パターン認識と機械学習
第4章 線形識別モデル

橋本龍二

January 26, 2021

分類問題に関する3つのアプローチ：

1. 識別関数
   入力ベクトル $\boldsymbol{x}$ から直接クラスを推定する関する $f(\boldsymbol{x})$ を見つける。2クラスなら $f(\cdot)$ は2値を出力する。

2. 推論段階で条件付き確率 $p(\mathcal{C}_k|\boldsymbol{x})$ をモデル化したのち、$p(\mathcal{C}_k|\boldsymbol{x})$ を利用して決定理論を適用する。決定理論は、例えば

   $$\min_j \sum_k L_{kj} p(\mathcal{C}_k|\boldsymbol{x})$$

   となるように j を決める期待損失最小化や、

   $$\begin{aligned} \max \sum_k p(\boldsymbol{x} \in \mathcal{R}_k, C_k) &= \max \sum_k \int_{\mathcal{R}_k} p(\boldsymbol{x}, \mathcal{C}_k) d\boldsymbol{x} \\ &= \max \sum_k \int_{\mathcal{R}_k} p(\boldsymbol{x}) p(\mathcal{C}_k|\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

   となるように決定領域 $\mathcal{R}_k$ を決める誤識別率最小化がある。
   推論段階において、$p(\mathcal{C}_k|\boldsymbol{x})$ を決める方法には次の二通りがある。

   (a) パラメトリックに分布を表現する。(確率的識別モデル)
   例えば、パラメータ $\boldsymbol{w}$ に分布を導入し、

   $$y(\boldsymbol{x}) = f(\ ^T\boldsymbol{w}\boldsymbol{x} + w_0)$$

   から各クラスの事後確率を並べたベクトルを求める。

(b) クラスの事前確率 $p(\mathcal{C}_k)$ とクラスで条件づけられた確率密度 $p(\boldsymbol{x}|\mathcal{C}_k)$ から各クラスの事後確率：

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\mathcal{C}_k)p(\boldsymbol{x}|\mathcal{C}_k)}{p(\boldsymbol{x})}$$

を求める。(確率的生成モデル)

以上の 1、2(a)、2(b) の方法について詳しく見ていく。

1 識別関数

1.1 2 クラス

D 次元の入力ベクトル $\boldsymbol{x}$ が与えられた時、線形識別関数

$$y(\boldsymbol{x}) = {}^T\boldsymbol{w}\boldsymbol{x} + w_0$$

が 0 以上であればクラス $\mathcal{C}_1$ に割り当てることにする (決定境界：$y(\boldsymbol{x}) = 0$ を満たす点 $\boldsymbol{x}$ の集合)。

$${}^T\boldsymbol{w}\boldsymbol{x} + w_0 = 0$$

を満たす決定境界は $D-1$ 次元の超平面である ($D-1$ 個の値が決まれば残りは方程式より決定する)。
決定境界上の任意の 2 点 $\boldsymbol{x}_A, \boldsymbol{x}_B$ を考える。

$${}^T\boldsymbol{w}(\boldsymbol{x}_A - \boldsymbol{x}_B) = 0 \tag{1}$$

より、ベクトル $\boldsymbol{w}$ は決定面上の任意のベクトルと直交する。(w は平面の法線ベクトル、みたいなこと)
今、決定面上の任意の点 $\boldsymbol{x}$ をとり、原点から決定面への垂線をひく。この垂線は $\boldsymbol{w}$ と平行なので、定数 t を用いて $t\boldsymbol{w}$ と表すことにする。$t\boldsymbol{w}$ の長さが原点から決定面への距離である。(1) より

$$\begin{aligned}
(\boldsymbol{x} - t\boldsymbol{w}, \boldsymbol{w}) &= 0 \\
(\boldsymbol{x}, \boldsymbol{w}) - t\|\boldsymbol{w}\|^2 &= 0 \\
\therefore t &= \frac{(\boldsymbol{x}, \boldsymbol{w})}{\|\boldsymbol{w}\|^2}
\end{aligned}$$

