PRML 第４章

宮澤　彬

総合研究大学院大学 博士前期

miyazawa-a@nii.ac.jp

July 16, 2015
(modified: May 15, 2015)

はじめに

- 計算過程を別ファイルに分けようと思いましたが面倒なので分けませんでした．そのためスライドがすこしごちゃごちゃしています．すみません．
- このスライドの LuaLaTeX のソースコードは `https://github.com/pecorarista/documents` にあります．
- 自分の好みにより教科書とは違う表記をしている箇所があります．
  - 転置を $^\top$ ではなく $'$ で表しています．
  - $\mathrm{tr}(A'B)$ は内積の公理を満たすので $\langle A, B \rangle$ と書きます．
  - 勾配の代わりに導関数を使っている箇所があります．
  - 誤解の恐れがない限りベクトルを太字で書きません．
  - 左辺を右辺で定義するとき (左辺) $:=$ (右辺) と書きます．

分類問題を解く

入力 x に対して，離散クラス $\mathcal{C}_1, \ldots, \mathcal{C}_k$ のただ 1 つを割り当てたい．つまり入力空間をいくつかの**決定領域** (decision region) に分離したい．

決定領域の境界は**決定境界** (decision boundary)，あるいは**決定面** (decision surface) などと呼ばれる．

しばらくは**線形識別モデル**を考える．

線形識別モデルとは

線形識別モデルは決定面が，x の線形関数であり，D 次元入力空間に対して，その決定面が $D-1$ 次元の超平面で定義されるもの.

線形決定面で正しく各クラスに分類できるデータの集合は**線形分離可能**であると言われる.

目的変数の表し方について

目的変数の表し方はいろいろあるが，2 変数の表し方で一般的なのは目的変数を $t \in \{0, 1\}$ として $t = 1$ で $\mathcal{C}_1$ を，$t = 0$ で $\mathcal{C}_2$ を表す方法である.

$K > 2$ クラスの場合は 1-of-K 表記法を使用する．この表記法では

$$t = (0 \ \cdots \ 0 \ \overset{\overset{i}{\smile}}{1} \ 0 \ \cdots \ 0)' \in \{0, 1\}^K$$

がクラス $\mathcal{C}_i$ に対応する.

分類問題に対する３つのアプローチ

分類問題に対するアプローチは大きく 3 つに分けられる.

1. **識別関数** (discriminant function) を構築する方法.
   - パーセプトロンや SVM など
2. 事後確率 $p(\mathcal{C}_k|x)$ を直接モデル化する方法（**識別モデル**）.
   - ロジスティック回帰モデルなど
3. $p(\mathcal{C}_k)$ と $p(x|\mathcal{C}_k)$ を生成し，これらから $p(\mathcal{C}_k|x)$ を求める方法（**生成モデル**）.
   - ナイーブベイズなど

これらの比較は 1.5.4 で詳しく扱った [1].

1 Murphy (2012) の 8.6 も詳しい.

第３章との違い

第 3 章の線形回帰モデルでは $y : x \in \mathbb{R}^D \mapsto w'x + w_0 \in \mathbb{R}$ のような関数を使ってきた．しかし分類問題では離散値であるクラスラベルを出力してほしいので，非線形関数 f を噛ませる．

$$y(x) = f(w'x + b) \tag{4.3}$$

f は**活性化関数** (activation function) と呼ばれる．f は非線形であっても，決定面は線形であるから，(4.3) で表現されるモデルのクラスを**一般化線形モデル** (generalized linear model) と呼ぶ．

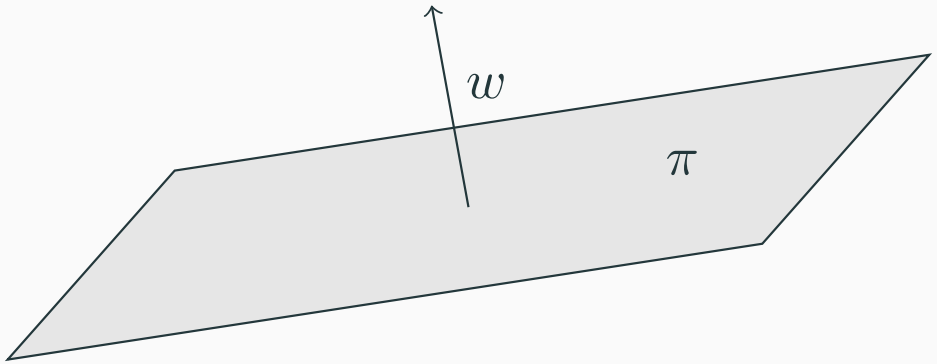
2クラスの線形識別関数

最も簡単な線形識別関数は，入力ベクトル x の線形関数

$$y(x) = w'x + w_0 \tag{4.4}$$

で与えられる．w を**重みベクトル**，w_0 を**バイアスパラメータ**と呼ぶ．

このとき決定境界は超平面 $\pi : y(x) = 0$ となる．



$y(x)$ の性質

$y(x)/\|w\|$ は x から平面までの（符号付き）距離を与える.

証明　Hilbert 空間の射影定理より，任意の点 x から超平面 π への最短距離を与える点 $x_\perp \in \pi$ がただ 1 つ存在し，$(x - x_\perp) \perp \pi$ を満たす．したがって適当な実数 r を使って

$$x - x_\perp = r\frac{w}{\|w\|}$$

と書ける．両辺と w の標準内積をとって，$w'x_\perp + w_0 = 0$ を使えば,

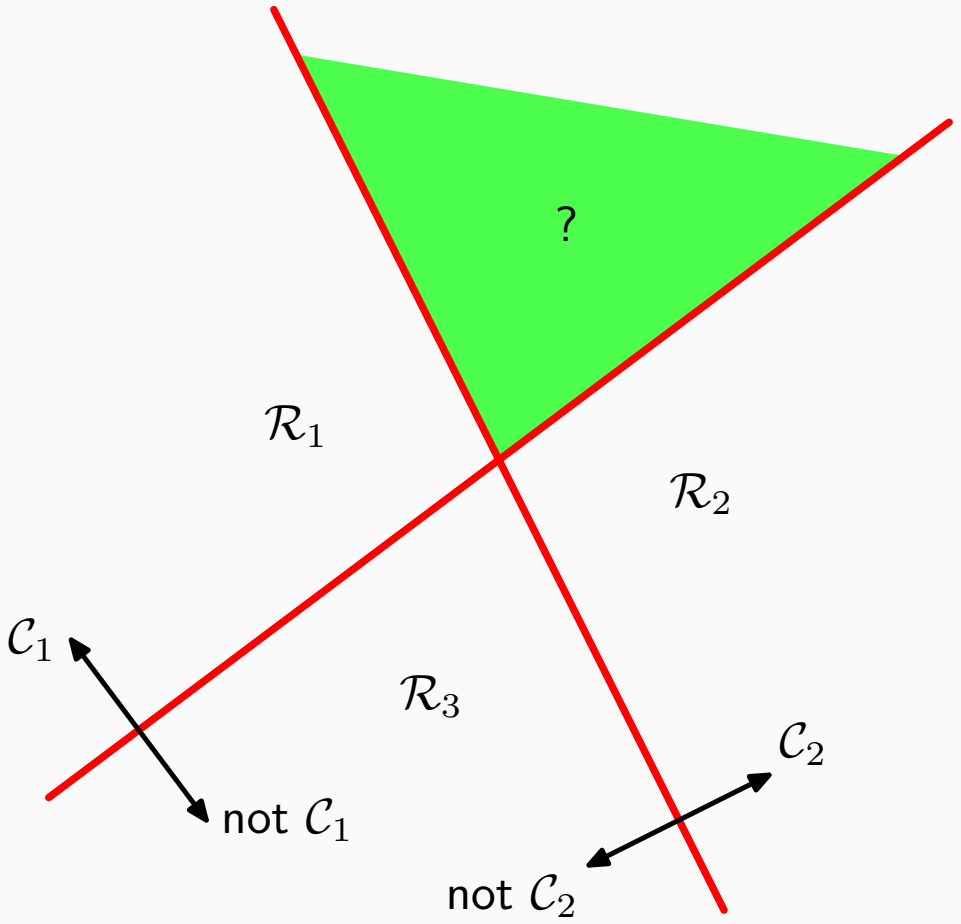
$$w'x - w'x_\perp = r\|w\|$$
$$r = \frac{y(x)}{\|w\|}$$

を得る.

多クラスに拡張すると生じる問題 (1)

$K = 2$ の線形識別を $K > 2$ に拡張する．単純に 2 クラス識別関数を組み合わせると問題が生じる．

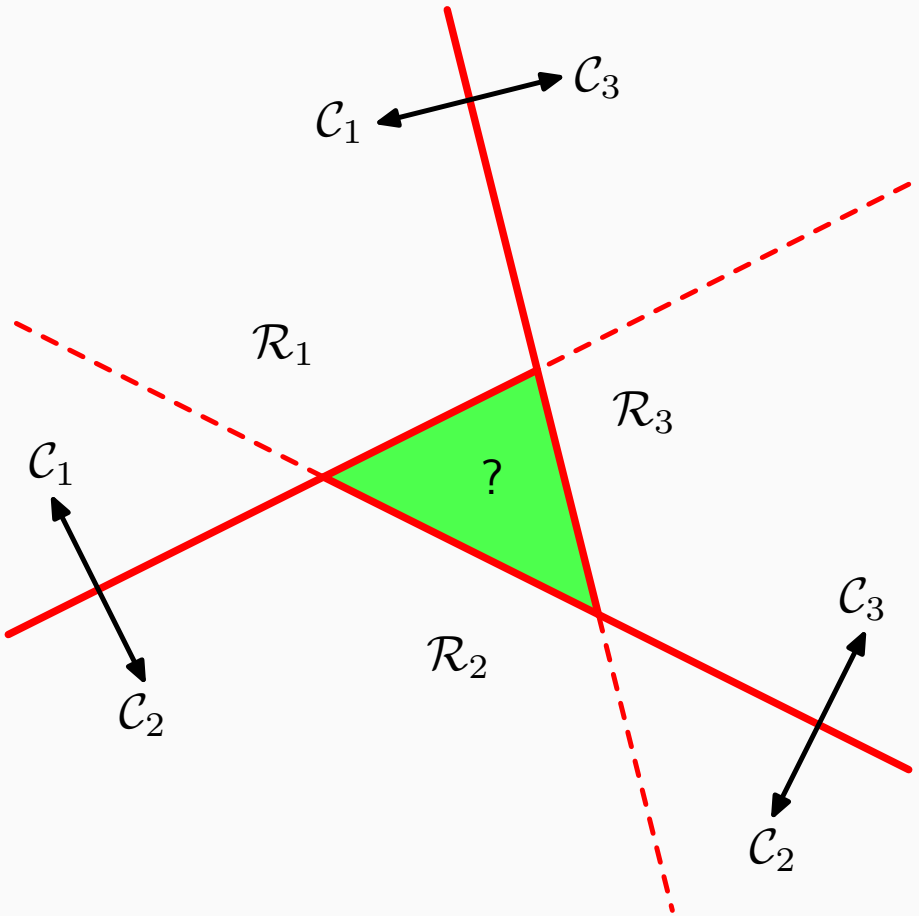
1 対他分類器 (one-versus-the-rest classifier) を使う方法．



多クラスに拡張すると生じる問題 (2)

1 対 1 分類器 (one-versus-on classifier)

