4.2.3 離散特徴

4.2.1、4.2.2などでは、特徴量 $\mathbf{x}$ をガウス分布と仮定した際に、事後分布が一般化線形モデルで表されることを示したが、特徴量 $\mathbf{x}$ が離散値の場合でも一般化線形モデルとなることを、$x_i \in \{0, 1\}$ の例で示す。

特徴量の数がD個であるとすると、 2^D 個の要素の表のすべての要素の確率を考えることになる。ナイーブベイズを仮定すると、各特徴量がクラスに対して条件付き独立とすることができ、独立変数の数を減らすことができる。

$$
\begin{aligned}
p(\mathbf{x} \mid C_k) &= p(x_1 \mid C_k)\, p(x_2 \mid C_k) \cdots p(x_D \mid C_k) \quad (\because \text{条件付き独立}) \\
&= \prod_{i=1}^{D} p(x_i \mid C_k) \\
&= \prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{1-x_i}
\end{aligned}
$$

ここで、 $p(x_i \mid C_k)$ にはベルヌーイ分布を仮定している。
これを、4.63式に入れると、ソフトマックス関数を使ったこれまでの議論通りに適用できる。

$$
a_k = \ln\left(p(\mathbf{x} \mid C_k)\, p(C_k) \right) \qquad (4.63)
$$

↑代入する。

$$
\begin{aligned}
&= \quad (\text{計算}) \\
&= \sum_{i=1}^{D} \left\{ x_i \ln \mu_{ki} + (1-x_i) \ln (1-\mu_{ki}) \right\} + p(C_k)
\end{aligned}
$$

これは、 $\mathbf{x}$ の線形関数である。

4.2.4 指数型分布族

ここまでで、ガウス分布と離散値入力において、事後確率 $p(x|C_k)$が一般化線形モデルで表されることを示した。
本項ではより一般化して、クラスの条件付き確率 $p(x|C_k)$ が指数型分布族であるならば、事後分布が一般化線形モデルとなることに拡張する。
指数型分布族は、尺度パラメータ s を導入して次のように表すことができる。

$$p(x|\lambda_k) = h(x)\, g(\lambda_k) \exp\left\{ \lambda_k^T u(x) \right\}$$

$$= h(x)\, g(\lambda_k) \exp\left\{ \lambda_k^T x \right\}$$

（相談）
「$u(x)=x$ となるような分布の部分クラスに注目する」
とあるが、条件付けて良いのか。任意の指数型分布族について
議論したかったのではなかったか。

$$= \frac{1}{s} h\left(\frac{1}{s}x\right) g(\lambda_k) \exp\left\{ \frac{1}{s} \lambda_k^T x \right\}$$

（相談）
ここで「(2.236) P116 の尺度パラメータを導入すると」
とあるが、これを使える根拠を良く分かっていない。

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

これを、先ほど同様に4.63式に代入すると、ソフトマックス関数を使った事後分布を構成することができる。その際の a_k は、

$$a_k(x) = \ln\left(p(x|C_k)\,p(C_k)\right) \quad (4.63)$$

$$= \underbrace{-\ln(s) + \ln\left(h\left(\frac{x}{s}\right)\right)} + \ln\left(g(\lambda_k)\right) + \frac{1}{s}\lambda^T x + \ln(p(C_k))$$

xに依らない部分なので、

$$P(C_k|x) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \quad (4.62)$$

に入れたときに消失する。

赤線の部分を除いたものを新たに a_k' とおくと、

$$a_k' = \frac{1}{s}\lambda^T x + \ln\left(g(\lambda_k)\right) + \ln p(C_k)$$

これは、x の線形関数なので、事後分布は一般化線形モデルとなる。

（所感）

1. $u(x)=x$ の部分クラスについて論じている点で、指数型分布族一般における議論ではないのではないか。現に、ガウス分布は $u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$ p.113 (2.221) であり、共分散行列が共通でない限り線形な境界面とならない。