したがって、原点から決定面の距離は、

$$\begin{aligned}
\|t\boldsymbol{w}\| = \|\frac{(\boldsymbol{x}, \boldsymbol{w})}{\|\boldsymbol{w}\|^2}\boldsymbol{w}\| &= \frac{(\boldsymbol{x}, \boldsymbol{w})}{\|\boldsymbol{w}\|} \\
&= -\frac{w_0}{\|\boldsymbol{w}\|}
\end{aligned}$$

(ただし、$\boldsymbol{w}$ がどちら向きでも r は変わらない訳だから厳密に距離を求めたくなったら絶対値を取る必要がある。)
次に、任意の点 $\boldsymbol{x}$ と決定面との距離を求める。$\boldsymbol{x}$ から決定面へ下ろした垂線と決定面との交点を $\boldsymbol{x}_\perp$($\boldsymbol{x}$ の決定面への直交射影) と置くと、垂線は $\boldsymbol{w}$ と平行なので、定数 s を用いて $s\boldsymbol{w}$ と表すことにする。すると $\boldsymbol{x}$ は、

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{x}_\perp + s\boldsymbol{w} \\
&= \boldsymbol{x}_\perp + \frac{r}{\|\boldsymbol{w}\|}\boldsymbol{w} \qquad (2) \\
&\quad (r = \|\boldsymbol{w}\|s)
\end{aligned}$$

と書ける。r が $\boldsymbol{x}$ と決定面の距離である。今、(2) の両辺に ${}^{T}\boldsymbol{w}$ を掛け w_0 を加えることで、

$$\begin{aligned}
{}^{T}\boldsymbol{w}\boldsymbol{x} + w_0 &= {}^{T}\boldsymbol{w}(\boldsymbol{x}_\perp + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}) + w_0 \\
&= {}^{T}\boldsymbol{w}\boldsymbol{x}_\perp + w_0 + r\|\boldsymbol{w}\| \\
\therefore y(\boldsymbol{x}) &= r\|\boldsymbol{w}\| \\
\therefore r &= \frac{y(\boldsymbol{x})}{\|\boldsymbol{w}\|}
\end{aligned}$$

まとめ

$$y(\boldsymbol{x}) = {}^{T}\boldsymbol{w}\boldsymbol{x} + w_0$$

なる線形識別関数において、

- $\boldsymbol{w}$ は決定面と直交する。すなわち $\boldsymbol{w}$ は決定面の方向を決定する。
- 原点から決定面までの距離は $-\frac{w_0}{\|\boldsymbol{w}\|}$ である。すなわち w_0 は決定面の位置を決定する。
- 任意の点 $\boldsymbol{x}$ と決定面との距離は $\frac{y(\boldsymbol{x})}{\|\boldsymbol{w}\|}$ である。すなわち $y(\boldsymbol{x})$ は決定面からの点 $\boldsymbol{x}$ の距離を与える。

1.2 多クラス

D 次元入力ベクトル $\boldsymbol{x}$ に対し、

$$y_k(\boldsymbol{x}) = {}^{T}\boldsymbol{w}_k\boldsymbol{x} + w_{k0}$$

を定義し、全ての $j \neq k$ に対し $y_k(\boldsymbol{x}) > y_j(\boldsymbol{x})$ ならば $\boldsymbol{x}$ をクラス $\mathcal{C}_k$ に分類することを考える。この時、クラス $\mathcal{C}_k$ とクラス $\mathcal{C}_j$ の決定境界は $y_k(\boldsymbol{x}) = y_j(\boldsymbol{x})$ である。すなわち、

$$^T(\boldsymbol{w}_k - \boldsymbol{w}_j)\boldsymbol{x} = -(w_{k0} - w_{j0})$$

なる $D-1$ 次元超平面。
識別器 $y_k(\boldsymbol{x})$ の決定領域は必ず凸領域である。つまり、任意の $\boldsymbol{x}_A, \boldsymbol{x}_B \in \mathcal{R}_k$ を取れば、線分 $\boldsymbol{x}_A\boldsymbol{x}_B$ 上の任意の点 $\hat{\boldsymbol{x}}$ は必ず $\mathcal{R}_k$ に属する。なぜなら、

$$\hat{\boldsymbol{x}} = \lambda\boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B, \quad 0 \leq \lambda \leq 1$$