$\binom{K}{2} = K(K-1)/2$ 個の 2 クラス識別関数を導入する．各点は識別関数の多数決で分類される．



曖昧な領域をなくしたい (1)

以下の K 個の線形関数

$$y_k(x) = w_k' x + w_{k0} \quad (k = 0, \ldots, K) \tag{4.9}$$

を作り，すべての $j\,(\neq k)$ で

$$y_k(x) > y_j(x)$$

すなわち

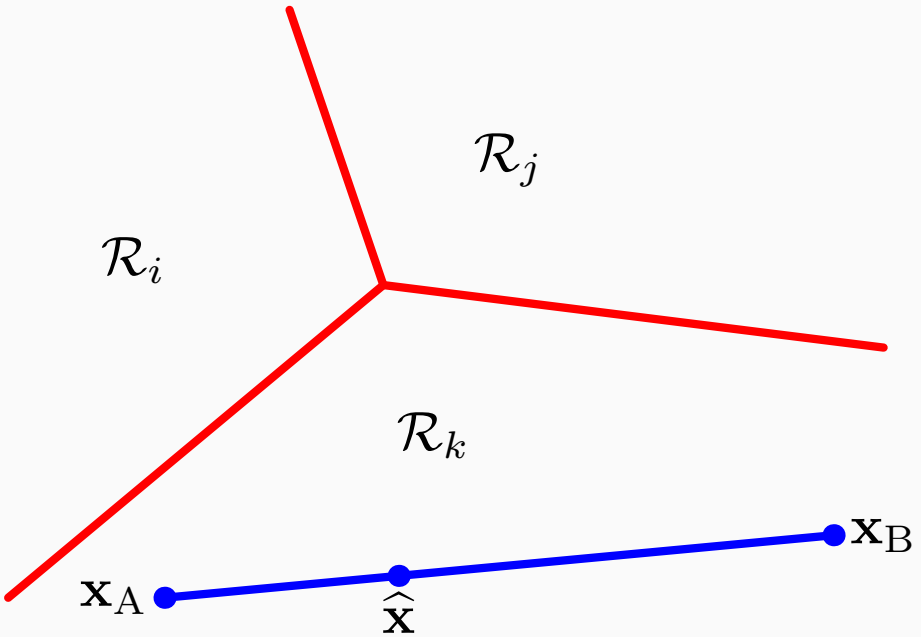
$$(w_k - w_j)' x + (w_{k0} - w_{j0}) > 0$$

が成り立つ場合に x は $\mathcal{C}_k$ に分類されるとする.

曖昧な領域をなくしたい (2)

前のページで紹介した基準で「謎の領域」が存在しないことを示す.

任意の点 x をとる. $I := \arg\max_i \{y_i(x)\} (\neq \emptyset)$ としたとき $|I| = 1$ ならば $\mathcal{C}_{i_0}$ $(i_0 \in I)$ に一意に分類できる. $|I| > 1$ ならば, 任意の異なる $i, j \in I$ について $y_i(x) = y_j(x)$ となっているので x が $\mathcal{C}_i$ と $\mathcal{C}_j$ の決定境界上にあることが分かる.



決定領域の凸性

決定領域は凸集合である.

証明　$x_{\mathrm{A}}, x_{\mathrm{B}}$ が共に $\mathcal{C}_k$ に分類されているとする. これらの凸結合 $\hat{x} := \lambda x_{\mathrm{A}} + (1-\lambda) x_{\mathrm{B}}$ について

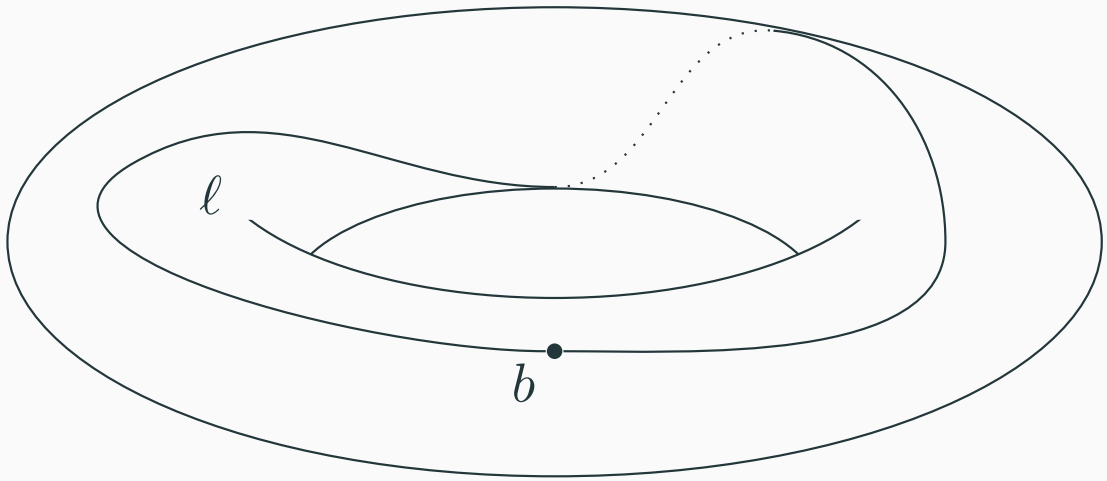
$$
\begin{aligned}
&(w_k - w_j)' \hat{x} + (w_{k0} - w_{j0}) \\
&= \lambda (w_k - w_j)' x_{\mathrm{A}} + \lambda (w_{k0} - w_{j0}) \\
&\qquad + (1-\lambda)(w_k - w_j)' x_{\mathrm{B}} + (1-\lambda)(w_{k0} - w_{j0}) > 0
\end{aligned}
$$

が成り立つので $\hat{x}$ も $\mathcal{C}_k$ に分類される. ∎

決定領域の単連結性 (1)

決定領域 $\mathcal{R}_k\,(k = 1, \ldots, K)$ は弧状連結（任意の 2 点をとったとき，それら結ぶ連続曲線が存在する）であり，更に単連結（ループを連続的に変形して 1 点に縮められる）である.

弧状連結であるが単連結ではない



決定領域の単連結性 (2)

証明　凸性から弧状連結であることは明らかである.
$\mathcal{R}_k$ の任意の点 p_0 をとり，$\ell : [0, 1] \to \mathcal{R}_k$ を p_0 を基点とした $\mathcal{R}_k$ 内の任意のループとする．p_0 に留まり続けるループを $\tilde{p}_0$ と表すことにする.
このとき

$$H(t, s) := s\tilde{p}_0(t) + (1 - s)\ell(t)$$

と定めれば，凸性から任意の (t, s) で $H(t, s) \in \mathcal{R}_k$ である．H は

$$H(0, s) = H(1, s) = p_0,$$
$$H(t, 1) = \ell(t),\ H(t, 1) = \tilde{p}_0(t)$$

を満たしているので $\ell \simeq \tilde{p}_0$ であり，したがって $\mathcal{R}_k$ は単連結である.

最適なパラメータを求める（最小二乗法）(1)

まずは二乗和誤差を最小化する方法を使ってみる．識別関数は

$$y_k(x) = w_k' x + w_{k0}, \quad k = 1, \ldots, K$$

と表されている．これはベクトルを使って

$$y(x) = \widetilde{W}' \tilde{x}$$

と書ける．ただし，

$$\widetilde{W} := \begin{pmatrix} w_{10} & w_{20} & \cdots & w_{K0} \\ w_{11} & w_{21} & \cdots & w_{K1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1D} & w_{2D} & \cdots & w_{KD} \end{pmatrix}, \tilde{x} := \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_D \end{pmatrix}, x_0 := 1$$

とした．

二乗和誤差の最小化 (1)

学習データ $\{(x_n, t_n)\}_{n=1, \ldots, N}$ が与えられたとして,

$$\widetilde{X} := \begin{pmatrix} x'_1 \\ \vdots \\ x'_N \end{pmatrix}, \quad \widetilde{T} := \begin{pmatrix} t'_1 \\ \vdots \\ t'_N \end{pmatrix}$$

と定める. このとき二乗和誤差関数は

$$\begin{aligned} E_{\mathrm{D}}(\widetilde{W}) &:= \frac{1}{2} \operatorname{tr}((\widetilde{X}\widetilde{W} - \widetilde{T})'(\widetilde{X}\widetilde{W} - \widetilde{T})) \\ &= \frac{1}{2} \|\widetilde{X}\widetilde{W} - \widetilde{T}\|^2 \end{aligned}$$

である.

二乗和誤差関数の最小化 (2)

$E_{\mathrm{D}}(\widetilde{W})$ を微分する．成分ごとに偏微分してもよいが Fréchet 導関数を求めるほうが簡単である．

$$\begin{aligned}
&\langle \widetilde{X}(\widetilde{W}+H)-\widetilde{T}, \widetilde{X}(\widetilde{W}+H)-\widetilde{T}\rangle - \langle \widetilde{X}\widetilde{W}-\widetilde{T}, \widetilde{X}\widetilde{W}-\widetilde{T}\rangle \\
&= 2\langle \widetilde{X}\widetilde{W}-\widetilde{T}, \widetilde{X}\widetilde{H}\rangle + \|\widetilde{X}H\|^2 \\
&= 2\langle \widetilde{X}'(\widetilde{X}\widetilde{W}-\widetilde{T}), H\rangle + \|\widetilde{X}H\|^2
\end{aligned}$$

より $DE_{\mathrm{D}}(\widetilde{W})(H) = \langle \widetilde{X}'(\widetilde{X}\widetilde{W}-\widetilde{T}), H\rangle$ である．したがって最小値を与える $\widetilde{W}$ は

$$\widetilde{X}'(\widetilde{X}\widetilde{W}-\widetilde{T}) = O$$

の解である．これを解いて

$$\widetilde{W} = (X'X)^{-1}\widetilde{X}'T = \widetilde{X}^{\dagger}T$$

を得る．

最小二乗解の実用性

以下のような問題がある.

1. 外れ値に敏感である（頑健でない）

   決定境界から遠く離れた「正しすぎる」予測にペナルティを与えてしまう.
   → 7.1.2 節で別の誤差関数を紹介

2. 2 値目的変数と，最小二乗法が仮定するガウス分布との相性の悪さ

   最小二乗法は条件付き確率分布にガウス分布を仮定した場合の最尤法であり，一方 2 値目的変数ベクトルは明らかにガウス分布からかけ離れているので，最小二乗法が使えないのは当たり前のことである.
   →適切な確率モデルを採用すれば，最小二乗法よりもよい特性を持つ分類法が得られる.

フィッシャーの線形判別 (1)

線形識別モデルは，D 次元の入力ベクトルを 1 次元空間に写すので情報の損失が発生する.

損失が起こるのは仕方ないが，D 次元空間では分離されていたクラスが 1 次元空間で重なり合ってしまうことは極力避けたい.

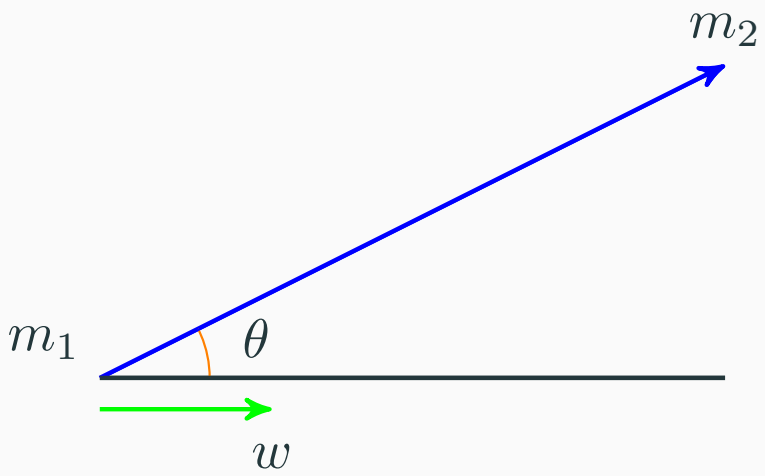
そこでクラスの平均を結ぶ直線の正射影の長さが最大になるような w をとる.

