



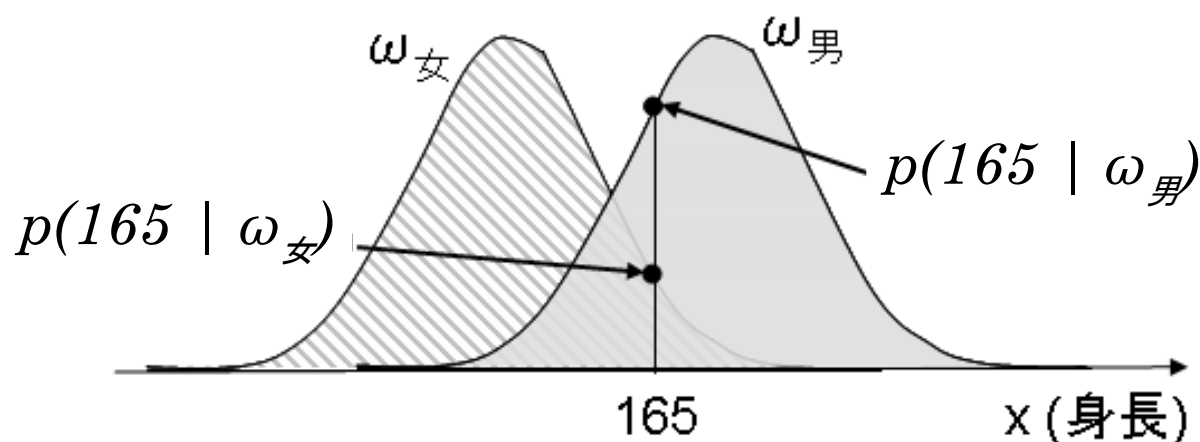
## 8. 未知データを推定しよう – 統計的方法 –

### 8.1 間違っ確率を最小にしたい

- 本当に作りたいシステムは？
  - 誤り0のシステム
    - 現実的には不可能
  - 出力誤差最小のシステム  Widrow-Hoffの学習規則  
ニューラルネットワーク
    - 学習データに対して最適化してしまうかもしれない
  - 誤り確率最小のシステム  **本日のテーマ**
    - 未知データに対する誤り確率最小
  - (期待損失最小のシステム)

