PRML 4.1.6~4.2.2, Fukunaga 1990

5501 酒井一徳
12/6/17

目次

4.1.6 多クラスにおけるフィッシャーの判別

Fisher判別の多クラスへの拡張

- (クラス数K)<(入力空間の次元D) であるとする.
  - 以下の線形特徴を導入.

$$y_k = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} \quad (k = 1, \ldots, D') \quad \text{where } D' > 1$$

- 上記をグループ化.

$$\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x} \quad \text{where} \quad \mathbf{y} = (y_1, \ldots, y_{D'})^{\mathrm{T}}$$
$$\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_{D'})$$

- ただしこの定義にバイアスパラメータは含まれていない.

クラス内共分散とクラス間共分散の多クラス拡張

- まずクラス内共分散,

$$\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} \mathbf{S}_k \quad \text{where} \quad \begin{cases} \mathbf{S}_k = \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}} \\ \mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n \end{cases}$$

N_kはクラスC_kに
含まれるパターンの個数

- 次に総共分散,

$$\mathbf{S}_{\mathrm{T}} = \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^{\mathrm{T}} \quad \text{where} \quad \begin{cases} \mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N} \sum_{k=1}^{K} N_k \mathbf{m}_k \\ N = \sum_k N_k \end{cases}$$

クラス間共分散

(総共分散)=(クラス内共分散)+(クラス間共分散)

$$\mathbf{S}_{\mathrm{T}} = \mathbf{S}_{\mathrm{W}} + \mathbf{S}_{\mathrm{B}}$$

- 上記から以下のクラス間共分散が導かれる.

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)^{\mathrm{T}}$$

- 導出は次ページ.

クラス間共分散の導出

使うもの

$$\mathbf{S}_{\mathrm{T}} = \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} - \sum_{k=1}^{K} N_k \mathbf{m}_k \mathbf{m}^{\mathrm{T}} - \sum_{k=1}^{K} N_k \mathbf{m} \mathbf{m}_k^{\mathrm{T}} + \sum_{n=1}^{N} \mathbf{m} \mathbf{m}^{\mathrm{T}}$$

$$\sum_{n=1}^{N} \mathbf{x}_n = \sum_{k=1}^{K} N_k \mathbf{m}_k$$

$$\mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} \sum_{n \in C_k} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} - \sum_{k=1}^{K} N_k \mathbf{m}_k \mathbf{m}_k^{\mathrm{T}}$$

$$\sum_{k=1}^{K} \sum_{n \in C_k} \mathbf{x}_n \mathbf{m}_k^{\mathrm{T}} = \sum_{k=1}^{K} N_k \mathbf{m}_k \mathbf{m}_k^{\mathrm{T}}$$

- 上記から,

$$N = \sum_{k} N_k$$

$$\mathbf{S}_{\mathrm{B}} = \mathbf{S}_{\mathrm{T}} - \mathbf{S}_{\mathrm{W}} = \sum_{k=1}^{K} N_k \left(\mathbf{m}_k \mathbf{m}_k^{\mathrm{T}} - \mathbf{m}_k \mathbf{m}^{\mathrm{T}} - \mathbf{m} \mathbf{m}_k^{\mathrm{T}} + \mathbf{m} \mathbf{m}^{\mathrm{T}} \right)$$

射影空間での定義

$$\mathbf{S}_{\mathrm{Wy}} = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} (\mathbf{y}_n - \boldsymbol{\mu}_k) (\mathbf{y}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\mathbf{S}_{\mathrm{By}} = \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu}) (\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathrm{T}}$$

where

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{y}_n, \quad \boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{\mu}_k$$

フィッシャーの線形判別基準

- クラス間分散を大きく, クラス内分散を小さくするための基準.
  - 例えば以下.

$$
\begin{aligned}
J(\mathbf{W}) &= \mathrm{Tr}\left\{\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\right\} \\
&= \mathrm{Tr}\left\{\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_{\mathrm{Wx}}\mathbf{W}\right)^{-1}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_{\mathrm{Bx}}\mathbf{W}\right)\right\}
\end{aligned}
$$

- このような基準は複数考えられる(Fukunaga, 1990).
  - 今回は上記基準の最大化を行う.

最適なWを見つけよう！

フィッシャーの線形判別基準の最大化

- $J(\mathbf{W})$を$\mathbf{W}$について最大化する.

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}} = -2\mathbf{S}_{\mathrm{Wx}}\mathbf{W}\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\mathbf{S}_{\mathrm{Wy}}^{-1} + 2\mathbf{S}_{\mathrm{Bx}}\mathbf{W}\mathbf{S}_{\mathrm{Wy}}^{-1}$$

- 上記を 0 とすると,

$$\left(\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}\right)\mathbf{W} = \mathbf{W}\left(\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\right)$$

最適なWがなんなのか
さっぱりわからない….

もうちょっと式変形👉

同時対角化による式変形

$$\left(\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}\right)\mathbf{W} = \mathbf{W}\left(\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\right)$$

- 右辺の二つの対称行列を次の非特異な線形変換行列によって同時対角化する.

$$\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{y} \qquad \text{s.t.} \qquad \begin{aligned} \mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathrm{By}}\mathbf{A} &= \mathbf{D} \\ \mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{A} &= \mathbf{I} \end{aligned}$$

- $\mathbf{D}$ は $\mathbf{S}_{\mathrm{By}}$ の固有値を要素にもつ対角行列.
- $\mathbf{A}$ は逆行列 $\mathbf{A}^{-1}$ を持つ D' 次正方行列.