とすれば、

$$y_k(\hat{\boldsymbol{x}}) = \lambda y_k(\boldsymbol{x}_A) + (1-\lambda)y_k(\boldsymbol{x}_B)$$

今、任意の $j \neq k$ について $y_k(\boldsymbol{x}_A) > y_j(\boldsymbol{x}_A),\ y_k(\boldsymbol{x}_B) > y_j(\boldsymbol{x}_B)$ が成り立つのだから、$\hat{\boldsymbol{x}} \in \mathcal{R}_k$
2 クラスの場合でも、$y_1(\boldsymbol{x}), y_2(\boldsymbol{x})$ の 2 つの識別関数によって定式化することも可能である。

1.3 分類における最小二乗

D 次元入力 $\boldsymbol{x}$ が与えられた時、2 値表記法の目標ベクトル $\boldsymbol{t}$ を $\boldsymbol{y}(\boldsymbol{x})$ で推定する。最小二乗法で $\boldsymbol{y}(\boldsymbol{x}) = E[\boldsymbol{t}|\boldsymbol{x}]$ とすれば、$\boldsymbol{y}$ の第 k 要素は、クラス $\mathcal{C}_k$ に割り当てる確率である (識別関数なので厳密には確率ではないことに注意。確率っぽく扱う)。今、n 個目のデータ $\boldsymbol{x}_n$ を $\mathcal{C}_k$ に割り当てる確率を

$$y_k(\boldsymbol{x}) \quad = \quad {}^T\boldsymbol{w}_k\boldsymbol{x} + w_{k0}, \quad k = 1, ..., K$$

とすれば、$\boldsymbol{x}_n$ に対する推定値は

$$\begin{aligned}\boldsymbol{y}(\boldsymbol{x}_n) &= \begin{pmatrix} y_1(\boldsymbol{x}_n) \\ \vdots \\ y_K(\boldsymbol{x}_n) \end{pmatrix} \\ &= \begin{pmatrix} {}^T\tilde{\boldsymbol{w}}_1\tilde{\boldsymbol{x}}_n \\ \vdots \\ {}^T\tilde{\boldsymbol{w}}_K\tilde{\boldsymbol{x}}_n \end{pmatrix} \\ &= {}^T\tilde{W}\tilde{\boldsymbol{x}}_n\end{aligned}$$

ただし、

$$\tilde{\boldsymbol{w}}_k = \begin{pmatrix} w_0 \\ \boldsymbol{w}_k \end{pmatrix} \quad \tilde{\boldsymbol{x}}_n = \begin{pmatrix} 1 \\ \boldsymbol{x}_n \end{pmatrix},\ \tilde{W} = \begin{pmatrix} \tilde{\boldsymbol{w}}_1 \\ \vdots \\ \tilde{\boldsymbol{w}}_K \end{pmatrix}$$

とできる。$\boldsymbol{y}(\boldsymbol{x}_n),\quad n = 1, ..., N$ を N 個縦に並べた行列を Y とすれば、

$$\begin{aligned} Y &= \begin{pmatrix} {}^T\boldsymbol{y}(\boldsymbol{x}_1) \\ \vdots \\ {}^T\boldsymbol{y}(\boldsymbol{x}_N) \end{pmatrix} \\ &= \begin{pmatrix} y_1(\boldsymbol{x}_1) & \dots & y_K(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ y_1(\boldsymbol{x}_N) & \dots & y_K(\boldsymbol{x}_N) \end{pmatrix} \end{aligned}$$