フィッシャーの線形判別 (2)

$\|w\| = 1$, $m_k := (1/N_k) \sum_{n \in \mathcal{C}_k} x_n$ とすると，正射影の長さ ℓ は

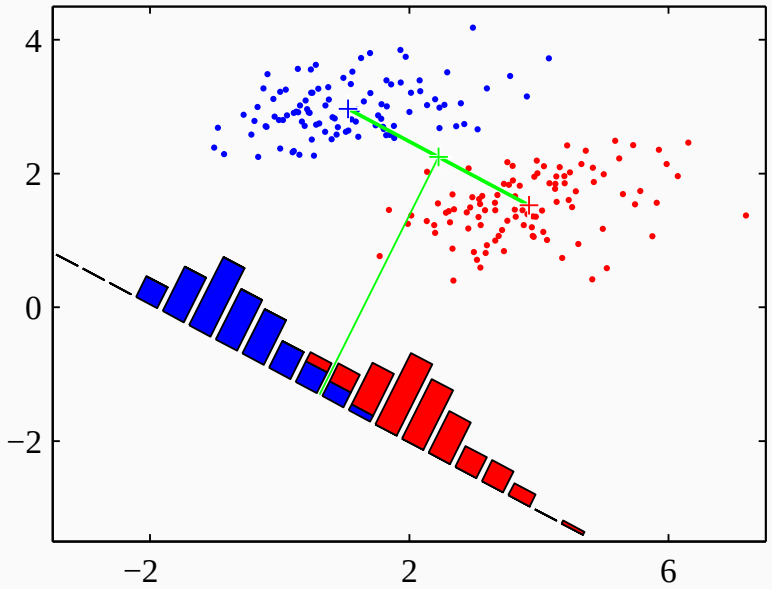
$$\begin{aligned} \ell &= \left| \|m_2 - m_1\| \cos\theta \right| \\ &= |w'(m_2 - m_1)| \end{aligned}$$

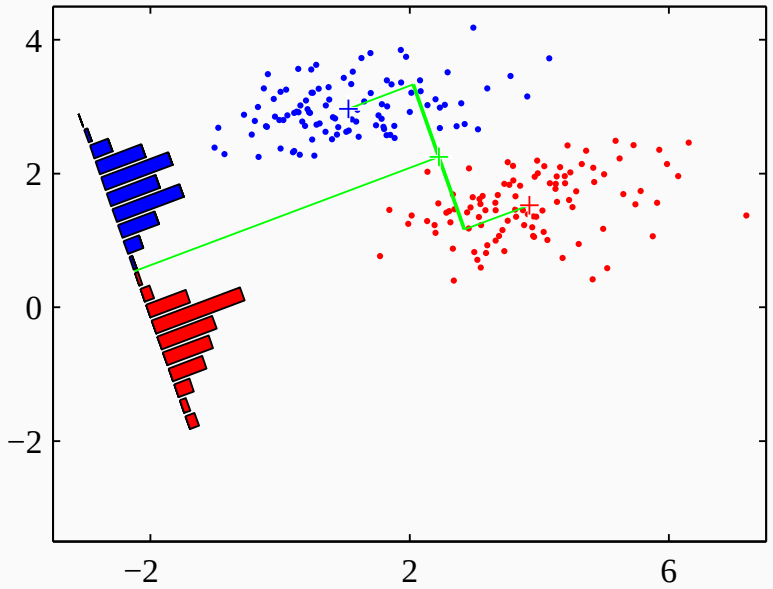
である．Cauchy-Schwarz の不等式より ℓ は $w /\!\!/ (m_2 - m_1)$ のとき最大になる．



フィッシャーの判別基準 (1)

クラスの平均が離れたところに射影されても，クラス全体が重なり合うように射影されてしまっては意味がない．





右の図のようになってほしい．

フィッシャーの判別基準 (2)

同一クラス内にあるデータは射影された後もまとまっていてほしいので

$$s_k^2 := \sum_{n \in \mathcal{C}_k} (w' x_n - w' m_k)^2 \tag{4.24}$$

を小さく抑えなければならない.

以上の考察をまとめると

$$J(w) := \frac{(w' m_2 - w' m_1)^2}{s_1^2 + s_2^2} \tag{4.25}$$

を最大化すればよいことが分かる. $J(w)$ はフィッシャーの判別基準と呼ばれる.

フィッシャーの判別基準の最大化 (1)

クラス間共分散行列 (between-class covariance matrix) [2]

$$S_{\mathrm{B}} = (m_2 - m_1)(m_2 - m_1)' \tag{4.27}$$

と総**クラス内共分散行列** (within-class covariance matrix)

$$S_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (x_n - m_1)(x_n - m_1)' + \sum_{n \in \mathcal{C}_2} (x_n - m_2)(x_n - m_2)' \tag{4.28}$$

を使って $J(w)$ を書き直すと

$$J(w) = \frac{w' S_{\mathrm{B}} w}{w' S_{\mathrm{W}} w} \tag{4.26}$$

となる [3].

2 フィッシャー判別基準に関してぐぐってみたところどちらかというと“covariance”ではなく“scatter”と呼ばれているほうが多い印象を受けた.
3 レイリー商 (Rayleigh quotient) の形だ.

フィッシャーの判別基準の最大化 (2)

J を微分すると

$$DJ(w) = (2w'S_{\mathrm{B}}(w'S_{\mathrm{W}}w) - 2w'S_{\mathrm{B}}w(w'S_{\mathrm{W}}))\big/(w'S_{\mathrm{W}}w)^2$$

であるから，$DJ(w) = O$ として

$$(w'S_{\mathrm{W}}w)S_{\mathrm{B}}w = (w'S_{\mathrm{B}}w)S_{\mathrm{W}}w \tag{4.29}$$

を解く．向きさえ分かればよいので $w'S_{\mathrm{B}}w = 1$, $w'S_{\mathrm{W}}w = 1$ となるように w を置き換えていくと

$$S_{\mathrm{W}}^{-1}S_{\mathrm{B}}w = w$$

となる．$S_{\mathrm{B}}w = (m_2 - m_1)((m_2 - m_1)'w) /\!/ (m_2 - m_1)$ より

$$w /\!/ S_{\mathrm{W}}^{-1}(m_2 - m_1) \tag{4.30}$$

を得る．(4.30) は**フィッシャーの線形判別** (Fisher's linear discriminant) と呼ばれる．

最小二乗との関連 (1)

この節では，前節で見たフィッシャーの判別基準が最小二乗の特殊な場合になっていることを示す.

この節では目的変数 t は 1-of-K はなく，$t = N/N_1$ で $\mathcal{C}_1$ を，$t = -N/N_2$ で $\mathcal{C}_2$ を表すことにする.

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

w_0 と w について微分して，その導関数を 0 とすると

$$\sum_{n=1}^{N}(w'x_n + w_0 - t_n) = 0 \tag{4.32}$$

$$\sum_{n=1}^{N}(w'x_n + w_0 - t_n)x_n = 0 \tag{4.33}$$

が得られる.

最小二乗との関連 (2)

(4.32) からバイアスが

$$\sum_{n=1}^{N} w' x_n + \sum_{n=1}^{N} w_0 - \sum_{n=1}^{N} t_n = 0$$

$$w' \left(\sum_{n=1}^{N} x_n \right) + N w_0 - N_1 \frac{N}{N_1} - N_2 \frac{N}{N_2} = 0$$

$$w_0 = -w' m \tag{4.34}$$

を満たすことが分かる．ただし

$$m := \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{1}{N} (N_1 m_1 + N_2 m_2) \tag{4.36}$$

とした．

最小二乗との関連 (3)

(4.33) を変形すると（演習 4.6）

$$\left(S_{\mathrm{W}} + \frac{N_1 N_2}{N} S_{\mathrm{B}}\right) = N(m_1 - m_2) \tag{4.37}$$

となる．$S_{\mathrm{B}} w = (m_2 - m_1)((m_2 - m_1)' w) /\!/ (m_2 - m_1)$ より

$$w /\!/ S_{\mathrm{W}}^{-1}(m_2 - m_1) \tag{4.38}$$

である．

このように最小二乗から出発してフィッシャーの線形判別を導出することができた．

多クラスにおけるフィッシャーの判別 (1)

フィッシャーの線形判別を $K > 2$ の場合に一般化する．入力空間の次元を $D > K$ とし，D' 次元の特徴量のベクトル $y = W'x$ に変換する．クラス内共分散行列は

$$S_{\mathrm{W}} := \sum_{k=1}^{K} S_k \tag{4.40}$$

とする．ここで

$$S_k := \sum_{n \in \mathcal{C}_k} (x_n - m_k)(x_n - m_k)', \tag{4.41}$$

$$m_k := \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} x_k \tag{4.42}$$

である．

多クラスにおけるフィッシャーの判別 (2)

クラス間共分散行列 S_{B} は

$$S_{\mathrm{T}} := \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)' \tag{4.46}$$

とする．ここで

$$m := \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{1}{N} \sum_{k=1}^{K} N_k m_k \tag{4.44}$$

である．

多クラスにおけるフィッシャーの判別 (2)

S_{W} と S_{B} を D' 次元の特徴空間に射影して

$$s_{\mathrm{W}} = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} (y_n - \mu_k)(y_n - \mu_k)' \tag{4.47}$$

$$s_{\mathrm{B}} = \sum_{k=1} N_k (\mu_k - \mu)(\mu_k - \mu)'$$

とする．ここで

$$\mu_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} y_n, \quad \mu = \frac{1}{N} \sum_{k=1} N_k \mu_k \tag{4.49}$$

である．

多クラスにおけるフィッシャーの判別 (3)

2 クラスのときと同様，クラス間共分散行列が大きく，クラス内共分散が小さいときに大きくなるスカラーを構成する．そのようなスカラーの 1 つの例は

$$J(W) = \mathrm{tr}(s_{\mathrm{W}}^{-1} s_{\mathrm{B}}) \tag{4.50}$$

$$= \mathrm{tr}((W S_{\mathrm{W}} W')^{-1} (W S_{\mathrm{B}} W')) \tag{4.51}$$

である．

(4.51) の最大化については Fukunaga(1990) に書かれているらしい [4]．結論として，最適な W は，一般化された固有方程式

$$S_{\mathrm{B}} v = \lambda S_{\mathrm{W}} v$$

を解いて得られる固有ベクトルにより決定される．

4 見てない．

フィッシャーの線形判別の制限

フィッシャー線形判別の制限について説明するため S_{B} の次元を求めておく.