同時対角化については後述

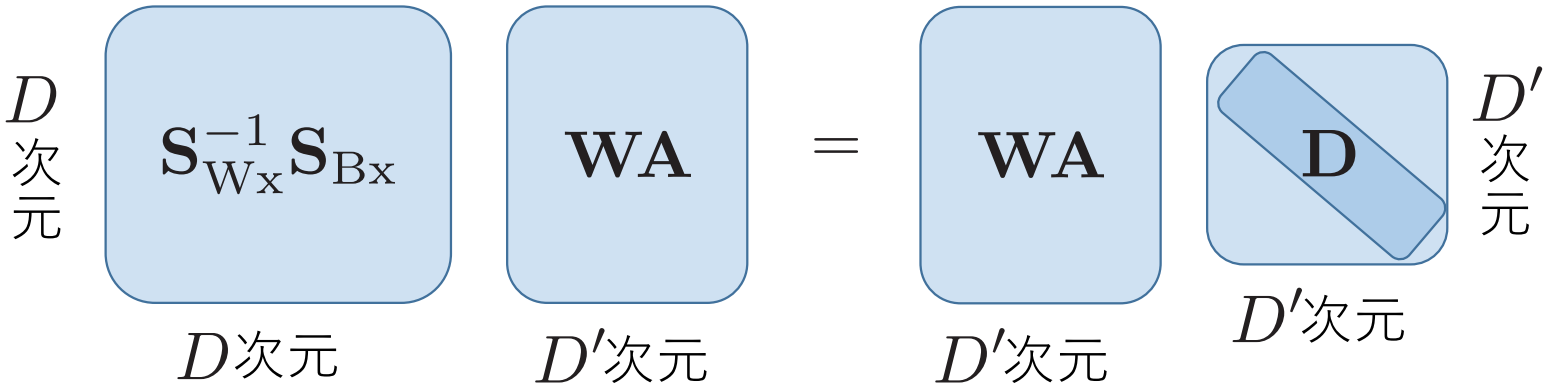
フィッシャーの線形判別基準の最大化

$$\left(\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}\right)\mathbf{W} = \mathbf{W}\left(\mathbf{ADA}^{-1}\right)$$

$$\left(\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}\right)\mathbf{WA} = \mathbf{WA}\left(\mathbf{D}\right)$$

$$\mathbf{S}_{\mathrm{By}} = \mathbf{ADA}^{-1}$$
$$\mathbf{S}_{\mathrm{Wy}}^{-1} = \mathbf{AIA}^{-1} = \mathbf{I}$$

- $\mathbf{WA}$ の列ベクトルは $\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}$ の固有ベクトル.
- $\mathbf{D}$ の要素は $\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}$ の固有値.
- D 次正方行列 $\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}$ の固有方程式が D' 個ある.



フィッシャーの判別基準の不変性

- 先の変換においてフィッシャーの判別基準は不変である.

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\right) &= \mathrm{Tr}\left(\mathbf{A}^{-1}\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\mathbf{A}\right)\\ &= \mathrm{Tr}\left(\left(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathrm{Wy}}\mathbf{A}\right)^{-1}\left(\mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathrm{By}}\mathbf{A}\right)\right)\\ &= \mathrm{Tr}\left(\mathbf{I}^{-1}\mathbf{D}\right)\end{aligned}$$

- D 次正方行列 $\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}$ の固有値 D' 個の総和.
  - 大きい順に D' 個の固有値を選べばいい.
    - 対応する固有ベクトルを列にもつ行列が最適な $\mathbf{W}$.

判別基準の別の側面

$$\mathrm{Tr}\left(\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\right) = \mathrm{Tr}(\mathbf{D}) = d_1 + \ldots + d_{D'}$$

- 判別基準は以下の同時対角化によって上記に書き換えられる.

$$\mathbf{D} = \mathbf{A}^{\mathrm{T}}\mathbf{S}_{\mathrm{By}}\mathbf{A} = \hat{\mathbf{W}}^{\mathrm{T}}\mathbf{S}_{\mathrm{Bx}}\hat{\mathbf{W}} \quad \text{where} \quad \hat{\mathbf{W}} = \mathbf{W}\mathbf{A}$$

$$\text{s.t.} \quad \hat{\mathbf{W}}^{\mathrm{T}}\mathbf{S}_{\mathrm{Wx}}^{-1}\hat{\mathbf{W}} = \mathbf{I}$$

- 制約つきの最大化問題としてラグランジュの未定乗数法.

(K-1)個の線形特徴

- D'個の線形特徴量を先の方法で得られる「はず」.
- (4.44),(4.45)から固有値は(K-1)個までのみ.

$$\mathbf{m} = \frac{1}{N}\sum_{k=1}^{K} N_k \mathbf{m}_k$$

$$\mathbf{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)^{\mathrm{T}}$$

$$\mathrm{rank}\left(\mathbf{S}_{\mathrm{Bx}}\right) \leq K - 1$$

$$\Rightarrow \quad \mathrm{rank}\left(\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}\right) \leq K - 1$$

- $\mathbf{S}_{\mathrm{Bx}}$の固有ベクトルで張る(K-1)次元空間への射影は基準に影響を与えない.

(K-1)個の線形特徴

- 射影空間と入力空間での判別基準.

$$\mathrm{Tr}\left(\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\right) = \mathrm{Tr}(\mathbf{D}) = d_1 + \ldots + d_{D'}$$

$$\mathrm{Tr}\left(\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}\right) = d_1 + \ldots + d_{K-1} + \ldots + d_D$$

以降ゼロ

$D' \geq K-1$ であるとき $\mathrm{Tr}\left(\mathbf{S}_{\mathrm{Wy}}^{-1}\mathbf{S}_{\mathrm{By}}\right) = \mathrm{Tr}\left(\mathbf{S}_{\mathrm{Wx}}^{-1}\mathbf{S}_{\mathrm{Bx}}\right)$

- $\mathbf{S}_{\mathrm{Bx}}$の固有ベクトルで張る(K-1)次元空間への射影は基準に影響を与えない.
  - K個以上の線形特徴は発見できない.

