性を仮定せずに話を進める.

その前にまず A を固定したときの $\boldsymbol{\mu}$ に関する最尤推定の解を求めておこう. 式 (2.8) より

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu}, A) &= \frac{1}{2} \sum_{i=1}^N \left(A^{-1} + (A^{-1})^T \right) (\mathbf{x}_i - \boldsymbol{\mu}) \\ &= \frac{1}{2} \left(A^{-1} + (A^{-1})^T \right) \left(\left(\sum_{i=1}^N \mathbf{x}_i \right) - N\boldsymbol{\mu} \right).\end{aligned}$$

これが $\mathbf{0}$ なので

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_i \mathbf{x}_i.$$

さて, 本題に戻る. 再び A が対称行列でないという仮定に注意して式を変形する. $\mathbf{y}_i = \mathbf{x}_i - \boldsymbol{\mu}$ とおき

$$F(A) = -N \log |A| - \sum_i \mathbf{y}_i^T A^{-1} \mathbf{y}_i = -N \log |A| - \text{tr} \left(A^{-1} \sum_i \mathbf{y}_i \mathbf{y}_i^T \right)$$

とおく. 第2項の式変形には式 (2.7) を用いた. $B = \sum_i \mathbf{y}_i \mathbf{y}_i^T$ と置いて A で微分しよう. 第1項は式 (2.10) を使って $-N(A^{-1})^T$. 第2項を求めるには式 (2.9) を使って

$$\begin{aligned}\frac{\partial}{\partial a_{ij}} \text{tr} (A^{-1} B) &= \text{tr} \left(\left(\frac{\partial}{\partial a_{ij}} A^{-1} \right) B \right) = -\text{tr} \left(A^{-1} \left(\frac{\partial}{\partial a_{ij}} A \right) A^{-1} B \right) \\ &= -\text{tr} \left(\left(\frac{\partial}{\partial a_{ij}} A \right) A^{-1} B A^{-1} \right).\end{aligned}$$

最後の式変形では $\text{tr}(XY) = \text{tr}(YX)$ を使った. $C = A^{-1} B A^{-1}$ とおく.

$$\begin{aligned}\text{tr} \left(\left(\frac{\partial}{\partial a_{ij}} A \right) C \right) &= \sum_s \left(\left(\frac{\partial}{\partial a_{ij}} A \right) C \right)_{ss} = \sum_s \left(\sum_t \left(\frac{\partial}{\partial a_{ij}} A \right)_{st} c_{ts} \right) \\ &= \sum_{s,t} \delta_{is} \delta_{jt} c_{ts} = c_{ji}.\end{aligned}$$

つまり

$$\frac{\partial}{\partial A} \text{tr}(A^{-1} B) = -C^T = -(A^{-1} B A^{-1})^T. \quad (2.11)$$

よって

$$\frac{\partial}{\partial A} F(A) = -N(A^{-1})^T + (A^{-1} B A^{-1})^T.$$

これが 0 になるような A が $F(A)$ の最大値を与える. 転置をとって

$$-N A^{-1} + A^{-1} B A^{-1} = 0.$$

$$A = \frac{1}{N} B = \frac{1}{N} \sum_i \mathbf{y}_i \mathbf{y}_i^T = \frac{1}{N} \sum_i (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T.$$

この A は明らかに対称行列である．つまり A に関する対称性を仮定せずに最尤解を求めると A が対称行列となることが分かった．また， μ について特に条件も無いので，先に μ に関して最尤推定による解 μ_{ML} を代入すれば μ と A を同時に最大化したものの解となることが分かる．

第 3 章

「線形回帰モデル」のための数学

この文章は『パターン認識と機械学習』（PRML）の 3 章を理解するために必要な数学をまとめてみたものです。

3.1 微分の復習

\mathbf{x}, \mathbf{y} を縦ベクトルとして

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^T \mathbf{y}) = \mathbf{y}.$$

$$\frac{\partial}{\partial \mathbf{y}} (\mathbf{x}^T \mathbf{y}) = \mathbf{x}.$$

ここで $\frac{\partial}{\partial \mathbf{x}}$ は $\frac{\partial}{\partial x_i}$ を縦に並べた縦ベクトルとする。2 章でも述べたが $\frac{\partial}{\partial \mathbf{x}}$ を ∇ と書くこともあるが PRML では場所によって縦ベクトル (3.22) だったり、横ベクトル (3.13) だったりする。常に縦ベクトルとしたほうが混乱は少ない。

3.2 誤差関数の最小化

$$f(\mathbf{w}) = \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 + \lambda \mathbf{w}^T \mathbf{w}$$

とする。ここで \mathbf{w} と $\phi(\mathbf{x}_n)$ は M 次元縦ベクトルである。

$$\Phi^T = (\phi(\mathbf{x}_1) \cdots \phi(\mathbf{x}_N))$$

とおく。 Φ は N 行 M 列の行列である。 $f(\mathbf{w})$ を \mathbf{w} で微分しよう。

$$\frac{\partial}{\partial \mathbf{w}} f(\mathbf{w}) = 2 \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n)) (-\phi(\mathbf{x}_n)) + 2\lambda \mathbf{w}.$$

一般に縦ベクトル \mathbf{x}, \mathbf{y} に対して

$$(\mathbf{x}^T \mathbf{y}) \mathbf{y} = (\mathbf{y}^T \mathbf{x}) \mathbf{y} = \mathbf{y}(\mathbf{y}^T \mathbf{x}) = (\mathbf{y} \mathbf{y}^T) \mathbf{x}$$

だから $\mathbf{t} = (t_1, \dots, t_N)^T$ とおくと

$$\begin{aligned} \frac{1}{2} \frac{\partial}{\partial \mathbf{w}} f(\mathbf{w}) &= - \sum_n t_n \phi(\mathbf{x}_n) + \sum_n \left(\phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right) \mathbf{w} + \lambda \mathbf{w} \\ &= -\Phi^T \mathbf{t} + \Phi^T \Phi \mathbf{w} + \lambda \mathbf{w} \\ &= -\Phi^T \mathbf{t} + (\Phi^T \Phi + \lambda I) \mathbf{w} = 0. \end{aligned}$$

よって $\det(\lambda I + \Phi^T \Phi) \neq 0$ のとき

$$\mathbf{w}_{\text{ML}} = (\lambda I + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

が最尤解. $\mathbf{y} = \Phi \mathbf{w}$ が予測値である.

3.3 正射影

前節で $\lambda = 0$ のときを考える.

$$\mathbf{y} = \Phi(\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

となる. ここでこの式の幾何学的な解釈を考えてみよう. $\Phi = (a_1 \cdots a_M)$ と縦ベクトルの集まりで表す. $N - M$ 個のベクトル b_1, \dots, b_{N-M} を追加して, $\{a_1, \dots, a_M, b_1, \dots, b_{N-M}\}$ 全体で N 次元ベクトル空間の基底であるようにとる. その際 b_i を a_j と直交するようにとれる.

$$a_i^T b_j = 0.$$

さて $X = \Phi(\Phi^T \Phi)^{-1} \Phi^T$ とおくと, $X\Phi = \Phi$. これは $Xa_i = a_i$ を意味する. つまり X は a_1, \dots, a_M で生成される部分空間 $V = \langle a_1, \dots, a_M \rangle$ の点を動かさない. また b_j のとりかたから $Xb_j = 0$ も成り立つ. つまり X は部分空間 $\langle b_1, \dots, b_{N-M} \rangle$ の点を 0 につぶす.

二つ合わせると, X は任意の点を部分空間 V 方向につぶす写像, つまり V への正射影写像と解釈できる. 式で書くと任意の点 \mathbf{t} を $\mathbf{t} = \sum_i s_i a_i + \sum_i t_i b_j$ と表したとすると,

$$\mathbf{y} = X\mathbf{t} = \sum_i s_i a_i$$

となる. \mathbf{t} から \mathbf{y} への変換を係数だけを使って書いてみると

$$X : (s_1, \dots, s_N, t_1, \dots, t_{N-M}) \rightarrow (s_1, \dots, s_N, 0, \dots, 0).$$

これを見ると正射影のニュアンスがより明確になる.

3.4 行列での微分

x を n 次元ベクトル, A を m 行 n 列として $y = Ax$ とおく.

$$f(A) = \|y\|^2 = (Ax)^T Ax$$

を A で微分してみよう.

$$(Ax)^T Ax = \sum_s (Ax)_s (Ax)_s = \sum_s \left(\sum_t a_{st} x_t \right) \left(\sum_u a_{su} x_u \right) = \sum_{s,t,u} x_t x_u a_{st} a_{su}.$$

よって

$$\begin{aligned} \frac{\partial}{\partial a_{ij}} f(A) &= \sum_{s,t,u} x_t x_u \left(\left(\frac{\partial}{\partial a_{ij}} a_{st} \right) a_{su} + a_{st} \frac{\partial}{\partial a_{ij}} a_{su} \right) \\ &= \sum_{s,t,u} x_t x_u (\delta_{is} \delta_{jt} a_{su} + a_{st} \delta_{is} \delta_{ju}) = \left(\sum_u x_j x_u a_{iu} \right) + \left(\sum_t x_t x_j a_{it} \right) \\ &= 2 \sum_u x_j x_u a_{iu} = 2x_j (Ax)_i = 2(Axx^T)_{ij}. \end{aligned}$$

よって

$$\frac{\partial}{\partial A} \|Ax\|^2 = 2Axx^T.$$

3.5 Woodbury の逆行列の公式

行列 A, B, C, D について

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$$

が成り立つ.

(証明)

$$\begin{aligned} A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1} &= A^{-1}B((DB^{-1}A + C)A^{-1}B)^{-1}CA^{-1} \\ &= (DB^{-1}A + C)^{-1}CA^{-1} \\ &= (DB^{-1}(A + BD^{-1}C))^{-1}CA^{-1} \\ &= (A + BD^{-1}C)^{-1}BD^{-1}CA^{-1} \end{aligned}$$

よって

$$\begin{aligned} \text{左辺} &= \left(I - (A + BD^{-1}C)^{-1}BD^{-1}C \right) A^{-1} \\ &= (A + BD^{-1}C)^{-1}((A + BD^{-1}C) - BD^{-1}C) A^{-1} \\ &= \text{右辺}. \end{aligned}$$

特に, A が n 次正方行列で B を n 次縦ベクトル \mathbf{x} , $C = \mathbf{x}^T$, D を n 次単位行列とすると

$$(A + \mathbf{x}\mathbf{x}^T)^{-1} = A^{-1} - \frac{(A^{-1}\mathbf{x})(\mathbf{x}^T A^{-1})}{1 + \mathbf{x}^T A^{-1}\mathbf{x}} \quad (3.1)$$

が成り立つ.

3.6 正定値対称行列

n 次元実対称行列 A はある直行列 P を用いて常に対角化可能であった.

$$P^{-1}AP = \text{diag}(\lambda_1, \dots, \lambda_n).$$

全ての固有値が正であるとき A を正定値といい, $A > 0$ とかく. 全ての固有値が正または 0 であるとき, 半正定値といい, $A \geq 0$ とかく.

任意の実ベクトル \mathbf{x} について $\mathbf{y} = P\mathbf{x}$ とおくと \mathbf{x} が \mathbb{R}^n の全ての点をとるとき \mathbf{y} も全ての点を渡る.

$$\mathbf{x}^T A \mathbf{x} = \sum_i \lambda_i y_i^2$$

なので $A \geq 0$ ならば $\mathbf{x}^T A \mathbf{x} \geq 0$. $A > 0$ のときは等号が成り立つのは $\mathbf{x} = 0$ のときのみである.

逆に任意の \mathbf{x} について $\mathbf{x}^T A \mathbf{x} \geq 0$ とすると, \mathbf{y} として単位ベクトル \mathbf{e}_i を考えれば $\lambda_i \geq 0$. つまり $A \geq 0$. 更に等号は $\mathbf{x} = 0$ のときに限るためには $\lambda_i > 0$. つまり $A > 0$ であることが分かる. まとめると

$$A \geq 0 \iff \lambda_i \geq 0 \text{ for } \forall i.$$

$$A > 0 \iff \lambda_i > 0 \text{ for } \forall i.$$

この同値性から $A > 0$ のとき $A^{-1} > 0$ も分かる. 定義から $A > 0$, $B > 0$ なら $A + B > 0$ も成り立つ.

また実ベクトル \mathbf{v} に対して $A = \mathbf{v}\mathbf{v}^T$ とおくと, A は実対称であり, 任意の \mathbf{x} に対して

$$\mathbf{x}^T A \mathbf{x} = (\mathbf{v}^T \mathbf{x})^2 \geq 0$$

なので $A \geq 0$.

3.7 予測分布の分散

$S_N^{-1} = S_0^{-1} + \beta \Phi_N^T \Phi_N$ としたときの予測分布の分散

$$\sigma_N^2 = \frac{1}{\beta} + \phi^T S_N \phi$$

を考える. $\beta > 0$ であり, S_0 は共分散行列なので実正定値であることに注意する. まず計画行列 Φ_N は N が一つ増える毎に 1 行増える. \mathbf{v}_N (煩雑なので v と略記する) を M 次元縦ベクトルとして

$$\Phi_{N+1}^T = \begin{pmatrix} \Phi_N^T & v \end{pmatrix}$$

としよう. すると

$$S_{N+1}^{-1} = S_0^{-1} + \beta \left(\Phi_N^T \Phi_N + vv^T \right) = S_N^{-1} + \beta vv^T.$$

行列 βvv^T は正定値であり, S_N に関して帰納法を使うと全ての S_N は正定値であることが分かる.

式 (3.1) を使って

$$\begin{aligned} \sigma_{N+1}^2 &= \frac{1}{\beta} + \phi^T (S_N^{-1} + \beta vv^T)^{-1} \phi \\ &= \frac{1}{\beta} + \phi^T \left(S_N - \frac{(S_N v)(v^T S_N)}{1 + v^T S_N v} \right) \phi \\ &= \sigma_N^2 - z \end{aligned}$$

ここで S_N は対称なので

$$\begin{aligned} z &= \phi^T \frac{(S_N v)(v^T S_N)}{1 + v^T S_N v} \phi \\ &= \frac{1}{1 + v^T S_N v} (v^T S_N \phi)^2. \end{aligned}$$

S_N は正定値なので任意の v に対して $v^T S_N v \geq 0$. よって $z \geq 0$ となり

$$\sigma_{N+1}^2 \leq \sigma_N^2.$$

帰納法の流れを見ると,

$$\Phi_N^T = (\mathbf{v}_1 \cdots \mathbf{v}_N)$$

とおくと

$$S_N^{-1} = S_0^{-1} + \beta \sum_{i=1}^N \mathbf{v}_i \mathbf{v}_i^T$$

となることがわかる. \mathbf{v}_i が基底関数のベクトルに訓練データの値を代入したものであることを考えると, 0 ベクトルになることは殆ど無い. また $N \rightarrow \infty$ で 0 になるわけでもない. つまりそれらの和はどんどん大きくなる. そういう状況の元では $\phi^T S_N \phi$ は 0 に近づき,

$$\sigma_N^2 \rightarrow \frac{1}{\beta}$$

となる.

3.8 カルバック距離

$p(x)$, $q(x)$ を恒等的に 0 ではない確率密度関数とする. つまり $p(x), q(x) \geq 0$.

$$\text{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

をカルバック距離 (Kullback-Leibler 距離, 相対エントロピー) という.

距離といいつつ, $\text{KL}(p||q) = \text{KL}(q||p)$ とは限らないので距離の公理は満たさない. しかし, $\text{KL}(p||q) \geq 0$ であり, $\text{KL}(p||q) = 0 \iff p = q$ はいえる. これを示そう.

まず $S(x) = e^{-x} + x - 1$ について $S(x) \geq 0$ であり, $S(x) = 0 \iff x = 0$ である.

なぜなら $S'(x) = -e^{-x} + 1$. $S''(x) = e^{-x} \geq 0$ なので $S'(x)$ は単調増加. $S'(0) = 0$ より $x > 0$ なら $S'(x) > 0$, $x < 0$ なら $S'(x) < 0$. つまり $S(x)$ は 0 で最小値 0 をとる.

$$\begin{aligned} \int p(x) S\left(\log \frac{p(x)}{q(x)}\right) dx &= \int p(x) \left(\frac{q(x)}{p(x)} + \log \frac{p(x)}{q(x)} - 1 \right) dx \\ &= \text{KL}(p||q) + \int (q(x) - p(x)) dx \\ &= \text{KL}(p||q). \end{aligned}$$

ここで p, q が確率密度関数なので $\int p(x) dx = 1$, $\int q(x) dx = 1$ であることを使った.

この式の左辺の被積分関数は常に 0 以上. よって $\text{KL}(p||q) \geq 0$.

$\text{KL}(p||q) = 0$ ならば殆ど全ての x について

$$p(x) S\left(\log \frac{p(x)}{q(x)}\right) = 0.$$

$p = 0$ ではないので殆ど全ての x について

$$S\left(\log \frac{p(x)}{q(x)}\right) = 0.$$

$S(x) = 0$ となる x は 0 のときだけだから, 殆ど全ての x について $p(x) = q(x)$.

真のモデル $p(D|M)$ があったときに, モデルエビデンス $p(D|M')$ とのカルバック距離 $\text{KL}(p(D|M)||p(D|M'))$ は, 0 に近いほど真のモデルに近そうだということにする.

3.9 エビデンス関数の評価の式変形

$A = \alpha I + \beta \Phi^T \Phi$ とおくと

$$\begin{aligned} E(w) &= \frac{\beta}{2} \|t - \Phi w\|^2 + \frac{\alpha}{2} w^T w \\ &= \frac{1}{2} w^T (\alpha I + \beta \Phi^T \Phi) w - \beta t^T \Phi w + \frac{\beta}{2} \|t\|^2 \\ &= \frac{1}{2} w^T A w - \beta w^T \Phi^T t + \frac{\beta}{2} \|t\|^2. \end{aligned}$$

ここで一般に対称行列 A とベクトル w, m について

$$\frac{1}{2} (w - m)^T A (w - m) = \frac{1}{2} w^T A w - w^T A m + \frac{1}{2} m^T A m.$$

この関数は $w = m$ のとき最小値 0 をとる. 二つを比較することで $E(w)$ は $\beta \Phi^T t = A m$, つまり

$$w = m_N = \beta A^{-1} \Phi^T t$$

のとき最小となる. 最小値は元の $E(w)$ の式に $w = m_N$ を代入すれば得られ,

$$E(m_N) = \frac{\beta}{2} \|t - \Phi m_N\|^2 + \frac{\alpha}{2} m_N^T m_N.$$

つまり

$$E(w) = \frac{1}{2} (w - m_N)^T A (w - m_N) + E(m_N)$$

と平方完成できる.

よって

$$\begin{aligned} E(w) &= \int \exp(-E(w)) dw \\ &= \exp(-E(m_N)) \int \exp\left(-\frac{1}{2} (w - m_N)^T A (w - m_N)\right) dw \\ &= \exp(-E(m_N)) (2\pi)^{M/2} |A|^{-1/2}. \end{aligned}$$

従って

$$\begin{aligned} \log p(t|\alpha, \beta) &= (N/2) \log\left(\frac{\beta}{2\pi}\right) + (M/2) \log\left(\frac{\alpha}{2\pi}\right) \log\left(\int \exp(-E(w)) dw\right) \\ &= (M/2) \log \alpha + \frac{N}{2} \log \beta - E(m_N) - \frac{1}{2} \log |A| - \frac{N}{2} \log(2\pi). \quad (3.2) \end{aligned}$$

3.10 ヘッセ行列

x が n 次縦ベクトルのとき, $y = f(x)$ における 2 階微分の n 次正方行列

$$H(f) = \left(\frac{\partial^2}{\partial x_i \partial x_j} f(x) \right)$$

をヘッセ行列という. 通常偏微分は可換なので, これは対称行列である.

