PRML輪読会
第4章線形識別モデル

目次

決定問題を解く3つのアプローチ

識別関数

入力データをいくつかのクラスの1つに割り当てる関数（識別関数）を直接導出する（確率は登場しない）
（4.1で議論済）

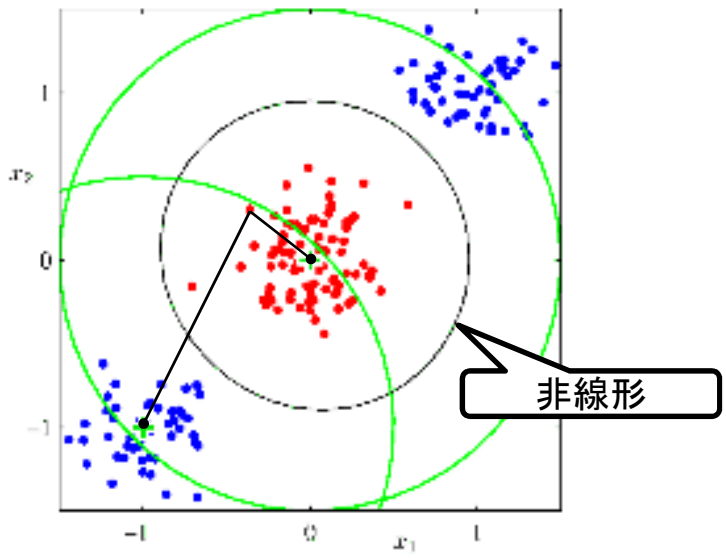
識別モデル

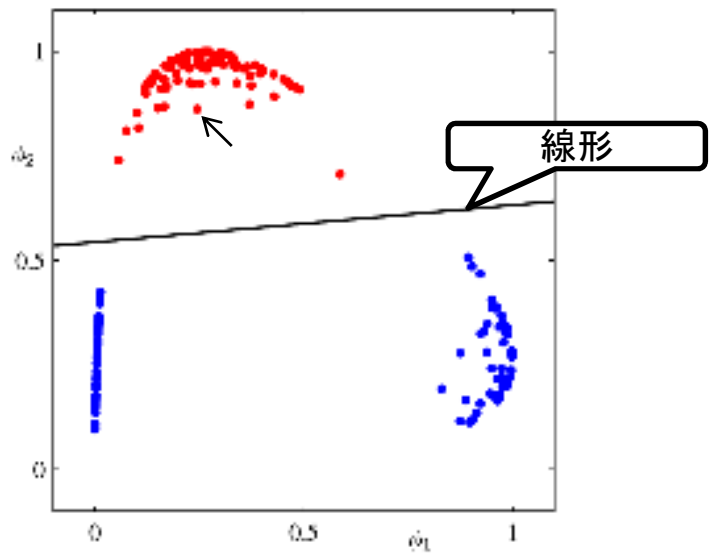
条件付き確率モデル$p(C_k \mid x)$を、データから直接モデル化する
（4.3で議論）

生成モデル

条件付き確率モデル$p(C_k \mid x)$を、$p(x|C_k)$と$p(C_k)$からベイズの定理を使って導出する（4.2で議論済）

4.3.1 固定基底関数





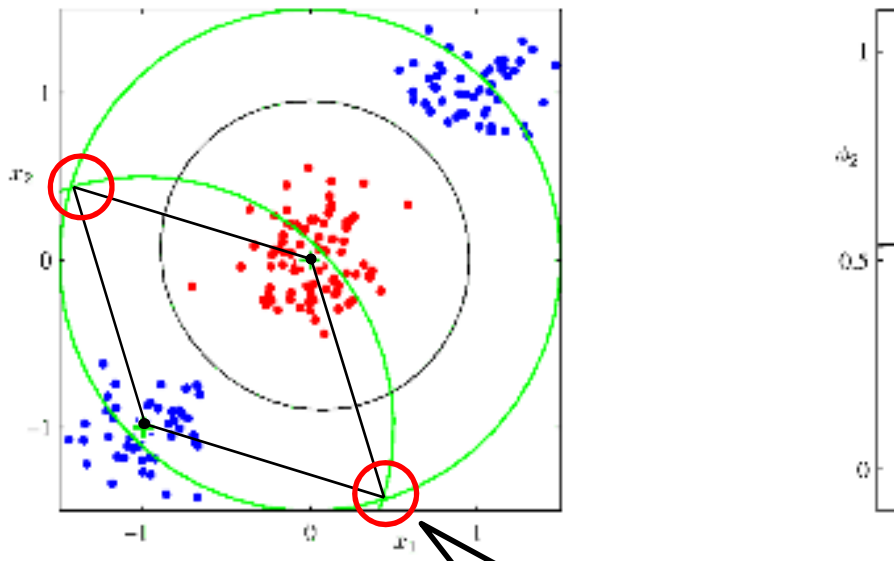
$$\phi_1(x) = \exp\left(-\left\|x - \begin{pmatrix}-1\\-1\end{pmatrix}\right\|\right)$$

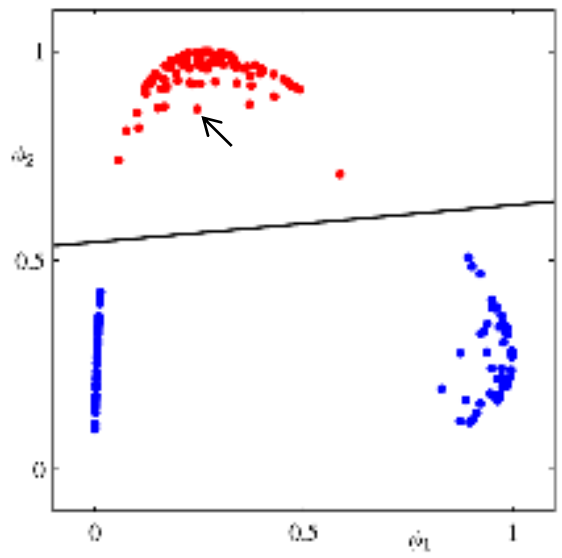
$$\phi_2(x) = \exp(-\|x\|)$$

（たぶんこんな感じ）

あらかじめ非線形変換を行っておけば、変換後の特徴空間においては線形決定境界を得られるかもしれない

4.3.1 固定基底関数





非線形変換を行うことで、もとの空間にはなかった重なりを作ってしまうこともある。

とはいえ有用なので固定基底関数使います。

4.3.2 ロジスティック回帰

ロジスティック回帰とは

（2クラス分類問題において）以下の式であらわされるモデル

$$p(C_1|\phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right) \quad (4.87)$$

ただし、$\sigma(a) = \frac{1}{1+\exp(-a)}$ (4.59)

「回帰」と言っているが分類のためのモデルである
→パラメータ$\mathbf{w}$を求める問題を解いていくものの、そうして得られた$p(C_1|\phi)$は、クラス分類問題で利用される。

識別モデルの場合

$$p(C_1|\phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right) \quad (4.87)$$

M個

生成モデルの場合(4.2.1)

$$p(C_1|\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right) \quad (4.87)$$

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad (4.66)$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)} \quad (4.67)$$