終わり

同時対角化について雑記

- 対称行列$\mathbf{A}, \mathbf{B}$について同時に対角化する正則行列$\mathbf{P}$が存在する.
  - まず行列$\mathbf{A}$に対し以下の線形変換行列で変換(白色化)を行う.

$$\mathbf{y} = \mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}}\mathbf{x} \qquad \text{s.t.} \qquad \mathbf{P}_{\mathrm{A}}^{\mathrm{T}}\mathbf{P}_{\mathrm{A}} = \mathbf{I}$$
$$\mathbf{P}_{A}^{\mathrm{T}}\mathbf{A}\mathbf{P}_{\mathrm{A}} = \mathbf{D}_{\mathrm{A}}$$

Aの固有値を
要素にもつ対角行列

- 行列$\mathbf{A}, \mathbf{B}$を変換.

$$\mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}}\mathbf{A}\mathbf{P}_{\mathrm{A}}\mathbf{D}_{\mathrm{A}}^{-1/2} = \mathbf{I}$$
$$\mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}}\mathbf{B}\mathbf{P}_{\mathrm{A}}\mathbf{D}_{\mathrm{A}}^{-1/2} = \mathbf{C}$$ 対角ではない

同時対角化について雑記

- 次に行列$\mathbf{B}$に対し以下の線形変換行列で変換(無相間化)を行う.

$$\mathbf{z} = \mathbf{P}_{\mathrm{C}}^{\mathrm{T}}\mathbf{y} \qquad \text{s.t.} \qquad \mathbf{P}_{\mathrm{C}}^{\mathrm{T}}\mathbf{P}_{\mathrm{C}} = \mathbf{I}$$
$$\mathbf{P}_{C}^{\mathrm{T}}\mathbf{C}\mathbf{P}_{\mathrm{C}} = \mathbf{D}_{\mathrm{C}}$$

Cの固有値を
要素にもつ対角行列

- 続けて変換.

$$\left(\mathbf{P}_{\mathrm{C}}^{\mathrm{T}}\mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}}\right)\mathbf{A}\left(\mathbf{P}_{\mathrm{A}}\mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{C}}\right) = \mathbf{I}$$
$$\left(\mathbf{P}_{\mathrm{C}}^{\mathrm{T}}\mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}}\right)\mathbf{B}\left(\mathbf{P}_{\mathrm{A}}\mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{C}}\right) = \mathbf{D}_{\mathrm{C}}$$

$\mathbf{P} = \mathbf{P}_{\mathrm{A}}\mathbf{D}_{\mathrm{A}}^{-1/2}\mathbf{P}_{\mathrm{C}}$ によってどちらも対角行列に.