1 階微分の行列 (ヤコビ行列) の行列式はその点の付近の拡大率を表していた. ヘッセ行列はその点の付近の関数の形を表す. たとえば正定値な場合は極小, 固有値が全て負の場合は極大, 固有値が正と負の両方の場合は鞍点となる.

$f = x^2 - y^2$, $g = x^2 + y^2$ というグラフを見てみよう. 図 3.1 は原点で鞍点, 図 3.2 は原点で極小である. それぞれヘッセ行列は

$$H(f) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}, \quad H(g) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

となり, ヘッセ行列が原点での形に対応していることが分かる.

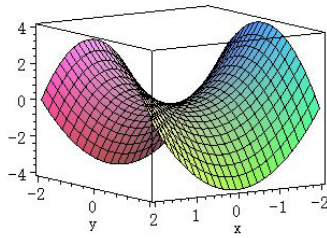


図 3.1 $f = x^2 - y^2$

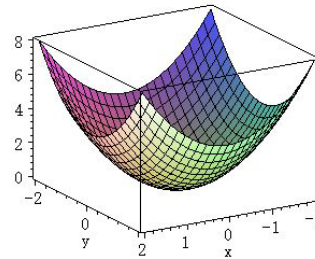


図 3.2 $g = x^2 + y^2$

3.11 エビデンス関数の最大化の式変形

行列 $\beta\Phi^T\Phi$ をある行列 P で対角化する.

$$P^{-1}(\beta\Phi^T\Phi)P = \text{diag}(\lambda_1, \dots, \lambda_M).$$

すると行列 $A = \alpha I + \beta\Phi^T\Phi$ も同じ P で対角化できて

$$P^{-1}AP = \text{diag}(\alpha + \lambda_1, \dots, \alpha + \lambda_M).$$

よって

$$|A| = \prod_{i=1}^M (\lambda_i + \alpha)$$

となる. α で微分すると

$$\frac{\partial}{\partial \alpha} \log |A| = \sum_{i=1}^M \frac{1}{\lambda_i + \alpha}.$$

式 (3.2) を α で微分すると

$$\frac{\partial}{\partial \alpha} \log p(\mathbf{t}|\alpha, \beta) = \frac{M}{2\alpha} - \frac{1}{2} m_N^T m_N - \frac{1}{2} \sum \frac{1}{\lambda_i + \alpha} = 0.$$

よって

$$\alpha m_N^T m_N = M - \sum_{i=1}^M \frac{\alpha}{\lambda_i + \alpha} = \sum_{i=1}^M \frac{\lambda_i}{\lambda_i + \alpha}.$$

これを γ とおくと

$$\alpha = \frac{\gamma}{m_N^T m_N}.$$

ただし, m_N は陰に α に依存しているのでこれは実は α を含む方程式である.

β についても同様にしてみる. $\beta \Phi^T \Phi$ の固有値が λ_i だから λ_i は β に比例する. つまり微分が比例係数に等しい.

$$\frac{\partial}{\partial \beta} \lambda_i = \lambda_i / \beta.$$

よって

$$\frac{\partial}{\partial \beta} \log |A| = \sum \frac{\lambda_i / \beta}{\lambda_i + \alpha} = \frac{\gamma}{\beta}.$$

式 (3.2) を β で微分すると

$$\frac{N}{2\beta} - \frac{1}{2} \|\mathbf{t} - \Phi m_N\|^2 - \frac{\gamma}{2\beta} = 0.$$

よって

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \|\mathbf{t} - \Phi m_N\|^2.$$

3.12 パラメータの関係

パラメータがたくさんできたのでそれらの関係を見直してみよう. まず線形基底モデルを考えた. $\phi(x)$ を M 個の基底関数からなるベクトルとする. x は観測値であり,

$$y(x, w) = w^T \phi(x)$$

とした。 \mathbf{t} を観測値に対する目標値で、それは x によらずに精度パラメータ β に従うガウス分布とした。

$$p(\mathbf{t}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

ベイズ的に扱うために \mathbf{w} に関して事前確率分布を与えたい。上式が \mathbf{w} に関する2次関数なので、共役事前分布としてハイパーパラメータ α を導入し、

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}I)$$

を仮定した。そうすることで事後分布は

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, S_N)$$

の形（ただし、 $\mathbf{m}_N = \beta S_N \Phi^T \mathbf{t}$, $S_N^{-1} = \alpha I + \beta \Phi^T \Phi$ ）になった。

さて、ここで α, β はハイパーパラメータではあるが、事前分布を入れて確率変数的に扱いたい。その上で最尤推定の手法を用いて実際のデータから値を決めるという枠組みを経験ベイズという。そのとき \mathbf{t} の予測分布は

$$p(\mathbf{t}|\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) p(\alpha, \beta|\mathbf{t}) d\mathbf{w} d\alpha d\beta$$

となる。とはいえ、そのまま扱うのは難しいのでまずデータが十分たくさんあるとき、 α, β は殆ど固定値、つまり α, β の分布はある特定の値 $\hat{\alpha}, \hat{\beta}$ にデルタ関数的に近づくと仮定しよう。

$$p(\alpha, \beta|\mathbf{t}) \approx \delta_{\alpha, \hat{\alpha}} \delta_{\beta, \hat{\beta}}.$$

そうすると

$$p(\mathbf{t}|\mathbf{t}) \approx \int p(\mathbf{t}|\mathbf{w}, \hat{\beta}) p(\mathbf{w}|\mathbf{t}, \hat{\alpha}, \hat{\beta}) d\mathbf{w}$$

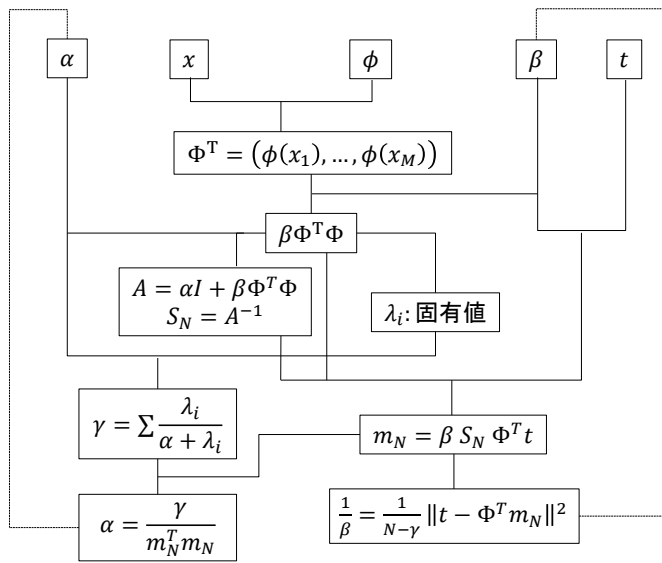
となり予測分布は $\hat{\alpha}, \hat{\beta}$ を求めればよいということになる。

次に α, β を求める方法を考える。ベイズの定理から

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta) p(\alpha, \beta)$$

となる。ここで $p(\alpha, \beta)$ はほぼ平坦、つまり α, β の値はどれも同じぐらいの可能性があるという仮定を置く。そうすると事後分布を最大化する α, β を求める最尤推定の問題は、尤度関数を最大化する問題に近似できる。この尤度関数をエビデンスといい、この手法をエビデンス近似という。そして、 $p(\mathbf{t}|\alpha, \beta)$ を最大化するための α, β の関係式を求めたのが前節であった。

以上のパラメータの関係を図3.3に示した。実際には、初期値 α, β を適当に決め、この図に従って計算して新しい α, β を求めたあと再度繰り返す。それが収束すればその値を採用する。ここではその収束性については議論しない。

図 3.3 $\alpha, x, \phi, t, \beta$ の関係図

第 4 章

「線形識別モデル」のための数学

この文章は『パターン認識と機械学習』（PRML）の 4 章の式変形を一部埋めたものです。

4.1 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。

4.2 行列の微分の復習

$A = (a_{ij})$ とかいた.

$$(AB)_{ij} = \sum_k a_{ik} b_{kj}, \quad \text{tr}(A) = \sum_i a_{ii}, \quad A^T = (a_{ji})$$

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 $\partial a_{st} / \partial a_{ij} = \delta_{is} \delta_{jt}$ 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,

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

4.3 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.

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

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.

領域内の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, $S^1 = \{(x, y) | x^2 + y^2 = 1\}$ を単位円とする.

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

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.

4.4 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, $\tilde{w}_k = (w_{k0}, w_k^T)^T$, $\tilde{x} = (1, x^T)^T$ と 1 次元増やすと $y_k(x) = \tilde{w}_k^T \tilde{x}$ 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

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

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 $X^T XW = X^T T$ より

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

4.5 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 $y = w^T x$ 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

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

$m_i = w^T \mathbf{m}_i$ 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. ラグランジュの未定乗数法を用いて

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

とおくと

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

よって

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

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

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

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

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。

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

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

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

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

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

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

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。

4.6 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

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

を最大化してみよう。

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

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。

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

よって

$$\begin{aligned} \left(S_W + \frac{N_1 N_2}{N} S_B \right) w &= N(\mathbf{m}_1 - \mathbf{m}_2) \\ &\quad + \left(N \mathbf{m} \mathbf{m}^T - N_1 \mathbf{m}_1 \mathbf{m}_1^T - N_2 \mathbf{m}_2 \mathbf{m}_2^T \right. \\ &\quad \left. + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \right) w. \end{aligned}$$

() 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) .

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

よって

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

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

4.7 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：

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

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.

4.8 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

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

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

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

4.9 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. $\sum p_i = 1$. 尤度関数は

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

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

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

μ_1 で微分して 0 とおくと

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

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

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

ここで最後の式変形に

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

を用いた。これを A で微分する。2 章で示した行列式の対数の微分の公式 (2) : 式 (2.10)

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

と式 (2.11) :

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

を使うと

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

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。

4.10 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

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

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

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

データ集合 $\{\phi_n, t_n\}$, $t_n \in \{0, 1\}$, $\phi_n = \phi(x_n)$, $n = 1, \dots, N$, $t = (t_1, \dots, t_N)^T$, $y_n = p(C_1|\phi_n) = \sigma(a_n)$, $a_n = w^T \phi_n$ 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との誤差なので、線形回帰モデルのときと同じ形。

4.11 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。

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

を w で微分して $\Phi = (\phi_1, \dots)^T$ とおくと

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

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

更新式に代入すると

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

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。

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

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

$y'_n = y_n(1 - y_n)\phi_n$ だったので

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

ここで $R = \text{diag}(R_n) = \text{diag}(y_n(1 - y_n))$.

$H > 0$ 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

$$u^T H u = v^T R v = \sum y_n (1 - y_n) v_n^2 > 0.$$

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.

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

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

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

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 $a = w^T \phi$, $y = \sigma(a)$ 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 $y_n = \sigma(w_{\text{old}}^T \phi)$ のまわりで一次近似を行うと

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

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.

4.12 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 $x, y, 0 \leq t \leq 1$ に対して

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

である.

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

このとき n 個の任意の実数 x_1, \dots, x_n に対して

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

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つつ. よって $q \neq 0$ とすると

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

$$\sum_{i=1}^{n+1} p_i f(x_i) = q \sum_{i=1}^n (p_i/q) f(x_i) + p_{n+1} f(x_{n+1})$$

帰納法の仮定を用いて

$$\geq q f\left(\sum_{i=1}^n (p_i/q) x_i\right) + p_{n+1} f(x_{n+1})$$

f が凸関数であることを用いて

$$\geq f\left(q \left(\sum_{i=1}^n (p_i/q) x_i\right) + p_{n+1} x_{n+1}\right) = f\left(\sum_{i=1}^{n+1} p_i x_i\right).$$

4.13 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

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

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

$$\frac{\partial}{\partial a_j} y_k = y_k (\delta_{kj} - y_j). \quad (4.1)$$

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

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

交差エントロピー誤差関数は

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

よって

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

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

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

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

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

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

$$u^T H u \geq 0.$$

4.14 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. $a = w^T \phi$, $f(a)$ 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

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

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

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

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.

$x \rightarrow \infty$ 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.

4.15 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.

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

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

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

η で微分して

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

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,

$$y = f(w^T \phi)$$

とかけるモデルを考える。対数尤度関数は

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

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

パラメータが多いので依存関係に注意して微分する。

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

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

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。

4.16 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

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

つまり

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

よって正規分布で近似すると

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

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

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

よって

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

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.

4.17 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

$$\begin{aligned} Z &= \int f(z) dx \approx \int f(z_0) \exp \left(-\frac{1}{2} (z - z_0)^T A (z - z_0) \right) dx \\ &= f(z_0) \sqrt{\frac{(2\pi)^M}{|A|}}. \end{aligned} \quad (4.2)$$

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 $\theta = \theta_{\text{MAP}}$ のときのラプラス近似を用いて

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

$$\begin{aligned} \log p(D) = \log Z &\approx \log f(\theta_{\text{MAP}}) + \log \sqrt{\frac{(2\pi)^M}{|A|}} \\ &= \log p(D|\theta_{\text{MAP}}) + \left(\log p(\theta_{\text{MAP}}) + \frac{M}{2} \log(2\pi) - \frac{1}{2} \log |A| \right). \end{aligned}$$

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.

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

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

とすると

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

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

を使って

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

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

$$\log |H| \approx \log \prod_{i=1}^M (NC) = M \log N + M \log C \approx M \log N.$$

よって

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

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.

4.18 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

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

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

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

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.

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

実数値関数 $f(x)$ に対して部分積分を適用すると

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) \delta(x) dx &= [H(x)f(x)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} H(x)f'(x) dx \\ &= f(\infty) - \int_0^{\infty} f'(x) dx = f(\infty) - (f(\infty) - f(0)) = f(0). \end{aligned}$$

4.19 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.

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

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

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

$$\int_{-\infty}^{\infty} \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da = \Phi\left(\frac{\lambda\mu}{\sqrt{1+\lambda^2\sigma^2}}\right). \quad (4.3)$$

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.

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

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

とおく.

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

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

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

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

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

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

$\exp()$ の内側を x について平方完成しよう:

$$\begin{aligned} & -\frac{1}{2}((\lambda\sigma)^2 + 1)x^2 + 2\lambda^2\mu\sigma x + \lambda^2\mu^2 \\ &= -\frac{1}{2}\left(\left(\sqrt{1+\lambda^2\sigma^2}x + \frac{\lambda^2\mu\sigma}{\sqrt{1+\lambda^2\sigma^2}}\right)^2 + \lambda^2\mu^2 - \frac{\lambda^4\mu^2\sigma^2}{1+\lambda^2\sigma^2}\right). \end{aligned}$$

定数項は

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

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

より

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

よって

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

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。

$\mu \rightarrow \infty$ で

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

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

よって $L = R$ が示された。

4.20 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: $y_n = \sigma(w^T \phi_n)$ 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 $t = (t_0, \dots, t_N)^T$ を用いて

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

よって

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

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

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

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．

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

デルタ関数の性質から

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

よって

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

ただし

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

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，

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

分散は

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

括弧内は

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

より

$$\sigma_a^2 = \phi^T S_N \phi.$$

これは線形回帰モデルの予測分布の分散（PRML 式 3.59）

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

でノイズを消したもの（ $(1/\beta) \rightarrow 0$ ）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 $w = (w_1, w_2)^T$ 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

$$w^T \phi = w_1 \|\phi\|.$$

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

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

$q(w)\mathcal{N}(w|m_N, S_N)$ で $w = (w_1, w_2)^T$ 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：

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

$a = w_1 \|\phi\|$ 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

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

ロジスティックモイド関数でのガウス分布の畳み込み積分は解析的には難しい．ここではロジスティック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畳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

$$\begin{aligned} \int \sigma(a) \mathcal{N}(a|\mu, \sigma^2) da &\approx \sigma\left(\frac{\mu}{\sqrt{1 + \lambda^2 \sigma^2}}\right) \\ &\quad (\kappa(\sigma^2) = 1/\sqrt{1 + (\pi/8)\sigma^2} \text{とおくと}) \\ &= \sigma(\kappa(\sigma^2)\mu). \end{aligned} \tag{4.4}$$

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．

第 5 章

「ニューラルネットワーク」の補足

PRML 5 章では、本来別の記号を割り当てるべきところに同じ記号を用いることがあり、初読時には混乱しやすい。慣れてしまえば読むのは難しくないが、ここでは出来るだけ区別してみる。

5.1 フィードフォワードネットワーク関数

3 章, 4 章でやったモデルは基底関数 ϕ_j とパラメータ w_j の線形和を非線形活性化関数に入れたものだった。ここではそれを拡張する。 x_1, \dots, x_D を入力変数とし $x_0 = 1$ をバイアス項（定数項）に対応する変数, $w_{ji}^{(1)}$ をパラメータとして

$$\hat{a}_j = \sum_{i=0}^D w_{ji}^{(1)} x_i$$

とする。PRML では上記の \hat{a}_j を a_j と書いているがすぐあとに出てくる a_k とは無関係である。ここでは異なることを強調するために \hat{a}_j とする。

\hat{a}_j を活性化関数 h で変換する。

$$z_j = h(\hat{a}_j).$$

h としてはロジスティックシグモイドなどのシグモイド関数を用いられる。これらの線形和をとって出力ユニット活性を求める。 $z_0 = 1$ をバイアス項に対応する変数として

$$a_k = \sum_{j=0}^M w_{kj}^{(2)} z_j.$$

この出力ユニット活性を活性化関数を通してネットワークの出力 y_k とする。2 クラス分類問題ならロジスティックシグモイド関数を使う。

$$y_k = y_k(x, w) = \sigma(a_k).$$

ここで w は $\{w_{ji}^{(1)}, w_{kj}^{(2)}\}$ をまとめたベクトルである。これらの式を組み合わせると

$$y_k = \sigma \left(\sum_{j=0}^M w_{kj}^{(2)} h \left(\sum_{i=0}^D w_{ji}^{(1)} x_i \right) \right).$$

5.2 ネットワーク訓練

回帰問題を考える。入力ベクトル x と K 次元の目標変数 t があり、 x, w における t の条件付き確率が精度 βI のガウス分布とする。 N 個の同時独立分布 $x = \{x_1, \dots, x_N\}$ と対応する目標値 $t = \{t_1, \dots, t_N\}$ を用意し、出力ユニットの活性化関数を恒等写像として $y_n = y(x_n, w)$ とする。

$$p(t|x, w) = \prod_{n=1}^N \mathcal{N}(t_n | y(x_n, w), \beta^{-1} I).$$

対数をとると

$$\begin{aligned} \log p(t|x, w) &= - \sum_n \left(\frac{1}{2} (t_n - y_n)^T (\beta I) (t_n - y_n) \right) - \sum_n \frac{K}{2} \log(2\pi) - \sum_n \frac{1}{2} \log |\beta^{-1} I| \\ &= - \frac{\beta}{2} \sum_n \|t_n - y(x_n, w)\|^2 - \frac{DN}{2} \log(2\pi) + \frac{NK}{2} \log \beta. \end{aligned} \quad (5.1)$$

そうするとこの関数の w についての最大化は最初の項の最小化、つまり

$$E(w) = \frac{1}{2} \sum_n \|y(x_n, w) - t_n\|^2$$

の最小化と同等である。最小値を与える $w (= w_{ML})$ をなんらかの方法で求める。その値を式 (5.1) に代入して β で微分して 0 とおくと

$$-\frac{1}{2} \sum_n \|t_n - y(x_n, w_{ML})\|^2 + \frac{NK}{2} \frac{1}{\beta} = 0.$$