M個　$\frac{M(M+1)}{2}$個　M個　1個　$p(C_2) = 1 - p(C_1)$

次元が大きい場合は識別モデルの方が有利

4.3.2 ロジスティック回帰

ロジスティック回帰モデルのパラメータを最尤法を用いて決定する。
そのための準備として、ロジスティックシグモイド関数の微分を求める。

（演習4.12）$\frac{d\sigma}{da} = \sigma(1-\sigma)$の証明

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

対数を取る

$$\ln \sigma = -\ln(1+\exp(-a))$$

合成関数の微分

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

対数の微分

$$\frac{d\sigma}{da} = \sigma(1-\sigma)$$

両辺σ倍

4.3.2 ロジスティック回帰

データ集合：

$$\{\phi_n, t_n\}$$

ϕ_nが、ロジスティック回帰モデルを用いてt_nに分類された、という訓練データ

ただし、

$$\phi_n = \phi(x_n)$$
$$t_n \in \{0,1\}$$
$$n = 1, \dots, N$$

t_nのそれぞれの尤度関数は、

$$p(t_n|\mathbf{w}) = \begin{cases} p(C_1|\phi_n) = y_n & (t_n = 1\text{のとき}) \\ p(c_2|\phi_n) = 1 - y_n & (t_n = 0\text{のとき}) \end{cases}$$

と書ける。これはさらに、t_nの離散性を用いて、以下のように表せる。

$$p(t_n|\mathbf{w}) = y_n^{t_n}\{1 - y_n\}^{1-t_n}$$

以上より、N個のデータに対する尤度関数は、

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} p(t_n|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1 - y_n\}^{1-t_n} \tag{4.89}$$

誤差関数は、負の対数を取り、

$$E(\mathbf{w}) = -\ln\big(p(\mathbf{t}|\mathbf{w})\big) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\} \tag{4.90}$$

となる。これを交差エントロピー誤差関数と呼ぶ。

4.3.2 ロジスティック回帰

$$y_n = \sigma(a_n)$$
$$a_n = \mathbf{w}^T \phi_n$$

より、

合成関数の微分

$$\nabla \ln y_n = \frac{\mathrm{d} \ln y_n}{\mathrm{d}\mathbf{w}} = \frac{\mathrm{d} \ln y_n}{\mathrm{d}y_n} \frac{\mathrm{d}y_n}{\mathrm{d}a_n} \frac{\mathrm{d}a_n}{\mathrm{d}\mathbf{w}}$$
$$= \frac{1}{y_n} y_n(1 - y_n)\phi_n$$
$$= (1 - y_n)\phi_n$$
$$\nabla \ln(1 - y_n) = -y_n \phi_n$$

これを使って、$E(\mathbf{w})$の勾配は、

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \tag{4.90}$$

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n(1 - y_n)\phi_n - (1 - t_n) y_n \phi_n\}$$
$$= \sum_{n=1}^{N} (y_n - t_n)\phi_n \tag{4.91}$$

と表せる（演習4.13）。

4.3.2 ロジスティック回帰

(3.13)との比較

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n \tag{4.91}$$

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}\phi(\mathbf{x}_n)^{\mathrm{T}} \tag{3.13}$$

$E(\mathbf{w})$にだけ着目し、$\phi_n = \phi(\mathbf{x}_n)$、$y_n = \mathbf{w}^{\mathrm{T}}\phi_n$とすると、

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}\phi(\mathbf{x}_n)$$

$$= \sum_{n=1}^{N} (y_n - t_n)\phi_n$$

となる。これは(4.91)と同じ形をしており、y_nの中身だけが違う。
この違いは第4章の冒頭でしれっと登場した活性化関数（(4.3)式参照）の違いであり、決定面を求める際には結局y_n = 定数だから決定面は線形となり、以後は(3.13)と本質的には同じように議論ができる。

4.3.2 ロジスティック回帰

（演習4.14）最尤解の考察

データ集合が線形分離可能であるとは、ある$\mathbf{w}_0$があって、超平面$\mathbf{w}_0^{\mathrm{T}}\phi = 0$が空間を分離し、$\mathbf{w}_0^{\mathrm{T}}\phi > 0$の時$C_1$、$\mathbf{w}_0^{\mathrm{T}}\phi < 0$の時$C_2$という風に分類できることを言う。
このような$\mathbf{w}_0$に対し、

$$\sigma\left(\mathbf{w}_0^{\mathrm{T}}\phi\right) = \frac{1}{1+\exp\left(-\mathbf{w}_0^{\mathrm{T}}\phi\right)} \begin{cases} > 0.5 & (\phi \in C_1) \\ < 0.5 & (\phi \in C_2) \end{cases}$$

であるから、

$$\lim_{\|\mathbf{w}_0\|\to\infty} \exp\left(-\mathbf{w}_0^{\mathrm{T}}\phi\right) = \begin{cases} 0 & (\phi \in C_1) \\ \infty & (\phi \in C_2) \end{cases}$$

したがって、

$$\lim_{\|\mathbf{w}_0\|\to\infty} y_n = \lim_{\|\mathbf{w}_0\|\to\infty} \frac{1}{1+\exp\left(-\mathbf{w}_0^{\mathrm{T}}\phi_n\right)} = \left\{ \begin{array}{ll} 1 & (\phi \in C_1) \\ 0 & (\phi \in C_2) \end{array} \right\} = t_n$$

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n \tag{4.91}$$

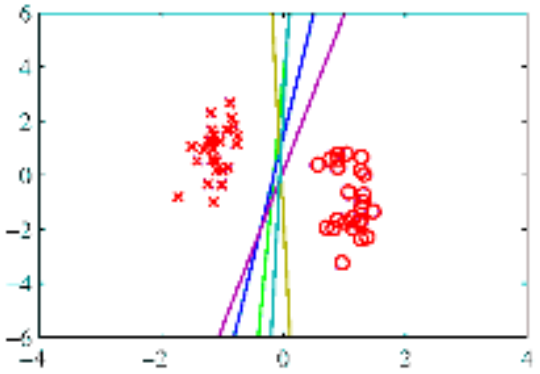
$$\therefore \lim_{\|\mathbf{w}_0\|\to\infty} \nabla E(\mathbf{w}_0) = 0$$

以上より、このような$\mathbf{w}_0$が誤差関数の最尤解になっていることが示された。

4.3.2 ロジスティック回帰

最尤解の考察続き

データ集合が線形分離可能である時、$\mathbf{w}^{\mathrm{T}}\phi = 0$が決定境界となっているような$\mathbf{w}$は無数に存在する（下図）。



最尤法では、解の中の1つを他の解と区別して好んで選別することはできない。→最適化アルゴリズムと初期値に依存し、場合によっては$\mathbf{w}$の大きさが発散してしまう。

4.3.2 ロジスティック回帰

最尤解の考察続き

p.205「データ数がパラメータ数と比べて大きくても、学習データ集合が線形分離可能である限りこの問題は発生してしまう」について

データ数がパラメータ数と比べて大きい場合、$\phi_1, \ldots, \phi_n$はベクトルの個数（＝データ数）が空間の次元（＝パラメータ数）より大きいため、線形独立ではない。つまり、

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n = 0$$

を満たし、かつ、いくつかのnについては

$$y_n - t_n \neq 0$$

であるような$\mathbf{w}$が存在するかもしれない。

しかし、そのような場合でも、線形分離可能な場合は、アルゴリズムの選び方によって$\mathbf{w}$の大きさが∞であるようなものを選んでしまう可能性がある。

1.1や1.2.5と同様に罰則項（$\mathbf{w}$が大きいほど$E(\mathbf{w})$が大きくなるようにする）を加え、正則化することで、この問題を避けることができる。