また、Y は

$$\begin{aligned} Y &= \begin{pmatrix} {}^T\boldsymbol{y}(\boldsymbol{x}_1) \\ \vdots \\ {}^T\boldsymbol{y}(\boldsymbol{x}_N) \end{pmatrix} \\ &= \begin{pmatrix} {}^T\tilde{\boldsymbol{x}}_1\tilde{W} \\ \vdots \\ {}^T\tilde{\boldsymbol{x}}_N\tilde{W} \end{pmatrix} \\ &= \tilde{X}\tilde{W}, \qquad (\tilde{X} = \begin{pmatrix} {}^T\tilde{\boldsymbol{x}}_1 \\ \vdots \\ {}^T\tilde{\boldsymbol{x}}_N \end{pmatrix}) \end{aligned}$$

である。二乗和誤差関数を求める。(二重和はトレースで書ける！)

$$
\begin{aligned}
& E_D(\tilde{W}) \\
= \; & \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\{y_k(\boldsymbol{x}_n) - t_{nk}\}^2 \\
= \; & \frac{1}{2}Tr\begin{pmatrix} (y_1(\boldsymbol{x}_1) - t_{11})^2 + ... + (y_1(\boldsymbol{x}_N) - t_{N1})^2 & & \\ & \ddots & \\ & & (y_K(\boldsymbol{x}_1) - t_{1K})^2 + ... + (y_K(\boldsymbol{x}_N) - t_{NK})^2 \end{pmatrix} \\
= \; & \frac{1}{2}Tr\begin{pmatrix} y_1(\boldsymbol{x}_1) - t_{11} & \dots & y_1(\boldsymbol{x}_N) - t_{N1} \\ \vdots & \ddots & \vdots \\ y_K(\boldsymbol{x}_1) - t_{1K} & \dots & y_K(\boldsymbol{x}_N) - t_{NK} \end{pmatrix}\begin{pmatrix} y_1(\boldsymbol{x}_1) - t_{11} & \dots & y_K(\boldsymbol{x}_1) - t_{1K} \\ \vdots & \ddots & \vdots \\ y_1(\boldsymbol{x}_N) - t_{N1} & \dots & y_K(\boldsymbol{x}_N) - t_{NK} \end{pmatrix} \\
= \; & \frac{1}{2}\{ \; {}^T(Y - T)(Y - T)\} \\
= \; & \frac{1}{2}Tr\{ \; {}^T(\tilde{X}\tilde{W} - T)(\tilde{X}\tilde{W} - T)\}
\end{aligned}
$$

二乗和誤差関数を $\tilde{W}$ で微分して 0 と置くと、

$$\tilde{W} = (\; {}^T\tilde{X}\tilde{X})^{-1} \; {}^T\tilde{X}T = \tilde{X}^{\dagger}T$$

なので、求める識別関数は

$$\boldsymbol{y}(\boldsymbol{x}) = \; {}^T\tilde{W}\tilde{x} = \; {}^TT \; {}^T\tilde{X}^{\dagger}\tilde{x}$$

この関数は任意の $\boldsymbol{x}$ に対し各要素の和が 1 になる、という性質を持つ (したがって確率かのように取り扱えるが、値域は $(0, 1)$ でない)。
最小二乗法の弱点 (回帰でも同じことが言える)：

- 最小二乗法は条件付きガウス分布のもとでの最尤推定に相当するので、外れ値に頑健でない。

- 最小二乗法は条件付き分布 $p(\boldsymbol{t}|\boldsymbol{x})$ が正規分布しない場合仮定を満たさないのだからうまくいかない。

1.4　フィッシャーの線形判別

D 次元入力ベクトルを

$$y = \; {}^T\boldsymbol{w}\boldsymbol{x}$$

によって 1 次元に射影して $y > -w_0$ の時クラス $\mathcal{C}_1$ に分類する二分類問題を考える。クラス $\mathcal{C}_1$ の点が N_1 個、クラス $\mathcal{C}_2$ の点が N_2 個あるとすると、各クラスの平均ベクトルは

$$\boldsymbol{m}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}\boldsymbol{x}_n,\quad \boldsymbol{m}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}\boldsymbol{x}_n$$