## 8.1 間違っ確率を最小にしたい

- 例) 身長を特徴量として成人男女を識別する
  - 身長が与えられたときの、確率の高い方 (= 誤り確率の低い方) を識別結果とすればよい



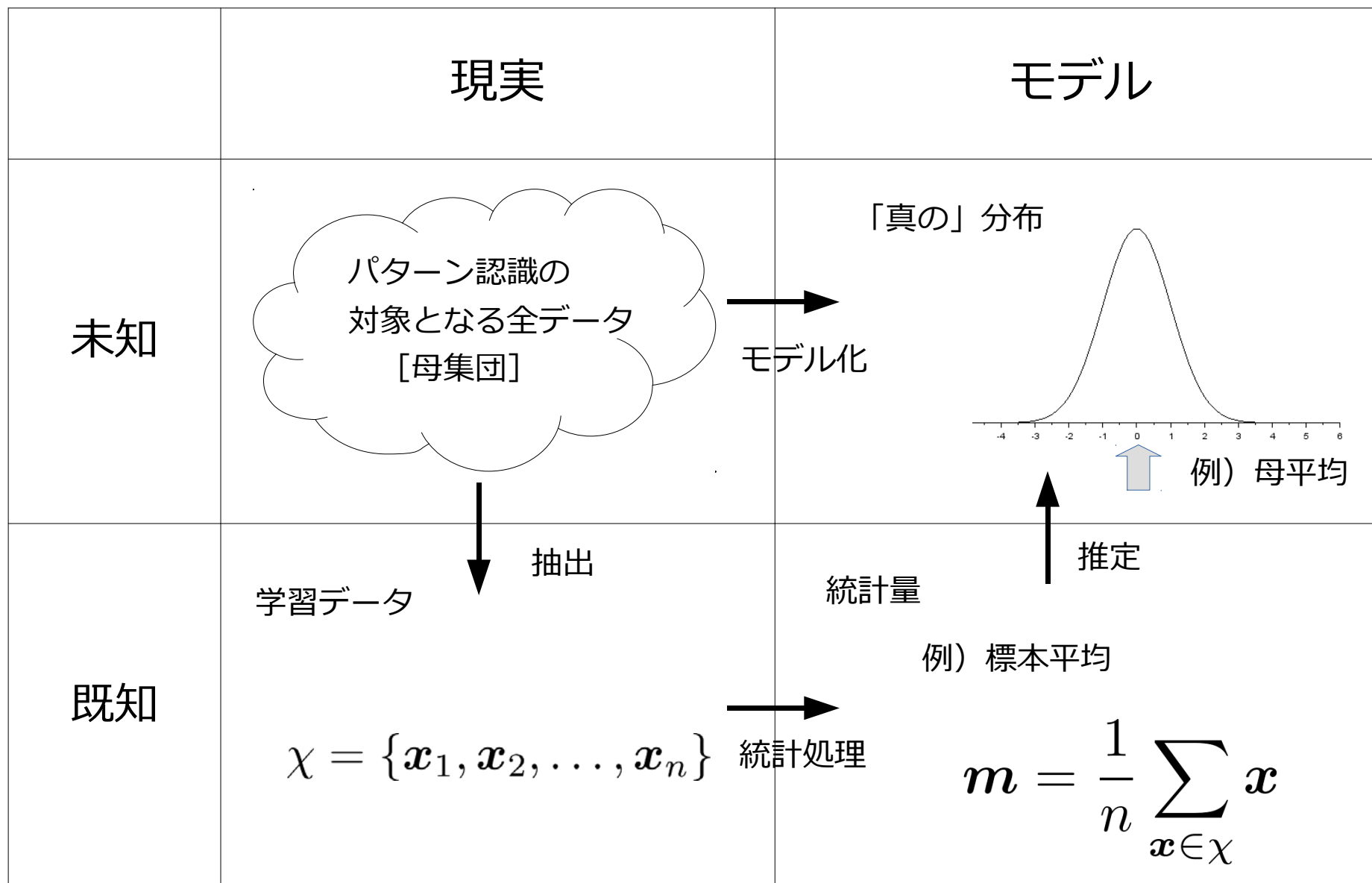
## 8.1.1 誤り確率最小の判定法

- 確率を用いたパターン認識
  - 事後確率最大法(ベイズ決定則)
    - $P(\omega_i|\mathbf{x})$  を最大にするクラス  $\omega_i$  を識別結果とする

$$\arg \max_{i=1,\dots,c} P(\omega_i|\mathbf{x}) = k \Rightarrow \mathbf{x} \in \omega_k$$

- 身長による成人男女の判別システムの場合
  - 入力が185.0cmの時、 $P(\text{男}|185.0)$  と  $P(\text{女}|185.0)$  の大きい方に判定する

## 8.1.2 事後確率の求め方



## 8.1.2 事後確率の求め方

- 一般に  $P(\omega_i|x)$  は直接求めることができない
  - 例) 185.0cmの人を何人集めれば  $P(\text{男}|185.0)$  の値が推定できる？

- 標本誤差 ( $p$ : 調査対象の比率、 $n$ : 標本数)

$$2 \cdot \sqrt{\frac{p(1-p)}{n}}$$

- $p=0.5$  のとき、真の値が95%の確率で存在する範囲
  - $n=100$ :  $50.0\% \pm 10.0$
  - $n=2,000$ :  $50.0\% \pm 2.2$
- それを130.0~200.0まで行おうと？

## 8.1.3 事後確率の間接的な求め方

- ベイズの定理

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})}$$

$P$ : 離散変数に対する確率関数

$p$ : 連続変数に対する確率密度関数

証明

$$\begin{aligned} P(A, B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

## 8.1.3 事後確率の間接的な求め方

- 事後確率  $P(\omega_i|x)$ 
  - $x$  が生起したとき、そのクラスが  $\omega_i$  である確率
- 事前確率  $P(\omega_i)$ 
  - クラス  $\omega_i$  の生起確率
- クラスによらない  $x$  の生起確率  $p(x)$
- 尤度
  - 認識対象としているパターンの生起確率を示したもの  
 $p(x|\omega_i)$  ( $i = 1, \dots, c$ )
  - クラス  $\omega_i$  に属する  $x$  が出現する確率