4.3.3 反復重み付け最小二乗

最小二乗法

誤差関数$E(w)$の値を最小にするようなパラメータ**w**を求める手法

重み付け最小二乗法

データごとに重要度が設定されている場合の最小二乗法

反復重み付け最小二乗法

データごとに重要度が設定され、それがパラメータ**w**に依存する場合の最小二乗法。この場合、重み付け行列とパラメータとを繰り返し更新しながら最尤解を求めていく。

線形回帰モデルでは、(ガウス分布雑音モデル、つまり雑音がガウス分布に従って現れるという仮定の下で)最尤解を解析的に導出できた。

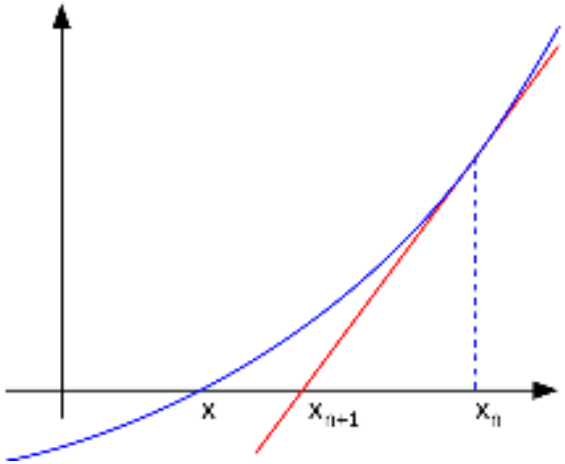
ロジスティック回帰モデルでは、ロジスティックシグモイド関数の非線形性のために最尤解を解析的に導出できない。

→ニュートン-ラフソン法を使って近似値を求めることは可能

4.3.3 反復重み付け最小二乗

ニュートン-ラフソン法（Newton-Raphson method）

入力値x_nにおける接線の零点を次の入力値x_{n+1}とし、これを繰り返すことで関数自体の零点を求める手法



接線の方程式：

$$y - f(x_n) = f'(x_n)(x - x_n)$$

に$(x, y) = (x_{n+1}, 0)$を代入し、

$$0 - f(x_n) = f'(x_n)(x_{n+1} - x_n)$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

を漸化式として値を目標値xに近づけていく。

関数がM次元ベクトルからM次元ベクトルへの写像の場合も同様にして

$$\mathbf{0} - f(\mathbf{x}_n) = \partial f(\mathbf{x}_n)(\mathbf{x}_{n+1} - \mathbf{x}_n)$$

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \partial f(\mathbf{x}_n)^{-1} f(\mathbf{x}_n)$$

となる。∂fはヤコビ行列と呼ばれ、以下で与えられる。

$$\partial f_{ij} = \frac{df(x_i)}{dx_j}$$

誤差関数の場合、$f = \nabla E(\mathbf{w})$であり、$\partial f = \mathbf{H} = \nabla\nabla E(\mathbf{w})$である。

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E\left(\mathbf{w}^{(\text{old})}\right) \tag{4.93}$$

$\mathbf{H}$はEのヘッセ行列と呼ばれ、$\mathbf{H} = \nabla\nabla E(\mathbf{w})$と定義される。

4.3.3 反復重み付け最小二乗

最小二乗法の場合

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 \tag{3.12}$$

これの勾配とヘッセ行列は

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right)\phi_n = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\phi}^{\mathrm{T}}\mathbf{t} \tag{4.93}$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} \tag{4.94}$$

で与えられる。これを式(4.92)に代入し、

$$\begin{aligned}\mathbf{w}^{(\mathrm{new})} &= \mathbf{w}^{(\mathrm{old})} - \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\left\{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \boldsymbol{\phi}^{\mathrm{T}}\mathbf{t}\right\} \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\end{aligned} \tag{4.95}$$

この式は、$\nabla E(\mathbf{w}) = 0$とした時の解と一致する。

4.3.3 反復重み付け最小二乗

重み付き最小二乗法の場合

（教科書には載っていないが）データ点t_nに重み$r_n > 0$が割り当てられている場合について考察する。この時の誤差関数は

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 \tag{3.104}$$

で与えられる。これの勾配とヘッセ行列は

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} r_n\left(\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right)\phi_n = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{t} \tag{4.93}$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi} \tag{4.94}$$

で与えられる。ただし、$\mathbf{R}$は$\mathbf{R}_{nn} = r_n$を満たす対角行列である。
これを式(4.92)に代入し、

$$\begin{aligned}\mathbf{w}^{(\mathrm{new})} &= \mathbf{w}^{(\mathrm{old})} - \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\left\{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{R}\mathbf{t}\right\} \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{R}\mathbf{t}\end{aligned} \tag{4.95}$$

この場合も、$\nabla E(\mathbf{w}) = 0$とした時の解と一致する。

4.3.3 反復重み付け最小二乗

重み付き最小二乗法の場合：r_nの解釈

誤差が正規分布に従い、その分散がデータに依存すると仮定した場合、
その尤度関数は、

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \boldsymbol{\beta}) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \middle| \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta_n^{-1}\right)$$

負の対数を取り、$\mathbf{w}$に関する部分だけ取ると、

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \beta_n\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2$$

重み付き最小二乗法の誤差関数と比較すると、結局、

$$r_n = \beta_n$$

であり、重み係数r_nは誤差の精度パラメータであると解釈できる。

oO（後の議論で「IRLSの対角行列の要素は、重み付き最小二乗法の場合と同様分散であると解釈できる」と出てくるが、これは間違いな気がしている。）

4.3.3 反復重み付け最小二乗

反復重み付き最小二乗法の場合

ロジスティック回帰における交差エントロピー誤差関数：

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \tag{4.90}$$

の勾配とヘッセ行列は

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n = \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \tag{4.96}$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi} \tag{4.97}$$

で与えられる。ただし、$\mathbf{R}$は$\mathbf{R}_{nn} = y_n(1-y_n)$を満たす対角行列である。
これを式(4.92)に代入し、

$$\begin{aligned}\mathbf{w}^{(\mathrm{new})} &= \mathbf{w}^{(\mathrm{old})} - \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\left\{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})\right\} \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\left\{\boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \mathbf{R}^{-1}(\mathbf{y}-\mathbf{t})\right\}\end{aligned} \tag{4.99}$$

ここで、

4.3.3 反復重み付け最小二乗

反復重み付き最小二乗法の場合

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t}) \tag{4.100}$$

とおくと、

$$\begin{aligned}\mathbf{w}^{(\mathrm{new})} &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\left\{\boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t})\right\} \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}\end{aligned} \tag{4.99}$$

さて、この式は、重み付き最小二乗法の場合の解：

$$\mathbf{w}^{(\mathrm{new})} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{R}\mathbf{t}$$

とよく似ている。違うのは、**R**が**w**の関数になっている点、**z**がなんだかごちゃごちゃしている点である。

Rが**w**の関数になっている点

このために、**w**が新たに求まるたびに**R**を再計算し直して、正規方程式を繰り返し解かなければならない。反復重み付き最小二乗法（IRLS）の名はここからきている。