ベクトル $m_k - m$ の第 i 成分を c_{ki} と書く. このとき S_{B} の第 i 列は

$$\sum_{k=1}^{K} N_k c_{ki}(m_k - m)$$

となるので列空間 $C(S_{\mathrm{B}})$ は $m_1 - m, \ldots, m_K - m$ で張られる. よって $\operatorname{rank} S_{\mathrm{B}} = \dim C(S_{\mathrm{B}}) \leq K$ である. 更に m は $N_k m_k \, (k = 1, \ldots, K)$ の重心であるから

$$N_1(m_1 - m) + \cdots + N_K(m_K - m) = 0$$

となり, $\operatorname{rank} S_{\mathrm{B}} = \dim C(S_{\mathrm{B}}) \leq K - 1$ が分かる. したがって, $S_{\mathrm{W}}^{-1} S_{\mathrm{B}}$ の固有空間は高々 $K - 1$ 次元であり, W で特徴をたくさん作ったところで $K - 1$ 個の特徴しか活用できないことになる.

パーセプトロン

線形識別モデルのもう 1 つの例として**パーセプトロン**が挙げられる.

2 クラスのモデルで入力ベクトル x を特徴ベクトルに変換する非線形関数 ϕ を用いて

$$y(x) = f(w'\phi(x)) \tag{4.52}$$

という形の一般化線形モデルを用いる. ここで f は符号関数

$$f(a) := \begin{cases} +1 & \text{if } a \geq 0, \\ -1 & \text{if } a < 0 \end{cases} \tag{4.53}$$

である.

パーセプトロンを使う場合は目的変数を $t \in \{0, 1\}$ をとし, $t = +1$ を $\mathcal{C}_1$ に, $t = -1$ を $\mathcal{C}_2$ に対応させる.

パーセプトロン基準 (1)

誤差関数を，誤識別したパターンの総数として最小化することで求められそうである．しかし，これでは誤差関数が w の区分的な定数関数になってしまい簡単なアルゴリズムが構築できない．そこで**パーセプトロン基準** (perceptron criterion) という誤差関数を考える．

パラメータ w を

- x_n がクラス $\mathcal{C}_1$ に分類されるならば $w'\phi(x_n) > 0$
- x_n がクラス $\mathcal{C}_2$ に分類されるならば $w'\phi(x_n) < 0$

となるようにとる．そのような w はすべての n で $t_n w'\phi(x_n) > 0$ を満たす．w を求めるため，正しく分類されたパターンに対しては何もせず，誤分類されたパターンに対しては $-t_n w'\phi(x_n)$ を最小化することにする．

パーセプトロン基準 (2)

パーセプトロン基準は

$$E_{\mathrm{P}}(w) = -\sum_{n \in \mathcal{M}} t_n w' \phi_n \tag{4.54}$$

で与えられる．ここで $\phi_n = \phi(x_n)$ であり，$\mathcal{M}$ は誤識別された (misclassified) パターン全体の集合である．最適なパラメータの候補 $w^{(\tau)}$ を以下の式を使って更新していく．

$$\begin{aligned} w^{(\tau+1)} &:= w^{(\tau)} - \eta \nabla E_{\mathrm{P}}(w) \\ &= w^{(\tau)} + \eta t_n \phi_n \end{aligned} \tag{4.55}$$

w を定数倍しても (4.52) は変化しないので，$\eta = 1$ としてよい．

パーセプトロンの収束定理

(4.55) の両辺と $-t_n\phi_n$ との標準内積をとると

$$-t_n w^{(\tau+1)\prime}\phi_n = -t_n w^{(\tau)\prime}\phi_n - \|t_n\phi_n\|^2 < -t_n w^{(\tau)\prime}\phi_n \qquad (4.56)$$

となるので更新により誤差が減少することが分かる.

パーセプトロンの収束定理から，学習データ集合が線形分離可能ならばパーセプトロンの学習アルゴリズムは有限回の繰り返しで厳密解（正しく分離できる w）に収束することが分かる [5].

5 証明は簡単だが省略する．たとえば海野ら (2015) を参照.

パーセプトロンの問題点

1. パーセプトロン学習アルゴリズムが収束するのに必要な繰り返し回数が多い．→分離できない問題なのか，単に収束が遅いのかの区別がわかりにくい．

2. 線形分離可能な場合でも，パラメータの初期値やデータの提示順に依存して様々な解に収束してしまう．

3. 線形分離不可能な場合に収束しない．

4. 確率的な出力を提供しない．

5. $K > 2$ クラスの場合への一般化が容易でない．

確率的生成モデル（2 クラス）

尤度 $p(x|\mathcal{C}_k)$ と事前確率 $p(\mathcal{C}_k)$ を使って事後確率 $p(\mathcal{C}_k|x)$ を生成するモデルを考える.

クラス $\mathcal{C}_1$ に対する事後確率は

$$p(\mathcal{C}_1|x) = \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_1)p(\mathcal{C}_1) + p(x|\mathcal{C}_2)p(\mathcal{C}_2)}$$

である. ロジスティックシグモイド関数 $\sigma(x) := 1/(1 + \exp(-x))$ を使って書き直しておくと便利なので書き直してみると

$$p(\mathcal{C}_1|x) = \sigma(a) \tag{4.57}$$

となる. ただし

$$a := \log \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{4.58}$$

とした.

確率的生成モデル（多クラス）

K クラスの場合，事後確率 $p(\mathcal{C}_k|x)$ は

$$p(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(x|\mathcal{C}_j)p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{4.62}$$

で与えられる．ただし

$$a_k := \log(p(x|\mathcal{C}_k)p(\mathcal{C}_k)) \tag{4.63}$$

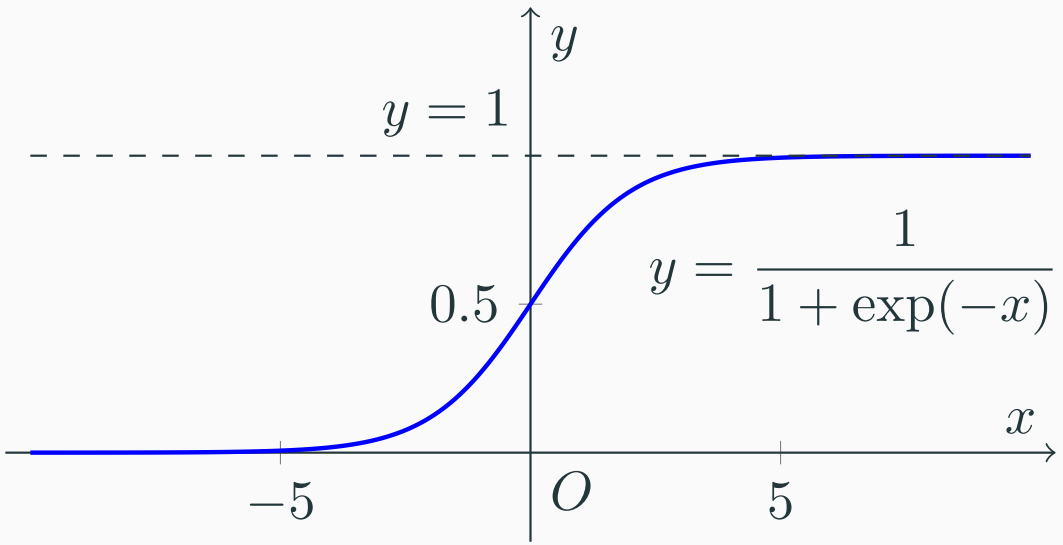
とした．(4.62) は**正規化指数関数** (normalized exponential function) または**ソフトマックス関数** (softmax function) と呼ばれる．

ソフトマックス関数の性質

正規化指数関数はロジスティックシグモイド関数の多変数への一般化と考えられる.

またこの関数の重要な性質として，任意の $j \neq k$ に対して $a_k \gg a_j$ となるような j について $p(\mathcal{C}_k|x) \simeq 1$ かつ $p(\mathcal{C}_j|x) \simeq 0$ となることがある. したがってこの関数を滑らかな最大値関数と捉えられ，そのことがソフトマックス関数と呼ばれる所以になっている.

ロジスティックシグモイド関数



連続値入力（2 クラス）(1)

仮定

1. クラスの条件付き確率密度がガウス分布である.
2. すべてのクラスが同じ共分散行列を共有する.

このとき尤度 $p(x|\mathcal{C}_k)$ は

$$p(x|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)\right) \qquad (4.64)$$

で与えられる.

連続値入力（2 クラス）(2)

クラス $\mathcal{C}_1$ の事後確率は

$$
\begin{aligned}
& p(\mathcal{C}_1|x) \\
&= \sigma\left(\log \frac{\exp\left(-(1/2)(x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right) p(\mathcal{C}_1)}{\exp\left(-(1/2)(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right) p(\mathcal{C}_2)}\right) \\
&= \sigma\left(\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \log\frac{\mathcal{C}_1}{\mathcal{C}_2}\right) \\
&= \sigma(w'x + w_0) \qquad (4.65)
\end{aligned}
$$

となる．ここで

$$
w := \Sigma^{-1}(\mu_1 - \mu_2) \qquad (4.66)
$$

$$
w_0 := \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1 + \log\frac{\mathcal{C}_1}{\mathcal{C}_2} \qquad (4.67)
$$

とした．

連続値入力（多クラス）

K クラスの場合を考える．このとき事後確率 $p(\mathcal{C}_k|x)$ は，

$$a_k(x) := w_k' x + w_{k0}, \tag{4.68}$$

$$w_k := \Sigma^{-1}\mu_k, \tag{4.69}$$

$$w_{k0} := \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log p(\mathcal{C}_k) \tag{4.70}$$

として

$$p(\mathcal{C}_k|x) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

で与えられる．

連続値入力のベイズ判別における決定境界

決定境界を $P(\mathcal{C}_i|x) \geq P(\mathcal{C}_j|x)$ で定めると，(4.65) から入力 x の線形関数になっていることが分かる.

共分散行列が共通であるという仮定を外すと，決定境界は x の 2 次関数となる.

最尤解（2 クラス）(1)

尤度 $p(x|\mathcal{C}_k)$ と事前確率 $p(\mathcal{C}_k)$ をパラメトリックな関数形で表現しておき，パラメータの最尤量を求めることを考える.

前節と同じく，各クラスの条件付き確率密度はガウス分布で，共分散行列は共通であると仮定する．またデータ $\{(x_n, t_n)\}_{n=1, \ldots, N}$ が与えられているとする．$t = 1$ で $\mathcal{C}_1$ を，$t = 0$ で $\mathcal{C}_2$ を表すことにする.

事前確率を $p(\mathcal{C}_1) = \alpha$, $p(\mathcal{C}_2) = 1 - \alpha$ とする．このとき,

$$
\begin{aligned}
p(x_n, \mathcal{C}_1) &= p(x_n|\mathcal{C}_1)p(\mathcal{C}_1) = \alpha \mathcal{N}(x_n|\mu_1, \Sigma), \\
p(x_n, \mathcal{C}_2) &= p(x_n|\mathcal{C}_2)p(\mathcal{C}_2) = (1 - \alpha)\mathcal{N}(x_n|\mu_2, \Sigma)
\end{aligned}
$$

である.

最尤解（2 クラス）(2)

尤度は以下のように計算できる.

$$
\begin{aligned}
&p(t, X|\alpha, \mu_1, \mu_2, \Sigma) \\
&= \prod_{n=1}^{N} (\alpha \mathcal{N}(x_n|\mu_1, \Sigma))^{t_n} ((1-\alpha)\mathcal{N}(x_n|\mu_2, \Sigma))^{1-t_n} \qquad (4.71)
\end{aligned}
$$

対数尤度を $\ell(\alpha, \mu, \Sigma)$ とすると

$$
\frac{\partial \ell(\alpha, \mu, \Sigma)}{\partial \alpha} = \frac{N_1}{\alpha} - \frac{N - N_1}{1-\alpha}
$$

となるので $\partial \ell / \partial \alpha = 0$ として

$$
\alpha = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} \qquad (4.73)
$$

を得る.