2. 尺度パラメータの導入は必要あったのか。最後の形を見る限り、sがなくとも最終形の線形は言えそう

4.3 確率的識別モデル

「4.1 識別関数」では、入力ベクトルから直接識別関数を構築する方法を学んだ。(ref. 最小二乗法、フィッシャーの判別、パーセプトロンアルゴリズム）「4.2 確率的生成モデル」では、クラスの条件付き確率密度と事前確率を最尤推定し、事後分布を導出するような、生成モデルのアプローチを学んだ。

本節では、$p(C_k|x)$を一般化線形モデルで陽に仮定し、最尤推定で直接パラメータを決定する手法を学ぶ。

おさらい：4.1、4.2、4.3の内容はざっくりと以下のイメージ

(4.1) $y(x) \geq$ (定数) なら C_1

$y(x) <$ (定数) なら C_2

(4.2) $p(x|C_k)$、$p(C_k)$ の分布 を仮定し、$p(C_k|x)$ を求める。

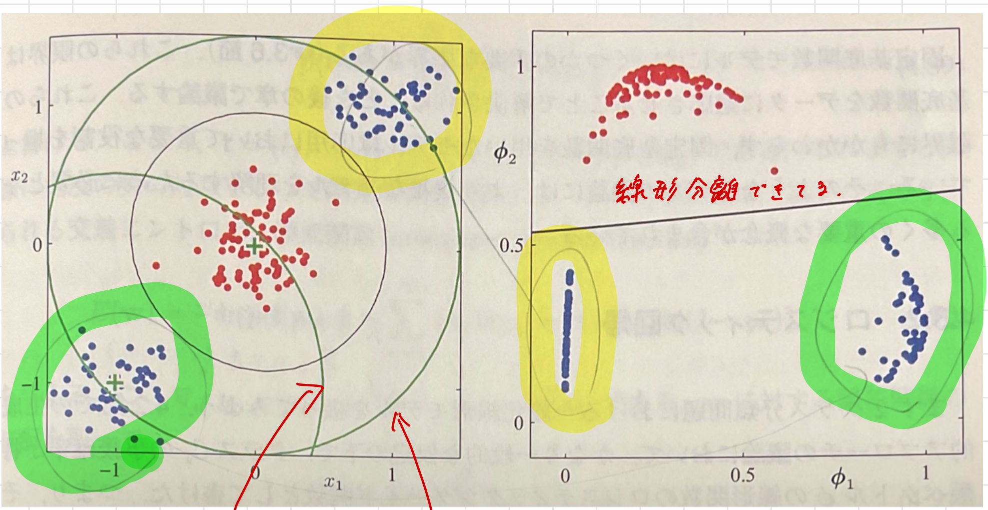
$p(x)$ を導出でき、データを生成できるので「生成モデル」

(4.3) $p(C_k|x)$ のモデル を仮定し、最尤推定で求める。

4.3.1 固定基底関数

本項以降では、入力ベクトル x の代わりに、基底関数ベクトル $\phi(x)$ を用いることとする。3章の回帰問題では基底関数を導入することで、x に対して非線形な目的変数も線形モデルで扱うことができたが、これと同様のアプローチである。

下図は、非線形な基底関数を用いることで線形な識別モデルで分離可能となるような例を示している。



ϕ_1 のガウス基底の等高線。緑の群は ϕ_1 が大きく、黄色の群は ϕ_1 が小さい。

ϕ_2 のガウス基底の等高線。赤のサンプルは ϕ_2 が大きい

4.3.2 ロジスティック回帰

2クラス分類問題において、4.2節ではいくつかの仮定の元で、事後確率がロジスティックシグモイド関数を用いた一般化線形モデルとなる例を見てきた。

本項では、特に仮定を置くことなく $p(C_1|\phi)$ をロジスティックシグモイド関数を用いた一般化線形モデルで表現し、最尤推定することを考える。この方法をロジスティック回帰と呼ぶ。

下記のモデル化を行う。

$$p(C_1|\phi) = y(\phi) = \underbrace{\sigma}_{\text{ロジスティック・シグモイド関数}}(w^T\phi)$$