各t_nの平均、分散は、

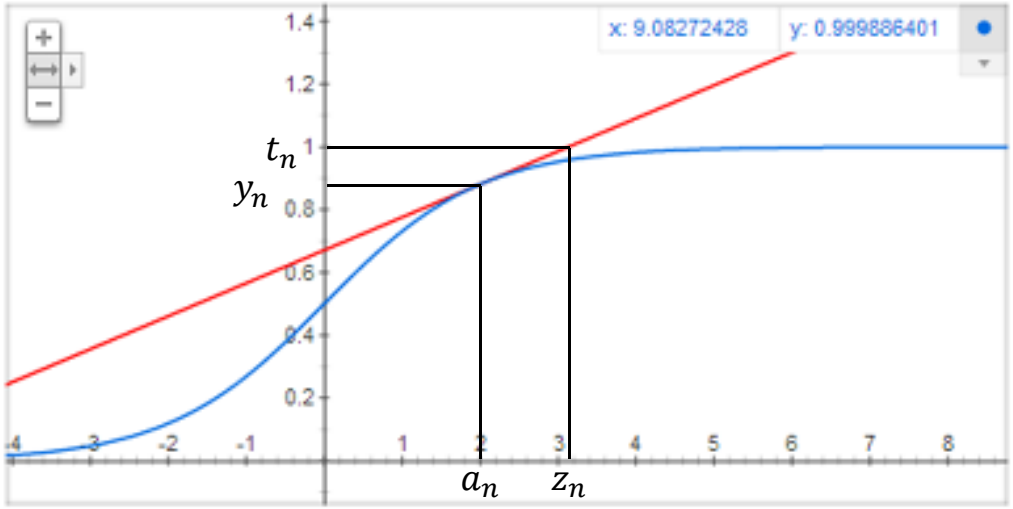
$$\mathbb{E}[t_n] = \sigma\left(\boldsymbol{w}^{\mathrm{T}}\phi_n\right) = y_n \tag{4.101}$$

$$\mathrm{var}[t_n] = \mathbb{E}[t_n^2] - \mathbb{E}[t_n]^2 = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi_n\right) - \sigma\left(\mathbf{w}^{\mathrm{T}}\phi_n\right)^2 = y_n(1 - y_n) \tag{4.102}$$

だから**R**の要素は分散だと解釈できる。

4.3.3 反復重み付け最小二乗

zがなんだかごちゃごちゃしている点



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

z_nとa_n、y_n、t_nは、上図のような位置関係にある。

$$t_n - y_n = \left.\frac{\mathrm{d}y}{\mathrm{d}a}\right|_{a=a_n} (z_n - a_n)$$

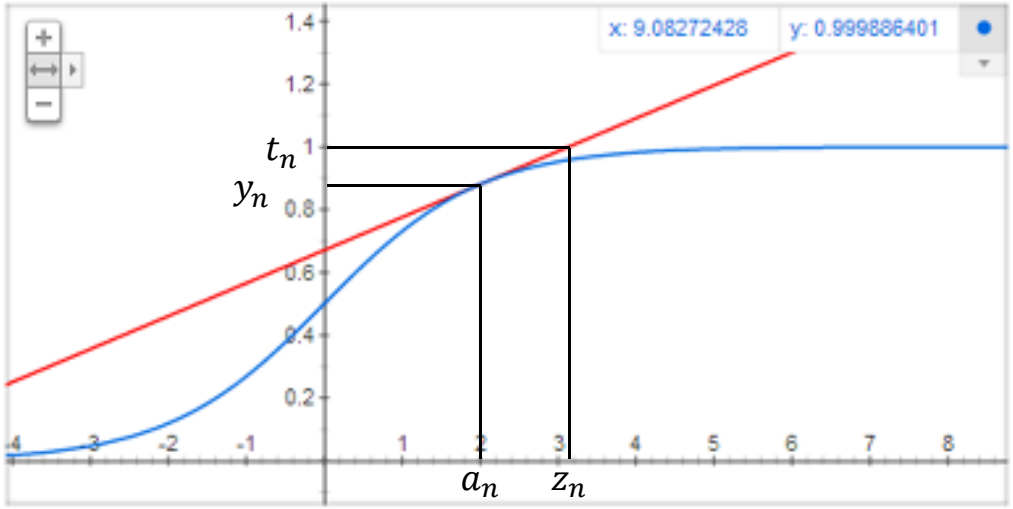
$\left.\frac{\mathrm{d}y}{\mathrm{d}a}\right|_{a=a_n} = y_n(1 - y_n) = r_n$、$a_n = \phi_n^{\mathrm{T}}\mathbf{w}^{(\mathrm{old})}$を使って上式を$z_n$について解くと、

$$t_n - y_n = r_n(z_n - a_n)$$

$$\begin{aligned} z_n &= a_n + r_n^{-1}(t_n - y_n) \\ &= \phi_n^{\mathrm{T}} w^{(\mathrm{old})} - r_n^{-1}(y_n - t_n) \end{aligned}$$

4.3.3 反復重み付け最小二乗

zがなんだかごちゃごちゃしている点



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

nについてまとめると、

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{w} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t})$$

となって、確かに(4.100)と一致する

線形回帰モデルの場合は、y_nが**w**の線形関数だったため、そのまま

$$\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}(\mathbf{y} - \mathbf{t}) = \mathbf{0}$$

としてこれを解くことができた。ロジスティック回帰モデルの場合、ロジスティックシグモイド関数を線形近似し、そこでの解を求めると解釈できる。

$$\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\left(\mathbf{a}^{(\mathrm{new})} - \mathbf{z}\right) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\left(\boldsymbol{\Phi}\mathbf{w}^{(\mathrm{new})} - \mathbf{z}\right) = \mathbf{0}$$

$$\mathbf{w}^{(\mathrm{new})} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}$$

4.3.4 多クラスロジスティック回帰

ソフトマックス変換

多クラスの場合、事後確率は特徴変数の線形関数のソフトマックス変換で与えられる:

$$p(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{4.104}$$

事後確率は

ソフトマックス変換で与えられる

ただし、

$$a_k = \mathbf{w}_k^{\mathrm{T}}\phi \tag{4.105}$$

特徴変数の線形関数の

2クラス分類と同じように、ベイズ定理を用いる生成モデルに対し、最尤法を用いてパラメータ$\{\mathbf{w}_k\}$を決定する。

4.3.4 多クラスロジスティック回帰

準備：(演習4.17) $\frac{\partial y_k}{\partial a_j} = y_k\left(I_{kj} - y_j\right)$の証明

(4.104)の両辺対数を取り、a_jで微分する（分母の総和変数をiとした）

$$\ln y_k = a_k - \ln \sum_i \exp a_i$$

$$\frac{1}{y_k}\frac{\partial y_k}{\partial a_j} = \frac{\partial a_k}{\partial a_j} - \frac{\exp a_j}{\sum_i \exp a_i}$$

a_jによる偏微分のため
添え字jのみが残る

合成関数の微分

ここで、

$$\frac{\partial a_k}{\partial a_j} = \left\{ \begin{array}{l} 1 \ \ (j = k) \\ 0 \ \ (j \neq k) \end{array} \right\} = I_{kj}$$

より、

$$\frac{1}{y_k}\frac{\partial y_k}{\partial a_j} = I_{kj} - y_j$$

$$\therefore \frac{\partial y_k}{\partial a_j} = y_k\left(I_{kj} - y_j\right)$$

4.3.4 多クラスロジスティック回帰

尤度関数の表現：1-of-K符号化法のうまみ

目的変数t_nはK次元ベクトルで、データ$\mathbf{x}_n$がクラスk_0に分類される場合、

$$t_{nk} = \begin{cases} 1 & (k = k_0) \\ 0 & (k \neq k_0) \end{cases}$$

という形をしている。これを使うと、尤度関数は簡単に表すことができて、

$$p(\mathbf{T}|\mathbf{w}_1, \dots \mathbf{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}} \qquad (4.107)$$

t_{nk}を要素とする$N \times K$行列

たくさんかけているように見えるが
実はK個のうち1つしか有効でない

$y_{nk} = y_k(\phi_n)$

誤差関数は、

$$E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \dots, \mathbf{w}_K) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk} \qquad (4.108)$$