よって

$$\frac{1}{\beta_{ML}} = \frac{1}{NK} \sum_n \|t_n - y(x_n, w_{ML})\|^2.$$

5.2.1 問題に応じた関数の選択

回帰問題を考える。 a_k の活性化関数を恒等写像にとる。すると二乗和誤差関数の微分は

$$\frac{\partial E}{\partial a_k} = y_k - t_k.$$

クラス分類問題でも同様の関係式が成り立つことを確認しよう。

目標変数 t が $t = 1$ でクラス C_1 , $t = 0$ でクラス C_2 を表す 2 クラス分類問題を考える。活性化関数をロジスティックシグモイド関数に選ぶ。

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

この微分は $dy/da = \sigma(a)(1 - \sigma(a)) = y(1 - y)$ であった。 $p(t = 1|x) = y(x, w)$, $p(t = 0|x) = 1 - y(x, w)$ なので

$$p(t|x, w) = y(x, w)^t (1 - y(x, w))^{1-t}.$$

よって 4 章と同様にして交差エントロピー誤差関数は

$$E(w) = - \sum_{n=1}^N (t_n \log y_n + (1 - t_n) \log(1 - y_n)).$$

これを a_k で微分すると

$$\begin{aligned} \frac{\partial E}{\partial a_k} &= - \left(t_k \frac{y_k(1 - y_k)}{y_k} + (1 - t_k) \frac{-y_k(1 - y_k)}{1 - y_k} \right) \\ &= -(t_k - t_k y_k - y_k + y_k t_k) \\ &= y_k - t_k. \end{aligned}$$

K 個の 2 クラス分類問題を考える。それぞれの活性化関数がロジスティックシグモイド関数とする。

$$p(t|x, w) = \prod_{k=1}^K y_k(x, w)^{t_k} (1 - y_k(x, w))^{1-t_k}.$$

$y_{nk} = y_k(x_n, w)$ とし n 番目の入力 x_n に対する目標変数を t_{nk} で表す。 $t_{nk} \in \{0, 1\}$ であり, $\sum_k t_{nk} = 1$ である。

$$E(w) = - \prod_{n,k} (t_{nk} \log y_{nk} + (1 - t_{nk}) \log(1 - y_{nk})).$$

y_{nj} に対応する a を a_{nj} とすると

$$\frac{\partial E(w)}{\partial a_{nj}} = -(t_{nj}(1 - y_{nj}) + (1 - t_{nj})(-y_{nj})) = y_{nj} - t_{nj}.$$

最後に K クラス分類問題を考える。同様に n 番目の入力 x_n に対する目標変数 t_{nk} で表す。 $y_k(x_n, w)$ を t_{nk} が 1 となる確率 $p(t_{nk} = 1|x_n)$ とみなす。

$$E(w) = - \log p(t|x, w) = - \sum_{n,k} t_{nk} \log y_k(x_n, w).$$

活性化関数はソフトマックス関数で

$$y_k(x, w) = \frac{\exp a_k(x, w)}{\sum_j \exp(a_j(x, w))}$$

のとき

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

よって $a_{nj} = a_j(x_n, w)$ とすると

$$\frac{\partial}{\partial a_{nj}} \log y_k(x_n, w) = \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} = \delta_{kj} - y_{nj}.$$

よって

$$\frac{\partial E}{\partial a_{nj}} = -\sum_k t_{nk}(\delta_{kj} - y_{nj}) = -t_{nj} + \left(\sum_k t_{nk}\right) y_{nj} = y_{nj} - t_{nj}.$$

5.3 局所二次近似

スカラー x についての関数 $E(x)$ の $x = a$ におけるテイラー展開を2次の項で打ち切った近似式は

$$E(x) \approx E(a) + E'(a)(x - a) + \frac{1}{2}E''(a)(x - a)^2$$

であった。これを n 変数関数 $E(x_1, \dots, x_n)$ に拡張する。 $x = (x_1, \dots, x_n)^T$, $a = (a_1, \dots, a_n)^T$ において

$$\begin{aligned} E(x) &\approx E(a) + (x - a)^T \left(\frac{\partial}{\partial x_i} E \right) + \frac{1}{2} (x - a)^T \left(\frac{\partial^2 E}{\partial x_i \partial x_j} \right) (x - a) \\ &= E(a) + (x - a)^T (\nabla E) + \frac{1}{2} (x - a)^T H(E) (x - a) \end{aligned}$$

となる。 $E(x)$ が $x = a$ の付近で極小ならば、そこでの勾配 ∇E は0なので

$$E(x) \approx E(a) + \frac{1}{2} (x - a)^T H(E) (x - a).$$

$H(E)$ は対称行列なので3.6節の議論より対角化することで

$$E(x) \approx E(a) + \frac{1}{2} \sum_i \lambda_i y_i^2$$

の形にできる。そして $E(x)$ が $x = a$ の付近で極小となるのは $H(E) > 0$ (正定値) であるときとわかる。

なお, $H(f) = \nabla^2 f = \nabla(\nabla f)$ という表記をすることがある。微分作用素 ∇ を2回するので2乗の形をしている。ただ ∇f が縦ベクトルならもう一度 ∇ をするときは結果が行列になるように、入力ベクトルの転置を取って作用するとみなす。 n^2 次元の長いベクトルになるわけではない。

5.4 誤差関数微分の評価

与えられたネットワークに対して誤差関数の変化の割合を調べる。この節ではどの変数がどの変数に依存しているか気をつけて微分する必要がある。誤差関数が訓練集合の各データに対する誤差の和で表せると仮定する：

$$E(w) = \sum_{n=1}^N E_n(w).$$

一般のフィードフォワードネットワークで

$$a_j = \sum_i w_{ji} z_i, \quad z_j = h(a_j) \quad (5.2)$$

とする。入力 z_i が出力ユニット a_j に影響を与え、その a_j が非線形活性化関数 $h()$ を通して z_j に影響を与える。ある特定のパターン E_n の重み w_{ji} に関する微分を考える。以下、特定のパターンを固定することで E_n 以外の添え字の n を省略する。式 (5.2) のように E_n は非線形活性化関数 h の変数 a_j を通して w_{ji} に依存している。

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}}.$$

a_j は w_{ji} に関しては線形なので

$$\frac{\partial a_j}{\partial w_{ji}} = z_i.$$

誤差と呼ばれる記号 $\delta_j = \partial E_n / \partial a_j$ を導入すると

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$

とかける。 δ_j は $h()$ が正準連結関数の場合は 5.2 節での考察により

$$\delta_j = \frac{\partial E_n}{\partial a_j} = y_j - t_j$$

で計算できる。ユニット j につながっているユニット k を通して E_n への a_j の影響があると考えたと

$$\frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial a_j}. \quad (5.3)$$

ここで a_j と a_k は z を経由して関係していると考えているので $\partial a_k / \partial a_j = \delta_{jk}$ (クロネッカーのデルタ) にはならないことに注意する。実際、

$$a_k = \sum_i w_{ki} h(a_i)$$

より

$$\frac{\partial a_k}{\partial a_j} = w_{kj} h'(a_j).$$

これを式 (5.3) に代入して

$$\delta_j = \frac{\partial E_n}{\partial a_j} = \sum_k \delta_k w_{kj} h'(a_j) = h'(a_j) \sum_k w_{kj} \delta_k.$$

5.5 外積による近似

$$E(w) = \frac{1}{2} \sum_{n=1}^N (y_n - t_n)^2$$

のときのヘッセ行列は

$$H = H(E) = \sum_n (\nabla y_n)(\nabla y_n)^T + \sum_n (y_n - t_n) H(y_n).$$

一般には成立しないがもしよく訓練された状態で y_n が目標値 t_n に十分近ければ第2項を無視できる. その場合 $b_n = \nabla y_n = \nabla a_n$ (活性化関数が恒等写像なので) とおくと

$$H \approx \sum_n b_n b_n^T.$$

これを Levenberg-Marquardt 近似という.

$$E = \frac{1}{2} \iint (y(x, w) - t)^2 p(x, t) dx dt$$

のときのヘッセ行列を考えてみると

$$\frac{\partial E}{\partial w_i} = \iint (y - t) \frac{\partial y}{\partial w_i} p(x, t) dx dt.$$

$$\frac{\partial^2 E}{\partial w_i \partial w_j} = \iint \left(\frac{\partial y}{\partial w_j} \frac{\partial y}{\partial w_i} + (y - t) \frac{\partial^2 y}{\partial w_i \partial w_j} \right) p(x, t) dx dt.$$

($p(x, t) = p(t|x)p(x)$ より)

$$= \int \frac{\partial y}{\partial w_j} \frac{\partial y}{\partial w_i} \left(\int p(t|x) dt \right) p(x) dx$$

$$+ \int \frac{\partial^2 y}{\partial w_i \partial w_j} \left(\int (y - t) p(t|x) dt \right) p(x) dx$$

(第1項のカッコ内は1. 第2項は $y(x) = \int t p(t|x) dt$ を使うと0)

$$= \int \frac{\partial y}{\partial w_j} \frac{\partial y}{\partial w_i} p(x) dx.$$

ロジスティックモイドのときは

$$\begin{aligned}\nabla E(w) &= \sum_n \frac{\partial E}{\partial a_n} \nabla a_n = - \sum_n \left(\frac{t_n y_n (1 - y_n)}{y_n} - \frac{(1 - t_n) y_n (1 - y_n)}{1 - y_n} \right) \nabla a_n \\ &= \sum_n (y_n - t_n) \nabla a_n.\end{aligned}$$

よって $y_n \approx t_n$ なら

$$\begin{aligned}\nabla^2 E(w) &= \sum_n \frac{\partial y_n}{\partial a_n} \nabla a_n \nabla a_n^T + \sum_n (y_n - t_n) \nabla^2 a_n \\ &\approx \sum_n y_n (1 - y_n) \nabla a_n \nabla a_n^T.\end{aligned}$$

5.6 ヘッセ行列の厳密な評価

この節は計算は難しくはないが、記号がややこしいので書いてみる。変数の関係式は $\hat{a}_j = \sum_i w_{ji}^{(1)} x_i$, $z_j = h(\hat{a}_j)$, $a_k = \sum_j w_{kj}^{(2)} z_j$, $y_k = a_k$ である。PRML の a_j と a_k は違う対象であることに注意する。ここでは a_j の代わりに \hat{a}_j を使う。

添え字の i, i' は入力, j, j' は隠れユニット, k, k' は出力である。また

$$\delta_k = \frac{\partial E_n}{\partial a_k}, \quad M_{kk'} = \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}}$$

という記号を導入する。 E_n 以外の添え字 n を省略する。

5.6.1 両方の重みが第2層にある

$$\frac{\partial a_k}{\partial w_{kj}^{(2)}} = z_j, \quad \frac{\partial E_n}{\partial w_{kj}^{(2)}} = \frac{\partial a_k}{\partial w_{kj}^{(2)}} \frac{\partial E_n}{\partial a_k} = z_j \delta_k.$$

よって

$$\frac{\partial^2 E_n}{\partial w_{kj}^{(2)} \partial w_{k'j'}^{(2)}} = \frac{\partial a'_k}{\partial w_{k'j'}^{(2)}} \frac{\partial}{\partial a_{k'}} \left(\frac{\partial E_n}{\partial w_{kj}^{(2)}} \right) = z_{j'} z_j \frac{\partial \delta_k}{\partial a_{k'}} = z_j z_{j'} M_{kk'}.$$

5.6.2 両方の重みが第1層にある

$$\frac{\partial a_k}{\partial \hat{a}_j} = w_{kj}^{(2)} h'(\hat{a}_j), \quad \frac{\partial \hat{a}_j}{\partial w_{ji}^{(1)}} = x_i$$

より

$$\frac{\partial E_n}{\partial w_{ji}^{(1)}} = \frac{\partial \hat{a}_j}{\partial w_{ji}^{(1)}} \frac{\partial E_n}{\partial \hat{a}_j} = x_i \sum_k \frac{\partial a_k}{\partial \hat{a}_j} \frac{\partial E_n}{\partial a_k} = x_i \sum_k w_{kj}^{(2)} h'(\hat{a}_j) \delta_k.$$

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} = \frac{\partial \hat{a}_{j'}}{\partial w_{j'i'}^{(1)}} \frac{\partial}{\partial \hat{a}_{j'}} \left(\frac{\partial E_n}{\partial w_{ji}^{(1)}} \right) = x_i x_{i'} \underbrace{\frac{\partial}{\partial \hat{a}_{j'}} \left(h'(\hat{a}_j) \sum_k w_{kj}^{(2)} \delta_k \right)}_{=:A}.$$

$j = j'$ のとき

$$A = h''(\hat{a}_{j'}) \sum_k w_{kj}^{(2)} \delta_k + B, \quad B := h'(\hat{a}_j) \frac{\partial}{\partial \hat{a}_{j'}} \left(\sum_k w_{kj}^{(2)} \delta_k \right).$$

$j \neq j'$ のとき

$$B = h'(\hat{a}_j) \sum_{k'} \frac{\partial a_{k'}}{\partial \hat{a}_j} \frac{\partial}{\partial a_{k'}} \left(\sum_k w_{kj}^{(2)} \delta_k \right) = \sum_{k,k'} h'(\hat{a}_j) h'(\hat{a}_{j'}) w_{k'j'}^{(2)} w_{kj}^{(2)} M_{kk'}.$$

二つをまとめて

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} = x_i x_{i'} \left\{ h''(\hat{a}_{j'}) \delta_{jj'} \sum_k w_{kj}^{(2)} \delta_k + h'(\hat{a}_j) h'(\hat{a}_{j'}) \sum_{k,k'} w_{kj}^{(2)} w_{k'j'}^{(2)} M_{kk'} \right\}.$$

$\delta_{jj'}$ はクロネッカーのデルタ.

5.6.3 重みが別々の層に一つずつある

$$\frac{\partial E_n}{\partial w_{kj'}^{(2)}} = z_{j'} \delta_k, \quad \frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} = \frac{\partial \hat{a}_j}{\partial w_{ji}^{(1)}} \underbrace{\frac{\partial}{\partial \hat{a}_j} (z_{j'} \delta_k)}_{=:A}, \quad \frac{\partial \hat{a}_j}{\partial w_{ji}^{(1)}} = x_i.$$

$j = j'$ のとき

$$A = h'(\hat{a}_{j'}) \delta_k + B, \quad B := z_{j'} \frac{\partial \delta_k}{\partial \hat{a}_j}.$$

$j \neq j'$ のとき

$$B = z_{j'} \sum_{k'} \frac{\partial a_{k'}}{\partial \hat{a}_j} \frac{\partial \delta_k}{\partial a_{k'}} = z_{j'} \sum_{k'} w_{k'j}^{(2)} h'(\hat{a}_j) M_{kk'}.$$

よって

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} = x_i h'(\hat{a}_j) \left\{ \delta_{jj'} \delta_k + z_{j'} \sum_{k'} w_{k'j}^{(2)} M_{kk'} \right\}.$$

5.7 ヘッセ行列の積の高速な計算

応用面を考えると最終的に必要なものはヘッセ行列 H そのものではなくあるベクトル v と H の積であることが多い. H を計算せず直接 $v^T H = v^T \nabla \nabla$ を計算するために, 左

半分だけを取り出して $\mathcal{R}\{\cdot\} = v^T \nabla$ という記法を導入する. 5.3 節の終わりに書いたようにこの ∇ は入力縦ベクトルなら転置を取ってから作用するとみなす. なお, v に依存するものをあたかも依存しないかのように $\mathcal{R}\{\cdot\}$ と書いてしまうのは筋がよいとは思わない.

簡単な例を見てみよう. 2 変数関数 $y = f(x_1, x_2)$ について

$$\mathcal{R}\{\cdot\} = (v_1, v_2) \nabla = (v_1, v_2) \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \end{pmatrix}.$$

よって

$$\mathcal{R}\{x_1\} = (v_1, v_2) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = v_1,$$

$$\mathcal{R}\{x_2\} = (v_1, v_2) \begin{pmatrix} 0 \\ 1 \end{pmatrix} = v_2,$$

これを, $\mathcal{R}\{\cdot\}$ は入力値の x_i をその添え字に対応する v_i に置き換える作用と考えることにする. $\mathcal{R}\{\cdot\}$ は x_i について明らかに線形, つまり

$$\mathcal{R}\{ax_1 + bx_2\} = av_1 + bv_2 = a\mathcal{R}\{x_1\} + b\mathcal{R}\{x_2\}.$$

前節と同じ 2 層ネットワークで考えてみる. $a_j = \sum_i w_{ji}^{(1)} x_i$, $z_j = h(a_j)$, $y_k = \sum_j w_{kj}^{(2)} z_j$ である. a_k の代わりに y_k を使うので \hat{a}_j ではなく PRML と同じ a_j にする. PRML では w_{ji} の肩の添え字を省略しているが念のためここではつけておく.

$w_{ji}^{(1)}$ に対応する値を v_{ji} とすると $w_{ji}^{(1)}$ の線形和である a_j について

$$\mathcal{R}\{a_j\} = \sum_i x_i \mathcal{R}\{w_{ji}^{(1)}\} = \sum_i v_{ji} x_i.$$

$$\mathcal{R}\{z_j\} = v^T \nabla \left(\sum_i w_{ji}^{(1)} h(a_j) \right) = v^T \left(\frac{\partial h(a_j)}{\partial a_j} \nabla a_j \right) = h'(a_j) \mathcal{R}\{a_j\}.$$

$$\begin{aligned} \mathcal{R}\{y_k\} &= v^{(2)T} \left(\nabla \left(\sum_j w_{kj}^{(2)} z_j \right) \right) = v^{(2)T} \left(\sum_j \left(\nabla w_{kj}^{(2)} \right) z_j + \sum_j w_{kj}^{(2)} \nabla z_j \right) \\ &= \sum_j v_{kj}^{(2)} z_j + \sum_j w_{kj}^{(2)} \mathcal{R}\{z_j\}. \end{aligned}$$

なんとなくルールが見えてきたであろう. $\mathcal{R}\{\cdot\}$ は $\mathcal{R}\{w\} = v$ という記号の置き換え以外は積や合成関数の微分のルールの形に従っている (もともと微分作用素を用いて定義しているのが当然ではあるが).