である。射影先で分離度が最大になるように $\boldsymbol{w}$ を定めたい。

$$\max_{\boldsymbol{w}}\quad m_2 - m_1 = {}^T\boldsymbol{w}(\boldsymbol{m}_2 - \boldsymbol{m}_1)\quad s.t.\quad \sum_i w_i^2 = 1$$

一階条件より、

$$\boldsymbol{w} \propto \boldsymbol{m}_2 - \boldsymbol{m}_1 \qquad (\sum_i w_i^2 = 1)$$

より、$\boldsymbol{x}$ 一次元数ベクトル空間 $< \boldsymbol{w} >=< \boldsymbol{m}_2 - \boldsymbol{m}_1 >$ へと射影 $(\boldsymbol{w}, \boldsymbol{x})\boldsymbol{w}$ してそこで閾値を決めれば良いのだが、図 4.6 左のようにこれではまだ射影先でクラスの被りが見られる。これを防ぐために、射影先での各クラスでのクラス内分散を

$$s_k = \sum_{n\in\mathcal{C}_k}(y_n - m_k)^2$$

と定義して、

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

を最大化する問題を考える。今、

$$\begin{aligned}
(m_2 - m_1)^2 &= ({}^T\boldsymbol{w}\boldsymbol{m}_2 - {}^T\boldsymbol{w}\boldsymbol{m}_1)^2 \\
&= {}^T\boldsymbol{w}(\boldsymbol{m}_2 - \boldsymbol{m}_1)\,{}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)\boldsymbol{w} \\
s_1^2 + s_2^2 &= \sum_{n\in\mathcal{C}_1}({}^T\boldsymbol{w}\boldsymbol{x}_n - {}^T\boldsymbol{w}\boldsymbol{m}_1)^2 + \sum_{n\in\mathcal{C}_2}({}^T\boldsymbol{w}\boldsymbol{x}_n - {}^T\boldsymbol{w}\boldsymbol{m}_2)^2 \\
&= {}^T\boldsymbol{w}\{\sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n - \boldsymbol{m}_1)\,{}^T(\boldsymbol{x}_n - \boldsymbol{m}_1) + \sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n - \boldsymbol{m}_2)\,{}^T(\boldsymbol{x}_n - \boldsymbol{m}_2)\}\boldsymbol{w}
\end{aligned}$$

なので、

$$\begin{aligned}
S_B &= (\boldsymbol{m}_2 - \boldsymbol{m}_1)\,{}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1) \\
S_W &= \sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n - \boldsymbol{m}_1)\,{}^T(\boldsymbol{x}_n - \boldsymbol{m}_1) + \sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n - \boldsymbol{m}_2)\,{}^T(\boldsymbol{x}_n - \boldsymbol{m}_2)
\end{aligned}$$

とおけば、最大化問題は

$$\max_{\boldsymbol{w}} \ J(\boldsymbol{w}) = \frac{{}^T\boldsymbol{w}S_B\boldsymbol{w}}{{}^T\boldsymbol{w}S_W\boldsymbol{w}} \quad s.t. \quad \sum_i w_i^2 = 1$$

と書ける。一階条件を解くことで、

$$\begin{aligned}
\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{\partial}{\partial \boldsymbol{w}} \frac{{}^T\boldsymbol{w}S_B\boldsymbol{w}}{{}^T\boldsymbol{w}S_W\boldsymbol{w}} \\
&= \frac{2({}^T\boldsymbol{w}S_W\boldsymbol{w})S_B\boldsymbol{w} - 2({}^T\boldsymbol{w}S_B\boldsymbol{w})S_W\boldsymbol{w}}{({}^T\boldsymbol{w}S_W\boldsymbol{w})^2} = \boldsymbol{0} \\
\therefore ({}^T\boldsymbol{w}S_B\boldsymbol{w})S_W\boldsymbol{w} &= ({}^T\boldsymbol{w}S_W\boldsymbol{w})S_B\boldsymbol{w} \qquad (1)
\end{aligned}$$