と表せる。これを（多クラス分類問題の）交差エントロピー関数と呼ぶ。

4.3.4 多クラスロジスティック回帰

演習4.18の証明

$$\nabla_{\mathbf{w}_j} \ln y_{nk} = \frac{\mathrm{d} \ln y_{nk}}{\mathrm{d} y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} \frac{\mathrm{d} a_{nj}}{\mathrm{d} \mathbf{w}_j}$$

$$= \frac{1}{y_{nk}} y_{nk}(I_{kj} - y_{nj})\phi_n = (I_{kj} - y_{nj})\phi_n$$

$$\sum_{k=1}^{K} t_{nk} I_{kj} = t_{nj} I_{jj} = t_{nj}$$

$$\sum_{k=1}^{K} t_{nk} y_{nj} = y_{nj} \sum_{k=1}^{K} t_{nk} = y_{nj} \cdot 1 = y_{nj}$$

よって、

$$\nabla_{\mathbf{w}_j} \mathrm{E}(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \nabla_{\mathbf{w}_j} \left(- \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk} \right) = - \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \nabla_{\mathbf{w}_j} \ln y_{nk}$$

$$= - \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} (I_{kj} - y_{nj}) \phi_n$$

$$= \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n$$

4.3.4 多クラスロジスティック回帰

ニュートン-ラフソン法の適用

多クラスの場合、パラメータ数は、M次元ベクトル$\mathbf{w}_k$がK個でMK個ある。
これらを縦に1列に並べて、

$$\mathbf{v} = \left(\mathbf{w}_1^{\mathrm{T}}, \ldots, \mathbf{w}_K^{\mathrm{T}}\right)^{\mathrm{T}}$$

と置くと、2クラスの場合と同様にニュートン-ラフソン法を適用できる。

$$\mathbf{v}^{(\mathrm{new})} = \mathbf{v}^{(\mathrm{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{v})$$

$\mathbf{H}$は$MK \times MK$行列で、次のようにあらわされる。

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}_{11} & \cdots & \mathbf{h}_{K1} \\ \vdots & \ddots & \vdots \\ \mathbf{h}_{1K} & \cdots & \mathbf{h}_{KK} \end{pmatrix}$$

$$\mathbf{h}_{jk} = \nabla_{\mathbf{w}_k}\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} y_{nk}\left(I_{kj} - y_{nk}\right)\phi_n\phi_n^{\mathrm{T}}$$

4.3.4 多クラスロジスティック回帰

演習4.20の証明

$$\left(\mathbf{R}_{jk}\right)_{nn} = y_{nk}\left(I_{kj} - y_{nj}\right)$$

と置くと、

$$\mathbf{h}_{jk} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{R}_{jk}\mathbf{\Phi}$$

と書ける。

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \cdots & \mathbf{R}_{K1} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{1K} & \cdots & \mathbf{R}_{KK} \end{pmatrix}, \qquad \mathbf{\Psi} = \begin{pmatrix} \mathbf{\Phi} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{\Phi} \end{pmatrix}$$

と置くと、$\mathbf{R}$は$NK \times NK$行列、$\mathbf{\Psi}$は$NK \times MK$行列になり、

$$\mathbf{H} = \mathbf{\Psi}^{\mathrm{T}}\mathbf{R}\mathbf{\Psi}$$

と書ける。これが正定値であることを証明する。まず、$j \neq k$の時

$$\left(\mathbf{R}_{jk}\right)_{nn} = -y_{nk}y_{nj} = \left(\mathbf{R}_{kj}\right)_{nn}$$

より、$\mathbf{R}$は対称行列である。

MK次元ベクトル$\mathbf{u}$に対し、

$$\mathbf{v} = \mathbf{\Psi}\mathbf{u} = \left(\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_K^{\mathrm{T}}\right)^{\mathrm{T}}$$

と置くと、

$$\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathbf{R}\mathbf{v} = \left(\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_K^{\mathrm{T}}\right)^{\mathrm{T}} \begin{pmatrix} \mathbf{R}_{11} & \cdots & \mathbf{R}_{K1} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{1K} & \cdots & \mathbf{R}_{KK} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_K \end{pmatrix}$$

4.3.4 多クラスロジスティック回帰

演習4.20の証明

$$\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathbf{R}\mathbf{v} = \left(\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_K^{\mathrm{T}}\right)^{\mathrm{T}} \begin{pmatrix} \mathbf{R}_{11} & \cdots & \mathbf{R}_{K1} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{1K} & \cdots & \mathbf{R}_{KK} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_K \end{pmatrix}$$

$$= \left(\mathbf{v}_1^{\mathrm{T}}, \ldots, \mathbf{v}_K^{\mathrm{T}}\right)^{\mathrm{T}} \begin{pmatrix} \mathbf{R}_{11}\mathbf{v}_1 + \cdots + \mathbf{R}_{K1}\mathbf{v}_K \\ \vdots \\ \mathbf{R}_{K1}\mathbf{v}_1 + \cdots + \mathbf{R}_{KK}\mathbf{v}_K \end{pmatrix}$$

$$= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{v}_j^{\mathrm{T}}\mathbf{R}_{jk}\mathbf{v}_k$$

$$= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} y_{nk}\left(I_{jk} - y_{nj}\right)v_{jn}v_{kn}$$

$$= \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} y_{nk}\left(I_{jk} - y_{nj}\right)v_{jn}v_{kn}$$

$$= \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} \left(y_{nk}I_{jk}v_{jn}v_{kn} - y_{nk}y_{nj}v_{jn}v_{kn}\right)$$

4.3.4 多クラスロジスティック回帰

演習4.20の証明

$$\sum_{j=1}^{K}\sum_{k=1}^{K} y_{nk} I_{jk} v_{jn} v_{kn} = \sum_{j=1}^{K} y_{nj} v_{jn} v_{jn} = \sum_{j=1}^{K}\sum_{k=1}^{K} y_{nk} y_{nj} v_{jn} v_{jn}$$

より、

$$\begin{aligned}
\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} &= \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}\left(y_{nk} I_{jk} v_{jn} v_{kn} - y_{nk} y_{nj} v_{jn} v_{kn}\right) \\
&= \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}\left(y_{nk} y_{nj} v_{jn} v_{jn} - y_{nk} y_{nj} v_{jn} v_{kn}\right) \\
&= \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} y_{nk} y_{nj} v_{jn}\left(v_{jn} - v_{kn}\right)
\end{aligned}$$

$a_{jk} = y_{nk} y_{nj} v_{jn}\left(v_{jn} - v_{kn}\right)$とおくと、

$$\begin{aligned}
a_{jj} &= 0, \\
a_{jk} + a_{kj} &= y_{nk} y_{nj} v_{jn}\left(v_{jn} - v_{kn}\right) + y_{nj} y_{nk} v_{kn}\left(v_{kn} - v_{jn}\right) \\
&= y_{nk} y_{nj}\left(v_{jn} - v_{kn}\right)^2
\end{aligned}$$

以上により、

4.3.4 多クラスロジスティック回帰

演習4.20の証明

$$\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} = \sum_{n=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} y_{nk}y_{nj}v_{jn}\left(v_{jn} - v_{kn}\right)$$

$$= \sum_{n=1}^{N}\sum_{1\le j<k\le K} y_{nk}y_{nj}\left(v_{jn} - v_{kn}\right)^2 > 0$$