最尤解（2 クラス）(3)

次に μ_1 の最尤量を求める．ℓ の μ_1 に関係する項を取り出すと

$$\begin{aligned}
&\sum_{n=1}^{N} t_n \log \mathcal{N}(x_n|\mu_1, \Sigma) \\
&= -\frac{1}{2}\sum_{n=1}^{N} t_n (x_n - \mu_1)' \Sigma^{-1} (x_n - \mu_1) + \text{const.} \qquad (4.74)
\end{aligned}$$

となっているので

$$D_{\mu_1}\ell(\mu_1) = \sum_{n=1}^{N} t_n (x_n - \mu_1)' \Sigma^{-1}$$

である．

最尤解（2 クラス）(4)

$D_{\mu_1}\ell(\mu_1) = 0$ を解いて

$$\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n x_n \tag{4.75}$$

を得る．同様にして

$$\mu_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n) x_n \tag{4.76}$$

も得られる．

最尤解（2 クラス）(5)

最後に Σ の最尤量を求める．ℓ の Σ に関係する項を取り出すと

$$
\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N} t_n \log|\Sigma| - \frac{1}{2}\sum_{n=1}^{N} t_n (x_n-\mu_1)'\Sigma^{-1}(x_n-\mu_1) \\
&-\frac{1}{2}\sum_{n=1}^{N} (1-t_n) \log|\Sigma| - \frac{1}{2}\sum_{n=1}^{N} (1-t_n)(x_n-\mu_2)'\Sigma^{-1}(x_n-\mu_2) \\
&= -\frac{N}{2}\log|\Sigma| \\
&-\frac{1}{2}\left\langle \sum_{n\in\mathcal{C}_1}(x_n-\mu_1)(x_n-\mu_1)' + \sum_{n\in\mathcal{C}_2}(x_n-\mu_2)(x_n-\mu_2)', \Sigma^{-1} \right\rangle
\end{aligned}
$$

である．

最尤解（2 クラス）(6)

$B := \Sigma^{-1}$ で微分して [6]

$$
\begin{aligned}
& D_B \ell(B)(H) \\
&= \frac{N}{2} \langle B^{-1\prime}, H \rangle \\
&- \frac{1}{2} \left\langle \sum_{n \in \mathcal{C}_1} (x_n - \mu_1)(x_n - \mu_1)' + \sum_{n \in \mathcal{C}_2} (x_n - \mu_2)(x_n - \mu_2)', H \right\rangle
\end{aligned}
$$

となるから，Σ の最尤量は

$$
\Sigma = \frac{1}{N} \left(\sum_{n \in \mathcal{C}_1} (x_n - \mu_1)(x_n - \mu_1)' + \sum_{n \in \mathcal{C}_2} (x_n - \mu_2)(x_n - \mu_2)' \right)
$$

である.

---

6 $f(X) = \log |X|$ の微分についてはスライドの最後の補足を参照.

離散特徴

入力が $x \in \{0,1\}^D$ で与えられる場合を考える．一般的な分布は $2^D - 1$ 個の独立な変数を含んでいるため扱いにくい．そこで各特徴値 x_i は互いに独立であり，その分布はクラスに依存したパラメータ μ_{ki} によって定まるベルヌーイ分布に従うと仮定する．すなわち

$$p(x|\mathcal{C}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i} \tag{4.81}$$

とする．このとき (4.63) に代入して

$$\begin{aligned} a_k(x) &= \log(p(x|\mathcal{C}_k)p(\mathcal{C}_k)) \\ &= \sum_{i=1}^{D}(x_i \log \mu_{ki} + (1-x_i)\log(1-\mu_{ki})) + \log p(\mathcal{C}_k) \end{aligned} \tag{4.82}$$

となる．これは入力 x_i の線形関数になので，決定境界が超平面になっていることが分かる．

指数型分布族 (1)

今まで入力がガウス分布の場合，離散値の場合の事後確率を求めてきた．どちらも $K = 2$ ならばロジスティックシグモイド，$K \geq 2$ ならばソフトマックスを活性化関数とする，一般化線形モデルで与えられることが分かった．この結果を尤度 $p(x|\mathcal{C}_k)$ が指数型分布族の場合に一般化する．

$$p(x|\lambda_k) = h(x)g(\lambda_k)\exp\left(\lambda_k' u(x)\right) \tag{4.83}$$

特に $u(x) = x$ であるような分布で $x \mapsto x/s$ と変数変換を行うと

$$p(x|\lambda_k) = \frac{1}{s}h\left(\frac{x}{s}\right)g(\lambda_k)\exp\left(\frac{1}{s}\lambda_k' x\right) \tag{4.84}$$

となる．

指数型分布族 (2)

2 クラスのとき，(4.58) の a は

$$\begin{aligned} a(x) &= \log \frac{p(x|\mathcal{C}_1)p(\mathcal{C}_1)}{p(x|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \log \left(\frac{g(\lambda_1)\exp((1/s)\lambda_1' x)p(\mathcal{C}_1)}{g(\lambda_2)\exp((1/s)\lambda_2' x)p(\mathcal{C}_2)} \right) \\ &= (\lambda_1 - \lambda_2)' x \\ &\quad + \log g(\lambda_1) - \log g(\lambda_2) + \log p(\mathcal{C}_1) - \log p(\mathcal{C}_2) \end{aligned} \tag{4.85}$$

となる．よって $p(\mathcal{C}_1|x) \geq p(\mathcal{C}_2|x)$ となる境界は x の線形関数になっていることがわかった．

指数型分布族 (3)

K クラスのときは,

$$
\begin{aligned}
p(\mathcal{C}_k|x) &= \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(x|\mathcal{C}_j)p(\mathcal{C}_j)} \\
&= \frac{h(x)g(\lambda_k)\exp(\lambda_k' x)p(\mathcal{C}_k)}{\sum_j h(x)g(\lambda_j)\exp(\lambda_j' x)p(\mathcal{C}_j)} \\
&= \frac{\exp\left(\lambda_k' x + \log g(\lambda_k) + \log p(\mathcal{C}_k)\right)}{\sum_j \exp\left(\lambda_k' x + \log g(\lambda_k) + \log p(\mathcal{C}_k)\right)}
\end{aligned}
$$

となる. よって $p(\mathcal{C}_i|x) \geq p(\mathcal{C}_j|x)$ となる境界は x の線形関数になっていることがわかった.

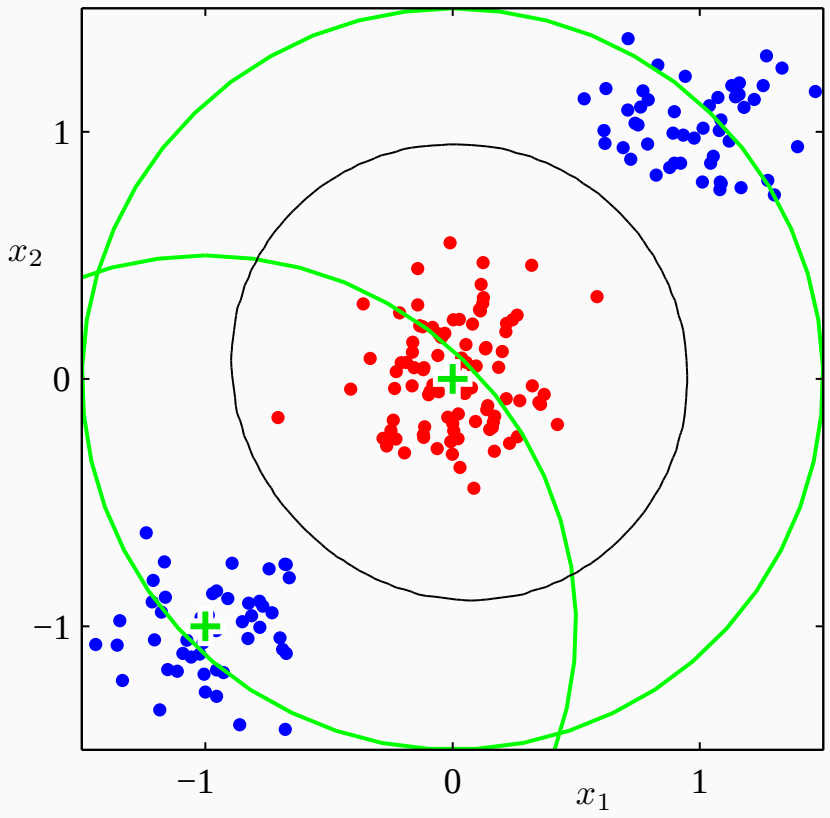
確率的識別モデル

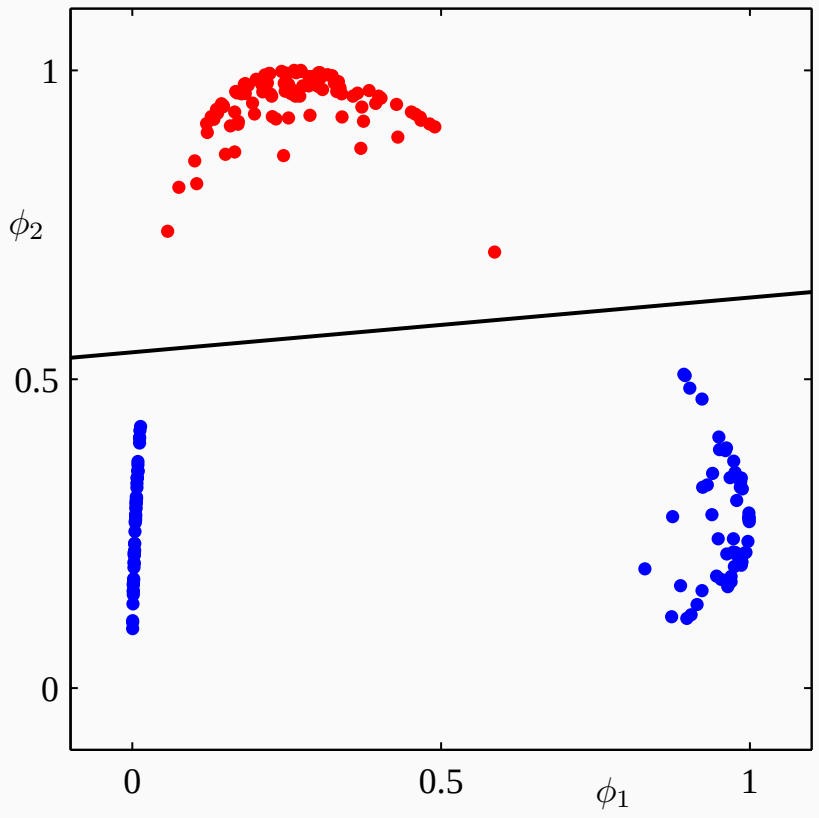
以下では，一般化線形モデルの関数形式を陽に仮定し，最尤法を利用してパラメータを決定する方法を扱う．このときに用いられるアルゴリズムとして反復再重み付け最小二乗 (iterative reweighted least squares, IRLS) がある.

固定基底関数 (1)

- 入力空間が線形分離不可能であっても，非線形変換 ϕ をうまくとれれば，特徴空間上で線形分離が可能になる.
- 一般に $\phi_0(x) = 1$ として，w_0 をバイアスとする.
- 非線形変換を行なってもクラス間の重なりは取り除くことができない.
- 固定基底関数の限界については 3.6 節を参照.

