はじめてのパターン認識
第6章　線形識別関数

山木翔馬

2015/06/16

2クラス問題の線形識別関数

線形識別関数の定義

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

$$\begin{aligned} \boldsymbol{x} &= (x_1, ..., x_d)^T && d\text{次元入力ベクトル} \\ \boldsymbol{w} &= (w_1, ..., w_d)^T && \text{係数ベクトル} \\ w_0 & && \text{バイアス項} \end{aligned}$$

識別境界を $f(\boldsymbol{x}) = 0$ とすれば

識別規則

$$f(\boldsymbol{x}) = \begin{cases} C_1 & (f(\boldsymbol{x}) \geq 0) \\ C_2 & (f(\boldsymbol{x}) < 0) \end{cases}$$

超平面の方程式 (1/3)

識別境界では $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$ より

$$f(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{w}^T\boldsymbol{x} = -w_0$$

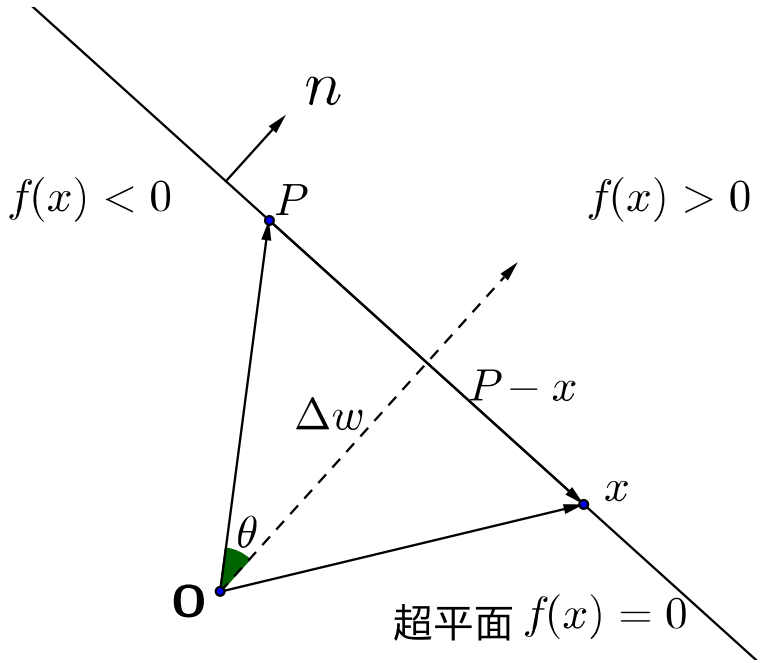
両辺を $\|\boldsymbol{w}\|$ で正規化, $\boldsymbol{n} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \Delta_w = -\frac{w_0}{\|\boldsymbol{w}\|}$

$$\frac{\boldsymbol{w}^T\boldsymbol{x}}{\|\boldsymbol{w}\|} = -\frac{w_0}{\|\boldsymbol{w}\|} \Leftrightarrow \boldsymbol{n}^T\boldsymbol{x} = \Delta_w$$

識別境界

$$f(\boldsymbol{x}) = \boldsymbol{n}^T\boldsymbol{x} - \Delta_w = 0$$

超平面の方程式 (2/3)



識別境界上の任意の点の位置ベクトル $\boldsymbol{P}$ について

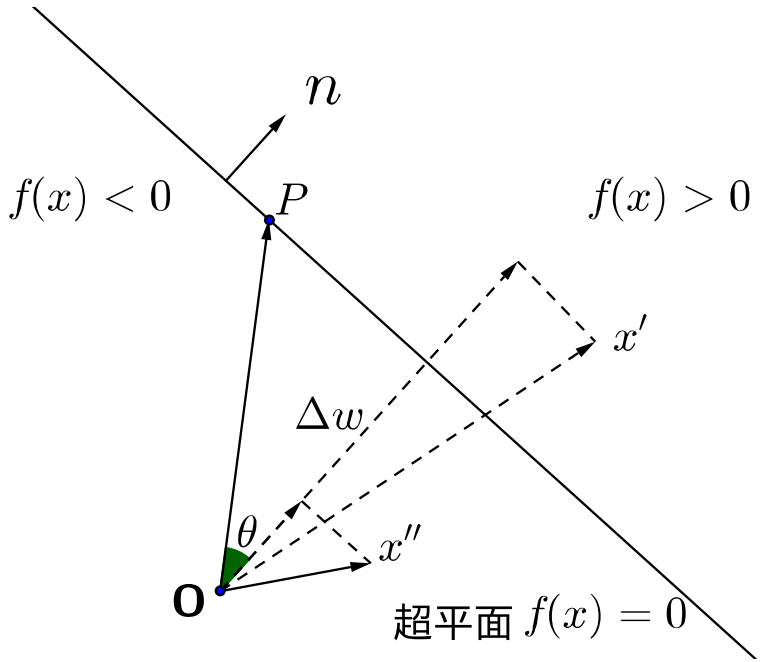
$$f(\boldsymbol{P}) = \boldsymbol{n}^T \boldsymbol{x} - \Delta_w = 0$$