したがって、Hは正定値行列であり、誤差関数は唯一の最小解を持つ。

4.3.5 プロビット回帰

もう少し一般化した仮定のもとで議論する

活性化関数をもう少し一般的に$f(a)$とし、$a = \mathbf{w}^{\mathrm{T}}\phi$として

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

と書ける場合を考える。例として、雑音しきい値モデルについて考察する。

雑音しきい値モデル

各入力ϕ_nに対して$a_n = \mathbf{w}^{\mathrm{T}}\phi_n$を評価し、あるしきい値$\theta$との大小で目標編数値を設定する。

$$\begin{cases} t_n = 1 & a_n \geq \theta \text{のとき} \\ t_n = 0 & a_n < \theta \text{のとき} \end{cases}$$

θがある確率密度$p(\theta)$から得られると仮定すると、活性化関数は

$$f(a) = p(t = 1|a) = p(a \geq \theta) = \int_{-\infty}^{a} p(\theta)\mathrm{d}\theta$$

と累積分布関数の形に書ける。

プロビット関数

$p(\theta)$の例として、標準正規分布を考える

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

これの逆関数をプロビット関数と言う。形は図4.9を参照

4.3.5 プロビット回帰

演習4.21：erf関数との関係

以下の関数をerf関数または誤差関数という。

$$\mathrm{erf}(a) = \frac{1}{\sqrt{\pi}} \int_0^a \exp(-\theta^2)\,\mathrm{d}\theta$$

プロビット関数の逆関数との関係は、

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\mathrm{d}\theta = \left(\int_{-\infty}^{0} + \int_0^a \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) \mathrm{d}\theta$$

右辺第1項は$\frac{1}{2}$、第2項は$\phi = \frac{\theta}{\sqrt{2}}$と変数変換し、

$$\begin{aligned}
\Phi(a) &= \frac{1}{2} + \int_0^{\frac{a}{\sqrt{2}}} \frac{1}{\sqrt{\pi}} \exp(-\phi^2)\,\mathrm{d}\theta \\
&= \frac{1}{2} + \frac{1}{2} \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right) \\
&= \frac{1}{2} \left\{ 1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right) \right\}
\end{aligned}$$

4.3.5 プロビット回帰

外れ値について

$$\sigma'(a) = \frac{\exp(-a)}{\left(1+\exp(-\mathrm{a})\right)^2} \cong \exp(-a)$$

$$\Phi'(a) = \mathcal{N}(a|0,1) \cong \exp(-a^2)$$

より、プロビット関数の方が外れ値に敏感である。

たとえば、$a_n = \mathbf{w}^{\mathrm{T}}\phi_n = \sqrt{2}, t_n = 0$となるデータ点に対しては、

$$\sigma(a_n) \fallingdotseq 0.804$$

$$\Phi(a_n) = \frac{1}{2}\{1+\mathrm{erf}(1)\} \fallingdotseq 0.921$$

（google電卓機能を利用）

となって、プロビット関数の方が外れ値として大きく加算されてしまう。

目的変数値tが間違った値に反転する確率をϵとすると、データ点$\mathbf{x}$における目的変数値の確率分布は

$$p(t=1|\mathbf{x}) = (1-\epsilon)\sigma(\mathbf{x}) + \epsilon\left(1-\sigma(\mathbf{x})\right)$$

$$= \epsilon + (1-2\epsilon)\sigma(\mathbf{x})$$

ただし、ϵは事前に設定されるか、データ点から推定されるハイパーパラメータとして扱われる

4.3.6 正準連結関数

より一般的な議論

これまでに取り扱ったモデルはすべて、誤差関数の微分が同じ形式：

$$\sum_{n=1}^{N}\{y_n - t_n\}\phi_n$$

を取った。
この節では、上記の結論が、次の仮定のもとに得られることを示す

（仮定1）目的変数に対する条件付き確率分布を指数型分布族の中から選ぶ。
（仮定2）期待値yと線形予測子$\mathbf{w}^{\mathrm{T}}\phi$との連結関数に、正準連結関数を選ぶ。

線形予測子（linear predictor）

独立変数の一次式（$\mathbf{w}^{\mathrm{T}}\phi$）

連結関数

$\mathbf{w}^{\mathrm{T}}\phi$を、目的変数$t$の期待値$y$の関数で表したもの（活性化関数$y = f\left(\mathbf{w}^{\mathrm{T}}\phi\right)$の逆関数）

正準連結関数

連結関数の中で特に、目的変数の分布を表すモデルのパラメータηと期待値yとの関数になっているもの

4.3.6 正準連結関数

より一般的な議論

仮定1より、目的変数tの条件付き確率分布は以下で与えられる。

$$p(t|\eta, s) = \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta) \exp\left\{\frac{\eta t}{s}\right\} \tag{4.118}$$

ただし、sは尺度パラメータで、全ての観測点で共通だと仮定する。
2.4と同様の議論をしてtの期待値を求めたい。(2.194),(2.226):

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\} \tag{2.194}$$

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{2.226}$$

と比較して、

$$\mathbb{E}\left[\frac{t}{s}\middle|\eta\right] = -\frac{d}{d\eta} \ln g(\eta)$$

$$y \equiv \mathbb{E}[t|\eta] = -s \frac{d}{d\eta} \ln g(\eta) \tag{4.119}$$

を得る。yがηの関数になっている。この逆関数が存在するとしてそれを
$\eta = \psi(y)$とする。
これの対数尤度関数は、

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

と書ける。

4.3.6 正準連結関数

より一般的な議論

一般化線形モデルの場合、目的変数tの期待値yが、特徴ベクトルの一次結合の非線形変換で与えられるとしている。つまり、

$$y = f\left(\mathbf{w}^{\mathrm{T}}\phi\right)$$

と書ける。この場合、yの勾配は、$a = \mathbf{w}^{\mathrm{T}}\phi$

$$\nabla_{\mathbf{w}} y = \frac{dy}{da}\frac{da}{d\mathbf{w}} = f'(a)\phi$$

と書ける。これを使って、対数尤度関数の$\mathbf{w}$に関する微分を求めると、

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\boldsymbol{\eta}, s) = \frac{d}{d\mathbf{w}}\left(\sum_{n=1}^{N}\left\{\ln g(\eta_n) + \frac{\eta_n t_n}{s}\right\}\right)$$

$$= \sum_{n=1}^{N}\left(\frac{d}{d\eta_n}\left\{\ln g(\eta_n) + \frac{\eta_n t_n}{s}\right\}\right)\frac{d\eta_n}{dy_n}\frac{dy_n}{da_n}\frac{da_n}{d\mathbf{w}}$$

$$= \sum_{n=1}^{N}\left\{-\frac{y_n}{s} + \frac{t_n}{s}\right\}\psi'(y_n)f'(a_n)\phi_n$$