尤度関数は次のようになる。

$$p(\underbrace{\mathbf{t}}_{\text{正解ラベルベクトル}}|w) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n}$$

これを最大化するのは、負の対数尤度を誤差関数として最小化するのと等価である。この時の誤差関数を交差エントロピー誤差関数と呼ぶ。

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

これの勾配を考える。4.3.4で述べるが解析的に「勾配＝0」の解は得られないので、ここで求める勾配は反復更新に使われる（後述）。

$$\frac{\partial E(w)}{\partial w} = -\sum_{n=1}^{N}\left\{t_n \frac{1}{y_n}\frac{\partial y_n}{\partial w} + \frac{1-t_n}{1-y_n}\left(-\frac{\partial y_n}{\partial w}\right)\right\}$$

$$= -\sum_{n=1}^{N}\left\{\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right\}\frac{\partial\sigma(w^T\phi)}{\partial w}$$

$$= +\sum_{n=1}^{N}\left\{-\frac{t_n}{y_n} + \frac{1-t_n}{1-y_n}\right\}\underbrace{\sigma(w^T\phi)(1-\sigma(w^T\phi))}_{\left(\text{演習4.12}\right)\ \frac{d\sigma(a)}{da} = \sigma(a)\{1-\sigma(a)\}\text{を使用}}\frac{\partial(w^T\phi)}{\partial w}$$

$$= -\sum_{n=1}^{N} \left\{ -\frac{t_n}{y_n} + \frac{1-t_n}{1-y_n} \right\} y_n(1-y_n)\, \phi$$

$$= -\sum_{n=1}^{N} \left\{ -t_n + t_n y_n + y_n - t_n y_n \right\} \phi$$

$$= \sum_{n=1}^{N} (y_n - t_n)\, \phi \qquad (4.91)$$

この勾配を用いて逐次アルゴリズムを構築することで誤差の最小化（つまりここでは最尤推定）を行う。

線形分離可能なデータ集合に対しては、 w の大きさが無限に発散することになる。これを防ぐために w のMAP解を見つければよく、これは誤差関数に正則化項を付加することと等価である。(ref. p30 (1.67)式)

MAP推定では、$p(t|w)$ の代わりに、$p(t|w)\,p(w)$ を最大化する。$p(w)$ をガウス分布としたとき、2次の正則化項が出てきてた。p30、復習

(演習 4.12)

$$\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a)) \quad \text{を求める。}$$

$$\frac{d}{da}\left(\frac{1}{1+e^{-a}}\right) = \frac{-(-e^{-a})}{(1+e^{-a})^2}$$

$$= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}}$$

$$= \frac{1}{1+e^{-a}} \cdot \frac{1+e^{-a}-1}{1+e^{-a}}$$

$$= \sigma(a)\{1-\sigma(a)\}$$

4.3.3 反復重み付け最小二乗

ロジスティック回帰の誤差関数は、これまでの二次の誤差関数と異なり解析的に最小解が得られない。ただ、誤差関数は凸関数であるため、反復最適化手順によって最小解を得ることができる。

ニュートン法は次の式によってパラメータを更新する。

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

線形回帰の最小二乗法の場合、一度の更新で解が得られる。

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\boldsymbol{\phi}_n\}^2 \\
\nabla E(\mathbf{w}) &= \sum_{n=1}^{N}(\mathbf{w}^T\boldsymbol{\phi}_n - t_n)\boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N}\boldsymbol{\phi}_n(\boldsymbol{\phi}_n^T\mathbf{w} - t_n) \\
&= \boldsymbol{\phi}_1\boldsymbol{\phi}_1^T\mathbf{w} - \boldsymbol{\phi}_1 t_1 \\
&\quad + \boldsymbol{\phi}_2\boldsymbol{\phi}_2^T\mathbf{w} - \boldsymbol{\phi}_2 t_2 \\
&\quad \vdots \\
&\quad + \boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{w} - \boldsymbol{\phi}_n t_n \\
&= (\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_n)\underbrace{\begin{pmatrix}\boldsymbol{\phi}_1^T \\ \vdots \\ \boldsymbol{\phi}_n^T\end{pmatrix}}_{\equiv \boldsymbol{\Phi}}\mathbf{w} - (\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_n)\underbrace{\begin{pmatrix}t_1 \\ \vdots \\ t_n\end{pmatrix}}_{\equiv \mathbf{t}} \\
&= \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Phi}\mathbf{t}
\end{aligned}
$$