今、

$$S_B\boldsymbol{w} = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1, \boldsymbol{w})$$

より、$S_B\boldsymbol{w} \propto \boldsymbol{m}_2 - \boldsymbol{m}_1$ である。また (1) においてスカラー部分 ${}^T\boldsymbol{w}S_B\boldsymbol{w}$, ${}^T\boldsymbol{w}S_W\boldsymbol{w}$ を無視すれば、

$$\begin{aligned}
\boldsymbol{w} &\propto S_W^{-1}S_B\boldsymbol{w} \\
&\propto S_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)
\end{aligned}$$

クラス内共分散が等方的 (きれいな円形であればどの断面で切っても分散は変わらない) であれば S_W は単位行列に比例し、$\boldsymbol{w} \propto \boldsymbol{m}_2 - \boldsymbol{m}_1$ となる。

2 確率的生成モデル

二クラス分類においてデータ $\boldsymbol{x}$ が与えられたときクラス $\mathcal{C}_1$ の事後確率を計算する。

$$\begin{aligned}
p(\mathcal{C}_1|\boldsymbol{x}) &= \frac{p(\mathcal{C}_1 \wedge \boldsymbol{x})}{p(\boldsymbol{x})} \\
&= \frac{p(\mathcal{C}_1 \wedge \boldsymbol{x})}{p(\mathcal{C}_1 \wedge \boldsymbol{x}) + p(\mathcal{C}_2 \wedge \boldsymbol{x})} \\
&= \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_1) + p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \frac{1}{1 + \frac{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}} \qquad (1)
\end{aligned}$$

ここで

$$\begin{aligned}\frac{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)} &= \exp\{\log p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2) - \log p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)\} \\ &= \exp\{-\log \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\}\end{aligned}$$

なので、

$$\begin{aligned}(1) &= \frac{1}{1+\exp(-a)}, \qquad a = \log \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \sigma(a)\end{aligned}$$

このように、二クラス分類の事後確率はロジスティックシグモイド関数 σ で書ける。σ の逆関数：

$$a = \log(\frac{\sigma}{1-\sigma}) = \log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_2|\boldsymbol{x})}$$

をロジット関数という。$p(\mathcal{C}_1|\boldsymbol{x}) \geq p(\mathcal{C}_2|\boldsymbol{x})$ の時 $a > 0$ となり、$0.5 \leq \sigma \leq 1$ となる。

2.1 連続値入力

クラス $\mathcal{C}_k$ の確率密度を

$$\begin{aligned}p(\boldsymbol{x}|\mathcal{C}_k) &= N(\boldsymbol{\mu}_k, \Sigma) \\ &= \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}\ {}^{T}(\boldsymbol{x}-\boldsymbol{\mu}_k)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\}\end{aligned}$$

と定義する。2 クラス分類においてロジット関数は

$$\begin{aligned}a &= \log \frac{p(\boldsymbol{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= -\frac{1}{2}\{\ {}^{T}(\boldsymbol{x}-\boldsymbol{\mu}_1)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) -\ {}^{T}(\boldsymbol{x}-\boldsymbol{\mu}_1)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\} + \log\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= -\frac{1}{2}\{\ {}^{T}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 -\ {}^{T}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 - 2\ {}^{T}\boldsymbol{x}\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\} + \log\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= {}^{T}\boldsymbol{x}\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \frac{1}{2}\ {}^{T}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\ {}^{T}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 + \log\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\end{aligned}$$

となるので、クラス $\mathcal{C}_1$ に対する事後確率は

$$\begin{aligned}p(\mathcal{C}_1|\boldsymbol{x}) &= \sigma(a) \\ &= \sigma(\ {}^{T}\boldsymbol{w}\boldsymbol{x} + w_0)\end{aligned}$$

ただし、

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \quad w_0 = -\frac{1}{2}\ {}^{T}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\ {}^{T}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 + \log\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

3　確率的識別モデル