同時対角化について雑記

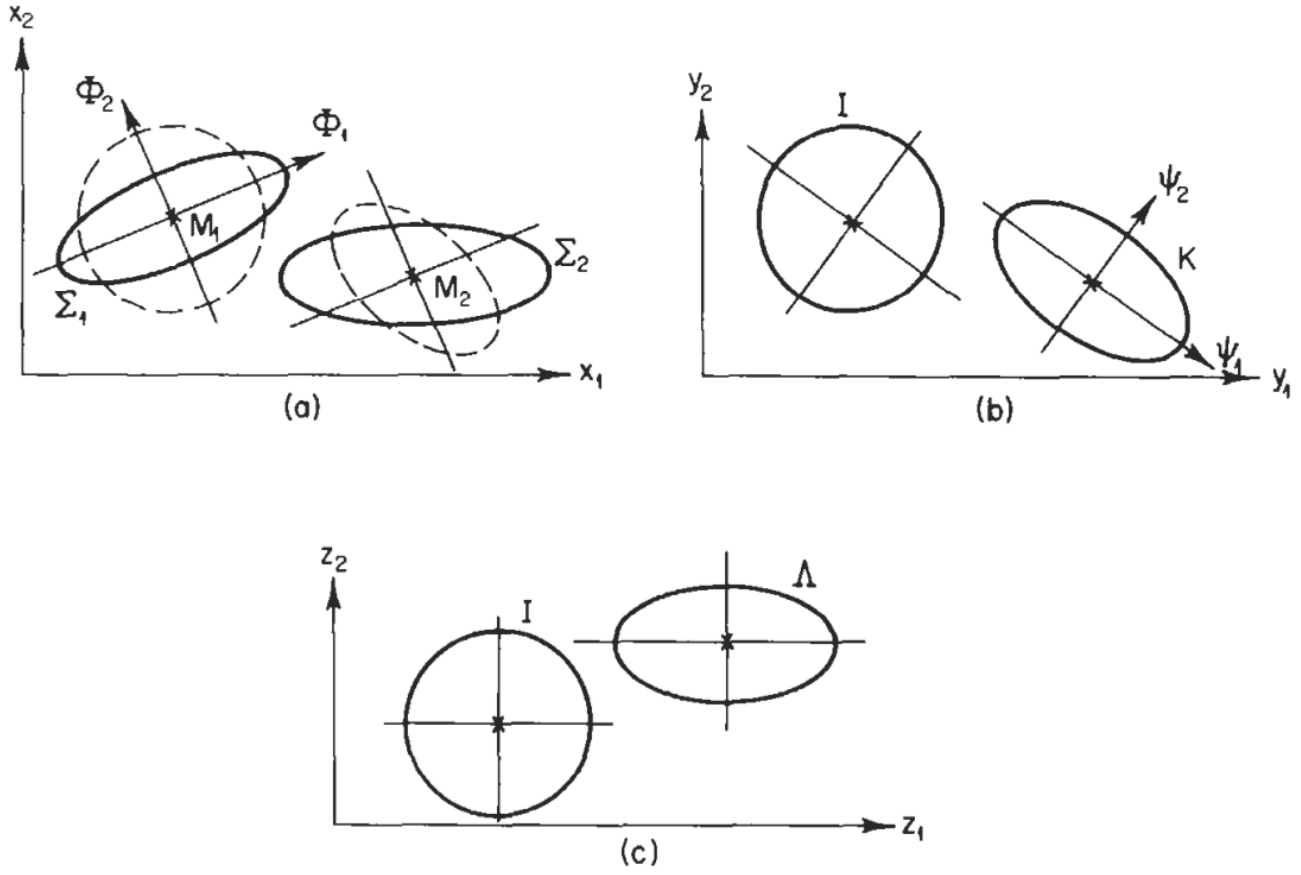


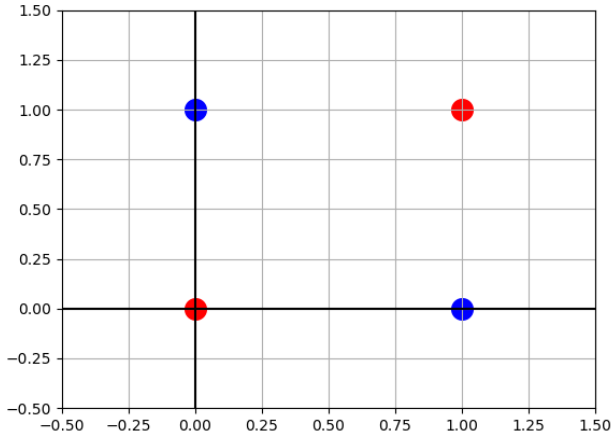
Fig. 2-3 Simultaneous diagonalization.

4.1.7 パーセプトロンアルゴリズム

パーセプトロン

- 2クラスのモデル，多クラスへの拡張は容易でない
- 線形分離可能なデータ集合にしか収束しない
- ニューラルネットワーク(多層パーセプトロン(詳しくは５章))においても線形分離可能なデータ集合にしかうまく動作しないと誤解されニューラルネットの発展は著しく遅れた.

Mark1パーセプトロンハードウェア



線形分離不可能な一例

パーセプトロン

入力ベクトル: $\mathbf{x}$

非線形変換: $\boldsymbol{\phi}(\cdot)$

特徴ベクトル: $\boldsymbol{\phi}(\mathbf{x})$

一般化線形モデル

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

ただし非線形活性化関数 $f(\cdot)$ はステップ関数

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

一般に $\boldsymbol{\phi}(\mathbf{x})$ はバイアス成分 $\boldsymbol{\phi}_0(\mathbf{x}) = 1$ を含む

目的変数に関して注意

パーセプトロンでは活性化関数との適合から目的変数値 $t \in \{-1,\ +1\}$ を使う.
つまりクラス C_1 に対して $t = +1$,クラス C_2 に対して $t = -1$ を使う.

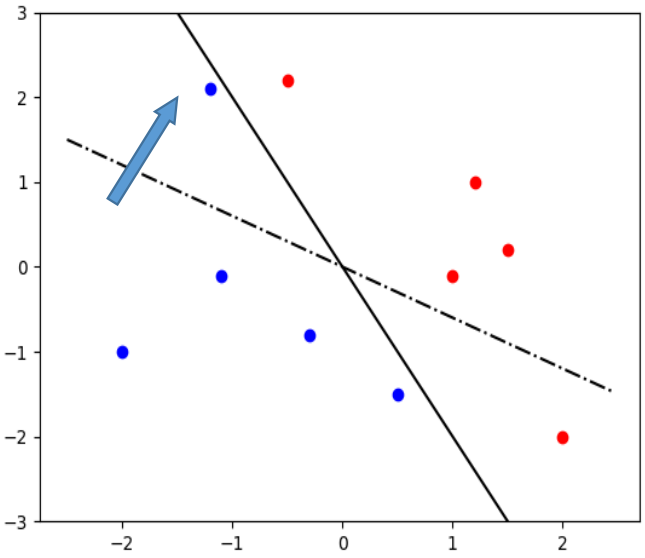
パーセプトロン基準

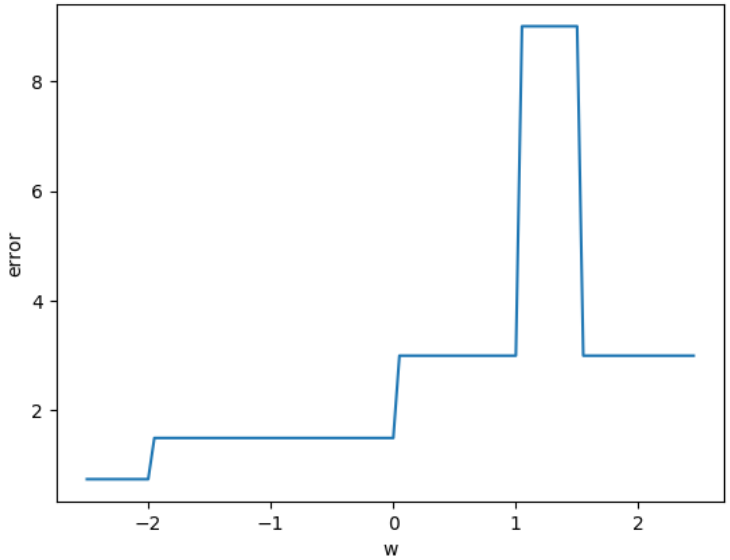
パラメータ**w** の決定は誤差関数が最小になるよう選ぶ

誤差関数はどう選ぶ？

誤認識したパターン総数とするのが自然

簡単な学習アルゴリズムが導出できない





パラメータwによって変化する決定境界がデータ点を跨ぐごとに誤差は変動する, しかも誤差関数は大部分が定数で不連続の関数となる

違う基準が必要！

パーセプトロン基準

今, クラス C_1 のパターン $\mathbf{x}_n$ に対して $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) > 0$,
クラス C_2 のパターン $\mathbf{x}_n$ に対して $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) < 0$ となる$\mathbf{w}$ を求めている