固定基底関数 (2)





ロジスティック回帰

識別モデルの例として**ロジスティック回帰** (logistic regression) と呼ばれるモデルを扱う．名前に「回帰」とあるが，これは回帰よりむしろ分類のためのモデルである．

2 クラスの場合を考える．事後確率 $p(\mathcal{C}_k|\phi)$ がロジスティックシグモイド関数で書けたことを思い出して

$$p(\mathcal{C}_1|\phi) = \sigma(w'\phi) \tag{4.87}$$

とする．このとき，$p(\mathcal{C}_2|\phi) = 1 - p(\mathcal{C}_1|\phi)$ である．

特徴空間の次元を M とすると，最尤法を使う場合，パラメータ数が $O(M^2)$ であるのに対し，ロジスティック回帰ならば M で済む．

ロジスティック回帰のパラメータ決定 (1)

$t_n \in \{0, 1\}$, $\phi_n := \phi(x_n)$, $y_n := p(\mathcal{C}_1|\phi_n) = \sigma(w'\phi_n)$ とする．データ集合 $\{(\phi_n, t_n)\}_{n=1,\ldots,N}$ が与えられたとき尤度は

$$\begin{aligned} p(t|w) &= \prod_{n=1}^{N} p(\mathcal{C}_1|\phi_n)^{t_n} p(\mathcal{C}_2|\phi_n)^{1-t_n} \\ &= \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n} \end{aligned} \tag{4.89}$$

となる．誤差関数としては**交差エントロピー誤差関数** (cross-entropy error function)

$$\begin{aligned} E(w) &:= -\log p(t|w) \\ &= -\sum_{n=1}^{N} (t_n \log y_n + (1 - t_n) \log(1 - y_n)) \end{aligned} \tag{4.90}$$

を用いる．

ロジスティック回帰のパラメータ決定

交差エントロピー誤差関数の勾配を求めるにあたり，ロジスティックシグモイドの導関数を求めておく．

$$\begin{aligned}\frac{d\sigma(a)}{da} &= \frac{d}{da}\frac{1}{1+\exp(-a)} = -\frac{-\exp(-a)}{(1+\exp(-a))^2} \\ &= \sigma(a)(1-\sigma(a)) \end{aligned} \tag{4.88}$$

これを使えば

$$\begin{aligned}\nabla E(w) &= -\sum_{n=1}^{N}\left(t_n\frac{1}{y_n}y_n(1-y_n)\phi_n + (1-t_n)\frac{1}{1-y_n}(-y_n)\phi_n\right) \\ &= \sum_{n=1}^{N}(y_n - t_n)\phi_n \end{aligned} \tag{4.91}$$

が得られる．$\nabla E(w) = 0$ の解をニュートン法で求めることを考える．

反復再重み付け最小二乗（二乗和誤差）

先に第 3 章で使った二乗和誤差関数

$$E(w) = \frac{1}{2}\|t - \Phi w\|^2$$

にニュートン法をする方法について紹介する．導関数を求めると $DE(w) = (t - \Phi w)'(-\Phi) = (\Phi w - t)'\Phi$ であるから,

$$\nabla E(w) = \Phi'\Phi w - \Phi' t, \tag{4.93}$$

$$H = D\nabla E(w) = \Phi'\Phi \tag{4.94}$$

となる．ゆえに w の更新は以下で与えられる.

$$\begin{aligned} w^{(\mathrm{new})} &= w^{(\mathrm{old})} - (\Phi'\Phi)^{-1}(\Phi'\Phi w^{(\mathrm{old})} - \Phi t) \\ &= (\Phi'\Phi)^{-1}\Phi' t \end{aligned} \tag{4.95}$$

よって二乗和誤差関数に関しては 1 度の更新で正確な解が求められることが分かる.

反復再重み付け最小二乗（交差エントロピー誤差）(1)

ロジスティック回帰の交差エントロピー誤差関数にニュートン法を適用することを考える.

$$\nabla E(w) = \sum_{n=1}^{N}(y_n - t_n)\phi_n = \Phi'(y - t) \tag{4.96}$$

これの導関数を求める. $D\nabla E(w) = \Phi' Dy$ なので $Dy(w)$ を求めなくてはならない. 成分ごとに偏微分すると

$$\frac{\partial y_i}{\partial w_j} = \delta_{ij} y_i(1 - y_i)\phi_i(x_j)$$

である. よって $R := (r_{ij}) = \delta_{ij} y_i(1 - y_i)\,(i = 1, \ldots, N,\, j = 1, \ldots, N)$ とすれば,

$$H = D\nabla E(w) = \Phi' R \Phi \tag{4.97}$$

となる.

反復再重み付け最小二乗（交差エントロピー誤差）(2)

H は正定値である．v を非零のベクトルとし，$u := \Phi v = (u_1 \ \cdots \ u_N)'$ とする．$0 < y_n < 1$ であるから

$$\begin{aligned} v'Hv &= v'\Phi' R\Phi v \\ &= u'Ru \\ &= \sum_{i,j} \delta_{ij} y_i (1 - y_i) u_i u_j \\ &= \sum_{i,j} y_i (1 - y_i) u_i^2 > 0 \end{aligned}$$

である[7]．ゆえに E は唯一の最小解を持つ．

7 $\Phi \neq O$ を仮定した．

反復再重み付け最小二乗（交差エントロピー誤差）(3)

よって w の更新は以下のようになる.

$$\begin{aligned} w^{(\mathrm{new})} &= w^{(\mathrm{old})} - (\Phi' R \Phi)^{-1} \Phi' (y - t) \\ &= (\Phi' R \Phi)^{-1} ((\Phi' R \Phi) w^{(\mathrm{old})} - \Phi' (y - t)) \\ &= (\Phi' R \Phi)^{-1} \Phi' R z \end{aligned} \tag{4.99}$$

ここで

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

とした．注意しなければならないのは R が w に依存していることである．ゆえに更新の度に R を計算し直さなければならない．このために反復重み付き最小二乗法と呼ばれているのである.

多クラスロジスティック回帰 (1)

多クラスの場合，事後確率 $p(\mathcal{C}_k|\phi)$ は

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

と計算できた．これに最尤法を用いてパラメータ w を求める．$\partial y_k/\partial a_j$ が必要になるので，これを計算しておく．

$$\begin{aligned}
\frac{\partial y_k}{\partial a_j} &= \frac{\delta_{kj}\exp(a_k)(\sum_i \exp(a_i)) - \exp(a_k)\exp(a_j)}{(\sum_i \exp(a_i))^2} \\
&= \frac{\exp(a_k)}{(\sum_i \exp(a_i))}\left(\delta_{kj} - \frac{\exp(a_j)}{\sum_i \exp(a_i)}\right) \\
&= y_k(\delta_{kj} - y_j)
\end{aligned} \tag{4.105}$$

多クラスロジスティック回帰 (2)

次に尤度を求める．1-of-K 表記を使って

$$t_{nk} = \texttt{if } t_n == (0 \ \cdots \ 0 \ \overset{k}{\overset{\smile}{1}} \ 0 \ \cdots \ 0) \texttt{ then } 1 \texttt{ else } 0$$

と表す．行列 T を (n, k) 成分が t_{nk} で与えられる行列とすると

$$p(T|w_1, \ldots, w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathcal{C}_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}} \tag{4.107}$$

となる[8]．ここで $y_{nk} = y_k(\phi_n)$ とした．負の対数をとると

$$E(w_1, \ldots, w_K) = -\log p(T|w_1, \ldots, w_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk} \tag{4.108}$$

となる．

8 これは Categorical distribution という Bernoulli 分布の自然な拡張になっている．

プロビット回帰 (1)

クラスの条件付き確率密度が指数型分布族のときを見てきたが，混合ガウス分布のときは今までのように計算できない．そこで別なタイプの識別確率モデルを考える．以下では 2 クラスの場合のみを扱う．

雑音閾値モデル (noisy threshold model) を考える．入力 ϕ_n に対して $a_n := w'\phi_n$ を評価し，以下の式にしたがって目的変数を設定する．

$$t_n = \begin{cases} 1 & \text{if } a_n \geq \theta, \\ 0 & \text{otherwise} \end{cases} \tag{4.112}$$

ここで θ は確率的な項で，確率密度関数 g にしたがっているとする [9]．このとき識別確率が

$$p(t = 1|a) = \int_{-\infty}^{a} g(u)du \tag{4.113}$$

で与えられるものとする．

9 個人的には，閾値が確率的に変動すると捉えるよりも，閾値に，入力に含まれる確率的誤差を含めたものが θ と捉えるほうが分かりやすいと思った．

プロビット回帰 (2)

特に g が標準ガウス分布にしたがう，すなわち識別確率が

$$\Phi(a) := \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \tag{4.114}$$

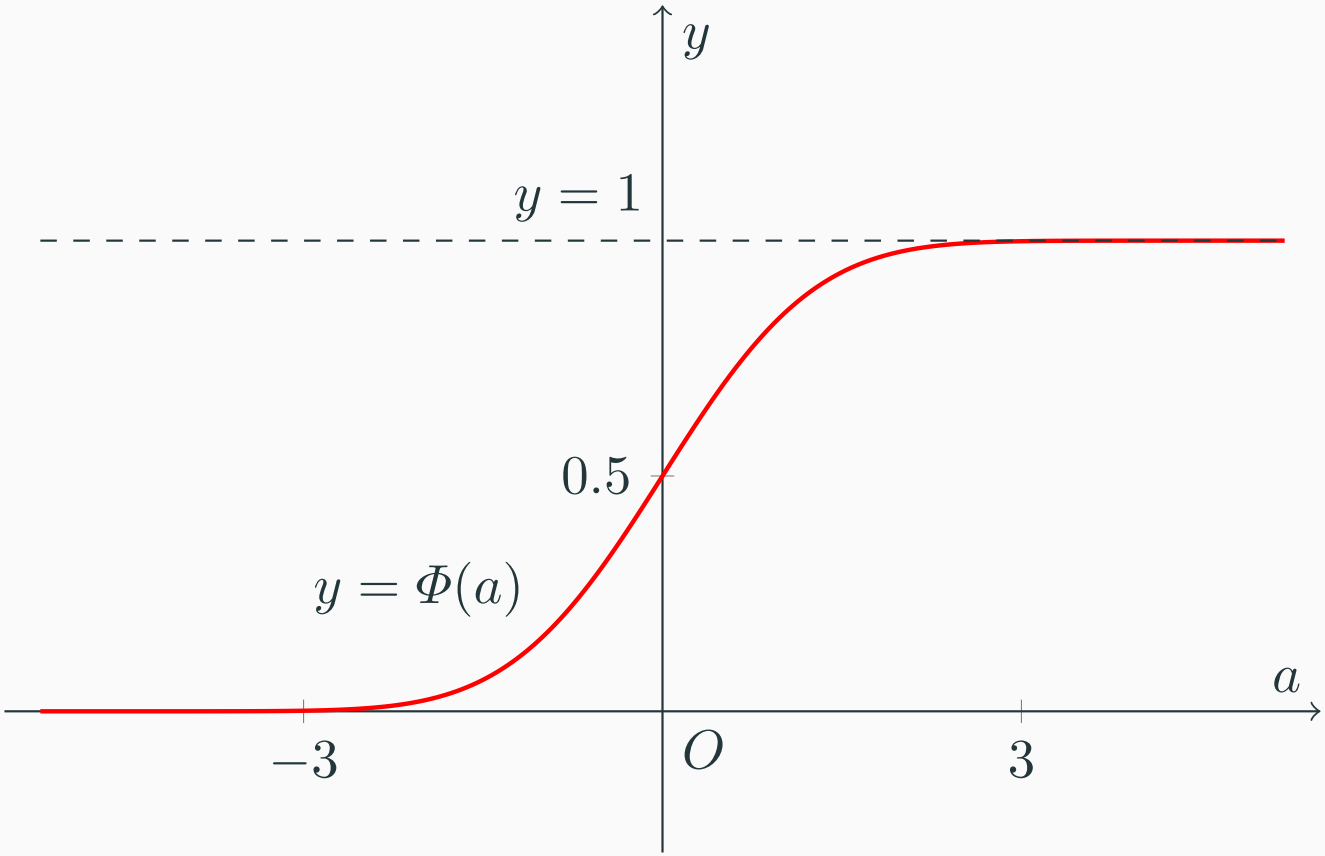
で与えられるとき，このモデルは**プロビット回帰** (probit regression) と呼ばれる [10].