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

よって更新式は、

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - (\Phi^T\Phi)^{-1}\{\Phi^T\Phi\mathbf{w}^{(old)} - \Phi \mathbf{t}\}$$

$$= \mathbf{w}^{(old)} - \mathbf{w}^{(old)} + (\Phi^T\Phi)^{-1}\Phi \mathbf{t}$$

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

（正規方程式の解）

ロジスティック回帰の場合は最終的に次のような更新式になる。

$$\mathbf{w}^{(new)} = (\Phi^T \mathbb{R} \Phi)^{-1}\Phi^T \mathbb{R}\, \mathbf{z}$$

ここで、$\mathbb{R}$ は $R_{nn} = y_n(1-y_n)$ の対角行列

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

これを以下より示す。

まず、$\nabla E(\mathbf{w})$ と $\mathbb{H}$ を求める。

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

（前項で求めた）

$$= \phi_1(y_1 - t_1) + \phi_2(y_2 - t_2) + \cdots + \phi_n(y_n - t_n)$$

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

$$\mathbb{H} = \nabla\nabla E(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}}\left\{\sum_{n=1}^{N}(y_n - t_n)\phi_n^{T}\right\}$$

$$\begin{pmatrix} \frac{\partial^2}{\partial w_1 \partial w_1} & \cdots & \frac{\partial^2}{\partial w_1 \partial w_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial w_n \partial w_1} & \cdots & \frac{\partial^2}{\partial w_n \partial w_n} \end{pmatrix} = \frac{\partial}{\partial \mathbf{w}}\left(\frac{\partial}{\partial w_1}\ \frac{\partial}{\partial w_2} \cdots \frac{\partial}{\partial w_n}\right)$$

なので、ϕ_n に転置を付けた。

$$
\begin{aligned}
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}} y_n \boldsymbol{\phi}_n^T \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}} \left\{ \sigma(\mathbf{w}^T \boldsymbol{\phi}) \right\} \boldsymbol{\phi}_n^T \\
&= \sum_{n=1}^{N} \sigma(\mathbf{w}^T \boldsymbol{\phi}_n) \left\{ 1 - \sigma(\mathbf{w}^T \boldsymbol{\phi}_n) \right\} \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \\
&= \sum_{n=1}^{N} y_n (1 - y_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \\
&= y_1 (1 - y_1) \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^T \\
&\quad + y_2 (1 - y_2) \boldsymbol{\phi}_2 \boldsymbol{\phi}_2^T \\
&\qquad \vdots \\
&\quad + y_n (1 - y_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \\
&= \underbrace{(\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_n)}_{= \boldsymbol{\Phi}^T} \underbrace{\begin{pmatrix} y_1(1-y_1) & & 0 \\ & \ddots & \\ 0 & & y_n(1-y_n) \end{pmatrix}}_{= \mathbf{R}} \underbrace{\begin{pmatrix} \boldsymbol{\phi}_1^T \\ \vdots \\ \boldsymbol{\phi}_n^T \end{pmatrix}}_{= \boldsymbol{\Phi}} \\
&= \boldsymbol{\Phi}^T \mathbf{R} \boldsymbol{\Phi}
\end{aligned}
$$

ここでのヘッセ行列は、先ほどと違い $\mathbf{w}$ に依存する行列となっている。つまり $\mathbf{w}$ を更新するたびに $\mathbf{H}$ も更新する必要がある。これが、反復「再重み付け」最小二乗と呼ばれる所以である。
また、誤差の2次微分 $\mathbf{H}$ が正定値行列であることから、誤差が凸関数であることも示される。（演習4.15）