目的変数値 $t \in \{-1, +1\}$ を用いると正しく分類される
すべてのパターンは

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n > 0$$

を満たす

上記からパーセプトロン基準とは

$$E_p(\mathbf{w}) = -\sum_{n\in\mathcal{M}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n$$

正しく分類されたパターン: 誤差0
誤分類されたパターン: $-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n > 0$

ただし　$\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$

$\mathcal{M}$ はご分類されたすべてのパターンの集合

重みベクトルと学習パラメータ

誤差関数の最小化には確率的最急降下アルゴリズムを適用する

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_p(\mathbf{w}) = \mathbf{w}^{(\tau)} - \eta \boldsymbol{\phi}_n t_n$$

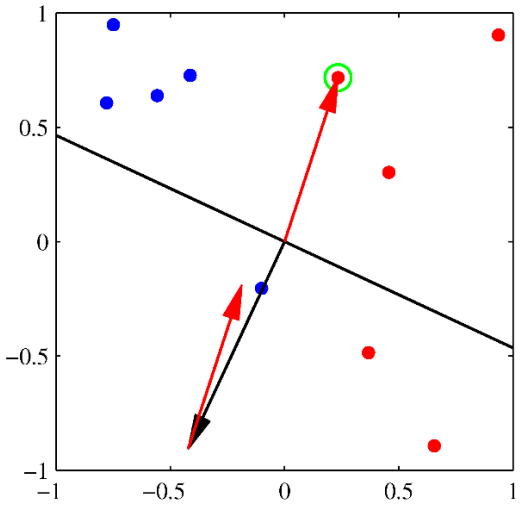
学習パラメータ: η
ステップ数: τ

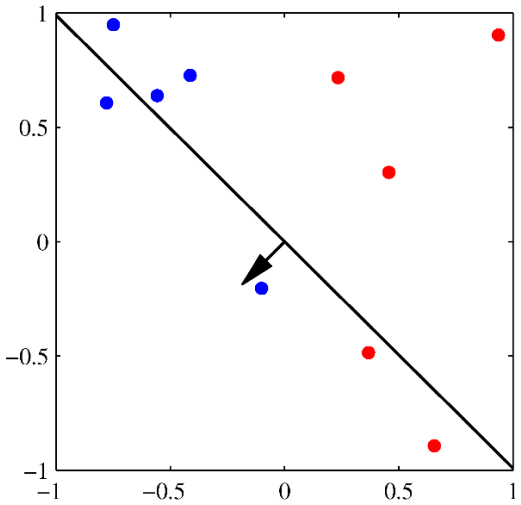
重みベクトル$\mathbf{w}$を定数倍しても判別境界には影響しない($\mathbf{w}\boldsymbol{\phi}_n$の正負が逆転しない)ので, 一般性を失うことなく学習パラメータは 1 にできる($\eta > 0$ならなんでも?)

一般性を失わないってなんじゃい

パーセプトロン学習の様子

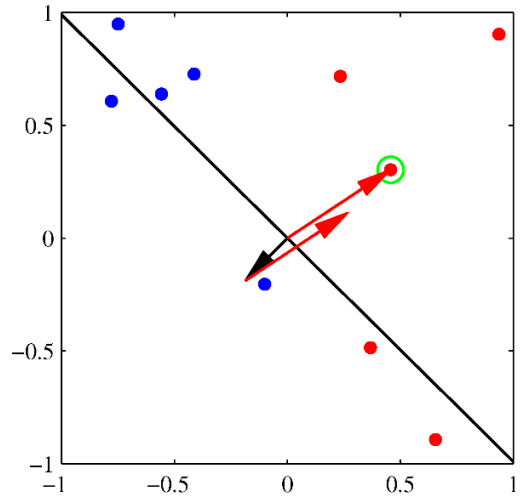
黒の矢印は赤のクラスが分類されるべき決定境界の方向を示していることに注意

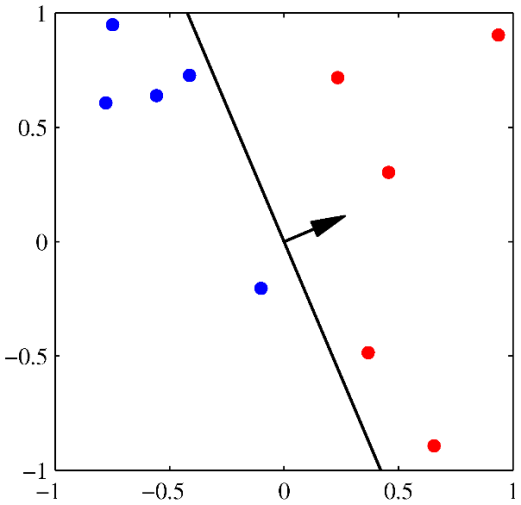




学習率パラメータ η は赤色の矢印の大きさに影響する

正しく分類された





パーセプトロン学習の各ステップ

一回一回の更新には誤分類されたパターンからの誤差への寄与を減らす効果がある.

$$-\mathbf{w}^{(\tau+1)T}\boldsymbol{\phi}_n t_n = -\mathbf{w}^{(\tau+1)T}\boldsymbol{\phi}_n t_n - (\boldsymbol{\phi}_n t_n)^T \boldsymbol{\phi}_n t_n$$
$$< -\mathbf{w}^{(\tau)T}\boldsymbol{\phi}_n t_n$$

しかし, このことは更新対象である誤分類された入力パターン $\mathbf{x}_n$ 一つについての
誤差への寄与の減少を意味する.

また, 決定境界が変化することで今まで正しく分類されていたパターンが
誤分類されることも起こりうる.

パーセプトロンの各ステップは総誤差関数の減少を保証しない

パーセプトロンの収束定理とその限界

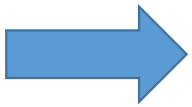
パーセプトロンの収束定理(perceptron convergence theorem)

厳密解が存在するとき(線形分離可能な場合), パーセプトロン学習アルゴリズムは有限回の繰り返しで厳密解に収束する.