最尤法でプロビット回帰のパラメータを決定できるらしいが省略する.

10標準正規分布の累積分布関数 Φ の逆関数をプロビット関数と呼ぶ.

プロビット回帰 (3)

標準正規分布の累積分布関数のグラフ



正準連結関数 (1)

今まで何度か

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

という式が出てきた．目的変数 t の分布が指数型分布族のときに，上の式を一般化した結果が得られることを示す．

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

(2.226) の導出と同様にして

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

が得られる．(4.119) が η について解くことができて，$\eta = \psi(y)$ と表せると仮定する．

正準連結関数 (2)

f を非線形関数として $y = f(w'\phi)$ というモデルを考える. f^{-1} は**連結関数** (link function) と呼ばれる.

対数尤度は

$$\begin{aligned}\log p(t|\eta, s) &= \sum_{n=1}^{N} \log p(t_n|\eta, s) \\ &= \sum_{n=1}^{N} \left(\log g(\eta_n) + \frac{\eta_n t_n}{s} \right) + \text{const.} \qquad (4.121)\end{aligned}$$

である.

正準連結関数 (3)

対数尤度の勾配は

$$\nabla_w \log p(t|\eta, s) = \sum_{n=1}^{N} \left(\frac{d}{d\eta_n} \log g(\eta_n) + \frac{t_n}{s} \right) \frac{d\eta_n}{dy_n} \frac{dy_n}{da_n} \nabla a_n$$
$$= \sum_{n=1}^{N} \frac{1}{s}(t_n - y_n)\psi'(y_n)f'(a_n)\phi_n \qquad (4.122)$$

と計算できる.

連結関数 $f^{-1} = \psi$ となるような連結関数をとっておけば,

$$\nabla E(w) = \frac{1}{s} \sum_{n=1}^{N} (y_n - t_n)\phi_n \qquad (4.124)$$

となる. ガウス分布のときは $s = \beta^{-1}$, ロジスティックモデルのときは $s = 1$ である.

1 次元のラプラス近似 (1)

4.5 節でロジスティック回帰のベイズ的な取り扱いをするために必要となる**ラプラス近似**という手法について説明する.

ラプラス近似の目的は，分布 $p(z)$ が与えられたとき，そのモード z_0 を中心とするガウス分布で $p(z)$ を近似することである．まず 1 変数の場合を考える．分布 $p(z)$ を仮定する.

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

ここで $Z := \int f(z)dz$ は正規化のための定数である.

1 次元のラプラス近似 (2)

z_0 を中心として $\log f(z)$ を Taylor 展開し，2 次の項までをとると

$$\log f(z) \simeq \log f(z_0) - \frac{1}{2}A(z - z_0)^2 \tag{4.127}$$

となる.

$$A = -\frac{d^2}{dz^2} \log f(z)\Big|_{z=z_0} (z - z_0)^2 \tag{4.128}$$

とした．f が z_0 で極大となるので Taylor 展開の 1 次の項が現れていないことに注意する．指数をとると

$$f(z) \simeq f(z_0) \exp\left(-\frac{A}{2}(z - z_0)^2\right) \tag{4.129}$$

である.

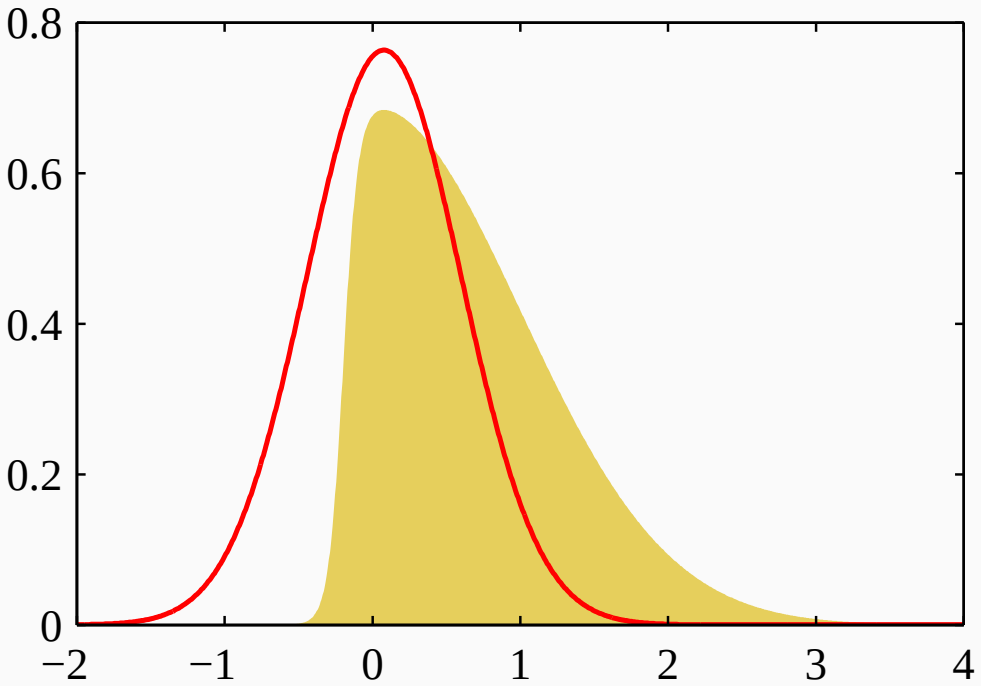
1 次元のラプラス近似 (3)

正規化して p を近似する分布

$$q(z) := \left(\frac{A}{2\pi}\right)^{1/2} \exp\left(-\frac{A}{2}(z - z_0)^2\right) \tag{4.130}$$

が得られる.

ラプラス近似



多次元のラプラス近似 (1)

M 次元空間上の分布 $p(z) = f(z)/Z$ を近似する．1 次元の場合と同様に

$$\log f(z) \simeq \log f(z_0) - \frac{1}{2}(z - z_0)' A (z - z_0) \tag{4.131}$$

とできる．A は $\log f(z)$ の z_0 におけるヘッセ行列を -1 倍したものである．両辺の指数をとると

$$f(z) \simeq f(z_0) \exp\left(-\frac{1}{2}(z - z_0)' A (z - z_0)\right) \tag{4.133}$$

となる．正規化して

$$\begin{aligned} q(z) &= \frac{|A|^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}(z - z_0)' A (z - z_0)\right) \\ &= \mathcal{N}(z|z_0, A^{-1}) \end{aligned} \tag{4.134}$$

を得る．

モデルの比較と BIC

分布 $p(z)$ を近似したときと同様にして，正規化係数 Z は

$$\begin{aligned} Z &= \int f(z)dz \\ &\simeq f(z_0) \int \exp\left(-\frac{1}{2}(z-z_0)'A(z-z_0)\right) dz \\ &= f(z_0)\frac{(2\pi)^{M/2}}{|A|^{1/2}} \end{aligned} \tag{4.135}$$

と近似できる．ただし $A := -D_z^2(\log f)(z_0)$ である．

モデルの比較と BIC

モデルエビデンスは

$$p(\mathcal{D}) = \int p(D|\theta)p(\theta)d\theta \tag{4.136}$$

で与えられる[11]．$f(\theta) = p(\mathcal{D}|\theta)p(\theta)$, $Z = p(\mathcal{D})$ として，(4.135) を適用すると

$$p(\mathcal{D}) \simeq f(\theta_{\mathrm{MAP}})\frac{(2\pi)^{M/2}}{|A|^{1/2}} \tag{ME}$$

となる．ただし

$$A := -D_\theta^2(\log f)(\theta_{\mathrm{MAP}}) = -D_\theta^2(p(\mathcal{D}|\theta)p(\theta))(\theta_{\mathrm{MAP}})$$

である．

11 $\mathcal{M}$ の条件付けを省略した．

モデルの比較と BIC

(ME) の両辺の対数をとって

$$\log p(\mathcal{D}) \simeq \log p(\mathcal{D}|\theta_{\mathrm{MAP}}) \overbrace{+ \log p(\theta_{\mathrm{MAP}}) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|A|}^{\text{オッカム係数 (Occam factor) と呼ばれる.}} \tag{4.137}$$

を得る.

もしも事前確率 $p(\theta)$ がガウス分布で，ヘッセ行列が非退化 [12] ならば,

$$\log p(\mathcal{D}) \simeq \log p(\mathcal{D}|\theta_{\mathrm{MAP}}) - \frac{1}{2}M\log N \tag{4.139}$$

と近似できる．(4.139) は**ベイズ情報量基準** (Bayesian Information Criterion, BIC)，あるいは**シュワルツ基準** (Schwarz criterion) と呼ばれる．(1.73) の AIC と比較して，BIC はモデルの複雑さに，より重いペナルティーを科している.

12この仮定が満たされていないことが多い.

ベイズロジスティック回帰のラプラス近似

ロジスティック回帰モデルをベイズ的に扱いたい．しかしそれを厳密に行うのは困難である．そこでラプラス近似を使って計算を行う方法を考える．事後分布をガウス分布で表現することがこの節での目標である．

事後分布がガウス分布ならば，事前分布もガウス分布とするのが自然である．そこで事前分布を

$$p(w) = \mathcal{N}(w|m_0, S_0) \tag{4.140}$$

とする．(4.89) と (4.140) を用いると事後確率の対数は

$$\begin{aligned}\log p(w|t) &= \log p(w)p(t|w) + \text{const.} \\ &= -\frac{1}{2}(w - m_0)' S_0 (w - m_0) \\ &\quad - \sum_{n=1}^{N} (t_n \log y_n + (1 - t_n) \log(1 - y_n)) + \text{const.} \end{aligned} \tag{4.142}$$

と計算できる．

ラプラス近似

事後確率を最大化して最大事後確率解 w_{MAP} を求める．共分散行列は

$$\begin{aligned} S_N^{-1} &= -D_w^2 \log p(w|t)(w_{\mathrm{MAP}}) \\ &= S_0^{-1} + \Phi' R \Phi \end{aligned} \tag{4.143}$$

で与えられる．よって事後確率分布のガウス分布による近似

$$q(w) = \mathcal{N}(w|w_{\mathrm{MAP}}, S_N) \tag{4.144}$$

を得る．

予測分布 (1)

前節で事後確率 $p(w|t)$ のガウス分布表現が得られた．最後に行うのは，新しく特徴ベクトル $\phi(x)$ が与えられたときに，予測分布 $p(\mathcal{C}_1|\phi, t)$ と $p(\mathcal{C}_2|\phi, t)$ を求めることである．

前節の結果 $p(w|t) = q(w)$ を用いると，$\mathcal{C}_1$ に対する予測分布は

$$\begin{aligned} p(\mathcal{C}_1|\phi, t) &= \int p(\mathcal{C}_1|\phi, w)p(w|t)dw \\ &\simeq \int \sigma(w'\phi)q(w)dw \end{aligned} \tag{4.145}$$

と書ける．このとき $\mathcal{C}_2$ に対する予測分布は $p(\mathcal{C}_2|\phi, t) = 1 - p(\mathcal{C}_1|\phi, t)$ で与えられる．

予測分布 (2)

$\hat{\phi} := \phi/\|\phi\|$ として V と N をそれぞれ

$$V := \{v \in \mathbb{R}^N;\ v'\phi = 0\},\ N := \{a\hat{\phi};\ a \in \mathbb{R}\}$$

と定義する．このとき任意の w は $v \in V$ と $a\hat{\phi} \in N$ を用いて
$w = v + a\hat{\phi}$ と一意に表すことができる（ベクトル空間の直交分解）.