ここでは、 $\mathbf{H}$ が正定値行列であることを示す。
零ベクトルでない任意のベクトル $\mathbf{u}$ に対して、

$$
\begin{aligned}
\mathbf{u}^T \mathbf{H} \mathbf{u} &= \mathbf{u}^T \boldsymbol{\Phi}^T \mathbf{R} \boldsymbol{\Phi} \mathbf{u} \\
&= \boldsymbol{v}^T \mathbf{R} \boldsymbol{v} \\
&= (v_1 \cdots v_n) \begin{pmatrix} y_1(1-y_1) & & \\ & \ddots & \\ & & y_n(1-y_n) \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \\
&= \sum_{n=1}^{N} v_n^2 y_n (1 - y_n) \\
&> 0 \qquad (\because \ 0 < y_n < 1)
\end{aligned}
$$

更新式は次のようになる。

$$
\begin{aligned}
w^{(new)} &= w^{(old)} - (\Phi^T R \Phi)^{-1} \Phi^T (y - t) \\
&= (\Phi^T R \Phi)^{-1} \left\{ \Phi^T R \Phi w^{(old)} - \Phi^T (y - t) \right\} \\
&= (\Phi^T R \Phi)^{-1} \Phi^T R \underbrace{\left\{ \Phi w^{(old)} - R^{-1} (y - t) \right\}}_{= z} \\
&= (\Phi^T R \Phi)^{-1} \Phi^T R z
\end{aligned}
$$

この更新を適用する手法を、反復再重み付け最小二乗法（ iterative reweighted least squares method, IRLS）と呼ぶ。

（余談）
ロジスティック回帰をニュートン法で解くという話は、実は「自然科学の統計学」でも少し登場していた（p.238）。ロジスティック回帰はニュートン法などの反復法で解くこと、対数尤度が凹関数であることについて触れている。

演習4.15

$\mathbb{H} = \nabla\nabla E(\mathbf{w})$ が正定値行列ならば、$E(\mathbf{w})$ が凸関数であることを示す。

任意の $\mathbf{w}_a, \mathbf{w}_b \in \mathbb{R}^n$, $\lambda \in (0, 1)$ に対して.

$$f(\lambda) = E(\lambda \mathbf{w}_a + (1-\lambda)\mathbf{w}_b)$$

$$> \lambda E(\mathbf{w}_a) + (1-\lambda) E(\mathbf{w}_b) \qquad (*)$$

を言えれば良い. これは f の凸性を示すのと等しい.

$$\frac{df}{d\lambda} = \nabla E(\lambda \mathbf{w}_a + (1-\lambda)\mathbf{w}_b)^T (\mathbf{w}_a - \mathbf{w}_b)$$

$$\frac{df}{d\lambda^2} = (\mathbf{w}_a - \mathbf{w}_b)^T \nabla\nabla E(\lambda \mathbf{w}_a + (1-\lambda)\mathbf{w}_b)(\mathbf{w}_a - \mathbf{w}_b)$$

$$> 0$$

よって f の凸性が言えたので. $(*)$ を示せた.

4.3.4 多クラスロジスティック回帰

多クラスに拡張した場合も、2クラス同様にモデルを構築することができる。尤度から交差エントロピー誤差関数を定義して、勾配とヘッセ行列から反復更新の手順を得る。

まずは、データが得られた際の各クラスの確率 $p(C_k|\phi)$ を次のようにモデル化する。

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

ここで、 $a_k = w_k^T \phi$

負の対数尤度（交差エントロピー誤差関数）は次のようになる。

$$\begin{aligned}
E(w_1, \cdots\cdots, w_K) &= -\ln p(T | w_1, \cdots, w_K) \\
&= -\ln \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} \\
&= -\ln \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}} \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}
\end{aligned}$$