しかし収束に必要な繰り返し回数はかなり多く, 実用では分離不可能なのか,
収束に時間がかかっているのか収束するまでわからない.
線形分離可能な場合でもデータの提示順によって収束解は変動する.

パーセプトロンの限界

- 確率的な出力がない
- K>2クラスへの拡張が難しい
- 線形分離でないデータ集合に対して決してパーセプトロン学習アルゴリズムは収束しない.



5章を待て

4.2 確率的生成モデル

確率的生成モデル

話はがらりと変わりまして

分類問題に対する３つのアプローチ

- 入力ベクトルから直接クラス推定する識別関数を作る
- 条件付き確率分布 $p(C_k|\mathbf{x})$ を直接モデル化する
- クラスの事前確率 $p(C_\mathrm{k})$ とクラスで条件付き確率密度 $p(\mathbf{x}|C_\mathrm{k})$ を使ってクラス事後確率 $p(C_\mathrm{k}|\mathbf{x})$ を計算する

２クラスの生成的アプローチ

2クラスのとき

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1) + p(\mathbf{x}|C_2)p(C_2)}$$

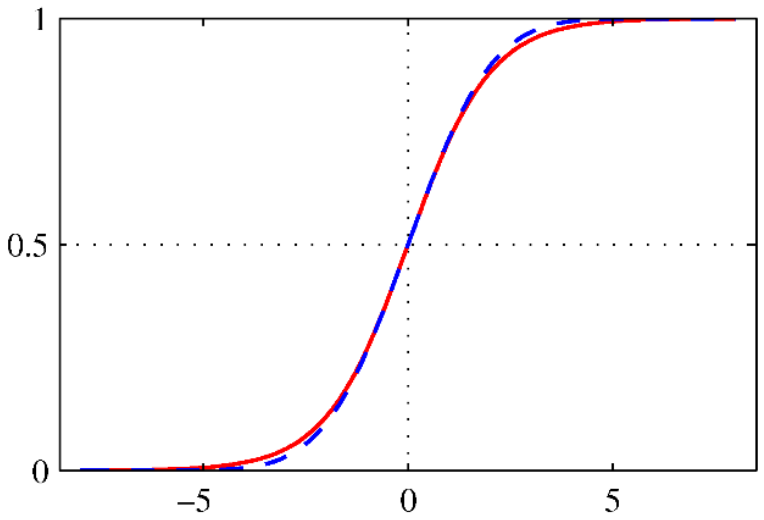
$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

ただし $a = \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}$

ロジスティックシグモイド関数: $\sigma(a) = \frac{1}{1+\exp(-a)}$

ロジスティックシグモイド関数は対称性を持つ.

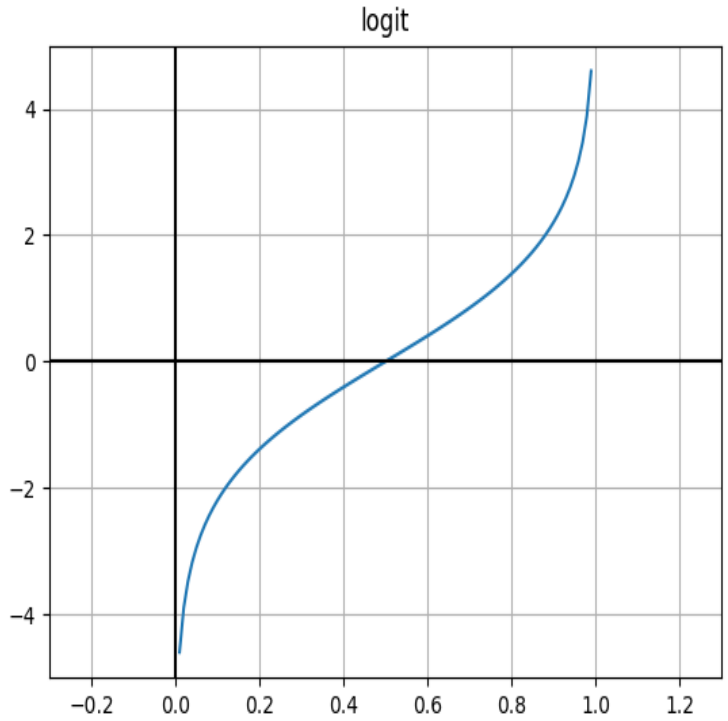
$$\sigma(-a) = 1 - \sigma(a)$$



ロジット関数

ロジット関数: ロジスティックシグモイド関数の逆関数

$$a = \ln\left(\frac{\sigma}{1-\sigma}\right)$$



２クラスにおける確率の対数比とみることができる.

$$\ln[p(C_1|\mathbf{x})/p(C_2|\mathbf{x})]$$

対数オッズとも.

ロジスティックシグモイド関数の恩恵

2 クラスのとき, 次を示した.

$$p(C_1|\mathbf{x}) = \sigma\big(a(\mathbf{x})\big)$$

$a(\mathbf{x})$ が**簡単な関数形をとる場合**, ロジスティックシグモイド関数の利用は重要になる.

今は, $a(\mathbf{x})$ が$\mathbf{x}$ の線形関数だとする.
そのとき事後確率は一般化線形モデルに支配されることに注意する.

K>2クラスのとき

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_j p\big(\mathbf{x}|C_j\big)p\big(C_j\big)} = \frac{\exp(a_k)}{\sum_j \exp\big(a_j\big)}$$

正規化関数もしくはソフトマックス関数
ただし,

$$a_k = \ln\big(p(\mathbf{x}|C_k)p(C_k)\big)$$

連続で滑らかなマックス関数

4.2.1 連続値入力

連続値入力

仮定 クラスの条件付き確率密度(生成モデル)がガウス分布
- 全てのクラスの共分散行列が同じ

事後確率の形はどうなるのか

クラス C_k の確率密度は

$$p(\mathbf{x}|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

で与えられる.