逆伝播の式：

$$\begin{aligned}\delta_k^{(2)} &= y_k - t_k, \\ \delta_j^{(1)} &= h'(a_j) \sum_k w_{kj}^{(2)} \delta_k^{(2)}\end{aligned}$$

で考えてみると

$$\mathcal{R}\{\delta_k^{(2)}\} = \mathcal{R}\{y_k\}.$$

$$\mathcal{R}\{\delta_j^{(1)}\} = h''(a_j) \mathcal{R}\{a_j\} \left(\sum_k w_{kj}^{(2)} \delta_k^{(2)} \right) + h'(a_j) \left(\sum_k v_{kj}^{(2)} \delta_k^{(2)} + \sum_k w_{kj}^{(2)} \mathcal{R}\{\delta_k^{(2)}\} \right).$$

誤差の微分の式:

$$\begin{aligned}\frac{\partial E}{\partial w_{kj}^{(2)}} &= \delta_k^{(2)} z_j, \\ \frac{\partial E}{\partial w_{jk}^{(1)}} &= \delta_j^{(1)} x_i.\end{aligned}$$

より

$$\begin{aligned}\mathcal{R}\left\{\frac{\partial E}{\partial w_{kj}^{(2)}}\right\} &= \mathcal{R}\{\delta_k^{(2)}\} z_j + \delta_k^{(2)} \mathcal{R}\{z_j\}, \\ \mathcal{R}\left\{\frac{\partial E}{\partial w_{jk}^{(1)}}\right\} &= x_i \mathcal{R}\{\delta_j^{(1)}\}.\end{aligned}$$

5.8 ソフト重み共有

ネットワークの、あるグループに属する重みを等しくすることで複雑さを減らす手法がある。しかし重みが等しいという制約は厳しい。ソフト重み共有はその制約を外し、代わりに正則化項を追加することで、あるグループに属する重みが似た値をとれるようにする手法である。 π_k を混合係数として確率密度関数は

$$p(w) = \prod_i p(w_i), \quad p(w_i) = \sum_{k=1}^M \pi_k \mathcal{N}(w_i | \mu_k, \sigma_k^2).$$

$p(w_i)$ が確率分布なので混合係数は $\sum_k \pi_k = 1$, $0 \leq \pi_k \leq 1$ を満たす。2乗ノルムの正規化項は平均0のガウス事前分布の負の対数尤度関数とみなせた。ここでは複数個の重みに対応させるため混合ガウス分布を用いてみる。

$$\Omega(w) = -\log p(w) = -\sum_i \log \left(\sum_{k=1}^M \pi_k \mathcal{N}(w_i | \mu_k, \sigma_k^2) \right).$$

最小化したい目的関数は誤差関数と正則化項の和で

$$\tilde{E}(w) = E(w) + \Omega(w).$$

$p(j) = \pi_j$ において負担率を導入する.

$$\gamma_j(w_i) = p(j|w_i) = \frac{p(j)p(w_i|j)}{p(w_i)} = \frac{\pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)}{p(w_i)}.$$

正規分布の微分

$$\begin{aligned} \frac{\partial}{\partial x} \mathcal{N}(x|\mu, \sigma) &= \mathcal{N}(x|\mu, \sigma) \left(-\frac{x - \mu}{\sigma^2} \right), \\ \frac{\partial}{\partial \mu} \mathcal{N}(x|\mu, \sigma) &= \mathcal{N}(x|\mu, \sigma) \left(\frac{x - \mu}{\sigma^2} \right) \\ \frac{\partial}{\partial \sigma} \mathcal{N}(x|\mu, \sigma) &= \mathcal{N}(x|\mu, \sigma) \left(-\frac{1}{\sigma} + \frac{(x - \mu)^2}{\sigma^3} \right) \end{aligned}$$

を思い出しておく. $\log p(w_i)$ を w_i で微分すると

$$\begin{aligned} \frac{\partial}{\partial w_i} \log p(w_i) &= \frac{1}{p(w_i)} \left(\sum_k \pi_k \frac{\partial}{\partial w_i} \mathcal{N}(w_i|\mu_k, \sigma_k^2) \right) \\ &= \frac{1}{p(w_i)} \left(\sum_k \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2) \left(-\frac{w_i - \mu_k}{\sigma_k^2} \right) \right) \\ &= - \sum_k \frac{\pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)}{p(w_i)} \frac{w_i - \mu_k}{\sigma_k^2} \\ &= - \sum_k \gamma_k(w_i) \frac{w_i - \mu_k}{\sigma_k^2}. \end{aligned}$$

よって

$$\frac{\partial \tilde{E}}{\partial w_i} = \frac{\partial E}{\partial w_i} + \sum_k \gamma_k(w_i) \frac{w_i - \mu_k}{\sigma_k^2}.$$

同様に

$$\begin{aligned} \frac{\partial}{\partial \mu_k} \log p(w) &= \sum_j \frac{\partial}{\partial \mu_k} \log p(w_j) = \sum_j \frac{\pi_k \mathcal{N}(w_j|\mu_k, \sigma_k^2)}{p(w_j)} \frac{w_j - \mu_k}{\sigma_k^2} \\ &= \sum_j \gamma_k(w_j) \frac{w_j - \mu_k}{\sigma_k^2}. \end{aligned}$$

よって

$$\frac{\partial \tilde{E}}{\partial \mu_k} = \sum_j \gamma_k(w_j) \frac{\mu_k - w_j}{\sigma_k^2}.$$

$$\begin{aligned}
\frac{\partial \tilde{E}}{\partial \sigma_k} &= - \sum_j \frac{\partial}{\partial \sigma_k} \log p(w_j) = - \sum_j \frac{\pi_k \mathcal{N}(w_j | \mu_k, \sigma_k^2)}{p(w_j)} \left(-\frac{1}{\sigma_k} + \frac{(w_j - \mu_k)^2}{\sigma_k^3} \right) \\
&= \sum_j \gamma_k(w_j) \left(\frac{1}{\sigma_k} - \frac{(w_j - \mu_k)^2}{\sigma_k^3} \right).
\end{aligned}$$

π_j に関する制約より補助変数 η_j を用いて

$$\pi_j = \frac{\exp(\eta_j)}{\sum_k \exp(\eta_k)}$$

と表すと 4.13 節式 (4.1) より

$$\frac{\partial \pi_k}{\partial \eta_j} = \pi_k (\delta_{kj} - \pi_j).$$

よって

$$\begin{aligned}
\frac{\partial \tilde{E}}{\partial \eta_j} &= - \sum_i \frac{\partial}{\partial \eta_j} \log p(w_i) = - \sum_i \frac{\partial}{\partial \eta_j} \log \left(\sum_k \pi_k \mathcal{N}(w_i | \mu_k, \sigma_k^2) \right) \\
&= - \sum_{i,k} \frac{\mathcal{N}(w_i | \mu_k, \sigma_k^2)}{p(w_i)} \frac{\partial \pi_k}{\partial \eta_j} \\
&= - \sum_{i,k} \frac{\mathcal{N}(w_i | \mu_k, \sigma_k^2)}{p(w_i)} \pi_k (\delta_{kj} - \pi_j) \\
&= - \sum_i \left(\frac{\pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)}{p(w_i)} - \frac{\pi_j \sum_k \pi_k \mathcal{N}(w_i | \mu_k, \sigma_k^2)}{p(w_i)} \right) \\
&= - \sum_i (\gamma_j(w_i) - \pi_j) = \sum_i (\pi_j - \gamma_j(w_i)).
\end{aligned}$$

5.9 混合密度ネットワーク

$$p(t|x) = \sum_{k=1}^K \pi_k(x) \mathcal{N}(t | \mu_k(x), \sigma_k^2(x) I)$$

という分布のモデルを考える。このモデルのパラメータを、 x を入力としてえられるニューラルネットワークの出力となるようにとることで推論する。前節と同様 $\sum_k \pi_k(x) = 1$, $0 \leq \pi_k(x) \leq 1$ という制約があるので変数 a_l^π を導入し

$$\pi_k(x) = \frac{\exp(a_k^\pi)}{\sum_l \exp(a_l^\pi)}$$

とする。分散は 0 以上という制約があるので変数 a_k^σ を導入し

$$\sigma_k(x) = \exp(a_k^\sigma)$$

とする. 平均は特に制約がないので

$$\mu_{kj}(x) = a_{kj}^\mu$$

とする.

$$\mathcal{N}_{nk} = \mathcal{N}(t_n | \mu_k(x_n), \sigma_k^2(x_n)I)$$

とおくとデータが独立の場合, 誤差関数は

$$E(w) = - \sum_n \log \left(\sum_k \pi_k(x_n) \mathcal{N}_{nk} \right).$$

前節と同様 $p(k|x) = \pi_k(x)$ において負担率を

$$\gamma_{nk}(t_n|x_n) = p(k|t_n, x_n) = \frac{p(k|x_n)p(t_n|k)}{p(t_n|x_n)} = \frac{\pi_k \mathcal{N}_{nk}}{\sum_l \pi_l \mathcal{N}_{nl}}$$

とする.

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j(\delta_{kj} - \pi_k)$$

より

$$\frac{\partial E_n}{\partial a_k^\pi} = - \frac{\sum_j \pi_j(\delta_{kj} - \pi_k) \mathcal{N}_{nj}}{\sum_j \pi_j \mathcal{N}_{nj}} = -(\gamma_{nk} - \pi_k) = \pi_k - \gamma_{nk}.$$

$$\mathcal{N}(t|\mu, \sigma^2 I) = \frac{1}{(2\pi)^{L/2}} \frac{1}{\sigma^L} \exp \left(-\frac{1}{2\sigma^2} \sum_{l=1}^L (t_l - \mu_l)^2 \right)$$

なので

$$\frac{\partial}{\partial \mu_l} \mathcal{N}(t|\mu, \sigma^2 I) = \mathcal{N}(t|\mu, \sigma^2 I) \left(-\frac{t_l - \mu_l}{\sigma^2} \right).$$

よって

$$\frac{\partial E_n}{\partial a_{kl}^\mu} = - \frac{\mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} \frac{t_{nl} - \mu_{kl}}{\sigma_k^2} = \gamma_{nk} \left(\frac{\mu_{kl} - t_{nl}}{\sigma_k^2} \right).$$

同様に

$$\frac{\partial}{\partial \sigma} \mathcal{N}(t|\mu, \sigma^2 I) = \mathcal{N}(t|\mu, \sigma^2 I) \left(-\frac{L}{\sigma} + \frac{\|t - \mu\|^2}{\sigma^3} \right)$$

より

$$\frac{\partial}{\partial a_{kl}^\mu} \mathcal{N}_{nj} = \delta_{jk} \mathcal{N}_{nj} \left(\frac{t_{nl} - \mu_{kl}}{\sigma_k^2} \right).$$

よって

$$\frac{\partial \mathcal{N}_{nk}}{\partial a_k^\sigma} = \frac{\partial \sigma_k}{\partial a_k^\sigma} \frac{\partial \mathcal{N}_{nk}}{\partial \sigma_k} = \sigma_k \mathcal{N}_{nk} \left(-\frac{L}{\sigma_k} + \frac{\|t_n - \mu_k\|^2}{\sigma_k^3} \right) = \mathcal{N}_{nk} \left(-L + \frac{\|t_n - \mu_k\|^2}{\sigma_k^2} \right).$$

よって

$$\frac{\partial E_n}{\partial a_k^\sigma} = \gamma_{nk} \left(L - \frac{\|t_n - \mu_k\|^2}{\sigma_k^2} \right).$$

条件付き平均についての密度関数の分散は

$$\begin{aligned} s^2(x) &= E[\|t - E[t|x]\|^2|x] \\ &= \sum_k \pi_k \int (\|t\|^2 - 2t^T E[t|x] + \|E[t|x]\|^2) \mathcal{N}(t|\mu_k, \sigma_k^2 I) dt \\ &= \sum_k \pi_k (\sigma_k^2 + \|\mu_k\|^2 - 2\mu_k^T E[t|x] + \|E[t|x]\|^2) \\ &= \sum_k \pi_k(x) \left(\sigma_k(x)^2 + \|\mu_k - \sum_j \pi_j(x) \mu_j(x)\|^2 \right). \end{aligned}$$

5.10 クラス分類のためのベイズニューラルネットワーク

ロジスティックシグモイド出力を一つ持つネットワークによる2クラス分類問題を考える。そのモデルの対数尤度関数は $t_n \in \{0, 1\}$, $y_n = y(x_n, w)$ として

$$\log p(\mathcal{D}|w) = \sum_n (t_n \log y_n + (1 - t_n) \log(1 - y_n)).$$

事前分布を

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) = \frac{1}{\sqrt{2\pi}^W |\alpha^{-1}|^{W/2}} \exp\left(-\frac{1}{2}\alpha w^T w\right)$$

とする (W は w に含まれるパラメータの総数)。ノイズがないので β を含まない。

$$E(w) = \log p(\mathcal{D}|w) + \frac{\alpha}{2} w^T w$$

の最小化で w_{MAP} を求め、 $A = -\nabla^2 \log p(w|\mathcal{D})|_{w=w_{\text{MAP}}}$ を何らかの方法で求める。ラプラス近似を使って事後分布をガウス近似すると

$$q(w|\mathcal{D}) = \mathcal{N}(w|w_{\text{MAP}}, A^{-1}).$$

正規化項を求める 4.17 節式 (4.2) を使って

$$\begin{aligned} \log p(\mathcal{D}|\alpha) &\approx \log \left(p(\mathcal{D}|w_{\text{MAP}}) p(w_{\text{MAP}}|\alpha) \sqrt{\frac{(2\pi)^W}{|A|}} \right) \\ &= \log p(\mathcal{D}|w_{\text{MAP}}) - \frac{W}{2} \log(2\pi) + \frac{W}{2} \log \alpha - \frac{1}{2} \alpha w_{\text{MAP}}^T w_{\text{MAP}} \\ &\quad + \frac{W}{2} \log(2\pi) - \frac{1}{2} \log |A| \\ &= -E(w_{\text{MAP}}) - \frac{1}{2} \log |A| + \frac{W}{2} \log \alpha. \end{aligned}$$

$$E(w_{\text{MAP}}) = - \sum_n (t_n \log y_n + (1 - t_n) \log(1 - y_n)) + \frac{1}{2} \alpha w_{\text{MAP}}^T w_{\text{MAP}}.$$

予測分布を考える。出力ユニットの活性化関数を線形近似する。

$$\begin{aligned} a(x, w) &\approx a(x, w_{\text{MAP}}) + \nabla a(x, w_{\text{MAP}})^T (w - w_{\text{MAP}}) \\ &\quad (a_{\text{MAP}}(x) = a(x, w_{\text{MAP}}), \quad b = \nabla a(x, w_{\text{MAP}}) \text{ として}) \\ &= a_{\text{MAP}}(x) + b^T (w - w_{\text{MAP}}). \end{aligned}$$

$$\begin{aligned} p(a|x, \mathcal{D}) &= \int \delta(a - a(x, w)) q(w|\mathcal{D}) dw \\ &= \int \delta(a - a_{\text{MAP}}(x) - w_{\text{MAP}}^T b + w^T b) q(w|\mathcal{D}) dw. \end{aligned}$$

平均は

$$\begin{aligned} E[a] &= \int a p(a|x, \mathcal{D}) da = \int \delta(a - a(x, w)) w(w) a da dw \\ &= \int a(x, w) q(w) dw = (a_{\text{MAP}}(x) - w_{\text{MAP}}^T b) + \int b^T w q(w) dw \\ &= a_{\text{MAP}}(x) - w_{\text{MAP}}^T b + b^T w_{\text{MAP}} = a_{\text{MAP}}(x). \end{aligned}$$

分散は $w^T b$ が効くので

$$\sigma_a^2(x) = b^T A^{-1} b(x)$$

予測分布は 4.20 節式 (4.4) の近似式を使って

$$p(t = 1|x, \mathcal{D}) = \int \sigma(a) p(a|x, \mathcal{D}) da \approx \sigma(\kappa(\sigma_a^2) a_{\text{MAP}}(x)).$$

第 9 章

「混合モデルと EM」の数式の補足

この文章は『パターン認識と機械学習』（PRML）の 9 章の式変形を一部埋めたものです。面倒なので特に紛らわしいと思わない限り \boldsymbol{x} を x と書いたりします。また対数尤度関数を F と書くことが多いです。

9.1 復習

よく使ういくつかの式を書いておく。どれも今までに既に示したものである。2 章や 3 章を参照。

9.1.1 行列の公式

$$\begin{aligned} x^T A x &= \text{tr}(A x x^T), \\ \frac{\partial}{\partial A} \log |A| &= (A^{-1})^T, \\ \frac{\partial}{\partial x} \log |A| &= \text{tr}\left(A^{-1} \frac{\partial}{\partial x} A\right), \\ \frac{\partial}{\partial A} \text{tr}(A^{-1} B) &= -(A^{-1} B A^{-1})^T. \end{aligned}$$

9.1.2 微分

関数 f に対して対数関数の微分は

$$(\log f)' = \frac{f'}{f}.$$

よって逆に

$$f' = f \cdot (\log f)'.$$

ガウス分布など対数の微分が分かりやすいときによく使う。

9.1.3 ガウス分布

$$\mathcal{N} = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right).$$

期待値と分散について

$$E[x] = \mu, \quad \text{cov}[x] = \Sigma, \quad E[xx^T] = \mu\mu^T + \Sigma, \quad E[x^T x] = \mu^T \mu + \text{tr}(\Sigma).$$

最後の式は3番目から出る。

$$E[x_i^2] = (\mu\mu^T)_{ii} + \Sigma_{ii} = \mu_i^2 + \Sigma_{ii}.$$

よって

$$E[x^T x] = \sum_i E[x_i^2] = \mu^T \mu + \text{tr}(\Sigma).$$

9.2 混合ガウス分布

離散的な潜在変数を用いた混合ガウス分布の定式化. K 次元 2 値確率変数 z を考える (どれか一つの成分のみが 1 であとは 0). つまり

$$\sum_k z_k = 1.$$

z の種類は K 個である. $0 \leq \pi_k \leq 1$ という係数を用いて

$$p(z_k = 1) = \pi_k$$

という確率分布を与える.

$$p(z) = \prod_k \pi_k^{z_k}, \quad p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$

なので

$$P(x|z) = \prod_k \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}.$$

これらを合わせて

$$\begin{aligned} p(x) &= \sum_z p(z)p(x|z) \\ &= \sum_z \prod_k (\pi_k \mathcal{N}(x|\mu_k, \Sigma_k))^{z_k} \\ &\quad (z_k \text{ はどれか一つのみが 1 (そのとき } \pi_k) \text{ であとは 0 なので}) \\ &= \sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k). \end{aligned}$$

x が与えられたときの z の条件付き確率 $p(z_k = 1|x)$ を $\gamma(z_k)$ とする.

$$\gamma(z_k) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_j p(z_j = 1)p(x|z_j = 1)} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}.$$

これを混合要素 k が観測値 x に対する負担率という.

9.3 混合ガウス分布の EM アルゴリズム

混合ガウス分布において観測したデータ集合を $X^T = \{x_1, \dots, x_N\}$, 対応する潜在変数を $Z^T = \{z_1, \dots, z_N\}$ とする. X は $N \times D$ 行列で Z は $N \times K$ 行列.