これの w_j に関する勾配は次のようになる。

$$\begin{aligned}
\frac{\partial}{\partial w_j} E(w_1, \cdots, w_K) &= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial a_j} \frac{\partial a_j}{\partial w_j} \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) \phi_n
\end{aligned}$$

（演習 4.17）

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

$$= -\sum_{n=1}^{N}\left(\underbrace{\sum_{k=1}^{K} t_{nk} I_{kj}}_{k=j\text{のみ残る。}} - \underbrace{\sum_{k=1}^{K} t_{nk}}_{=1} y_{nj}\right)\phi_n$$

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

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

(演習4.17)

次式で示される、ソフトマックス関数の偏微分を求める。

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

(解)

$$\frac{\partial y_k}{\partial a_j} = \frac{\partial}{\partial a_j}\frac{\exp(a_k)}{\sum_i \exp(a_i)}$$

$\frac{\partial}{\partial a_j}\exp(a_k)$ だけ丁寧にやると。

$$\frac{\partial}{\partial a_j}\exp(a_k) = \frac{\partial a_k}{\partial a_j}\cdot\frac{\partial}{\partial a_k}\exp(a_k) = I_{kj}\exp(a_k)$$

k=jのときのみ残る。　　　　i=jの項のみ残る。

$$= \frac{I_{kj}\exp(a_k)\sum_i \exp(a_i) - \exp(a_k)\exp(a_j)}{\left(\sum_i \exp(a_i)\right)^2}$$

$$= \frac{\exp(a_k)}{\sum_i \exp(a_i)}\cdot\frac{I_{kj}\sum_i \exp(a_i) - \exp(a_j)}{\sum_i \exp(a_i)}$$

$$= y_k \cdot (I_{kj} - y_j)$$

ヘッセ行列について考える前に、多クラスロジスティック回帰におけるニュートン法の更新式を考える。PRMLには詳しく載っていないので、こちらを参考にした：
https://www.iwanttobeacat.com/entry/2018/09/08/010518

結論としては、次の更新式を使う。

$$\underset{(KM,\,1)}{\begin{pmatrix} w_1^{(new)} \\ w_2^{(new)} \\ \vdots \\ w_K^{(new)} \end{pmatrix}} = \underset{(KM,\,1)}{\begin{pmatrix} w_1^{(old)} \\ w_2^{(old)} \\ \vdots \\ w_K^{(old)} \end{pmatrix}} - \underset{(KM,\,KM)}{\begin{pmatrix} H_{11} & \cdots & H_{1K} \\ \vdots & & \vdots \\ H_{K1} & \cdots & H_{KK} \end{pmatrix}^{-1}} \underset{(KM,\,1)}{\begin{pmatrix} \frac{\partial E}{\partial w_1^{(old)}} \\ \vdots \\ \frac{\partial E}{\partial w_K^{(old)}} \end{pmatrix}}$$

1個1個が $M \times M$ 行列
M は w の次元 (P.204)

ここで、ヘッセ行列のうち、ブロック j,k について次のように求めることができる。

$$\nabla_{w_j} \nabla_{w_k} E(w_1, \cdots, w_K) = \frac{\partial}{\partial w_j} \frac{\partial}{\partial w_k^T} E$$

教科書と演算子の順序違うけど、こうだよね？？

$$= \frac{\partial}{\partial w_j} \sum_{n=1}^{N} (y_{nk} - t_{nk}) \phi_n^T$$

$$= \sum_{n=1}^{N} \frac{\partial y_{nk}}{\partial a_j} \frac{\partial a_j}{\partial w_j} \phi_n^T$$

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

$$= (\phi_1 \cdots \phi_n) \begin{pmatrix} y_{1k}(I_{kj} - y_{1j}) & & 0 \\ & \ddots & \\ 0 & & y_{nk}(I_{kj} - y_{nj}) \end{pmatrix} \begin{pmatrix} \phi_1^T \\ \vdots \\ \phi_n^T \end{pmatrix}$$

$$= \Phi^T R_{k,j} \Phi$$