が成り 立つので

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{n}^T \boldsymbol{x} - \Delta_w \\ &= \boldsymbol{n}^T (\boldsymbol{x} - \boldsymbol{P}) = 0 \end{aligned}$$

- $\boldsymbol{x} - \boldsymbol{P} \perp \boldsymbol{n}$
- 識別境界は単位法線ベクトル $\boldsymbol{n}$ をもつ超平面
- $\Delta_w = \boldsymbol{n}^T \boldsymbol{P}$ は原点から識別超平面までの距離
  $(\Delta_w = \boldsymbol{n}^T \boldsymbol{P} = \|\boldsymbol{n}\| \cdot \|\boldsymbol{P}\| cos\theta = \|\boldsymbol{P}\| cos\theta)$
- Δ_w:正規化されたバイアス

超平面の方程式 (3/3)



$$
\begin{aligned}
& \boldsymbol{n}^T \boldsymbol{x'} > \Delta_w \\
\Leftrightarrow \quad & f(\boldsymbol{x'}) = \boldsymbol{n}^T \boldsymbol{x'} - \Delta_w > 0 \\
\Rightarrow \quad & \boldsymbol{x'} \in C_1
\end{aligned}
$$

$$
\begin{aligned}
& \boldsymbol{n}^T \boldsymbol{x''} < \Delta_w \\
\Leftrightarrow \quad & f(\boldsymbol{x''}) = \boldsymbol{n}^T \boldsymbol{x''} - \Delta_w < 0 \\
\Rightarrow \quad & \boldsymbol{x''} \in C_2
\end{aligned}
$$

多クラス問題への拡張

- 一対他 (one-versus-the-rest)
- 一対一 (one-versus-one)

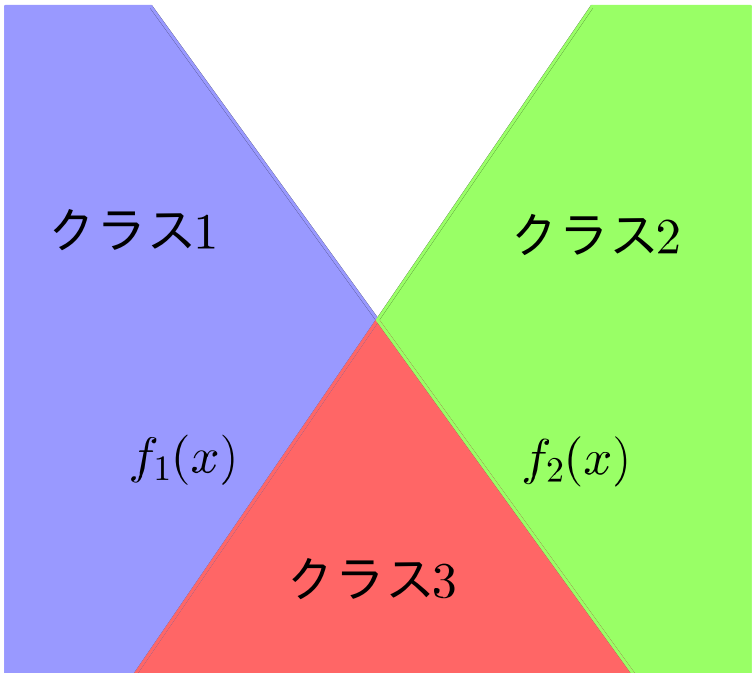
上 2 つの方法では、クラス決定ができない領域がある。
それを解消したのが

- 最大識別関数法

一対他 (one-versus-the-rest)