よって

$$\begin{aligned} p(\mathcal{C}_1|t) &\simeq \iint \sigma(a)(p(v|a)p(a))dvda \\ &= \int \sigma(a)p(a)\left(\int p(v|a)dv\right)da \\ &= \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_2^a)da \end{aligned} \tag{4.151}$$

と計算できる.

予測分布 (3)

μ_a, σ_a^2 はそれぞれ

$$\mu_a := \mathbb{E}(w'\phi) = \mathbb{E}(w)'\phi = w'_{\mathrm{MAP}}\phi,$$
$$\sigma_a := \mathrm{var}(w'\phi) = \phi' \,\mathrm{var}(w)\phi = \phi' S_N \phi$$

である.

予測分布 (4)

(4.151) のままでは計算が進められないので $\sigma(a)$ を標準正規分布の累積分布関数 $\Phi(a)$ を用いて近似する．$y = \sigma(a)$ と $y = \Phi(\lambda a)$ の原点における傾きが一致するような λ を求める．

$u(a) := \lambda a$ とすると

$$\left.\frac{d\Phi(u(a))}{da}\right|_{a=0} = \frac{du}{da}\frac{d}{du}\int_{-\infty}^{u}\frac{e^{-\theta^2/2}}{\sqrt{2\pi}}d\theta\Bigg|_{a=0} = \left.\frac{\lambda e^{-u^2/2}}{\sqrt{2\pi}}\right|_{a=0} = \frac{\lambda}{\sqrt{2\pi}}$$

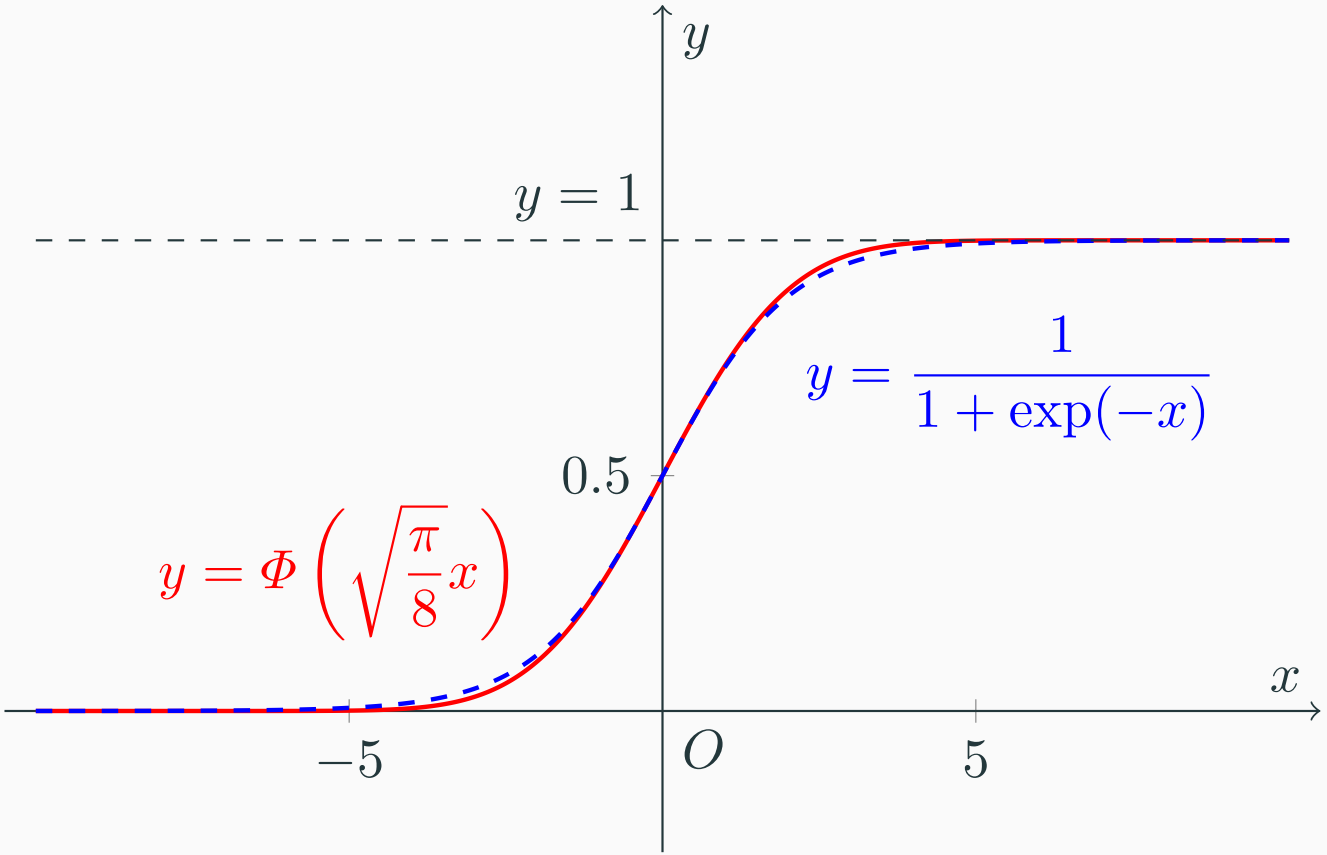
であり，一方

$$\left.\frac{d\sigma}{da}\right|_{a=0} = \sigma(0)(1-\sigma(0)) = \frac{1}{4}$$

であるから，$\lambda = \sqrt{\pi/8}$ がわかる．

Φ による σ の近似

赤い実線が標準正規分布の累積分布関数で，青い破線がロジスティックシグモイド関数である．



予測分布 (5)

σ を Φ で近似することで次の性質（演習 4.26）を使うことができる.

$$\int \Phi(\lambda a)\mathcal{N}(a|\mu, \sigma^2)da = \Phi\left(\frac{\mu}{\sqrt{\lambda^{-2} + \sigma^2}}\right) \tag{4.152}$$

これを使うと

$$\begin{aligned} p(\mathcal{C}_1|t) &\simeq \int \Phi(\lambda a)\mathcal{N}(a|\mu_a, \sigma_a^2)da \\ &= \Phi\left(\frac{\mu_a}{\sqrt{\lambda^{-2} + \sigma_a^2}}\right) \\ &\simeq \sigma\left(\frac{\mu_a}{\sqrt{1 + \pi\sigma_a^2/8}}\right) \end{aligned}$$

を得る.

2 次形式の Fréchet 導関数

$f(x) := x'Ax$ とすると $D_x f(x)(h) = x'(A + A')h$ である.

証明　$h'Ax$ はスカラーで $h'Ax = (h'Ax)' = x'A'h$ を満たすので

$$\begin{aligned}(x+h)'A(x+h) - x'Ax &= x'Ah + h'Ax + h'Ah \\ &= x'(A+A')h + h'Ah\end{aligned}$$

が成り立つ．Cauchy-Schwarz の不等式と，作用素ノルムの性質から

$$\begin{aligned}|(x+h)'A(x+h) - x'Ax - x'(A+A')h| &\leq |h'Ah| \\ &\leq \|h\|\|Ah\| \\ &\leq \|h\|\|A\|\|h\|\end{aligned}$$

である．よって任意の $\varepsilon > 0$ に対して，$\|h\| \leq \varepsilon/\|A\|$ とすれば
$D_x f(x)(h) = x(A + A')h$ であることが分かる. ∎

$\log|X|$ の導関数 (1)

X を $N \times N$ の実対称正定値行列とする．このとき $f(X) := \log|X|$ と定めると $D_X f(X)(H) = \langle X^{-1}, H \rangle$ である．

証明 X を H を変化させたときの f の変化量は

$$\begin{aligned} f(X+H) - f(X) &= \log|X(I + X^{-1}H)| - \log|X| \\ &= \log|I + X^{-1}H| \end{aligned}$$

と計算できる．$X^{-1}H$ のジョルダン標準形を J，変換行列を P とする．また $Y := X^{-1}H$ の固有値を $\lambda_i\,(i = 1, \ldots, n)$ とする．このとき

$$\begin{aligned} \log|I + X^{-1}H| &= \log|P^{-1}(I+J)P| = \log|I+J| \\ &= \log\left(\prod_{i=1}^{N}(1+\lambda_i)\right) = \sum_{i=1}^{N}\log(1+\lambda_i) \end{aligned}$$

である．

$\log|X|$ の導関数 (2)

$h \in \mathbb{C} \mapsto \log(1+h)$ について，どのような $\varepsilon > 0$ をとってもある $\delta > 0$ が存在して，$|h| < \delta$ ならば

$$\log(1+h) \leq h + \frac{\varepsilon}{N\|X^{-1}\|}|h|$$

とできる．したがって $\|H\| < \delta/\|X^{-1}\|$ ととれば $\max_i |\lambda_i| < \delta$ なので [13]

$$\begin{aligned}\sum_{i=1}^{N} \log(1+\lambda_i) &= \sum_{i=1}^{n} \lambda_i + \frac{\varepsilon}{N\|X^{-1}\|}\left(\sum_{i=1}^{N}|\lambda_i|\right) \\ &\leq \operatorname{tr}(X^{-1}H) + \varepsilon\|H\|\end{aligned}$$

となる．

13一般に行列 A の任意の固有値 $\lambda \neq 0$ と対応する固有ベクトル x について

$$\|A\| = \frac{\|A\|\|x\|}{\|x\|} \geq \frac{\|Ax\|}{\|x\|} = \frac{\|\lambda x\|}{\|x\|} \geq |\lambda|$$

が成り立つ．すなわちスペクトル半径 $\rho(A) := \max_i |\lambda_i|$ と行列のノルムの間に不等式 $\rho(A) \leq \|A\|$ が成り立つ．

$\log|X|$ の導関数 (3)

ゆえに $f(X) = \log|X|$ の Fréchet 導関数は

$$D_X f(X)(H) = \mathrm{tr}(X^{-1}H) = \langle X^{-1\prime}, H\rangle$$

である.

証明にあたり Manton (2010) と Manton (2012) を参考にした.

参考文献

Manton, J. H. (2010). `https://jmanton.wordpress.com/tag/frechet-derivative`.

Manton, J. H. (2012). Differential calculus, tensor products and the importance of notation. arXiv preprint arXiv:1208.0197.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

ビショップ, C. M. (2008). パターン認識と機械学習 上. シュプリンガージャパン.

海野裕也, 岡野原大輔, 得居誠也, and 徳永拓之 (2015). オンライン機械学習 (機械学習プロフェッショナルシリーズ). 講談社.