対数尤度関数の最大点の条件をもとめる.

$$F = \log p(X|\pi, \mu, \Sigma) = \sum_{n=1}^N \log \left(\sum_{j=1}^K \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j) \right)$$

とする.

$$\frac{\partial}{\partial \mu} \log \mathcal{N}(x|\mu, \Sigma) = \frac{\partial}{\partial \mu} \left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right) = \Sigma^{-1}(x - \mu)$$

より

$$\frac{\partial}{\partial \mu} \mathcal{N} = \mathcal{N} \cdot \left(\frac{\partial}{\partial \mu} \log \mathcal{N} \right) = \mathcal{N} \cdot \Sigma^{-1}(x - \mu).$$

もちろんガウス分布の微分は普通にそのまましてもよい. だが今回は対数をとってから微分をとった方が, 微分してでてくる \mathcal{N} が $\gamma(z_{nk})$ の一部となることを見通しやすいのでそうしてみた. さて $\mathcal{N}_{nk} = \mathcal{N}(x_n|\mu_k, \Sigma_k)$ とおいて

$$\begin{aligned} \frac{\partial}{\partial \mu_k} F &= \sum_n \frac{\pi_k \frac{\partial}{\partial \mu_k} \mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} = \sum_n \left(\frac{\pi_k \mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} \right) \frac{\partial}{\partial \mu_k} \log \mathcal{N}_{nk} \\ &= \sum_n \gamma(z_{nk}) \frac{\partial}{\partial \mu_k} \log \mathcal{N}_{nk} = \Sigma_k^{-1} \left(\sum_n \gamma(z_{nk})(x_n - \mu_k) \right) = 0. \end{aligned}$$

よって

$$\begin{aligned} \sum_n \gamma(z_{nk}) x_n - \left(\sum_n \gamma(z_{nk}) \right) \mu_k &= 0. \\ N_k &= \sum_n \gamma(z_{nk}) \end{aligned}$$

とおくと

$$\mu_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) x_n.$$

これは μ_k が X の重みつき平均であることを示している. 次に Σ_k に関する微分を考える.

$$\mathcal{N} = \mathcal{N}(x|\mu, \Sigma)$$

のとき

$$\log \mathcal{N} = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \text{tr} \left(\Sigma^{-1} (x - \mu)(x - \mu)^T \right)$$

なので $\Sigma^T = \Sigma$ だから

$$\frac{\partial}{\partial \Sigma} (\log \mathcal{N}) = -\frac{1}{2} (\Sigma^{-1}) + \frac{1}{2} \left(\Sigma^{-1} (x - \mu)(x - \mu)^T \Sigma^{-1} \right).$$

よって μ_k の微分と同様にして

$$\begin{aligned} \frac{\partial}{\partial \Sigma_k} F &= \sum_n \gamma(z_{nk}) \frac{\partial}{\partial \Sigma_k} \log \mathcal{N}_{nk} \\ &= \sum_n \gamma(z_{nk}) \left(-\frac{1}{2} (\Sigma_k^{-1}) + \frac{1}{2} \left(\Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right) \right) = 0. \end{aligned}$$

よって

$$\sum_n \gamma(z_{nk}) \left(I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right) = 0.$$

$$\Sigma_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T.$$

最後に π_k に関する微分を考える. $\sum_k \pi_k = 1$ の制約を入れる.

$$G = F + \lambda \left(\sum_k \pi_k - 1 \right)$$

とすると

$$\frac{\partial}{\partial \pi_k} G = \sum_n \frac{\mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} + \lambda = \sum_n \gamma(z_{nk}) / \pi_k + \lambda = N_k / \pi_k + \lambda = 0.$$

つまり $N_k = -\lambda \pi_k$. よって

$$N = \sum_k N_k = \sum_k (-\lambda \pi_k) = -\lambda.$$

よって

$$\pi_k = \frac{N_k}{-\lambda} = \frac{N_k}{N}.$$

9.4 混合ガウス分布再訪

$$p(z) = \prod_k \pi_k^{z_k}, \quad p(x|z) = \prod_k \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$$

より

$$\begin{aligned} F = \log p(X, Z|\mu, \Sigma, \boldsymbol{\pi}) &= \log \left(\prod_{n,k} \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}} \right) \\ &= \sum_{n,k} z_{nk} (\log \pi_k + \log \mathcal{N}_{nk}). \end{aligned}$$

z_n は $(0, 0, \dots, 1, 0, \dots, 0)$ の形で $\sum_k \pi_k = 1$ の制約条件を入れると上式の微分を考えると

$$G = F + \lambda \left(\sum_k \pi_k - 1 \right)$$

として

$$\frac{\partial}{\partial \pi_k} G = \sum_n z_{nk} \frac{1}{\pi_k} + \lambda = \left(\sum_n z_{nk} \right) / \pi_k + \lambda = 0.$$

よって

$$\begin{aligned} \pi_k &= -\frac{1}{\lambda} \sum_n z_{nk}. \\ \sum_k \pi_k &= -\frac{1}{\lambda} \sum_{n,k} z_{nk} = -\frac{N}{\lambda} = 1. \end{aligned}$$

よって $\lambda = -N$. つまり

$$\pi_k = \frac{1}{N} \sum_n z_{nk}.$$

完全データ集合についての対数尤度関数の最大化は解けるが、潜在変数が分からない場合の不完全データに関する対数尤度関数の最大化は困難。この場合は潜在変数の事後分布に関する完全データ尤度関数の期待値を考える。

$$p(Z|X, \mu, \Sigma, \boldsymbol{\pi}) = \frac{p(X, Z|\mu, \Sigma, \boldsymbol{\pi})}{p(X|\mu, \Sigma, \boldsymbol{\pi})} \propto \prod_{n,k} (\pi_k \mathcal{N}_{nk})^{z_{nk}}.$$

$$E[z_{nk}] = \frac{\sum_{z_n} z_{nk} \prod_j (\pi_j \mathcal{N}_{nj})^{z_{nj}}}{\sum_{z_n} \prod_j (\pi_j \mathcal{N}_{nj})^{z_{nj}}} = \frac{\pi_k \mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} = \gamma(z_{nk}).$$

よって

$$F = E_Z[\log p(X, Z|\mu, \Sigma, \boldsymbol{\pi})] = \sum_{n,k} \gamma(z_{nk}) (\log \pi_k + \log \mathcal{N}_{nk}).$$

まずパラメータ μ, Σ, π を適当に決めて負担率 $\gamma(z_{nk})$ を求め, それを fix して μ_k, Σ_k, π_k について F を最大化. 今までと同様にできる. $F' = F + \lambda(\sum_k \pi_k - 1)$ として

$$\frac{\partial}{\partial \pi_k} F' = \sum_n \gamma(z_{nk})(1/\pi_k) + \lambda = 0$$

より

$$\sum_n \gamma(z_{nk}) = \lambda \pi_k.$$

$$\sum_{n,k} \gamma(z_{nk}) = -\lambda \left(\sum_k \pi_k \right) = -\lambda = N$$

より

$$\pi_k = \frac{1}{N} \sum_n \gamma(z_{nk}) = \frac{N_k}{N}.$$

$$\frac{\partial}{\partial \mu_k} F = \sum_n \gamma(z_{nk}) (-\Sigma_k^{-1}(x_n - \mu_k)) = \Sigma_k^{-1} \left(\sum_n \gamma(z_{nk}) x_n - \left(\sum_n \gamma(z_{nk}) \right) \mu_k \right) = 0.$$

よって

$$\mu_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) x_n.$$

$$\frac{\partial}{\partial \Sigma_k} F = \sum_n \gamma(z_{nk}) \frac{\partial}{\partial \Sigma_k} \log \mathcal{N}_{nk} = 0$$

として同様（流石に略）.

9.5 K -means との関連

PRML 式 (9.43) は不正確. E ではなく ϵE を考えないと PRML 式 (9.43) の右辺にはならない. RPML 式 (9.40) を E とおく.

$$E = \sum_{n,k} \gamma(z_{nk}) (\log \pi_k + \log \mathcal{N}(x_n | \mu_k, \Sigma_k)).$$

ϵE に

$$\mathcal{N}(x | \mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp \left(-\frac{1}{2\epsilon} \|x - \mu_k\|^2 \right)$$

を代入する.

$$\epsilon E = \sum_{n,k} \gamma(z_{nk}) \left(\epsilon \log \pi_k - \frac{D}{2} \epsilon \log(2\pi\epsilon) - \frac{1}{2} \|x_n - \mu_k\|^2 \right).$$

$\epsilon \rightarrow 0$ で

$$\gamma(z_{nk}) \rightarrow r_{nk}, \quad \epsilon \log \pi_k \rightarrow 0, \quad \epsilon \log(2\pi\epsilon) \rightarrow 0$$

より

$$\epsilon E \rightarrow -\frac{1}{2} \sum_{n,k} r_{nk} \|x_n - \mu_k\|^2 = -J.$$

よって期待完全データ対数尤度の最大化は J の最小化と同等.

9.6 混合ベルヌーイ分布

$x = (x_1, \dots, x_D)^T$, $\mu = (\mu_1, \dots, \mu_D)^T$ とする.

$$p(x|\mu) = \prod_{i=1}^D \mu_i^{x_i} (1 - \mu_i)^{(1-x_i)}.$$

$E[x] = \mu$ は容易に分かる.

$$E[x_i x_j] = \begin{cases} \mu_i \mu_j & (i \neq j) \\ \mu_i & (i = j). \end{cases}$$

よって

$$\text{cov}[x]_{ij} = E[(x - \mu)(x - \mu)^T]_{ij} = E[x_i x_j] - (\mu \mu^T)_{ij} = (\mu_i - \mu_i^2) \delta_{ij}$$

より

$$\text{cov}[x] = \text{diag}(\mu_i(1 - \mu_i)).$$

$\mu = \{\mu_1, \dots, \mu_K\}$, $\pi = \{\pi_1, \dots, \pi_K\}$ として次の混合分布を考えよう.

$$p(x|\mu_k) = \prod_i \mu_{ki}^{x_i} (1 - \mu_{ki})^{(1-x_i)}.$$

$$E[x] = \int x p(x|\mu) dx = \sum_k \pi_k \int x p(x|\mu_k) dx = \sum_k \pi_k E_k[x] = \sum_k \pi_k \mu_k.$$

$$E_k[xx^T] = \text{cov}_k[x] + \mu_k \mu_k^T = \Sigma_k + \mu_k \mu_k^T$$

より

$$\begin{aligned} \text{cov}[x] &= E[(x - E[x])(x - E[x])^T] = E[xx^T] - E[x]E[x]^T \\ &= \sum_k \pi_k (\Sigma_k + \mu_k \mu_k^T) - E[x]E[x]^T. \end{aligned}$$

データ集合 $X = \{x_1, \dots, x_N\}$ が与えられたとき, 対数尤度関数は

$$\log p(X|\mu, \pi) = \sum_n \log \left(\sum_k \pi_k p(x_n|\mu_k) \right).$$

対数の中に和があるので解析的に最尤解をもとめられない．EM アルゴリズムを使う． x に対応する潜在変数を $z = (z_1, \dots, z_K)^T$ を導入する．どれか一つのみ 1 でその他は 0 のベクトルである． z の事前分布を

$$p(z|\pi) = \prod_k \pi_k^{z_k}$$

とする． z が与えられたときの条件付き確率は

$$p(x|z, \mu) = \prod_k p(x|\mu_k)^{z_k}.$$

$$p(x, z|\mu, \pi) = p(x|z, \mu)p(z|\pi) = \prod_k (\pi_k p(x|\mu_k))^{z_k}.$$

よって

$$p(x|\mu, \pi) = \sum_z p(x, z|\mu, \pi) = \sum_k \pi_k p(x|\mu_k).$$

完全データ対数尤度関数は $X = \{x_n\}$, $Z = \{z_n\}$ として

$$\begin{aligned} \log p(X, Z|\mu, \pi) &= \sum_{n,k} z_{nk} \underbrace{\left(\log \pi_k + \sum_i x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki}) \right)}_{=: A_{nk}} \\ &= \sum_{n,k} z_{nk} A_{nk}. \end{aligned}$$

$$\begin{aligned} E[z_{nk}] &= \frac{\sum_{z_n} z_{nk} \prod_j (\pi_j p(x_n|\mu_j))^{z_{nj}}}{\sum_{z_n} \prod_j (\pi_j p(x_n|\mu_j))^{z_{nj}}} \\ &\quad (z_{nk} = 1 \text{ となるものだけとるので}) \\ &= \frac{\pi_k p(x_n|\mu_k)}{\sum_j \pi_j p(x_n|\mu_j)} \end{aligned}$$

を $\gamma(z_{nk})$ とおく．すると

$$E_Z[\log p(X, Z|\mu, \pi)] = \sum_{n,k} \gamma(z_{nk}) A_{nk}.$$

$$N_k = \sum_n \gamma(z_{nk}), \quad \bar{x}_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) x_n$$

とおく.

$$\begin{aligned}
 F &= E_Z[\log p(X, Z|\mu, \pi)] \\
 &= \sum_k (\log \pi_k) \left(\sum_n \gamma(z_{nk}) \right) + \sum_{k,i} \log \mu_{ki} \left(\sum_n \gamma(z_{nk}) x_{ni} \right) \\
 &\quad + \sum_{k,i} \log(1 - \mu_{ki}) \left(\sum_n \gamma(z_{nk}) (1 - x_{ni}) \right) \\
 &= \sum_k N_k \log \pi_k + \sum_{k,i} N_k \bar{x}_{ki} \log \mu_{ki} + \sum_{k,i} \log(1 - \mu_{ki}) N_k (1 - \bar{x}_{ki}).
 \end{aligned}$$

μ_{ki} に関する最大化.

$$\frac{\partial}{\partial \mu_{ki}} F = N_k \bar{x}_{ki} \frac{1}{\mu_{ki}} + \frac{-1}{1 - \mu_{ki}} N_k (1 - \bar{x}_{ki}) = \frac{N_k}{\mu_{ki}(1 - \mu_{ki})} (\bar{x}_{ki}(1 - \mu_{ki}) - (1 - \bar{x}_{ki})\mu_{ki}) = 0.$$

よって

$$\bar{x}_{ki} - \bar{x}_{ki}\mu_{ki} - \mu_{ki} + \bar{x}_{ki}\mu_{ki} = \bar{x}_{ki} - \mu_{ki} = 0.$$

よって

$$\mu_k = \bar{x}_k.$$

π_k に関する最適化. $G = F + \lambda(\sum_k \pi_k - 1)$ とすると

$$\frac{\partial}{\partial \pi_k} G = \frac{N_k}{\pi_k} + \lambda = 0.$$

よって

$$N_k = -\lambda \pi_k, \quad N = \sum_k N_k = -\lambda \sum_k \pi_k = -\lambda.$$

つまり $\lambda = -N$ となり

$$\pi_k = \frac{N_k}{N}.$$

$0 \leq p(x_n|\mu_k) \leq 1$ より

$$\log p(X|\mu, \pi) = \sum_n \log \left(\sum_k \pi_k p(x_n|\mu_k) \right) \leq \sum \log \left(\sum_k \pi_k \right) = 0.$$

よって尤度関数が発散することはない.

9.7 ベイズ線形回帰に関する EM アルゴリズム

EM アルゴリズムに基づいてベイズ線形回帰を考えてみる. w を潜在関数と見なしてそれを最大化する方針を採る.

$$p(w|t) = \mathcal{N}(w|m_N, S_N)$$

で w の事後分布が求まっているとする.

$$p(t|w, \beta) = \prod_n \mathcal{N}(t_n | w^T \phi(x_n), \beta^{-1}), \quad p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I)$$

であった. このとき完全データ対数尤度関数は

$$\log p(t, w|\alpha, \beta) = \log p(t|w, \beta) + \log p(w|\alpha).$$

なので

$$\begin{aligned} F &= E[\log p(t, w|\alpha, \beta)] \\ &= E \left[\sum_n \left(\frac{1}{2} \log \left(\frac{\beta}{2\pi} \right) - \frac{\beta}{2} (t_n - w^T \phi_n)^2 \right) + \frac{M}{2} \log \left(\frac{\alpha}{2\pi} \right) - \frac{\alpha}{2} w^T w \right] \\ &= \frac{M}{2} \log \left(\frac{\alpha}{2\pi} \right) - \frac{\alpha}{2} E[w^T w] + \frac{N}{2} \log \left(\frac{\beta}{2\pi} \right) - \frac{\beta}{2} \sum_n E[(t_n - w^T \phi_n)^2]. \end{aligned}$$

α に関する最大化

$$\frac{\partial}{\partial \alpha} F = \frac{M}{2} \frac{1}{\alpha} - \frac{1}{2} E[w^T w] = 0.$$

よって

$$\alpha = \frac{M}{E[w^T w]} = \frac{M}{m_N^T m_N + \text{tr}(S_N)}.$$

β に関する最大化

$$\frac{\partial}{\partial \beta} F = \frac{N}{2} \frac{1}{\beta} - \frac{1}{2} \sum_n E[(t_n - w^T \phi_n)^2] = 0.$$

よって

$$\frac{1}{\beta} = \frac{1}{N} \sum_n E[(t_n - w^T \phi_n)^2].$$

9.8 一般の EM アルゴリズム

潜在変数をもつ確率モデルの最尤解を求めるための一般的手法. X を確率変数, Z を潜在変数, θ をパラメータとする. 完全データ対数尤度関数 $\log p(X, Z|\theta)$ の最適化は容易であるという仮定の元で目的は $p(X|\theta) = \sum_Z p(X, Z|\theta)$ の最大化.

Z に対する分布を $q(Z)$ とする

$$p(X, Z|\theta) = p(Z|X, \theta)p(X|\theta).$$

$$\mathcal{L}(q, \theta) = \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)}, \quad \text{KL}(q||p) = - \sum_Z q(Z) \log \frac{p(Z|X, \theta)}{q(Z)}$$

とおく. $\text{KL}(q||p)$ は $q(Z)$ と事後分布 $p(Z|X, \theta)$ との距離なので常に 0 以上 (3 章のカルバック距離を参照) .

$$\mathcal{L}(q, \theta) + \text{KL}(q||p) = \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{p(Z|X, \theta)} = \sum_Z q(Z) \log p(X|\theta) = \log p(X|\theta).$$

よって

$$\log p(X|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q||p) \geq \mathcal{L}(q, \theta).$$

したがって $\mathcal{L}(q, \theta)$ は $\log p(X|\theta)$ の下界. パラメータの現在の値が θ^o だったときに

E ステップでは θ^o を固定して $\mathcal{L}(q, \theta)$ を $q(Z)$ について最大化する. $\log p(X|\theta)$ は q によらないのでそれは $\text{KL} = 0$ のとき, つまり

$$q(Z) = p(Z|X, \theta^o)$$

のときである.

M ステップでは $q(Z)$ を固定して $\mathcal{L}(q, \theta)$ を θ について最大化する. その θ を θ^n とする. 最大値になっていなければ, 必ず \mathcal{L} が増加し, $\log p(X|\theta)$ も増える. このときの $\text{KL}(q||p)$ は θ^o を使って計算されていた (そして値は 0) ので新しい θ^n を使って計算し直すと通常正となる.

$q(Z) = p(Z|X, \theta^o)$ より

$$\begin{aligned} \mathcal{L}(q, \theta) &= \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} \\ &= \sum_Z p(Z|X, \theta^o) \log p(X, Z|\theta) - \sum_Z p(Z|X, \theta^o) \log p(Z|X, \theta^o) \\ &= (\mathcal{Q}(\theta, \theta^o) = \sum_Z p(Z|X, \theta^o) \log p(X, Z|\theta) \text{ において}) \\ &= \mathcal{Q}(\theta, \theta^o) + \theta \text{ に非依存.} \end{aligned}$$

つまり $\mathcal{L}(q, \theta)$ の最大化は $\mathcal{Q}(\theta, \theta^o)$ の最大化に等しい.

9.9 混合ガウス分布のオンライン版 EM アルゴリズム

各 EM のステップで一つのデータ点のみの更新を行うことを考える. これは m 番目のデータ以外を潜在変数とする EM アルゴリズムとみなすことができる.