## 8.1.4 厄介者 $p(\mathbf{x})$ を消そう

- $p(\mathbf{x})$  は全クラスに共通であり、 最大となる  $P(\omega_i|\mathbf{x})$  を決めるのに関与しない

$$\begin{aligned} & \arg \max_{i=1,\dots,c} P(\omega_i|\mathbf{x}) \\ &= \arg \max_{i=1,\dots,c} \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} \\ &= \arg \max_{i=1,\dots,c} p(\mathbf{x}|\omega_i)P(\omega_i) \end{aligned}$$



## 8.1.5 事前確率 $P(\omega_i)$ を求める

- 事前確率  $P(\omega_i)$  の求め方
  - 本当は全ての可能なデータを集めて、それぞれのクラスのデータ数を集計しなければ求まらないが...
- 最尤推定
  - 学習データ数:  $N$
  - クラス  $\omega_i$  のデータ数:  $n_i$
  - 事前確率の最尤推定値

$$P(\omega_i) = \frac{n_i}{N}$$

## 8.1.6 最後の難敵「クラス分布 $P(x|\omega_i)$ 」

- クラス分布  $p(x|\omega_i)$  の求め方
  - クラス分布とは
    - あるクラスのデータ集合からある特徴ベクトルが観測される確率をあらわす確率密度関数
    - ある値は観測されやすく、それから遠くなるに従って観測されにくくなるような性質を持つ
  - 確率分布の形を仮定して、そのパラメータを学習データから推定
  - 例) 正規分布：平均と共分散行列を推定

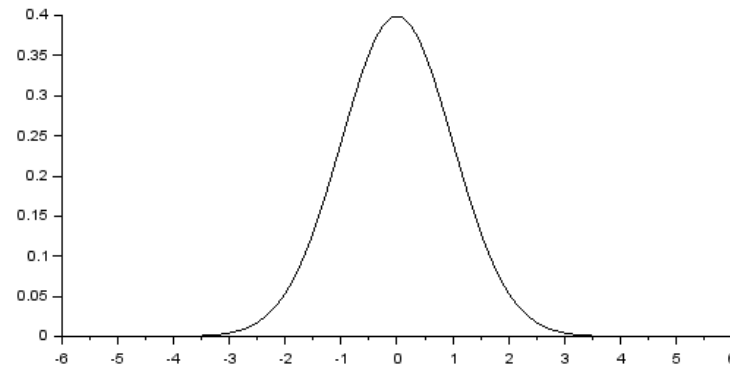
## 8.2 データの広がりを推定する

### 8.2.1 未知データの統計的性質を予測する

- 確率密度関数  $p(\boldsymbol{x}|\omega_i)$  の例
  - 正規分布

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^t \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{m}_i)\right\}$$