2クラスのとき, 生成的アプローチに則って計算(4.57, 4.58)すると

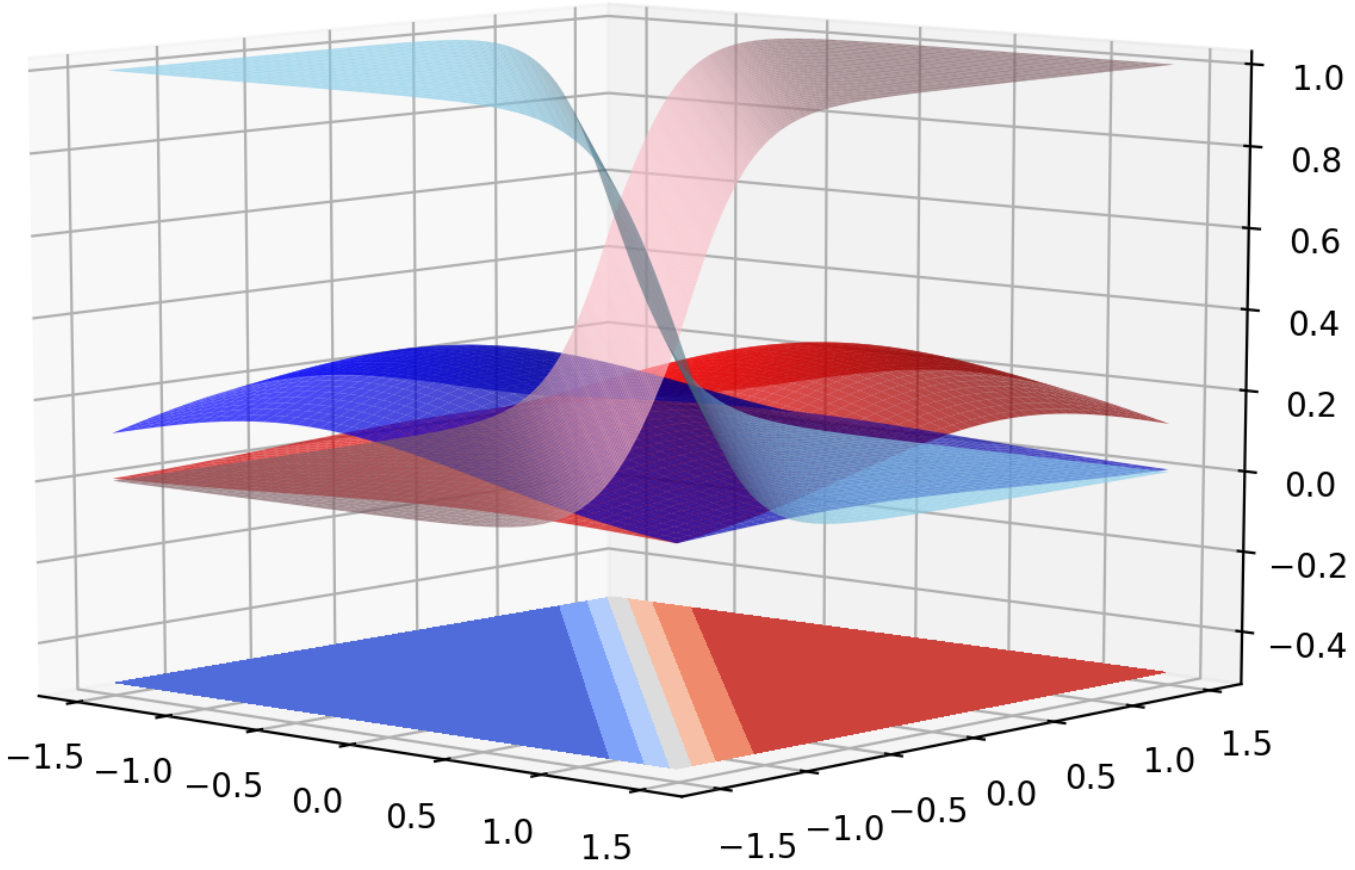
クラス C_1 に対する事後確率: $p(C_1|\mathbf{x}) = \sigma\left(\mathbf{w}^T\mathbf{x} + w_0\right)$

ただし,

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$$

$$w_0 = -\frac{1}{2}\mu_1^T\mathbf{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2^T\mathbf{\Sigma}\mu_2 + \ln\frac{p(C_1)}{p(C_2)}$$

2クラスの事後確率と識別関数の様子



K>2クラスへ一般化

K>2クラスのとき, クラスの事後確率は

$$p(C_k|\mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \ln\left(p(\mathbf{x}|C_k)p(C_k)\right)$$