一つのクラスと他の全てのクラスを識別する $K-1$ 個の
2クラス識別関数 $f_j(\boldsymbol{x})(j=1,\ldots,K-1)$ を用意し

$$\text{識別クラス} = \begin{cases} C_j & (\text{ある } j \text{ について } f_j(\boldsymbol{x}) > 0 \text{ の場合}) \\ C_K & (\text{すべての } j \neq K \text{ について} \\ & f_j(\boldsymbol{x}) < 0 \text{ の場合}) \end{cases}$$

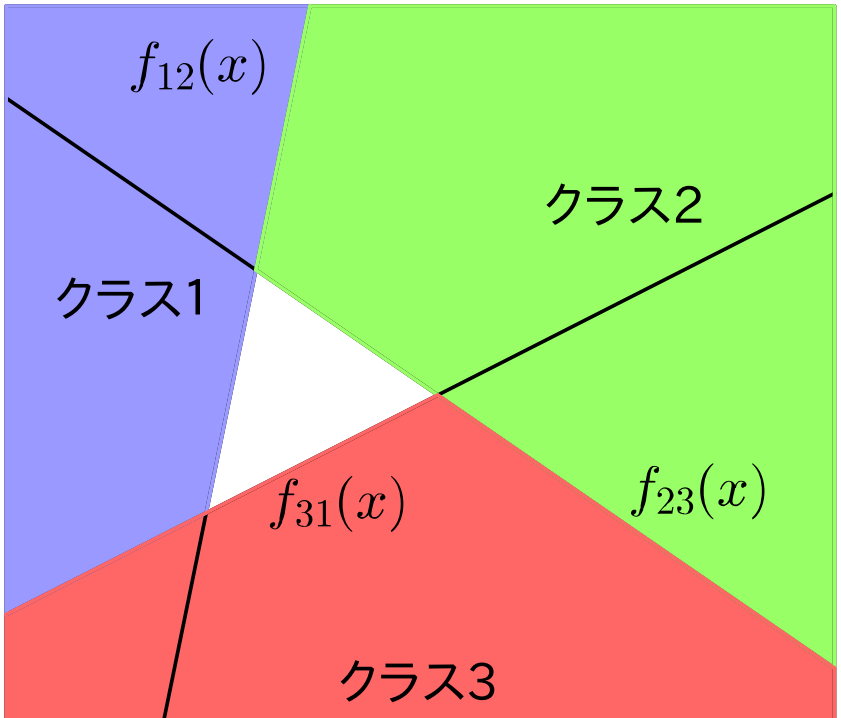


- 複数の識別関数が >0 となる空白領域のクラスが決定できない
- 正のクラスの学習データ数が負のクラスの学習データ数に比べて極端に少なくなる

一対一 (one-versus-one)

クラス i と j を識別する $K(K-1)/2$ 個のクラス識別関数 $f_{ij}(\boldsymbol{x})(1 \leq i < j \leq K)$ を用意し、 $K(K-1)/2$ 個の識別関数の多数決で識別クラスを決める。

$$\begin{cases} f_{ij}(\boldsymbol{x}) > 0 \Rightarrow C_i \text{ に一票} \\ f_{ij}(\boldsymbol{x}) < 0 \Rightarrow C_j \text{ に一票} \end{cases}$$



- 一対他と同じく空白領域が存在
- 手書き数字識別
  10 クラス→識別関数 45 個
  関係のある識別関数 9 個
  過半数を取れない可能性あり

最大識別関数法 (1/2)

K 個の線形識別関数を用意し

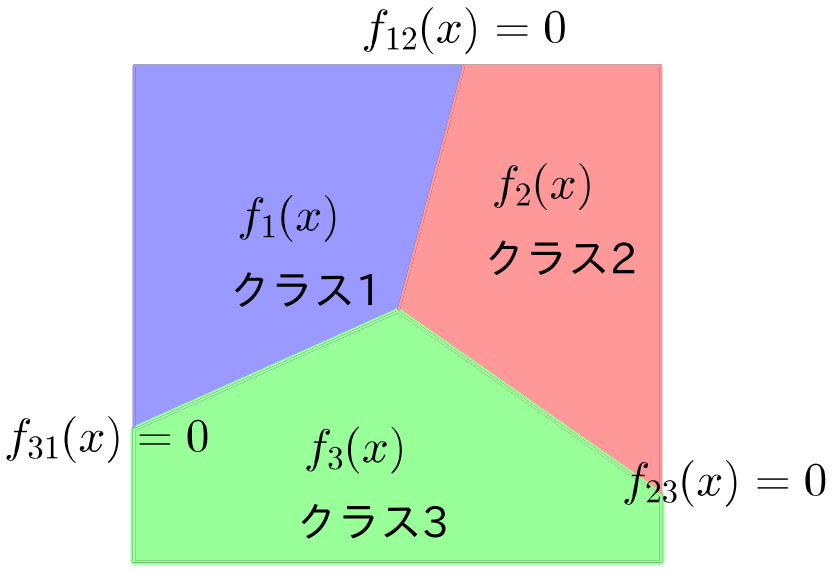
$$\text{識別クラス} = \arg\max_j f_j(\boldsymbol{x}) = \arg\max_j (\boldsymbol{w}_j^T \boldsymbol{x} + w_{j0})$$