E ステップでは分布 $p(Z|X, \theta^o)$ を求める必要があるが, M ステップに必要な μ_k, Σ_k, π_k の更新式の右辺を見ると必要なデータは $\gamma(z_{nk})$ のみであることが分かる. つまりそれらの差分さえ分かればアルゴリズムを書き下すことができる.

$$N_k = \sum_n \gamma(z_{nk})$$

を m 番目の値だけ更新する. 新しい値を N'_k とすると

$$N'_k = \sum_{n \neq m} \gamma(z_{nk}) + \gamma'(z_{mk}) = N_k + \gamma'(z_{mk}) - \gamma(z_{mk}).$$

$d := \gamma'(z_{mk}) - \gamma(z_{mk})$ とおくと $N'_k = N_k + d$. $\pi_k = N_k/N$ なので

$$\pi'_k = \frac{N'_k}{N} = \frac{N_k + d}{N_k} = \pi_k + \frac{d}{N}.$$

$\mu_k = (1/N_k) \sum_n \gamma(z_{nk}) x_n$ より

$$\mu'_k = \frac{1}{N'_k} \left(\sum_{n \neq m} \gamma(z_{nk}) x_n + \gamma'(z_{mk}) x_m \right).$$

よって

$$N'_k \mu'_k = N_k \mu_k - \gamma(z_{mk}) x_m + \gamma'(z_{mk}) x_m = (N'_k - d) \mu_k + d x_m = N'_k \mu_k + d(x_m - \mu_k).$$

よって

$$\mu'_k = \mu_k + \frac{d}{N'_k} (x_m - \mu_k).$$

$S := \Sigma_k = (1/N_k) \sum_n \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T$ より $S' := \Sigma'_k$ とすると

$$N'_k S' = \sum_{n \neq m} \gamma(z_{nk}) (x_n - \mu'_k)(x_n - \mu'_k)^T + \gamma'(z_{mk}) (x_m - \mu'_k)(x_m - \mu'_k)^T.$$

以下式変形をひたすら行う.

$$x_m - \mu'_k = x_m - \mu_k - \frac{d}{N'_k} (x_m - \mu_k) = \left(1 - \frac{d}{N'_k}\right) (x_m - \mu_k) = \frac{N_k}{N'_k} (x_m - \mu_k).$$

$$x_n - \mu'_k = (x_n - \mu_k) - \frac{d}{N'_k} (x_m - \mu_k).$$

$A := (x_m - \mu_k)(x_m - \mu_k)^T$ とおくと

$$(x_n - \mu'_k)(x_n - \mu'_k)^T = (x_n - \mu_k)(x_n - \mu_k)^T - 2 \frac{d}{N'_k} (x_n - \mu_k)(x_m - \mu_k)^T + \frac{d^2}{N_k'^2} A.$$

$$\sum_{n \neq m} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T = N_k S - \gamma(z_{mk}) (x_m - \mu_k)(x_m - \mu_k)^T = N_k S - \gamma(z_{mk}) A.$$

$$\begin{aligned}
\sum_{n \neq m} \gamma(z_{nk})(x_n - \mu_k) &= \sum_{n \neq m} \gamma(z_{nk})x_n - \sum_{n \neq m} \gamma(z_{nk})\mu_k \\
&= N_k \mu_k - \gamma(z_{mk})x_m - (N_k - \gamma(z_{mk}))\mu_k \\
&= -\gamma(z_{mk})(x_m - \mu_k).
\end{aligned}$$

よって $\gamma := \gamma(z_{mk})$ とおくと

$$\begin{aligned}
N'_k S' &= N_k S - \gamma A + 2 \frac{d}{N'_k} \gamma A + (N_k - \gamma) \frac{d^2}{N'^2_k} A + (\gamma + d) A \frac{N^2_k}{N'^2_k} \\
&= N_k S + \frac{A}{N'^2_k} (-\gamma(N_k + d)^2 + 2d\gamma(N_k + d) + (N_k - \gamma)d^2 + (\gamma + d)N^2_k) \\
&= N_k S + \frac{A}{N'^2_k} (-\gamma N^2_k - 2d\gamma N_k - \gamma d^2 + 2d\gamma N_k + 2d^2\gamma + N_k d^2 - \gamma d^2 + \gamma N^2_k + dN^2_k) \\
&= N_k S + \frac{A}{N'^2_k} N_k d(N_k + d) = N_k S + \frac{A N_k d}{N'_k}.
\end{aligned}$$

よって

$$S' = \frac{N_k}{N'_k} \left(S + \frac{d}{N'_k} (x_m - \mu_k)(x_m - \mu_k)^T \right).$$

第 10 章

「近似推論法」の数式の補足

この文章は『パターン認識と機械学習』（PRML）の 10 章の式変形を一部埋めたものです。

10.1 この章でよく使われる公式

9 章と同じようによく使う公式を列挙しておく。

10.1.1 ガンマ関数

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt, \quad \Gamma(x+1) = x\Gamma(x).$$

ディガンマ関数

$$\phi(x) = \frac{\partial}{\partial x} \log \Gamma(x).$$

10.1.2 ディリクレ分布

$0 \leq \mu_k \leq 1, \sum_k \mu_k = 1, \hat{\alpha} = \sum_k \alpha_k$ として

$$\text{Dir}(\mu|\alpha) = C(\alpha) \prod_{k=1}^K \mu_k^{\alpha_k-1}, \quad C(\alpha) = \frac{\Gamma(\hat{\alpha})}{\prod_k \Gamma(\alpha_k)}, \quad E[\mu_k] = \frac{\alpha_k}{\hat{\alpha}}.$$

10.1.3 ガンマ分布

$$\text{Gam}(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}, \quad E[\tau] = \frac{a}{b}, \quad \text{var}[\tau] = \frac{a}{b^2}, \quad E[\log \tau] = \phi(a) - \log b.$$

10.1.4 正規分布

$$\mathcal{N} = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} |\Sigma|^{-1/2} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

$$E[x] = \mu, \quad \text{cov}[x] = \Sigma, \quad E[xx^T] = \mu\mu^T + \Sigma, \quad E[x^T x] = \mu^T \mu + \text{tr}(\Sigma).$$

$p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})$, $p(y|x) = \mathcal{N}(x|Ax + b, L^{-1})$ のとき

$$p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T).$$

10.1.5 スチューデントの t 分布

$$\text{St}(x|\mu, \Lambda, \nu) = \frac{\Gamma(\frac{\nu+D}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\Lambda|^{1/2}}{(\pi\nu)^{1/2}} \left(1 + \frac{\Delta^2}{\nu} \right)^{-\frac{\nu+1}{2}}, \quad \Delta^2 = (x - \mu)^T \Lambda (x - \mu), \quad E[x] = \mu.$$

10.1.6 ウィシャート分布

$$\mathcal{W}(\Lambda, W, \nu) = B(W, \nu) |\Lambda|^{\frac{\nu-D-1}{2}} \exp \left(-\frac{1}{2} \text{tr}(W^{-1}\Lambda) \right).$$

$$B(W, \nu) = |W|^{\nu/2} \left(2^{\nu D/2} \pi^{D(D-1)/4} \prod_{i=1}^D \Gamma \left(\frac{\nu+1-i}{2} \right) \right)^{-1}.$$

$$E[\Lambda] = \nu W.$$

$$E[\log |\Lambda|] = \sum_{i=1}^D \phi \left(\frac{\nu+1-i}{2} \right) + D \log 2 + \log |W|.$$

$$H[\Lambda] = -\log B(W, \nu) - \frac{\nu-D-1}{2} E[\log |\Lambda|] + \frac{\nu D}{2}.$$

10.1.7 行列の公式

$$\begin{aligned}
x^T A x &= \text{tr}(A x x^T). \\
\frac{\partial}{\partial A} \log |A| &= (A^{-1})^T. \\
\frac{\partial}{\partial x} \log |A| &= \text{tr} \left(A^{-1} \frac{\partial}{\partial x} A \right). \\
\frac{\partial}{\partial A} \text{tr}(A^{-1} B) &= -(A^{-1} B A^{-1})^T. \\
|I + a b^T| &= 1 + a^T b.
\end{aligned}$$

10.1.8 カルバック距離

$$\text{KL}(q||p) = - \int q(Z) \log \frac{p(Z|X)}{q(Z)} dZ \geq 0.$$

10.2 下限と下界

下界 (lower bound) と下限 (inf, greatest lower bound) は意味が違う。できるだけ使い分けた方がよいだろう。一般には順序集合に対して定義するがここでは実数の話に制限する。 A が \mathbb{R} の部分集合とする。ある x に対して x が A の下界であるとは、全ての $a \in A$ に対して $x \leq a$ であることをいう。つまり x は A の中の一番小さい値よりも同じかより小さいときのことをいう。

たとえば $A = \{x | x \geq 0\}$ のとき、 -3 は A の下界である。 -5 も A の下界である。普通下界となる値はたくさんあるので主語と述語を入れ換えた「 A の下界は -3 である」という言い方はあまりしないと思う。これは「 $x^2 = 1$ の解は 1 である」とは言わないのと同じ感覚である（ -1 はどうしたの? と聞かれるだろう）。

たくさんある下界の中で一番大きい値を下限という。下限は存在すればただ一つである。上記 A の下限は 0 。一つしかないので「 A の下限は 0 である」ともいうし、「 0 は A の下限である」ともいう。

たとえば PRML 上巻 (4 刷) p.49 の一番下では「確率変数の状態を送るために必要なビット数の下限がエントロピーである」とありこれは正しい。あるいは「エントロピーは確率変数の状態を送るために必要なビット数の下限である」でもよい。しかし、これを「エントロピーは確率変数の状態を送るために必要なビット数の下界である」としてしまうと (2011/7/27 時点での日本語サポートの正誤表)、間違っていないが上の文章と全

然意味が違ってしまふ．これではエントロピーがぎりぎりの値であるという主張が消えたしょうもないものになっている．

10.3 分解による近似の持つ性質

ここで Λ_{ij} はスカラーで $\Lambda_{12} = \Lambda_{21}$. $E[z_1] = m_1$, $E[z_2] = m_2$ より

$$\begin{aligned} m_1 &= \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (m_2 - \mu_2) = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (\mu_2 - \Lambda_{22}^{-1} \Lambda_{21} (m_1 - \mu_1) - \mu_2) \\ &= \mu_1 + \Lambda_{11}^{-1} \Lambda_{22}^{-1} \Lambda_{12}^2 (m_1 - \mu_1). \end{aligned}$$

よって

$$(m_1 - \mu_1)(\Lambda_{11}^{-1} \Lambda_{22}^{-1} \Lambda_{12}^2 - 1) = 0.$$

分布が非特異なら $|\Lambda| = \Lambda_{11} \Lambda_{22} - \Lambda_{12}^2 \neq 0$ より $m_1 = \mu_1$. 同様に $m_2 = \mu_2$.

変数 $Z = \{z_1, \dots, z_N\}$ に対する分布 $q(Z)$ が

$$q(Z) = \prod_{i=1}^M q_i(Z_i)$$

と複数のグループの関数の積としてかけていると仮定する．ここで $\{Z_i\}$ は Z の disjoint-union である．(PRML p.182) $\text{KL}(p||q)$ を Z_j について最小化する問題を考える（以下、対象変数以外の項をまとめて C と略記する）．

$$\begin{aligned} \text{KL}(p||q) &= - \int p(Z) \left(\sum_i \log q_i(Z_i) \right) dZ + C \\ &= - \int \left(p(Z) \log q_j(Z_j) + p(Z) \sum_{i \neq j} \log q_i(Z_i) \right) dZ + C \\ &= - \int p(Z) \log q_j(Z_j) dZ + C = - \int \log q_j(Z_j) \underbrace{\left(\int p(Z) \prod_{i \neq j} dZ_i \right)}_{=: F_j(Z_j)} dZ_j \\ &= - \int F_j(Z_j) \log q_j(Z_j) dZ_j. \end{aligned}$$

$$\int q_j(Z_j) dZ_j = 1$$

の条件の元で

$$X = - \int F_j(Z_j) \log q_j(Z_j) dZ_j + \lambda \left(\int q_j(Z_j) dZ_j - 1 \right)$$

を最小化する.

$$\begin{aligned} \frac{\partial}{\partial q_j} X &= - \int F_j(Z_j) \log(q_j + \delta q_j) dZ_j + \lambda \left(\int (q_j + \delta q_j) dZ_j - 1 \right) \\ &= \left(- \int F_j(Z_j) \log q_j dZ_j + \lambda \left(\int q_j dZ_j - 1 \right) \right) - \left(\int F_j(Z_j)/q_j dZ_j - \lambda \right) \delta q_j = 0. \end{aligned}$$

$F_j/q_j - \lambda = 0$. よって $F_j = \lambda q_j$. 積分して

$$\int F_j dZ_j = \int \lambda q_j dZ_j = \lambda = 1.$$

よって

$$q_j^*(Z_j) = q_j = F_j = \int p(Z) \prod_{i \neq j} dZ_i.$$

10.4 α ダイバージェンス

α を実数として

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2} \left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2} dx \right)$$

を α ダイバージェンスという. $\alpha \rightarrow 1$ のとき $\text{KL}(p||q)$, $\alpha \rightarrow -1$ のとき $\text{KL}(q||p)$ になる.

(証明) $\alpha = 1 - 2\epsilon$ と置く. $\alpha \rightarrow 1$ で $\epsilon \rightarrow 0$ となる.

$$(q/p)^\epsilon = \exp(\epsilon \log(q/p)) \approx 1 + \epsilon \log(q/p)$$

より

$$\begin{aligned} D_\alpha(p||q) &= \frac{1}{\epsilon(1-\epsilon)} \left(1 - \int p(q/p)^\epsilon dx \right) \simeq \frac{1}{\epsilon} \left(1 - \int p \left(1 + \epsilon \log \frac{q}{p} \right) dx \right) \\ &= \frac{1}{\epsilon} \left(-\epsilon \int p \log \frac{q}{p} dx \right) = \text{KL}(p||q). \end{aligned}$$

$\alpha \rightarrow -1$ も同様.

10.5 例：一変数ガウス分布

ガウス分布から独立に発生した観測値 x のデータセットを $\mathcal{D} = \{x_1, \dots, x_N\}$ とする.
もとのガウス分布の平均 μ と精度 τ の事後分布をもとめる.

$$p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi} \right)^{N/2} \exp \left(-\frac{\tau}{2} \sum_n (x_n - \mu)^2 \right).$$

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}), \quad p(\tau) = \text{Gam}(\tau|a_0, b_0).$$

この問題は厳密にもとめられるがここでは事後分布が次のように分解できると仮定したときの変分近似を考える.

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau).$$

まず μ について

$$\begin{aligned} \log q_\mu^*(\mu) &= E_\tau[\log p(D, \mu, \tau)] = E_\tau[\log p(D|\mu, \tau) + \log p(\mu|\tau)] + (\mu \text{ に依存しない部分 } C) \\ &= \frac{E[\tau]}{2} \left(\sum_n (x_n - \mu)^2 \right) + E_\tau \left[-\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 \right] + C \\ &= -(E[\tau]/2) \left(\lambda_0 (\mu - \mu_0)^2 + \sum_n (x_n - \mu)^2 \right) + C. \end{aligned}$$

μ について平方完成すると

$$\begin{aligned} &-(E[\tau]/2) \left((\lambda_0 + N)\mu^2 - 2\mu \left(\lambda_0 \mu_0 + \sum_n x_n \right) + \lambda_0 \mu_0^2 + \sum_n x_n^2 \right) + C \\ &\sum_n x_n = N\bar{x} \text{ より} \\ &= -\frac{E[\tau](\lambda_0 + N)}{2} \left(\mu - \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N} \right)^2 + \dots \end{aligned}$$

よってこの分布はガウス分布であることが分かり,

$$\mu_N = \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N}, \quad \lambda_N = (\lambda_0 + N)E[\tau]$$

と置くと $\mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$ となることが分かる. $N \rightarrow \infty$ のとき $\mu_N \rightarrow \bar{x}$ で分散は 0 (精度は ∞). τ について

$$\begin{aligned} \log q_\tau^*(\tau) &= E_\mu[\log p(D, \tau|\mu)] = E_\mu[\log p(D|\mu, \tau) + \log p(\mu|\tau)] + \log p(\tau) \\ &= E_\mu \left[(N/2) \log \tau - (\tau/2) \sum_n (x_n - \mu)^2 \right] + E_\mu \left[(1/2) \log(\lambda_0 \tau) - (\lambda_0 \pi/2) (\mu - \mu_0)^2 \right] \\ &\quad + E_\mu [(a_0 - 1) \log \tau - b_0 \tau - \log \Gamma(a_0) + a_0 \log b_0] + C \\ &= (a_0 - 1) \log \tau - b_0 \tau + (N + 1)/2 \log \tau \\ &\quad - (\tau/2) E_\mu \left[\sum_n (x_n - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right] + C. \end{aligned}$$

よって $q_\tau(\tau)$ はガンマ分布となり

$$a_N = a_0 + \frac{N + 1}{2}, \quad b_N = b_0 + \frac{1}{2} E_\mu \left[\sum_n (x_n - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right]$$

と置くと $q_\tau(\tau) = \text{Gam}(\tau|a_N, b_N)$. $q_\mu(\mu)$, $q_\tau(\tau)$ の関数の形に何も仮定を置いていないのに, 尤度関数と共役事前分布の構造から決まったことに注意. $N \rightarrow \infty$ で

$$E[\text{Gam}(\tau|a_N, b_N)] = \frac{a_N}{b_N} \rightarrow 1/E_\mu \left[(1/N \sum_n (x_n - \mu)^2) \right] \rightarrow 1/\text{分散}.$$

$$\sigma[\text{Gam}] = a_N/b_N^2 \rightarrow 0.$$

$\mu_0 = a_0 = b_0 = \lambda_0 = 0$ という無情報事前分布を入れてみる.

$$a_N = \frac{N+1}{2}, \quad b_N = \frac{1}{2} E_\mu \left[\sum_n (x_n - \mu)^2 \right].$$

よって

$$E[\tau]^{-1} = \frac{b_N}{a_N} = E \left[\frac{1}{N+1} \sum_n (x_n - \mu)^2 \right] = \frac{N}{N+1} (\bar{x}^2 - 2\bar{x}E[\mu] + E[\mu^2]).$$

$$\mu_N = \frac{0 + N\bar{x}}{0 + N} = \bar{x}, \quad \lambda_N = NE[\tau].$$

よって

$$E[\mu] = \bar{x}, \quad E[\mu^2] = E[\mu]^2 + \frac{1}{\lambda_N} = \bar{x}^2 + \frac{1}{NE[\tau]}.$$

$$\frac{1}{E[\tau]} = \frac{N}{N+1} \left(\bar{x}^2 - 2\bar{x}^2 + \bar{x}^2 + \frac{1}{NE[\tau]} \right).$$

$$\frac{N}{N+1} (\bar{x}^2 - \bar{x}^2) = \frac{1}{E[\tau]} - \frac{1}{N+1} \frac{1}{E[\tau]} = \frac{N}{N+1} \frac{1}{E[\tau]}.$$

よって

$$\frac{1}{E[\tau]} = \bar{x}^2 - \bar{x}^2 = \frac{1}{N} \sum_n (x_n - \bar{x})^2.$$

10.6 モデル比較

事前確率 $p(m)$ を持つ複数のモデルの比較. 観測データ X の下で事後確率 $p(m|X)$ を近似したい.

$$q(Z, m) = q(Z|m)q(m), \quad p(X, Z, m) = p(X)p(Z, m|X).$$

$\sum_{m,Z} q(Z, m) = 1$ に注意して

$$\begin{aligned} \log p(X) &= \sum_{m,Z} q(Z, m) \log p(X) = \sum_{m,Z} q(Z, m) \log \frac{p(X, Z, m)}{q(Z, m)} \frac{q(Z, m)}{p(Z, m|X)} \\ &= \underbrace{\left(\sum_{m,Z} q(Z, m) \log \frac{p(X, Z, m)}{q(Z, m)} \right)}_{:=\mathcal{L}} + \left(\sum_{m,Z} q(Z, m) \log \frac{q(Z, m)}{p(Z, m|X)} \right) \\ &= \mathcal{L} - \sum_{m,Z} q(Z, m) \log \frac{q(Z, m|X)}{q(Z, m)} = \mathcal{L} - \sum_{m,Z} q(Z|m)q(m) \log \frac{p(Z, m|X)}{q(Z|m)q(m)}. \end{aligned}$$

\mathcal{L} を $q(m)$ について最大化する. $\sum_Z q(Z|m) = 1$ に注意して

$$\begin{aligned} \mathcal{L} &= \sum_{m,Z} q(Z|m)q(m) (\log p(Z, X|m) + \log p(m) - \log q(Z|m) - \log q(m)) \\ &= \sum_m q(m) \left((\log p(m) - \log q(m)) + \underbrace{\sum_Z q(Z|m) \log \frac{p(Z, X|m)}{q(Z|m)}}_{:=\mathcal{L}_m} \right), \\ &= \sum_m q(m) \log \frac{p(m) \exp \mathcal{L}_m}{q(m)} = - \sum_m \text{KL}(q(m) || p(m) \exp \mathcal{L}_m). \end{aligned}$$

よって \mathcal{L} が最大となるのはカルバック距離が最小となるとき, つまり $q(m) \propto p(m) \exp \mathcal{L}_m$ のときである ($p(m) \exp \mathcal{L}_m$ は正規化されているとは限らないので $=$ ではなく \propto).