- $\boldsymbol{m}_i$  : 平均ベクトル
- $\boldsymbol{\Sigma}_i$  : 共分散行列



# 正規分布とは

- 離散型二項分布の例

- $n$ 枚のコインを投げた時の、表の枚数の度数

- $n=1$                       1   1

- $n=2$                       1   2   1

- $n=3$                       1   3   3   1

- $n=4$                       1   4   6   4   1

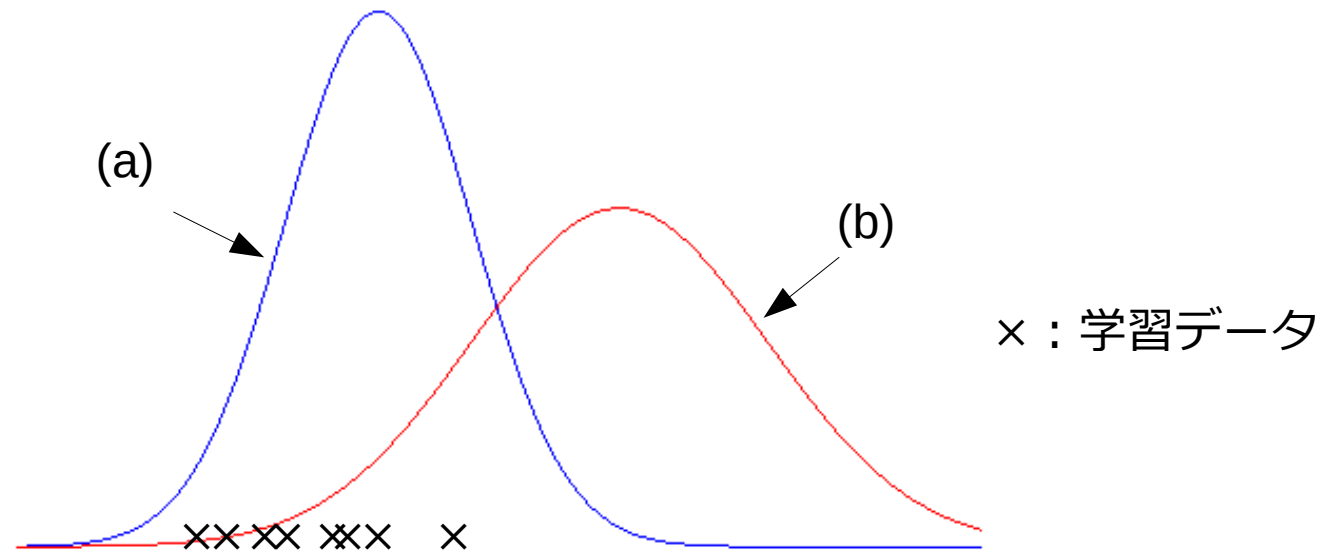
- $n=5$                       1   5   10   10   5   1

- ...

- $n \rightarrow \infty$  の時の分布が正規分布

## 8.2.2 最尤推定

- データを最もうまく説明できる分布を探す



= 対数尤度が最大となる分布を探す

$$p(\chi|\boldsymbol{\theta}) = \prod_{x_p \in \chi} p(x_p|\boldsymbol{\theta}) \quad \boldsymbol{\theta} : \text{分布のパラメータ}$$

## 8.2.2 最尤推定

- 最尤推定の結果

- 平均ベクトル

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \chi_i} \boldsymbol{x}$$

- 共分散行列

$$\boldsymbol{\Sigma}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \chi_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t$$

対角要素はその次元の分散を、それ以外の要素は交差する次元間の相関を表す

## 8.2.2 最尤推定

- 確率密度関数の平均と共分散行列を推定する学習法をパラメトリックな学習と呼ぶ

→8章の方法

- 確率密度関数の形を想定せずに、学習パターンから直接識別関数を求める学習法をノンパラメトリックな学習と呼ぶ

→4～7章の学習アルゴリズム

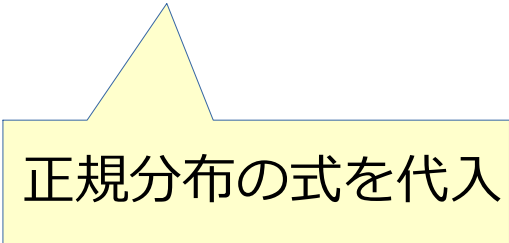
## 8.2.3 統計的な識別

- 識別関数の設定

$$g_i(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)P(\omega_i)$$

- アンダーフローを避けるため対数をとる

$$g_i(\boldsymbol{x}) = \log p(\boldsymbol{x}|\omega_i) + \log P(\omega_i)$$



正規分布の式を代入



## 8.2.3 統計的な識別

$$\begin{aligned} g_i(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^t \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{m}_i) \\ &\quad - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{d}{2} \log 2\pi + \log P(\omega_i) \\ &= -\frac{1}{2} \boldsymbol{x}^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{x} + \boldsymbol{x}^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{m}_i - \frac{1}{2} \boldsymbol{m}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{m}_i \\ &\quad - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{d}{2} \log 2\pi + \log P(\omega_i) \end{aligned}$$

- 確率密度関数  $p(\boldsymbol{x}|\omega_i)$  が正規分布で表される場合、識別関数は  $\boldsymbol{x}$  の2次関数となる。

## 8.2.3 統計的な識別

- 共分散行列が全クラスで等しい場合
  - $x$  の 2 次の係数は定数となるので、識別関数は線形( 1 次)式となる。

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_i - \frac{1}{2} \boldsymbol{m}_i^t \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_i + \log P(\omega_i)$$

## 8.2.3 統計的な識別

- 共分散行列を単位行列とする(特徴間の相関がなく、分散も等しい)
- 事前確率  $P(\omega_i)$  が全クラスで等しいとする  
→ 識別関数は最小距離識別法と同じになる。