で与えられるから, $a_k(\mathbf{x})$ は

$$a_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0}$$

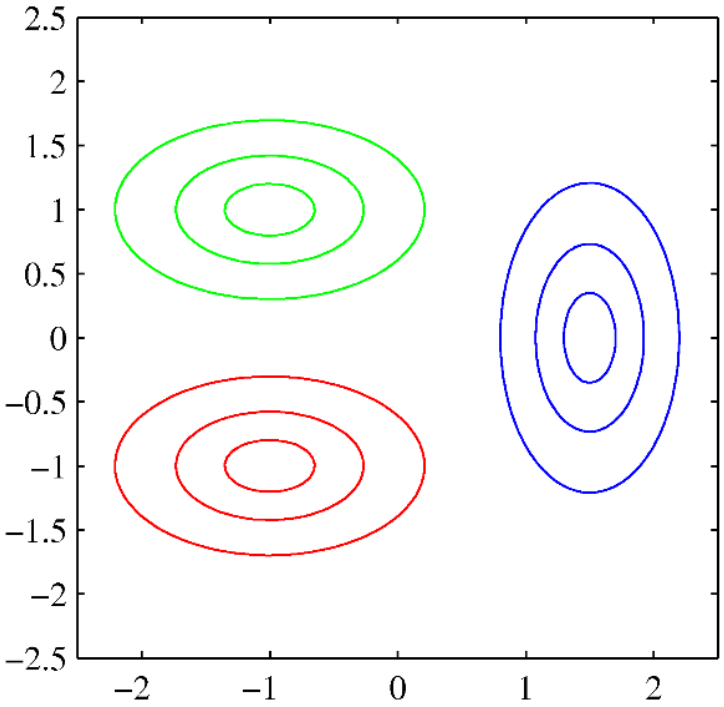
ただし,

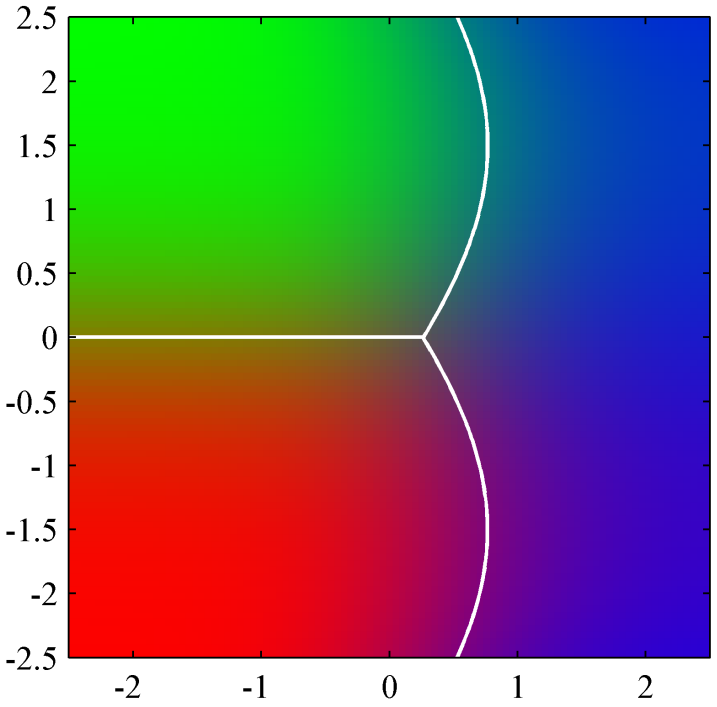
$$\mathbf{w}_k = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$$

$$w_{k0} = -\frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + \ln p(C_k)$$

これも二次の項がキャンセルされている.

共分散行列の条件緩和





4.2.2 最尤解

最尤法によるパラメータ決定

クラスの条件付き確率密度をパラメトリックな関数形でおけるならクラスの事前確率とともに, 最尤法でパラメータを決定できる.

まずは2クラス

仮定

- クラスの条件付き確率密度(生成モデル)がガウス分布
- 各クラスの共分散行列が同じ
- データ集合 $\{\mathbf{x}_n,\ t_n\}$ $(n = 1, \ldots, N)$が与えられている
- $t_n \in \{0,1\}$, 1はクラス C_1, 2はクラス C_2
- クラスの事前確率は $p(C_1) = \pi$, $(C_2) = 1 - \pi$

このとき

$$p(\mathbf{x}_n, C_1) = p(C_1)\,p(\mathbf{x}_n|C_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$p(\mathbf{x}_n, C_2) = p(C_2)\,p(\mathbf{x}_n|C_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

よって尤度関数は

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N}\left[\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]^{t_n}\left[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]^{1-t_n}$$

ただし

$$\mathbf{t} = (t_1, \ldots, t_N)^T$$

最尤法によるパラメータ決定

対数尤度関数の最大化に置き換える

まずは π について最大化.

π に依存する項は

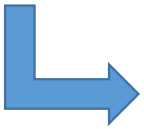
$$\sum_{n=1}^{N}\{t_n \ln \pi - (1-t_n)\ln(1-\pi)\}$$

π に関して微分を 0 として整理すると

$$\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

N_k: クラス C_k 内のデータの総数

予想を裏切らず, クラス内に含まれるデータの個数の割合になっている



多クラスへの拡張も容易(演習4.9)

次は μ について

最尤法によるパラメータ決定

μ_1 についても同様に対数尤度から

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_1, \boldsymbol{\Sigma}\right) = -\frac{1}{2}\sum_{n=1}^{N} t_n \left(\mathbf{x}_n - \boldsymbol{\mu}_1\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_1\right) + c$$

μ_1 について微分を 0 として

$$\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n$$

同様の計算から

$$\mu_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1 - t_n)\mathbf{x}_n$$

それぞれ、各々のクラスに割り当てられるすべての入力ベクトルの平均である

最後は共有している共分散行列

最尤法によるパラメータ決定

共有共分散についても同様に対数尤度から抽出し，整理すると

$$
\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \mu_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \mu_1) \\
&-\frac{1}{2}\sum_{n=1}^{N} (1 - t_n) \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (1 - t_n)(\mathbf{x}_n - \mu_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \mu_2) \\
&= \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2} Tr\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\}
\end{aligned}
$$

ただし，

$$
\begin{aligned}
\mathbf{S} &= \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 \\
\mathbf{S}_1 &= \frac{N}{N_1}\sum_{n \in C_1} (\mathbf{x}_n - \mu_1)(\mathbf{x}_n - \mu_1)^T \\
\mathbf{S}_2 &= \frac{N}{N_2}\sum_{n \in C_2} (\mathbf{x}_n - \mu_2)(\mathbf{x}_n - \mu_2)^T
\end{aligned}
$$

標準的な最尤解の結果を用いれば $\boldsymbol{\Sigma} = \mathbf{S}$
(2.3.4節)
K>2クラスへの拡張が容易

ガウス分布の最尤推定は外れ値に頑健でない
(2.3.7節)
このアプローチは外れ値に頑健でない