10.6.1 変分混合ガウス分布

ガウス混合モデルに変分推論法を適用してみる. x_n に対応する潜在変数 z_n . z_n は K 個の要素 z_{nk} からなる. $z_{nk} = 0$ または 1 で $\sum_k z_{nk} = 1$.

$X = \{x_1, \dots, x_N\}$, $Z = \{z_1, \dots, z_N\}$, 混合比は $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$.

$$p(z_n) = \prod_k \pi_k^{z_{nk}}, \quad p(x_n|z_n) = \prod_k \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}, \quad p(X|Z, \mu, \Lambda) = \prod_{n,k} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}.$$

$\boldsymbol{\pi}$ の事前分布はディリクレ分布とする.

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\alpha_0) = C(\alpha_0) \prod_k \pi_k^{\alpha_0-1}.$$

混合要素の持つガウス分布の事前分布はガウス・ウィシャート分布とする.

$$p(\mu, \Lambda) = p(\mu|\Lambda)p(\Lambda) = \prod_k \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k, |W_0, \nu_0).$$

10.6.2 変分事後分布

$$p(X, Z, \boldsymbol{\pi}, \mu, \Lambda) = p(X|Z, \mu, \Lambda)p(z|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\mu|\Lambda)p(\Lambda).$$

$q(Z, \boldsymbol{\pi}, \mu, \Lambda) = q(Z)q(\boldsymbol{\pi}, \mu, \Lambda)$ という変分近似を考える.

Z について (以後対象としている変数以外の項は無視する)

$$\begin{aligned} & \log q^*(Z) \\ &= E_{\boldsymbol{\pi}, \mu, \Lambda}[\log p(X, Z, \boldsymbol{\pi}, \mu, \Lambda)] = E_{\boldsymbol{\pi}}[\log p(Z|\boldsymbol{\pi})] + E_{\mu, \Lambda}[\log p(X|Z, \mu, \Lambda)] \\ &= \sum_{n,k} z_{nk} E_{\boldsymbol{\pi}}[\log \pi_k] + \sum_{n,k} z_{nk} E_{\mu, \Lambda} \left[\frac{1}{2} \log |\Lambda_k| - \frac{1}{2} (x_n - \mu_n)^T \Lambda_k (x_n - \mu_n) - \frac{D}{2} \log(2\pi) \right] \\ &= \sum_{n,k} z_{nk} \underbrace{\left(E_{\boldsymbol{\pi}}[\log \pi_k] + \frac{1}{2} E[\log |\Lambda_k|] - \frac{D}{2} \log(2\pi) - \frac{1}{2} E_{\mu_k, \Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] \right)}_{=:\log \rho_{nk}} \\ &= \sum_{n,k} z_{nk} \log \rho_{nk}. \end{aligned}$$

よって

$$q^*(Z) \propto \prod_{n,k} \rho_{nk}^{z_{nk}}.$$

Z について総和をとると $\sum_k z_{nk} = 1$ より

$$\sum_Z \prod_{n,k} \rho_{nk}^{z_{nk}} = \prod_n \left(\sum_k \rho_{nk} \right).$$

よって

$$r_{nk} = \rho_{nk} / \left(\sum_k \rho_{nk} \right)$$

と置くと

$$q^*(Z) = \prod_{n,k} r_{nk}^{z_{nk}}$$

とできる. $q(Z)$ の最適解は事前分布 $p(Z|\boldsymbol{\pi})$ と同じ形. $\rho_{nk} = e^?$ の形なので $\rho_{nk} \geq 0$. つ

まり $r_{nk} \geq 0$. 各 n について $\sum_k r_{nk} = 1$

$$E[z_{nk}] = r_{nk}.$$

次の値を定義する:

$$N_k = \sum_n r_{nk}, \quad \bar{x}_k = \frac{1}{N_k} \sum_n r_{nk} x_n, \quad S_k = \frac{1}{N_k} \sum_n r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T.$$

$q(\boldsymbol{\pi}, \mu, \Lambda)$ について考える.

$$\begin{aligned}\log q^*(\boldsymbol{\pi}, \mu, \Lambda) &= E_Z [\log p(X, Z, \boldsymbol{\pi}, \mu, \Lambda)] \\ &= \log p(\boldsymbol{\pi}) + \sum_k \log p(\mu_k, \Lambda_k) \\ &\quad + E_Z [\log p(Z|\boldsymbol{\pi})] + \sum_{n,k} E[z_{nk}] \log \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1}).\end{aligned}$$

この式は $\boldsymbol{\pi}$ だけを含む項とそれ以外の項に分かれている. 更に μ_k, Λ_k の積にもなっている. つまり $q(\boldsymbol{\pi}, \mu, \Lambda) = q(\boldsymbol{\pi}) \prod_k q(\mu_k, \Lambda_k)$ という形になっている. $\boldsymbol{\pi}$ に依存する部分を見る.

$$\begin{aligned}\log q^*(\boldsymbol{\pi}) &= \log \text{Dir}(\boldsymbol{\pi} | \alpha_0) + E_Z \left[\sum_{n,k} z_{nk} \log \pi_k \right] \\ &= (\alpha_0 - 1) \sum_k \log \pi_k + \sum_{n,k} r_{nk} \log \pi_k = \sum_k \left(\alpha_0 - 1 + \sum_n r_{nk} \right) \log \pi_k.\end{aligned}$$

よって $q^*(\boldsymbol{\pi})$ はディリクレ分布となる. その係数は $\alpha_k = \alpha_0 + N_k$ において $\alpha = (\alpha_k)$ とすると $q^*(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} | \alpha)$. $q^*(\mu_k, \Lambda_k) = q^*(\mu_k | \Lambda_k) q^*(\Lambda_k)$ を考える. まず

$$\begin{aligned}\log q^*(\mu_k, \Lambda_k) &= \log \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) + \log \mathcal{W}(\Lambda_k | W_0, \nu_0) + \sum_n r_{nk} \log \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1}) \\ &\quad (\mu_k, \Lambda_k \text{ の依存部分だけとりだして}) \\ &= \frac{1}{2} \log |\beta_0 \Lambda_k| - \frac{1}{2} (\mu_k - m_0)^T \beta_0 \Lambda_k (\mu_k - m_0) + \frac{\nu_0 - D - 1}{2} \log |\Lambda_k| \\ &\quad - \frac{1}{2} \text{tr}(W_0^{-1} \Lambda_k) + \sum_n r_{nk} \left(\frac{1}{2} \log |\Lambda_k| - \frac{1}{2} (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right).\end{aligned}$$

このうち更に μ_k に依存する部分を見る:

$$\begin{aligned}\log q^*(\mu_k | \Lambda_k) &= -\frac{1}{2} \mu_k^T \left(\beta_0 + \sum_n r_{nk} \right) \Lambda_k \mu_k + \mu_k^T \Lambda_k \left(\beta_0 m_0 + \sum_n r_{nk} x_n \right) \\ &\quad (\beta_k := \beta_0 + N_k, \quad m_k := \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) \text{ と置くと}) \\ &= -\frac{1}{2} \mu_k^T (\beta_k \Lambda_k) \mu_k + \mu_k^T (\beta_k \Lambda_k) m_k.\end{aligned}$$

よって

$$q^*(\mu_k | \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}).$$

残りを考える.

$$\begin{aligned}
\log q^*(\Lambda_k) &= \log q^*(\mu_k, \Lambda_k) - \log q^*(\mu_k | \Lambda_k) \\
&= \frac{1}{2} \log |\beta_0 \Lambda_k| - \frac{1}{2} (\mu_k - m_0)^T (\beta_0 \Lambda_k) (\mu_k - m_0) + \frac{\nu_0 - D - 1}{2} \log |\Lambda_k| \\
&\quad - \frac{1}{2} \text{tr}(W_0^{-1} \Lambda_k) + \frac{1}{2} N_k \log |\Lambda_k| - \frac{1}{2} \sum_n r_{nk} (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \\
&\quad - \frac{1}{2} \log |\beta_0 \Lambda_k| + \frac{1}{2} (\mu_k - m_k)^T (\beta_0 \Lambda_k) (\mu_k - m_k) \\
&= \frac{(\nu_0 + N_k) - D - 1}{2} \log |\Lambda_k| - \frac{1}{2} \text{tr} \left((\beta_0 \Lambda_k) (\mu_k - m_0) (\mu_k - m_0)^T \right. \\
&\quad \left. + \sum_n r_{nk} \Lambda_k (x_n - \mu_k) (x_n - \mu_k)^T - \beta_k \Lambda_k (\mu_k - m_k) (\mu_k - m_k)^T \right) - \frac{1}{2} (\Lambda_k W_0^{-1}) \\
&\quad (v_k := \nu_0 + N_k \text{ と置く}) \\
&= \frac{\nu_k - D - 1}{2} \log |\Lambda_k| - \frac{1}{2} \text{tr} \left(\Lambda_k (W_0^{-1} + \beta_0 (\mu_k - m_0) (\mu_k - m_0)^T \right. \\
&\quad \left. + \sum_n r_{nk} (x_n - \mu_k) (x_n - \mu_k)^T - \beta_k (\mu_k - m_k) (\mu_k - m_k)^T \right) \\
&= \frac{\nu_k - D - 1}{2} \log |\Lambda_k| - \frac{1}{2} \text{tr}(\Lambda_k W_k^{-1}) \text{ と置く.}
\end{aligned}$$

W_k を求めよう. まず

$$\begin{aligned}
\sum_n r_{nk} x_n x_n^T &= \sum_n r_{nk} \left((x_n - \bar{x}_k)(x_n - \bar{x}_k)^T + 2x_n \bar{x}_k - \bar{x}_k \bar{x}_k^T \right) \\
&= N_k S_k + 2N_k \bar{x}_k + 2N_k \bar{x}_k \bar{x}_k^T - N_k \bar{x}_k \bar{x}_k^T \\
&= N_k S_k + N_k \bar{x}_k \bar{x}_k^T.
\end{aligned} \tag{10.1}$$

よって

$$\begin{aligned}
W_k^{-1} &= W_0^{-1} + \beta_0 (\mu_k \mu_k^T - 2\mu_k m_0^T + m_0 m_0^T) + N_k S_k + N_k \bar{x}_k \bar{x}_k^T - 2 \sum_n r_{nk} x_n \mu_k^T \\
&\quad + \sum_n r_{nk} \mu_k \mu_k^T - (\beta_0 + N_k) \left(\mu_k \mu_k^T - 2\mu_k \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k)^T \right. \\
&\quad \left. + \frac{1}{\beta_k^2} (\beta_0 m_0 + N_k \bar{x}_k) (\beta_0 m_0 + N_k \bar{x}_k)^T \right) \\
&\quad \left(\sum_n r_{nk} = N_k \text{ に注意して} \right) \\
&= W_0^{-1} + N_k S_k + \beta_0 m_0 m_0^T + N_k \bar{x}_k \bar{x}_k^T \\
&\quad - \frac{1}{\beta_k} (\beta_0^2 m_0 m_0^T + 2\beta_0 N_k m_0 \bar{x}_k^T + N_k^2 \bar{x}_k \bar{x}_k^T) \\
&= W_0^{-1} + N_k S_k + \frac{(\beta_0 + N_k) \beta_0 - \beta_0^2}{\beta_k} m_0 m_0^T + \frac{(\beta_0 + N_k) N_k - N_k^2}{\beta_k} \bar{x}_k \bar{x}_k^T - \frac{2\beta_0 N_k}{\beta_k} m_0 \bar{x}_k^T \\
&= W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_k} (m_0 - \bar{x}_k) (m_0 - \bar{x}_k)^T.
\end{aligned}$$

よって

$$q^*(\Lambda_k) = \mathcal{W}(\Lambda_k | W_k, \nu_k), \quad q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k).$$

$\mathcal{N}(x | \mu, \Lambda^{-1})$ について $E[xx^T] = \mu\mu^T + \Lambda^{-1}$, $\mathcal{W}(\Lambda_k | W_k, \nu_k)$ について $E[\Lambda_k] = \nu_k W_k$ なので

$$\begin{aligned} & E_{\mu_k, \Lambda_k} [(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] \\ &= \text{tr} (E [\Lambda_k x_n x_n^T] - 2E [\Lambda_k x_n \mu_k^T] + E [\Lambda_k \mu_k \mu_k^T]) \\ &= \text{tr} E [\nu_k W_k x_n x_n^T] - 2 \text{tr} E [\nu_k W_k x_n m_k^T] + \text{tr} E [\Lambda_k (m_k m_k^T + (\beta_k \Lambda_k)^{-1})] \\ &= \nu_k \text{tr} (W_k x_n x_n^T) - 2\nu_k \text{tr} (W_k x_n m_k^T) + \text{tr} (\nu_k W_k m_k m_k^T) + D\beta_k^{-1} \\ &= D\beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k). \end{aligned} \quad (10.2)$$

ウィンシャート分布の公式から

$$\log \tilde{\Lambda}_k := E[\log |\Lambda_k|] = \sum_i \phi \left(\frac{\nu_k + 1 - i}{2} \right) + D \log 2 + \log |W_k|.$$

ディリクレ分布の公式から

$$\log \tilde{\pi}_k := E[\log \pi_k] = \phi(\alpha_k) - \phi(\hat{\alpha}), \quad \hat{\alpha} = \sum_k \alpha_k.$$

$$\begin{aligned} \log \rho_{nk} &= E[\log \pi_k] + \frac{1}{2} E[\log |\Lambda_k|] - \frac{D}{2} \log(2\pi) - \frac{1}{2} E[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)] \\ &= \log \tilde{\pi}_k + \frac{1}{2} \log \tilde{\Lambda}_k - \frac{D}{2} \log(2\pi) - \frac{1}{2} (D\beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k)). \end{aligned}$$

よって

$$r_{nk} \propto \rho_{nk} \propto \tilde{\pi}_k \Lambda_k^{1/2} \exp \left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2} (x_n - m_k)^T W_k (x_n - m_k) \right).$$

混合ガウス分布の EM アルゴリズムでの負担率は $\gamma(z_{nk}) \propto \pi_k \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})$ だったので

$$r_{nk} \propto \pi_k |\lambda_k|^{1/2} \exp \left(-\frac{1}{2} (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right).$$

これは上の式とよく似ている.

ディリクレ分布の平均の式 $E[\mu_k] = \alpha_k / \hat{\alpha}$ より

$$E[\pi_k] = \frac{\alpha_0 + N_k}{\sum_k \alpha_k} = \frac{\alpha_0 + N_k}{K\alpha_0 + \sum_k N_k} = \frac{\alpha_0 + N_k}{K\alpha_0 + N}.$$

ある混合要素 k について $r_{nk} \simeq 0$ なら $N_k \simeq 0$ (PRML p.193 はかつとなってるけど片方の条件から出る). このとき $\alpha_k \simeq \alpha_0$ となる. PRML10 章では分布が幅広いという状態を「なだらか」と表記しているようだ. ちょっとニュアンスが違う気もするけど.

事前分布で $\alpha_0 \rightarrow 0$ とすると $E[\pi_k] \rightarrow 0$. $\alpha_0 \rightarrow \infty$ なら $E[\pi_k] \rightarrow 1/K$.

10.7 変分下限

PRML 下巻 (3 刷) ではこの章は変分下限である。しかしここで計算する \mathcal{L} の値は $\log p(X)$ の下界の中で一番大きいものになるとは限らない。KL($q||p$) は一般に 0 ではない。その意味で 10.2 節で述べたように変分下界が適切と思われる。

それはさておき、 $q(Z, \boldsymbol{\pi}, \mu, \Lambda) = q(Z)q(\boldsymbol{\pi}, \mu, \Lambda)$ と分解できると仮定すると

$$\begin{aligned}\mathcal{L}(q) &= \int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ \\ &= \sum \int q(Z, \boldsymbol{\pi}, \mu, \Lambda) \log \frac{p(X, Z, \boldsymbol{\pi}, \mu, \Lambda)}{q(Z, \boldsymbol{\pi}, \mu, \Lambda)} d\boldsymbol{\pi} d\mu d\Lambda \\ &= E[\log p(X, Z, \boldsymbol{\pi}, \mu, \Lambda)] - E[\log q(Z, \boldsymbol{\pi}, \mu, \Lambda)] \\ &= E[\log p(X|Z, \mu, \Lambda)] + E[\log p(Z|\boldsymbol{\pi})] + E[\log p(\boldsymbol{\pi})] + E[\log p(\mu, \Lambda)] \\ &\quad - E[\log q(Z)] - E[\log q(\boldsymbol{\pi})] - E[\log q(\mu, \Lambda)].\end{aligned}$$

以下、ひたすら計算する。

$$\begin{aligned}E[\log p(X|Z, \mu, \Lambda)] &= E \left[\sum_{n,k} z_{nk} \log \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1}) \right] \\ &= \frac{1}{2} E \left[\sum_{n,k} z_{nk} \left(-D \log(2\pi) + \log |\Lambda_k| - (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right) \right] \\ &= \frac{1}{2} \sum_k E \left[-N_k D \log(2\pi) + N_k \log |\Lambda_k| - \sum_n z_{nk} (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right] \\ &= \frac{1}{2} \sum_k N_k \left(\log \tilde{\Lambda}_k - D \log(2\pi) \right) - \underbrace{\frac{1}{2} \sum_{n,k} r_{nk} \left(D \beta_k^{-1} + \nu_k (x_n - m_k)^T W_k (x_n - m_k) \right)}_{=: X}.\end{aligned}$$

$$X = \sum_k N_k D \beta_k^{-1} + \sum_k \nu_k \left(\underbrace{\sum_n r_{nk} (x_n - m_k)^T W_k (x_n - m_k)}_{=: Y} \right).$$

式 (10.1) より

$$\sum_n r_{nk} x_n x_n^T = N_k S_k + N_k \bar{x}_k \bar{x}_k^T.$$

よって

$$\begin{aligned}
Y &= \text{tr} \left(W_k \left(\sum_n r_{nk} x_n x_n^T - 2 \sum_n r_{nk} x_n m_k^T + \sum_n r_{nk} m_k m_k^T \right) \right) \\
&= \text{tr} (W_k (N_k S_k + N_k \bar{x}_k \bar{x}_k^T - 2 N_k \bar{x}_k m_k^T + N_k m_k m_k^T)) \\
&= N_k \text{tr} \left(W_k \left(S_k + (\bar{x}_k - m_k)(\bar{x}_k - m_k)^T \right) \right) \\
&= N_k \left(\text{tr}(S_k W_k) + (\bar{x}_k - m_k)^T W_k (\bar{x}_k - m_k) \right).
\end{aligned}$$