クラス i と j の識別境界は $f_i(\boldsymbol{x}) = f_j(\boldsymbol{x})$ となるので

$$f_{ij}(\boldsymbol{x}) = (\boldsymbol{w}_i - \boldsymbol{w}_j)^T \boldsymbol{x} + (w_{i0} - w_{j0}) = 0$$

この式を満たす $K-1$ 個の識別境界ができる。
この識別境界は 2 クラスの場合の識別境界と同じ。

最大識別関数法 (2/2)



- 識別境界上を除けばどの点でも必ずどれかの識別関数が最大となる
  - クラスが決定できない領域が存在しない
- 各クラスの占める領域は単連結で凸となる

単連結

穴の開いていない領域の性質

凸

領域内の任意の 2 点を結ぶ直線上のすべての点が、その領域に含まれている時の性質

最小２乗誤差基準によるパラメータの推定(1/4)

係数ベクトル、i 番目の学習用入力ベクトルはバイアスを含めて

$$\begin{aligned} \boldsymbol{w} &= (w_0, w_1, \ldots, w_d)^T \\ \boldsymbol{x}_i &= (x_{i0} = 1, x_{i1}, \ldots, x_{id})^T \end{aligned}$$

線形識別関数

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + \cdots + w_d x_d = \boldsymbol{w}^T \boldsymbol{x}$$

入力ベクトル $\boldsymbol{x}_i$ が所属するクラスは、教師入力 t_i により

$$t_i = \begin{cases} +1 & (\boldsymbol{x}_i \in C_1) \\ -1 & (\boldsymbol{x}_i \in C_2) \end{cases}$$

のように与えられるものとする。

最小2 乗誤差基準によるパラメータの推定 (2/4)

学習数を N とし、以下を定義.

$$\begin{aligned} \text{データ行列:} \quad \boldsymbol{X} &= (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T \\ \text{教師ベクトル:} \quad \boldsymbol{t} &= (t_i, \ldots, t_N)^T \end{aligned}$$

識別関数の出力値と教師入力の差を 2 乗誤差で評価

評価関数 $E(\boldsymbol{w})$

$$E(\boldsymbol{w}) = \sum_{i=1}^{N} (t_i - f(\boldsymbol{x}_i))^2 = (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})$$

$E(\boldsymbol{w})$ を最小にするパラメータ $\boldsymbol{w}$ は

$$\begin{aligned} \frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} &= -2\boldsymbol{X}^T(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}) = 0 \\ \Rightarrow \hat{\boldsymbol{w}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t} \end{aligned}$$

最小2 乗誤差基準によるパラメータの推定 (3/4)

正規方程式

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}$$

学習データに対する予測値

$$\hat{\boldsymbol{t}} = \boldsymbol{X}\hat{\boldsymbol{w}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}$$