$y_n = -s\frac{d}{d\eta_n}\ln g(\eta)$

$\eta = \psi(y)$

$y = f(a_n)$

4.3.6 正準連結関数

より一般的な議論

さて、連結関数f^{-1}は、一般化線形モデルの上では何でもよいとしてきたが、ここで、正準連結関数を選ぶ、と仮定する。つまり、

$$f^{-1}(y) = \psi(y) \tag{4.121}$$

この仮定はとても強力で、あとの議論がずっと楽になる。

$$y = f\big(\psi(y)\big)$$

の両辺を微分して、

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

$a = f^{-1}(y) = \psi$より、

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

これを誤差関数の勾配に代入して、

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = -\nabla_{\mathbf{w}} \ln p(t|\eta, s) = -\sum_{n=1}^{N} \left\{-\frac{y_n}{s} + \frac{t_n}{s}\right\} \psi'(y_n) f'(a_n) \phi_n$$

$$= \frac{1}{s} \sum_{n=1}^{N} \{y_n - t_n\} \phi_n$$

以上により、指数型分布族とそれに対応する正準連結関数を選べば、その誤差関数の勾配が誤差$y_n - t_n$と特徴ベクトルϕ_nとの積で表せることが分かった。

4.3.6 正準連結関数

指数分布族と正準連結関数の対照表（以下で、$a = \mathbf{w}^{\mathrm{T}}\phi$）

モデル名	構造	正準連結関数	活性化関数
正規分布	$\mathcal{N}(t\|y,\beta^{-1})$	$a = id_y = y$	$y = id_a = a$
ベルヌーイ分布	$\mathrm{Bern}(t\|y) = y^t(1-y)^{1-t}$	$a = \ln\frac{y}{1-y}$	$y = \sigma(a) = \frac{1}{1+\exp(-a)}$
多項分布	$\prod_{k=1}^{K} y_k^{t_k}$	$a_k - a_K = \ln\frac{y_k}{y_K}$ $(\Sigma_j y_j = 1)$	$y_k = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$
ポアソン分布	$\frac{\lambda^t e^{-\lambda}}{t!}$	$a = \ln y$	$y = \exp a$
指数分布	$\lambda e^{-\lambda t}\ (0 \le t)$	$a = \frac{1}{y}$	$y = \frac{1}{a}$

例：ポアソン分布

$$\frac{\lambda^t e^{-\lambda}}{t!} = \frac{1}{t!} e^{-\lambda} \exp(t \ln \lambda)$$

指数型分布族の形と照らし合わせて、

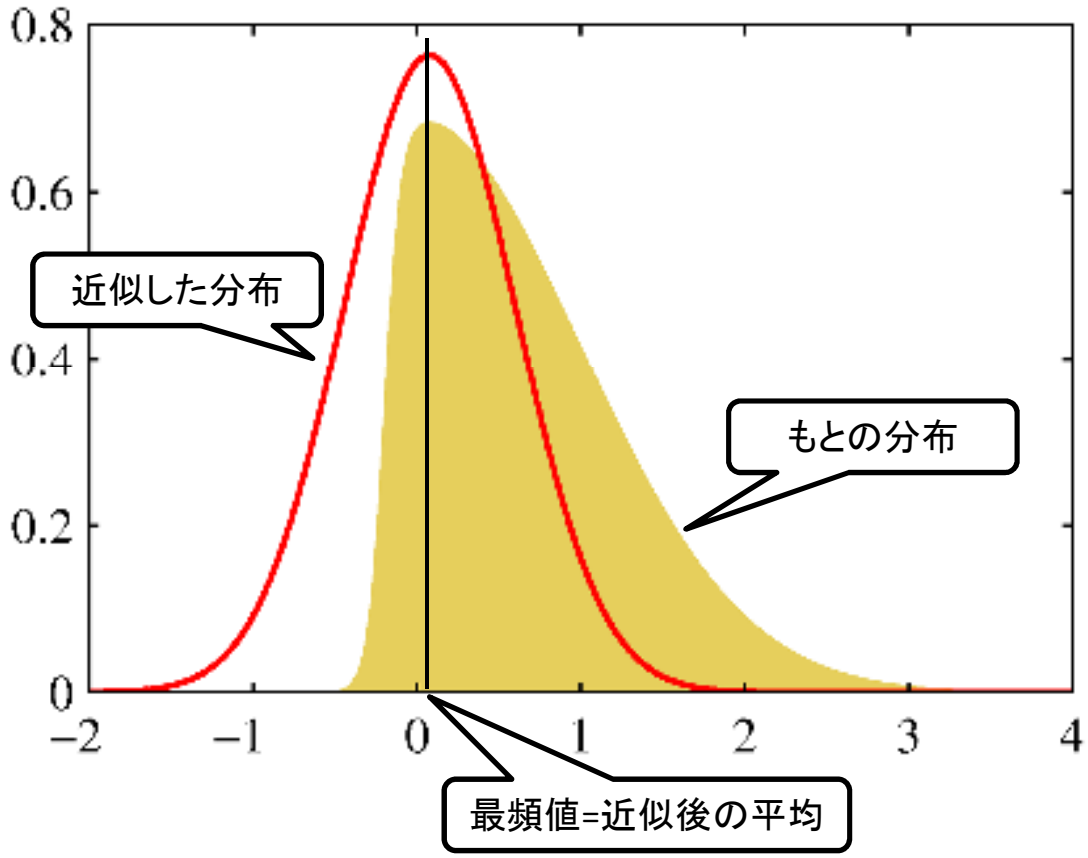
$$\eta = \ln\lambda, \qquad h(t) = \frac{1}{t!}, \qquad g(\eta) = e^{-\lambda} = e^{-\exp\eta}$$

$$y = -\frac{d}{d\lambda}\ln g(\eta) = -\frac{d}{d\lambda}(-\exp\eta) = \exp\eta$$

4.4 ラプラス近似

ラプラス近似とは

分布を正規分布で近似してしまおうという大胆な発想
最頻値を近似後の正規分布の平均値として、うまいこと近似をする。



4.4 ラプラス近似

近似のこころ

・最頻値における1次微分は0である
・正規分布の指数の肩には変数の2次関数が乗っている
→もとの分布の対数を取って2次で近似すればうまくいきそう

ごりごり計算

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

とする。($Z = \int f(z)\mathrm{d}z$は正規化係数、未知)
最頻値をz_0とすると、

$$\left.\frac{df}{dz}\right|_{z=z_0} = 0$$

が成り立つ。したがって、$\ln f$のテイラー展開には1次の項が現れず、

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A(z - z_0)^2$$

と書ける。ここで、

$$A = -\left.\frac{d^2}{dz^2} \ln f(z)\right|_{z=z_0}$$

である。

4.4 ラプラス近似

ごりごり計算

両辺の指数をとると、

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

と近似することができる。指数部分だけ取り出して、正規化した分布$q(z)$を作ると、これは、

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

と書ける。

この近似がうまくいくのは、$A > 0$の場合、つまり、定常点z_0が局所最大である場合である。

4.4 ラプラス近似

多次元の場合

今度は、多次元の場合：

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

の場合を考える。この場合も、最頻値$\mathbf{z}_0$のまわりでの勾配が0であること、
を利用して、$\ln f(\mathbf{z})$を$\mathbf{z}_0$の周りでテイラー展開すると、

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)$$

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}$$

ただし、

$$A = -\nabla\nabla \ln f(\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z}_0}$$

1次元の時と同様に正規化して、

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

を得る。ただし、$|\mathbf{A}|$は$\mathbf{A}$の行列式、Mは変数空間の次元である
この近似がうまくいくには、$\mathbf{A}$は正定値行列、つまり、定常点$\mathbf{z}_0$が局所最大
であって、局所最小であったり鞍点でないことが必要である。

4.4 ラプラス近似

ラプラス近似のメリット

中心極限定理より、観測データが増えるほどガウス分布による近似精度が良くなる。
近似をする際に、真の分布の正規化係数Zを求める必要がない
（$Z = \int f(z)\mathrm{d}z$は一般には計算することができない）。