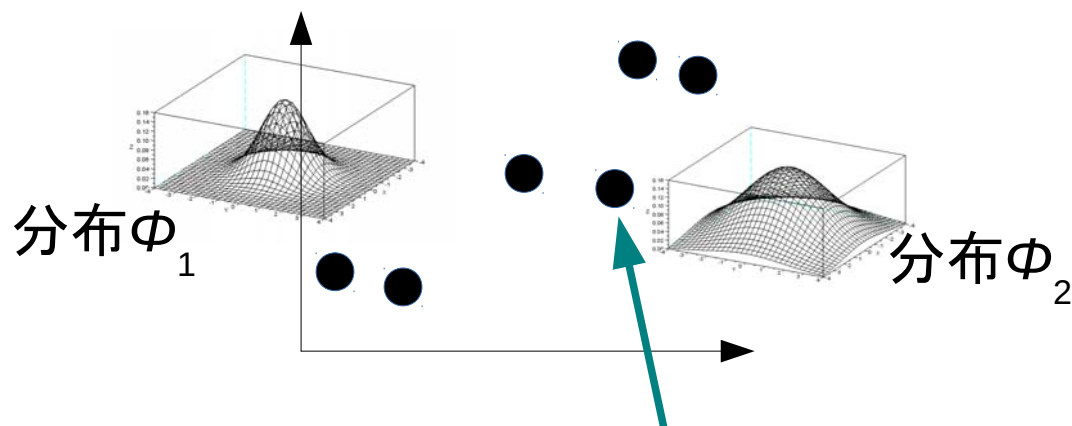
$$g_i(\mathbf{x}) = \mathbf{m}_i^t \mathbf{x} - \frac{1}{2} \|\mathbf{m}_i\|^2$$

## 8.3 実践的な統計的識別

- 単純ベイズ法
  - 特徴空間各次元の独立性を仮定
- ベイズ推定
  - 事前分布が観測によって事後分布に変わる
- 複雑な確率密度関数の推定
  - 複数の正規分布の重み付き和（混合分布）を用いる
  - 各正規分布のパラメータや重みをEMアルゴリズムで学習