ハット行列

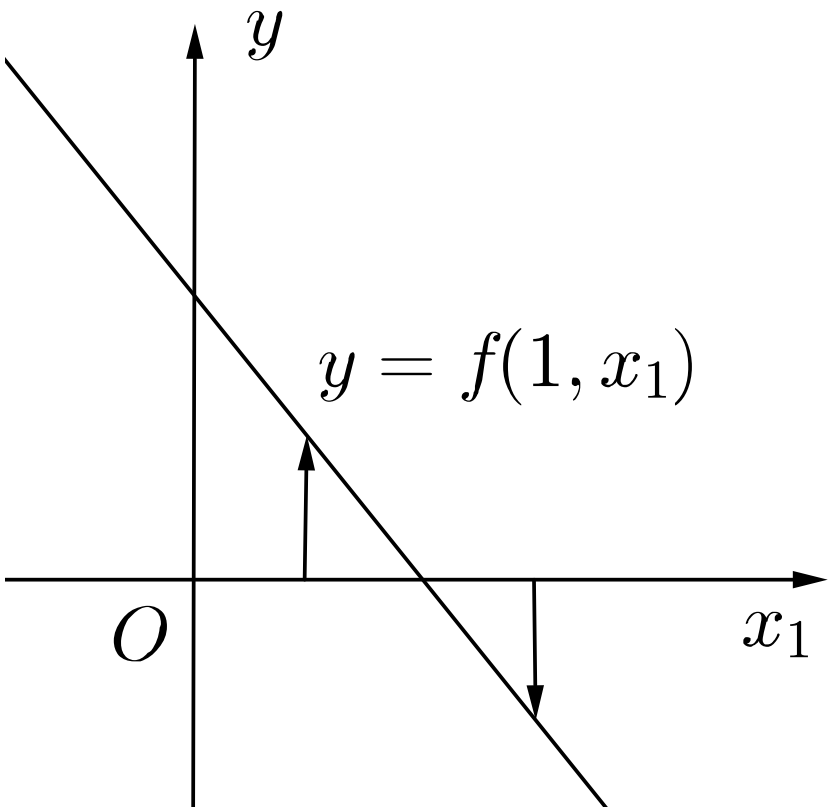
$$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

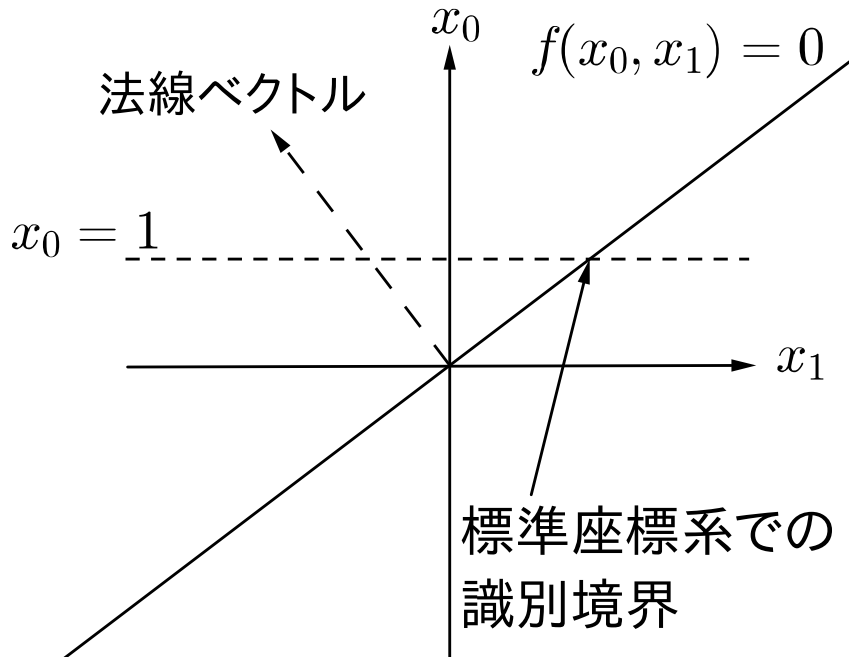
教師データ $\boldsymbol{t}$ を予測値 $\hat{\boldsymbol{t}}$ に変換する行列

最小2 乗誤差基準によるパラメータの推定 (4/4)

標準座標系

x_0 を 1 に固定し、$(x_1, f(1, x_1))$ 平面による表現





同次座標系

x_0 を固定せず独立した座標として扱うことで、2 次元空間 (x_0, x_1) 内の識別関数として表現する

多クラス問題への拡張 (1/2)

最大識別関数法

識別関数

$$f_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} \qquad (k = 1, \ldots, K)$$

N 個の学習データ　$\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x}_N)^T$

教師データの行列　$\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N)^T$,　$\boldsymbol{t_i} = \underbrace{(0, \ldots, 1, \ldots, 0)^T}_{K \text{ 個の要素}}$

2 乗誤差を最小にするパラメータは

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

多クラス問題への拡張 (2/2)

識別関数

$$G(\boldsymbol{x}) = \widehat{\boldsymbol{W}}^T \boldsymbol{x} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)^T \boldsymbol{x} = (g_1(\boldsymbol{x}), \cdots, g_K(\boldsymbol{x}))^T$$

識別規則

$$\text{識別クラス} = \arg\max_j g_j(\boldsymbol{x})$$

複数 (> 2) のクラスが一直線上に並んでいるような分布をしている場合、最大識別関数法ではうまく識別できない.