ラプラス近似のデメリット

実数変数の場合にしか適用できない（離散分布では直接適用できない）
真の分布のある1点（最頻値z_0）のまわりしか注目していないため、多峰的な分布の特性のすべてを捉えることはできない。

4.4.1 モデルの比較とBIC

正規化係数Zの近似

(4.133)の両辺をZで割ると、

$$p(\mathbf{z}) = \frac{f(\mathbf{z})}{Z} \simeq \frac{f(\mathbf{z}_0)}{Z} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}$$

これと(4.134):

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} \tag{4.134}$$

を比較して、

$$\frac{f(\mathbf{z}_0)}{Z} \simeq \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}$$

$$Z \simeq f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|A|^{1/2}} \tag{4.135}$$

を得る。

4.4.1 モデルの比較とBIC

ベイズモデル比較復習

$\mathcal{D}$：データ集合
$\{(\mathcal{M}_i, \boldsymbol{\theta}_i)\}$：モデルとパラメータの組
データ$\mathcal{D}$は、モデル$\mathcal{M}_i$のどれかに従って生成されたが、そのどれかはわからない。どのモデルが一番確からしいかは、モデルの事前分布$p(\mathcal{M}_i)$を仮定し、データ集合$\mathcal{D}$が与えられた時、

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i)$$

で評価される。右辺の$p(\mathcal{D}|\mathcal{M}_i)$をモデルエビデンスという。

モデルエビデンスのラプラス近似

モデルエビデンスは、周辺尤度とも呼ばれ、パラメータ$\boldsymbol{\theta}_i$で尤度関数を周辺化して得られる：

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)p(\boldsymbol{\theta}_i|\mathcal{M}_i)\mathrm{d}\boldsymbol{\theta}_i$$

以下、簡単のため$\mathcal{M}_i$を省略して、

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

と書く。

4.4.1 モデルの比較とBIC

演習4.22

$$f(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \qquad Z = p(\mathcal{D})$$

とする。事後確率$p(\theta|\mathcal{D})$は、ベイズの定理より、

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} = \frac{f(\boldsymbol{\theta})}{Z}$$

で与えられる。事後確率の最頻値を$\boldsymbol{\theta}_{MAP}$としてラプラス近似を適用すると、(4.135)より、

$$\begin{aligned} Z &= \int f(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &\simeq f(\boldsymbol{\theta}_{MAP}) \int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})^{\mathrm{T}}\mathbf{A}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})\right\}\mathrm{d}\boldsymbol{\theta} \\ &= f(\boldsymbol{\theta}_{MAP})\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \end{aligned}$$

両辺の対数を取って、

$$\begin{aligned} \ln p(\mathcal{D}) = \ln Z &\simeq \ln f(\boldsymbol{\theta}_{MAP}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \\ &= \ln p(\mathcal{D}|\theta_{MAP}) + \ln p(\theta_{MAP}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \end{aligned} \qquad (4.137)$$

となって、(4.137)を得る。

4.4.1 モデルの比較とBIC

演習4.22

ここで、$\mathbf{A}$は、(4.132)より、

$$\begin{aligned}\mathbf{A} &= -\nabla\nabla \ln f(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{MAP}} \\ &= -\nabla\nabla \ln p(\mathcal{D}|\theta_{MAP})p(\theta_{MAP}) \\ &= -\nabla\nabla \ln p(\theta_{MAP}|\mathcal{D})p(\mathcal{D}) \\ &= -\nabla\nabla \ln p(\theta_{MAP}|\mathcal{D}) \end{aligned} \tag{4.138}$$

である。

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\theta_{MAP}) + \ln p(\theta_{MAP}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \tag{4.137}$$

最適なパラメータを使用して評価した対数尤度

Occam係数

(4.137)で、右辺第1項は、最適なパラメータを使用して評価した対数尤度である。
残りの3項は「Occam係数」と呼ばれ、モデルの複雑さにペナルティーを科す役目を果たす。

4.4.1 モデルの比較とBIC

ベイズ情報量基準（演習4.23）

対称行列**A**が非退化であるとは、**A**の行列式が0でないことを言う。パラメータの事前分布がガウス分布に従う:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0) = \frac{1}{(2\pi)^{M/2}|\mathbf{V}_0|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\mathbf{m})\right\}$$

と仮定する。この両辺対数を取り、

$$\ln p(\boldsymbol{\theta}) = -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}_0| - \frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\mathbf{m})$$

2階微分は、

$$\nabla\nabla \ln p(\boldsymbol{\theta}) = -\mathbf{V}_0^{-1}$$

次に、$\boldsymbol{\theta}_{MAP}$で評価された負の対数尤度$\ln p(\mathcal{D}|\boldsymbol{\theta})$の2階微分を**H**とする:

$$\mathbf{H} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP})$$

(4.138)より、

$$\begin{aligned}\mathbf{A} &= -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP})\, p(\boldsymbol{\theta}_{MAP}) \\ &= -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \nabla\nabla \ln p(\boldsymbol{\theta}_{MAP}) = \mathbf{H} + \mathbf{V}_0^{-1}\end{aligned}$$

これらを式(4.137)に代入すると、

$$\begin{aligned}\ln p(\mathcal{D}) &\simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \\ &= \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}_0| \\ &\quad - \frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\mathbf{m}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{H} + \mathbf{V}_0^{-1}\right|\end{aligned}$$

4.4.1 モデルの比較とBIC

ベイズ情報量基準（演習4.23）

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}_0|$$
$$-\frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\mathbf{m}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{H}+\mathbf{V}_0^{-1}\right|$$
$$= \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\mathbf{m}) - \frac{1}{2}\ln\left|\mathbf{H}+\mathbf{V}_0^{-1}\right| + const.$$

ここで、事前確率が広い幅を持っている、つまり$\mathbf{V}_0^{-1}$が十分に小さいと仮定すると、右辺第3項の行列式は$|\mathbf{H}|$と近似でき、

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + const.$$

を得る（演習4.23の前半）

さらに、各データ点が独立同分布だと仮定すると、

$$p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) = \prod_{n=1}^{N} p(d_n|\boldsymbol{\theta}_{MAP})$$

$$\mathbf{H} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) = \sum_{n=1}^{N} -\nabla\nabla \ln p(d_n|\boldsymbol{\theta}_{MAP})$$

4.4.1 モデルの比較とBIC

ベイズ情報量基準（演習4.23）

$$\widehat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N} -\nabla\nabla \ln p(d_n|\boldsymbol{\theta}_{MAP})$$

とおくと、

$$\mathbf{H} = N\widehat{\mathbf{H}}$$

$$\ln|\mathbf{H}| = \ln\left|N\widehat{\mathbf{H}}\right| = \ln N^M\left|\widehat{\mathbf{H}}\right| = M\ln N + \ln\left|\widehat{\mathbf{H}}\right|$$

さて、V_0^{-1}が十分小さいことから

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + const.$$

の右辺第2項は無視され、さらに、Nが十分大きいと仮定すれば、$\ln\left|\widehat{\mathbf{H}}\right|$も無視できる。定数項を省略して、

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}M\ln N \qquad (4.137)$$

を得る。

この右辺の値をベイズ情報量規準（Bayesian Infomation Criterion,BIC）あるいは種ワルツ規準と呼ぶ。(1.73)のAIC:

$$\ln p(\mathcal{D}\,|\,\mathbf{w}_{ML}) - M$$

と比べると、BICの方が、罰金項が$\frac{N}{2}$倍大きく、モデルの複雑さにより重いペナルティを科していると言える。