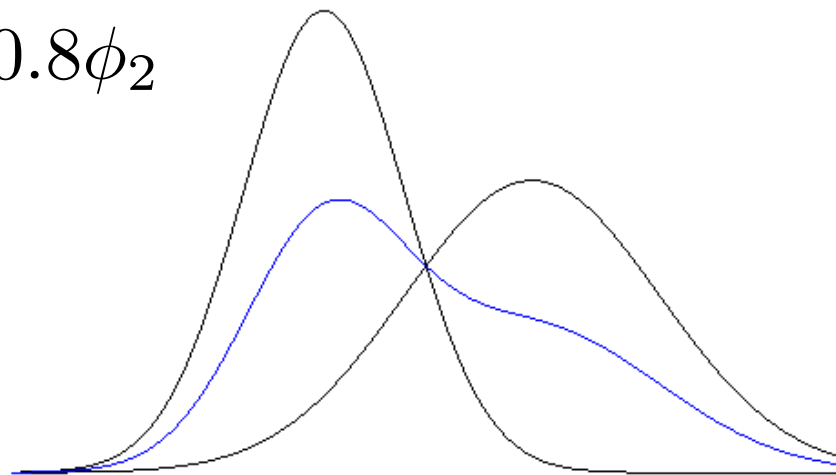
## 8.3 実践的な統計的識別

- EMアルゴリズム



分布 $\phi_1$ の再計算の際、  
重み0.2だけ寄与する

$$0.2\phi_1 + 0.8\phi_2$$



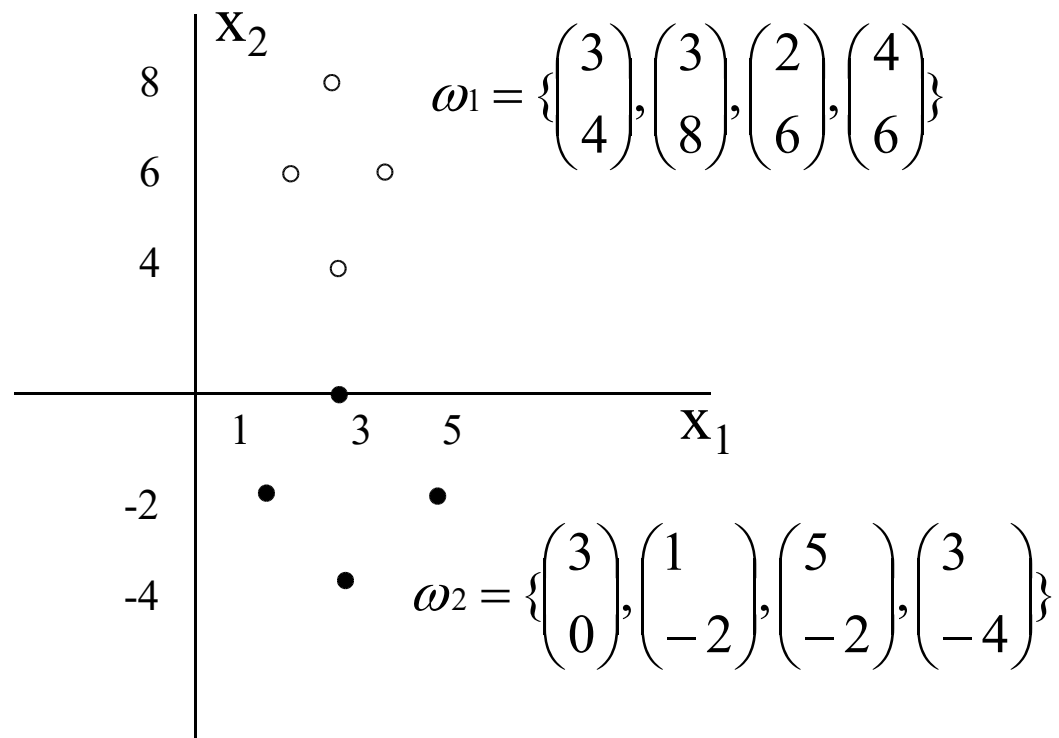
## 8.3 実践的な統計的識別

- EMアルゴリズム

1.  $k$ 個の正規分布を乱数で決める
2. 各データが各分布から生成された確率を計算し、各分布にゆるやかに帰属させる
3. 各データの分布への帰属度に基づき各分布のパラメータ（平均値、共分散行列）を再計算
4. 2, 3をパラメータの更新幅が閾値以下になるまで繰り返す

## 例題8.1

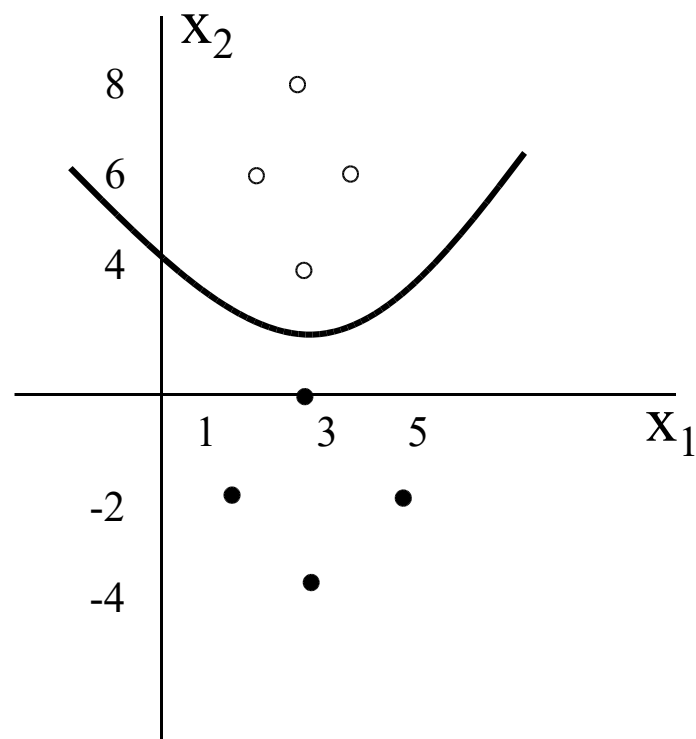
- 以下の2クラスの2次元データの平均・共分散を求めよ。また、識別面をおおまかに図示せよ。



# 解答例

$$\boldsymbol{m}_1 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \quad \boldsymbol{m}_2 = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$



共分散行列が異なるので、識別面は2次曲面