線形判別分析

- 線形識別関数: d 次元ベクトル $\boldsymbol{x}$ を、ベクトル $\boldsymbol{w}$ 上のスカラー関数 $f(\boldsymbol{x})$ に写像している
- 最小 2 乗誤差法: 教師データにできるだけ忠実になるよう線形識別関数を求めた

線形判別分析

1 次元に写像されたとき、クラス間の分布ができるだけ重ならないような写像方向を見つける.

フィッシャーの線形判別関数 (1/2)

2 クラス (C_1, C_2) 問題について

各クラスの学習データ数	N_1, N_2
線形識別関数	$y = \boldsymbol{w}^T \boldsymbol{x}$
平均ベクトル	$\boldsymbol{\mu}_k = \dfrac{1}{N_k} \displaystyle\sum_{i \in C_k} \boldsymbol{x}_i$
写像	$\boldsymbol{\mu}_k \mapsto m_k = \boldsymbol{w}^T \boldsymbol{\mu}_k$
平均の差	$m_1 - m_2 = \boldsymbol{w}^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

平均の差が大きいほど, クラスの分離が良くなる.

フィッシャーの線形判別関数 (2/2)

クラス間変動	平均の差の 2 乗	$(m_1 - m_2)$
クラス内変動	1 次元に写像後の分散	$S_k^2 = \sum_{i \in C_k} (y_i - m_k)^2$
全クラス内変動		$S_1^2 + S_2^2$

フィッシャーの基準

クラス間変動とクラス内変動の比

$$J(\boldsymbol{w}) = \frac{(m_1 - m_2)^2}{S_1^2 + S_2^2}$$

を最大にする $\boldsymbol{w}$ を見つけること.

多クラス問題への拡張

他クラス $K(>2)$ の場合， $d(>K)$ 次元のデータをたかだか $K-1$ 次元の特徴空間に写像する線形変換行列を 見つける問題になる.

$\Rightarrow$ 識別境界は計算できない.

ロジスティック回帰

線形識別関数 $y = \boldsymbol{w}^T \boldsymbol{x}$ は，識別境界から離れるに従って，関数値の大きさが線形に上昇し続ける．

ロジスティック回帰

関数値を区間 $(0, 1)$ に制限して確率的な解釈を可能にする．

2クラス問題において，クラス C_1 の事後確率 $\mathrm{P}(C_i|\boldsymbol{x})$ は

$$\mathrm{P}(C_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)\mathrm{P}(C_1)}{p(\boldsymbol{x}|C_1)\mathrm{P}(C_1) + p(\boldsymbol{x}|C_2)\mathrm{P}(C_2)}$$

$$a = \ln \frac{p(\boldsymbol{x}|C_1)\mathrm{P}(C_1)}{p(\boldsymbol{x}|C_2)\mathrm{P}(C_2)}$$

とおけば

$$\mathrm{P}(C_1|\boldsymbol{x}) = \frac{1}{1 + \exp(-a)} = \sigma(a)$$

ロジスティック回帰

ロジスティック関数

$\sigma(a)$ をロジスティック関数(シグモイド 関数) と呼び, $y = \sigma(x)$ は無限区間 $(-\infty, \infty)$ を 区間 $(0, 1)$ に写像する圧縮関数.
$\sigma(-a) = 1 - \sigma(a)$ のような対称性を 示す.

ロジスティック関数の逆関数 (ロジット 関数)

$$a = \ln\left(\frac{\sigma(a)}{1-\sigma(a)}\right) = \ln\frac{\mathrm{P}(C_1|\boldsymbol{x})}{\mathrm{P}(C_2|\boldsymbol{x})}$$

オッズ (事後確率の比) $\dfrac{\mathrm{P}(C_1|\boldsymbol{x})}{\mathrm{P}(C_2|\boldsymbol{x})}$

ログオッズ (オッズの対数) $\ln\dfrac{\mathrm{P}(C_1|\boldsymbol{x})}{\mathrm{P}(C_2|\boldsymbol{x})}$