よって

$$\begin{aligned}
E[\log p(X|Z, \mu, \Lambda)] &= \frac{1}{2} \sum_k N_k \left(\log \tilde{\Lambda}_k - D \beta_k^{-1} - \nu_k \text{tr}(S_k W_k) \right. \\
&\quad \left. - \nu_k (\bar{x}_k - m_k)^T W_k (\bar{x}_k - m_k) - D \log(2\pi) \right).
\end{aligned}$$

$$E[\log p(Z|\boldsymbol{\pi})] = E \left[\sum_{n,k} z_{nk} \log \pi_k \right] = \sum_{n,k} r_{nk} \log \tilde{\pi}_k.$$

$$E[\log p(\boldsymbol{\pi})] = E \left[\log C(\alpha_0) + \sum_k (\alpha_0 - 1) \log \pi_k \right] = \log C(\alpha_0) + (\alpha_0 - 1) \sum_k \log \tilde{\pi}_k.$$

$$E[\log q(Z)] = E \left[\sum_{n,k} z_{nk} \log r_{nk} \right] = \sum_{n,k} \log r_{nk}.$$

$$E[\log q(\boldsymbol{\pi})] = E \left[\log C(\alpha) + \sum_k (\alpha_k - 1) \log \pi_k \right] = \log C(\alpha) + \sum_k (\alpha_k - 1) \log \tilde{\pi}_k.$$

$$\begin{aligned}
&E[\log q(\mu, \Lambda)] \\
&= \sum_k E \left[\log \mathcal{N}(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) + \log \mathcal{W}(\lambda_k | W_k, \nu_k) \right] \\
&= \sum_k E \left[-\frac{D}{2} \log(2\pi) + \frac{1}{2} \log |\beta_k \Lambda_k| - \frac{1}{2} (\mu_k - m_k)^T (\beta_k \Lambda_k) (\mu_k - m_k) \right] + E[\log W] \\
&= \sum_k \frac{1}{2} \log \tilde{\Lambda}_k + \frac{D}{2} \log \left(\frac{\beta_k}{2\pi} \right) - \frac{1}{2} \underbrace{\text{tr} E \left[(\beta_k \Lambda_k) (\mu_k - m_k) (\mu_k - m_k)^T \right]}_{=:X} + \underbrace{E[\log W]}_{=:Y}.
\end{aligned}$$

$$\begin{aligned}
X &= \text{tr}(\beta_k \Lambda_k) (E[\mu_k \mu_k^T] - 2E[\mu_k] m_k^T + m_k m_k^T) \\
&= \text{tr}(\beta_k \Lambda_k) (m_k m_k^T + (\beta_k \Lambda_k)^{-1} - m_k m_k^T) = \text{tr} I = D.
\end{aligned}$$

$$\begin{aligned}
Y &= E[\log W(\Lambda_k | W_k, \nu_k)] \\
&= \log B(W_k, \nu_k) + \frac{\nu_k - D - 1}{2} E[\log |\Lambda_k|] - \frac{1}{2} \text{tr}(W_k^{-1} E[\Lambda_k]) \\
&= \log B(W_k, \nu_k) + \frac{\nu_k - D - 1}{2} E[\log |\Lambda_k|] - \frac{1}{2} \nu_k D = -H[\Lambda_k].
\end{aligned}$$

よって

$$E[\log q(\mu, \Lambda)] = \sum_k \left(\frac{1}{2} \log \tilde{\Lambda}_k + \frac{D}{2} \log \left(\frac{\beta_k}{2\pi} \right) - \frac{D}{2} - H[\Lambda_k] \right).$$

$$E[\log p(\mu, \Lambda)] = \sum_k \underbrace{E[\log \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1})]}_{=:A} + \underbrace{E[\log \mathcal{W}(\Lambda_k | W_0, \nu_0)]}_{=:B}.$$

$$\begin{aligned}
A &= E \left[-\frac{D}{2} \log(2\pi) + \frac{1}{2} \log |\beta_0 \Lambda_k| - \frac{1}{2} (\mu_k - m_0)^T (\beta_0 \Lambda_k) (\mu_k - m_0) \right] \\
&= \frac{D}{2} \log \left(\frac{\beta_0}{2\pi} \right) + \frac{1}{2} \log \tilde{\Lambda}_k - \frac{1}{2} \beta_0 E[(m_0 - \mu_k)^T \Lambda_k (m_0 - \mu_k)] \\
&\quad (\text{数ページ前の式 (10.2) で } x_n = m_0 \text{ として使うと}) \\
&= \frac{1}{2} \left(D \log \left(\frac{\beta_0}{2\pi} \right) + \log \tilde{\Lambda}_k - \beta_0 \left(D \beta_k^{-1} + \nu_k (m_0 - \mu_k)^T W_k (m_0 - \mu_k) \right) \right) \\
&= \frac{1}{2} \left(D \log \left(\frac{\beta_0}{2\pi} \right) + \log \tilde{\Lambda}_k - \frac{D \beta_0}{\beta_k} - \beta_0 \nu_k (m_k - m_0)^T W_k (m_k - m_0) \right).
\end{aligned}$$

$$\begin{aligned}
B &= E \left[\log B(W_0, \nu_0) + \frac{\nu_0 - D - 1}{2} \log |\Lambda_k| - \frac{1}{2} \text{tr}(W_0^{-1} \Lambda_k) \right] \\
&= \log B(W_0, \nu_0) + \frac{\nu_0 - D - 1}{2} \log \tilde{\Lambda}_k - \frac{1}{2} \text{tr}(W_0^{-1} \underbrace{E[\Lambda_k]}_{=: \nu_k W_k}).
\end{aligned}$$

よって

$$\begin{aligned}
E[\log p(\mu, \Lambda)] &= \frac{1}{2} \sum_k \left(D \log \left(\frac{\beta_0}{2\pi} \right) + \log \tilde{\Lambda}_k - \frac{D \beta_0}{\beta_k} - \beta_0 \nu_k (m_k - m_0)^T W_k (m_k - m_0) \right) \\
&\quad + K \log B(W_0, \nu_0) + \frac{\nu_0 - D - 1}{2} \sum_k \log \tilde{\Lambda}_k - \frac{1}{2} \sum_k \nu_k \text{tr}(W_0^{-1} W_k).
\end{aligned}$$

最後に \mathcal{L} を求めよう.

$$\sum_k N_k = N,$$

$$H[q(\Lambda_k)] = -\log B(W_k, \nu_k) - \frac{\nu_k - D - 1}{2} \log \tilde{\Lambda}_k + \frac{\nu_k D}{2}$$

に注意する.

$$\begin{aligned}
\mathcal{L} = & \frac{1}{2} \sum_k N_k \log \tilde{\Lambda}_k - \frac{1}{2} \sum_k N_k \frac{D}{\beta_k} - \frac{1}{2} \sum_k N_k \nu_k \operatorname{tr}(S_k W_k) \\
& - \frac{1}{2} \sum_k N_k \nu_k (\bar{x}_k - m_k)^T W_k (\bar{x}_k - m_k) \\
& - \frac{1}{2} \sum_k N_k D \log(2\pi) + \sum_k N_k \log \tilde{\pi}_k + \log C(\alpha_0) + (\alpha_0 - 1) \sum_k \log \tilde{\pi}_k + \frac{DK}{2} \log \left(\frac{\beta_0}{2\pi} \right) \\
& + \frac{1}{2} \sum_k \log \tilde{\Lambda}_k - \frac{1}{2} \sum_k \frac{D\beta_0}{\beta_k} - \frac{1}{2} \sum_k \beta_0 \nu_k (m_k - m_0)^T W_k (m_k - m_0) + K \log B(W_0, \nu_0) \\
& + \frac{\nu_0 - D - 1}{2} \sum_k \log \tilde{\Lambda}_k - \frac{1}{2} \sum_k \nu_k \operatorname{tr}(W_0^{-1} W_k) - \sum_{n,k} r_{nk} \log r_{nk} \\
& - \sum_k (\alpha_k - 1) \log \tilde{\pi}_k - \log C(\alpha) - \frac{1}{2} \sum_k \log \tilde{\Lambda}_k - \frac{D}{2} \sum_k \log \frac{\beta_k}{2\pi} + \frac{DK}{2} \\
& + \sum_k \left(-\log B(W_k, \nu_k) - \frac{\nu_k - D - 1}{2} \log \tilde{\Lambda}_k + \frac{\nu_k D}{2} \right).
\end{aligned}$$

$$\begin{aligned}
& \mathcal{L} \\
= & \log \frac{C(\alpha_0)}{C(\alpha)} - \sum_{n,k} r_{nk} \log r_{nk} + \frac{1}{2} \sum_k \log \tilde{\Lambda}(N_k + 1 - \nu_0 - D - 1 - 1 - \nu_k + D + 1) \\
& + \sum_k \log \tilde{\pi}_k (N_k + \alpha_0 - 1 - \alpha_k + 1) + K \log B(W_0, \nu_0) - \sum_k \log B(W_k, \nu_k) - \frac{DN}{2} \log(2\pi) \\
& + \underbrace{\frac{D}{2} \sum_k \left(\frac{N_k}{\beta_k} + \frac{\beta_0}{\beta_k} \right)}_{=K} + \frac{DK}{2} (\log \beta_0 - \log(2\pi)) - \frac{D}{2} \sum_k \log \beta_k + \frac{DK}{2} \log(2\pi) \\
& + \frac{DK}{2} + \frac{D}{2} \sum_k \nu_k \\
& - \frac{1}{2} \sum_k \nu_k \operatorname{tr} \left(W_k \underbrace{(N_k S_k + N_k (\bar{x}_k - m_k)(\bar{x}_k - m_k)^T + \beta_0 (m_k - m_0)(m_k - m_0)^T + W_0^{-1})}_{=:X} \right) \\
= & \log \frac{C(\alpha_0)}{C(\alpha)} - \sum_{n,k} r_{nk} \log r_{nk} + K \log B(W_0, \nu_0) - \sum_k \log B(W_k, \nu_k) + \frac{DK}{2} \log \beta_0 \\
& - \frac{D}{2} \sum_k \log \beta_k - \frac{DN}{2} \log(2\pi) + \frac{D}{2} \sum_k \nu_k - \frac{1}{2} \sum_k \nu_k \operatorname{tr}(W_k X). \\
& \bar{x}_k - m_k = \bar{x}_k - \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) = \frac{1}{\beta_k} (\beta_k \bar{x}_k - N_k \bar{x}_k - \beta_0 m_0) = \frac{\beta_0}{\beta_k} (\bar{x}_k - m_0). \\
& m_k - m_0 = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) - m_0 = \frac{1}{\beta_k} (N_k \bar{x}_k + \beta_0 m_0 - \beta_k m_0) = \frac{N_k}{\beta_k} (\bar{x}_k - m_0).
\end{aligned}$$

よって

$$\begin{aligned} N_k(\bar{x}_k - m_k)(\bar{x}_k - m_k)^T + \beta_0(m_k - m_0)(m_k - m_0)^T &= \left(\frac{N_k\beta_0^2}{\beta_k^2} + \frac{\beta_0 N_k^2}{\beta_k^2} \right) \\ &= \frac{\beta_0 N_k}{\beta_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T. \end{aligned}$$

よって

$$X = W_k^{-1}, \quad \sum_k \nu_k \operatorname{tr}(W_k W_k^{-1}) = D \sum_k \nu_k.$$

$$\mathcal{L} = \log \frac{C(\alpha_0)}{C(\alpha)} - \sum_{n,k} r_{nk} \log r_{nk} + \sum_k \log \frac{B(W_0, \nu_0)}{B(W_k, \nu_k)} + \frac{D}{2} \sum_k \log \frac{\beta_0}{\beta_k} - \frac{DN}{2} \log(2\pi).$$

10.8 予測分布

新しい観測値の予測分布を知りたい。

$$p(Z|\boldsymbol{\pi}) = \prod_{n,k} \pi_k^{z_{nk}}, \quad p(X|Z, \mu, \Lambda) = \prod_{n,k} \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})^{z_{nk}}$$

と $\sum_k z_{nk} = 1$ を使って

$$\begin{aligned} p(\hat{x}|X) &= \sum_{\hat{z}} \int p(\hat{x}|\hat{z}, \mu, \Lambda) p(\hat{z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}, \mu, \Lambda|X) d\boldsymbol{\pi} d\mu d\Lambda \\ &= \sum_k \pi_k \int \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) \underbrace{p(\boldsymbol{\pi}, \mu, \Lambda|X)}_{\simeq q(\boldsymbol{\pi})q(\mu, \Lambda)} d\boldsymbol{\pi} d\Lambda d\mu \end{aligned}$$

$$\begin{aligned} &p(\hat{x}|X) \\ &\simeq \sum_k \int \pi_k \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) q(\boldsymbol{\pi}) \prod_j q(\mu_j, \Lambda_j) d\boldsymbol{\pi} d\Lambda d\mu \\ &\quad (k \neq j \text{ なら積分して } 1 \text{ なので}) \\ &= \sum_k \int \pi_k \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) q(\boldsymbol{\pi}) q(\mu_k, \Lambda_k) d\boldsymbol{\pi} d\mu_k d\Lambda_k \\ &= \sum_k \underbrace{\left(\int \pi_k q(\boldsymbol{\pi}) d\boldsymbol{\pi} \right)}_{=:X} \underbrace{\left(\int \mathcal{N}(\hat{x}|\mu_k, \Lambda_k^{-1}) \mathcal{N}(\mu_k|m_k, (\beta_k \Lambda_k)^{-1}) d\mu_k \right)}_{=:Y} W(\Lambda_k|W_k, \nu_k) d\Lambda_k \end{aligned}$$

X, Y をそれぞれ計算する：

$$X = \int \pi_k \operatorname{Dir}(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi} = \frac{\alpha_k}{\hat{\alpha}}.$$

$$\begin{aligned}
A &:= \int \mathcal{N}(x|\mu, \Lambda^{-1}) \mathcal{N}(\mu|m, (\beta\Lambda)^{-1}) d\mu \\
&= \int \frac{1}{(2\pi)^D} |\Lambda|^{1/2} |\beta\Lambda|^{1/2} \exp \left(-\frac{1}{2} \Lambda \underbrace{\left((x-\mu)(x-\mu)^T + \beta(\mu-m)(\mu-m)^T \right)}_{=:B} \right) d\mu
\end{aligned}$$

$$\begin{aligned}
B &= (\beta+1)\mu\mu^T - 2\mu(x+\beta m)^T + xx^T + \beta mm^T \\
&= (\beta+1) \left(\mu - \frac{1}{\beta+1}(x+\beta m) \right) \left(\mu - \frac{1}{\beta+1}(x+\beta m) \right)^T \\
&\quad + \underbrace{xx^T + \beta mm^T - \frac{1}{\beta+1}(x+\beta m)(x+\beta m)^T}_{=:C}
\end{aligned}$$

$$C = \frac{\beta}{\beta+1} xx^T + \frac{\beta^2 + \beta - \beta^2}{\beta+1} mm^T - \frac{2\beta}{\beta+1} xm^T = \frac{\beta}{\beta+1} (x-m)(x-m)^T.$$

よって

$$\begin{aligned}
A &= \int \mathcal{N} \left(\mu \left| \frac{x+\beta m}{\beta+1}, ((\beta+1)\Lambda)^{-1} \right. \right) \frac{1}{(2\pi)^{D/2}} \frac{|\beta\Lambda^2|^{1/2}}{|(\beta+1)\Lambda|^{1/2}} \\
&\quad \cdot \exp \left(-\frac{1}{2} (x-m)^T \left(\frac{\beta}{\beta+1} \Lambda \right) (x-m) \right) d\mu = \mathcal{N} \left(x \left| m, \left(\frac{\beta}{\beta+1} \Lambda \right)^{-1} \right. \right).
\end{aligned}$$

つまり

$$Y = \mathcal{N} \left(\hat{x} \left| m_k, \left(\frac{\beta_k}{\beta_k+1} \Lambda_k \right)^{-1} \right. \right).$$

$$\begin{aligned}
D &:= \mathcal{N} \left(x \left| m, \left(\frac{\beta}{\beta+1} \Lambda \right)^{-1} \right. \right) W(\Lambda|W, \nu) \\
&= \frac{1}{(2\pi)^{D/2}} \left| \frac{\beta}{\beta+1} \Lambda \right|^{1/2} \exp \left(-\frac{1}{2} \text{tr} \left(\Lambda \frac{\beta}{\beta+1} (x-m)(x-m)^T \right) \right) \\
&\quad \cdot B(W, \nu) |\Lambda|^{\frac{\nu-D-1}{2}} \exp \left(-\frac{1}{2} \text{tr}(W^{-1}\Lambda) \right).
\end{aligned}$$

$$W'^{-1} := W^{-1} + \frac{\beta}{\beta+1} (x-m)(x-m)^T$$

とおく. $|I + ab^T| = 1 + a^T b$ より $W = W^T$ のとき

$$|W^{-1} + xx^T| = |W^{-1}| |I + Wxx^T| = |W|^{-1} (1 + x^T W x).$$

よって

$$|W'^{-1}| = |W|^{-1} \left(1 + \underbrace{\frac{\beta}{\beta+1} (x-m)^T W (x-m)}_{=: \lambda} \right) = |W|^{-1} (1 + \lambda).$$

$$|W'| = |W| \left(\frac{1}{1 + \lambda} \right).$$

よって

$$\begin{aligned} & \int D d\Lambda \\ &= \left(\frac{\beta}{2\pi(\beta+1)} \right)^{D/2} \frac{B(W, \nu)}{B(W', \nu+1)} \underbrace{\int B(W', \nu+1) |\Lambda|^{\frac{(\nu+1)-D-1}{2}} \exp\left(-\frac{1}{2} \text{tr}(W'^{-1} \Lambda)\right) d\Lambda}_{=1} \\ &= \left(\frac{\beta}{2\pi(\beta+1)} \right)^{D/2} \frac{|W'|^{(\nu+1)/2} 2^{(\nu+1)D/2} \pi^{D(D-1)/4} \prod_i \Gamma\left(\frac{\nu+2-i}{2}\right)}{|W|^{\nu/2} 2^{\nu D/2} \pi^{D(D-1)/4} \prod_i \Gamma\left(\frac{\nu+1-i}{2}\right)} \\ &= \left(\frac{\beta}{\pi(\beta+1)} \right)^{D/2} \frac{\Gamma((\nu+1)/2)}{\Gamma((\nu+1-D)/2)} \left(\frac{1}{1+\lambda} \right)^{\nu/2} |W|^{1/2} \left(\frac{1}{1+\lambda} \right)^{1/2} \\ &= \frac{\Gamma((\nu+1)/2)}{\Gamma((\nu+1-D)/2)} \left(\frac{\beta}{\pi(1+\beta)} \right)^{D/2} |W|^{1/2} (1+\lambda)^{-(\nu+1)/2} \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu+1-D}{2})} \frac{\left(\frac{\beta(\nu+1-D)}{1+\beta} \right)^{\frac{D}{2}}}{(\nu+1-D)^{\frac{D}{2}} \pi^{\frac{D}{2}}} |W|^{\frac{1}{2}} \left(1 + \frac{(x-m)^T \left(\frac{\beta(\nu+1-D)}{1+\beta} W \right) (x-m)}{\nu+1-D} \right)^{\frac{-(\nu+1)}{2}} \\ &= \text{St}(x|m, L, \nu+1-D), \quad L := \frac{\beta(\nu+1-D)}{1+\beta} W. \end{aligned}$$

よって

$$p(\hat{x}|X) \simeq \frac{1}{\hat{\alpha}} \sum_k \alpha_k \text{St}(\hat{x}|m_k, L_k, \nu_k + 1 - D).$$

だいたいこのあたりまで。あとはここよりは易しいので大丈夫だろう。

PRML 下巻 p.215 の下から 5 行目：「 $\lambda'(\xi)$ は $\xi \geq 0$ のとき単調関数」とあるが間違い。

$\lambda(\xi)$ は $\xi \geq 0$ のとき単調減少だが $\lambda'(\xi)$ はそうではない。また $\lambda(\xi)$ が単調減少だからと

いってそれだけで $\lambda'(\xi) \neq 0$ が言えるわけでもない。