ロジスティック回帰モデル

N 人の喫煙量 $\hat{\boldsymbol{x}} = (x_1, \ldots, x_N)^T$ を観測したとき，肺がんになる確率を

$$p(1|x_1, \ldots, x_N) = f(\boldsymbol{x}) = \frac{1}{1 + \exp(-(w_0 + w_1 x_1 + \cdots + w_N x_N))}$$

で表す. ($\hat{\boldsymbol{x}}$ にはバイアス項に対応する 1 を追加)

$$\begin{aligned} a &= \boldsymbol{w}^T \boldsymbol{x} \\ \Rightarrow \quad f(\boldsymbol{x}) &= \sigma(a) = \frac{1}{1 + \exp(-a)} = \frac{\exp a}{1 + \exp a} \end{aligned}$$

$f(\boldsymbol{x})$ は非線形 $\Rightarrow$ 学習データの線形関数である a をロジスティック関数で非線形変換したモデルで事象を表現していることになる.

ロジスティック回帰モデル

ロジスティック関数の逆関数であるロジット関数は，

$$a = \ln \frac{p(1|\boldsymbol{x})}{1 - p(1|\boldsymbol{x})} = \boldsymbol{w}^T \boldsymbol{x}$$

で，オッズは

$$\frac{p(1|\boldsymbol{x})}{1 - p(1|\boldsymbol{x})} = \frac{p(1|\boldsymbol{x})}{p(0|\boldsymbol{x})} = \exp(\boldsymbol{w}^T \boldsymbol{x})$$

で表される．

ロジスティック回帰モデル

x の中の x_1 が1 増えた状態 $\hat{\boldsymbol{x}} = (1, (x_1 + 1), x_2, \ldots, x_N)$ を考えると，x とのオッズ比は

$$\frac{\frac{p(1|\hat{\boldsymbol{x}})}{1-p(1|\hat{\boldsymbol{x}})}}{\frac{p(1|\boldsymbol{x})}{1-p(1|\boldsymbol{x})}} = \frac{\exp(w_0 + w_1(x_1 + 1) + w_2 x_2 + \cdots + w_N x_N)}{\exp(w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_N x_N)} = \exp w_1$$

x_1 が 1 単位分増えることによるオッズの増加分が $\exp w_1$

パラメータの最尤推定

2クラスロジスティック回帰モデルのパラメータの最尤推定を考える.

確率変数	t	:	モデルの出力
$t=1$ となる確率	$\mathrm{P}(t=1)$	$=$	α
$t=0$ となる確率	$\mathrm{P}(t=0)$	$=$	$1-\alpha$

確率変数 t はパラメータ α をもつベルヌーイ試行

$$f(t|\alpha) = \alpha^t(1-\alpha)^{1-t} \qquad (t=0 \text{ または } 1)$$

に従うので，N 回の試行に基づく尤度関数は

$$L(\alpha_1,\ldots,\alpha_N) = \prod_{i=1}^{N} f(t_i|\alpha_i) = \prod_{i=1}^{N} \alpha^{t_i}(1-\alpha_i)^{(1-t_i)}$$

パラメータの最尤推定

負の対数尤度関数

$$\zeta(\alpha_1, \ldots, \alpha_N) = -\sum_{i=1}^{N}(t_i \ln \alpha_i + (1 - t_i)\ln(1 - \alpha_i))$$

この評価関数は交差エントロピー型誤差関数とよばれる.
最尤推定とは，この交差エントロピー型誤差関数を最小にするパラメータ $\boldsymbol{w}$ を得ることである.

- 解析的に求められない
- 最急降下法やニュートン-ラフソン法などで数値的に解を求めることになる

多クラス問題への拡張

各クラスごとに線形変換

$$a_k = \boldsymbol{w}_k^T \boldsymbol{x} \qquad (k = 1, \ldots, K)$$

を求め，事後確率を

$$\mathrm{P}(C_k|\boldsymbol{x}) = \pi_k(\boldsymbol{x}) = \frac{\exp a_k}{\sum_{j=1}^{K} \exp a_j}$$

で計算して、最大事後確率を与えるクラスに分類する.
この関数をソフトマップ関数という.

多クラス問題への拡張

負の対数尤度関数

$$E(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K) = -\ln p(T|\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K) = -\sum_{i=1}^{K}\sum_{k=1}^{K} t_{ik}\pi_{ik}$$

となる.
各 $\boldsymbol{w}_j$ の最尤推定は評価関数を $\boldsymbol{w}_j$ で微分して 0 とおけば求められる. $\Rightarrow$ 2 クラス問題同様, 解析的には解けない.