

## 8. 未知データを推定しようー統計的方法ー

### 8.1 間違える確率を最小にしたい

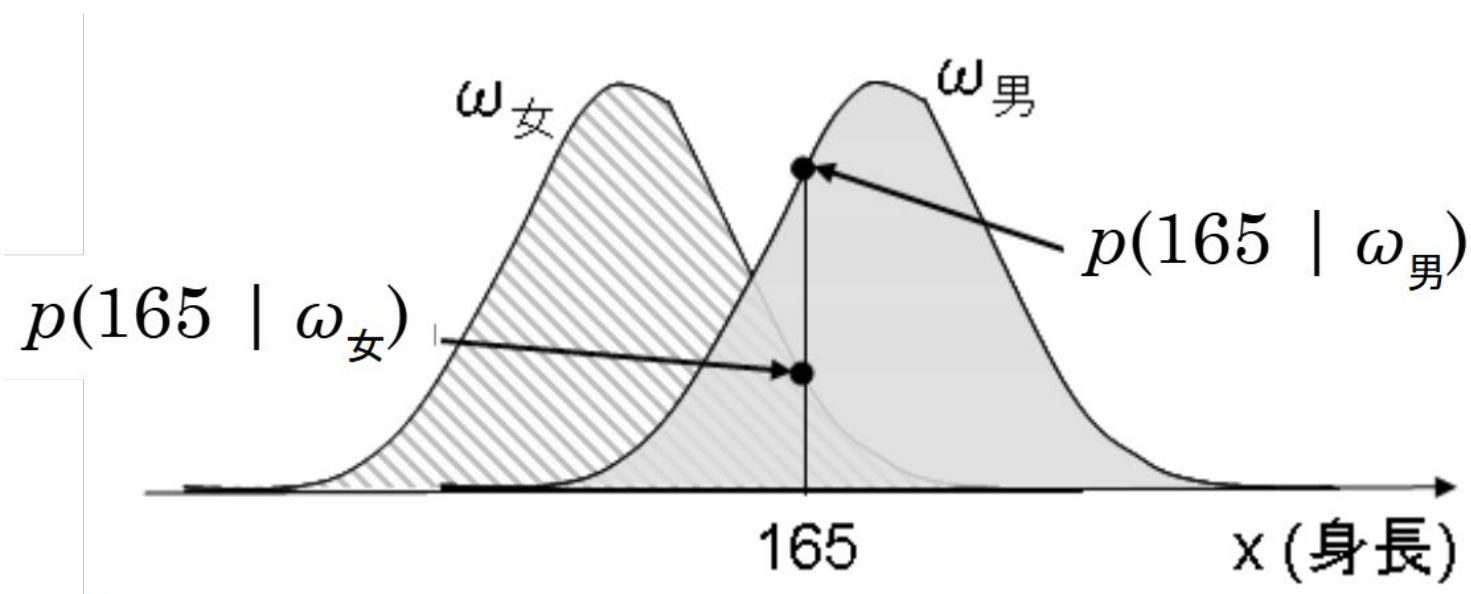
- 本当に作りたいシステムは？
  - ◆ 誤り0のシステム
    - 現実的には不可能
  - ◆ 出力誤差最小のシステム (Widrow-Hoffの学習規則など)
    - 学習データに対して最適化してしまうかもしれない
  - ◆ 誤り確率最小のシステム
    - 未知データに対する誤り確率最小
  - ◆ (期待損失最小のシステム)

## 8.1 間違っ確率を最小にしたい

- 誤り確率を最小にするには

- ◆ 例) 身長を特徴量として成人男女を識別するタスク

- 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
  - ◆ 事後確率最大化識別（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

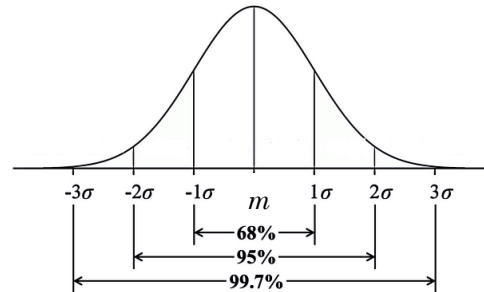
$$\arg \max_{i=1, \dots, c} P(\omega_i | \mathbf{x}) = k \Rightarrow \mathbf{x} \in \omega_k$$

- 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

- 一般に事後確率  $P(\omega_i | \mathbf{x})$  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?
- 比率の誤差 ( $p$ : 調査対象の比率 (正規分布を仮定)、 $n$ : 標本数)

$$1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$



- ◆  $p=0.5$ のときの95%信頼区間

- $n=100$ :  $50.0\% \pm 10.0$
- $n=2,000$ :  $50.0\% \pm 2.2$

- ◆ 2000人の調査を1mm刻みの全ての可能な値で行うと?

## 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

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

## 8.1.3 事後確率の間接的な求め方

- 事後確率  $P(\omega_i | \mathbf{x})$ 
  - ◆  $\mathbf{x}$  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  $\mathbf{x}$  の生起確率  $p(\mathbf{x})$
- クラス分布（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  
$$p(\mathbf{x} | \omega_i) \quad (i = 1, \dots, c)$$
  - ◆ クラス  $\omega_i$  の分布から  $\mathbf{x}$  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

$$\begin{aligned} & \arg \max_i P(\omega_i | \mathbf{x}) \\ &= \arg \max_i \frac{p(\mathbf{x} | \omega_i) P(\omega_i)}{p(\mathbf{x})} \\ &= \arg \max_i p(\mathbf{x} | \omega_i) P(\omega_i) \end{aligned}$$

## 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

- ◆ 本当はすべ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 $p(\mathbf{x} | \omega_i)$ 」

- クラス分布  $p(\mathbf{x} | \omega_i)$  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
- いくつかの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

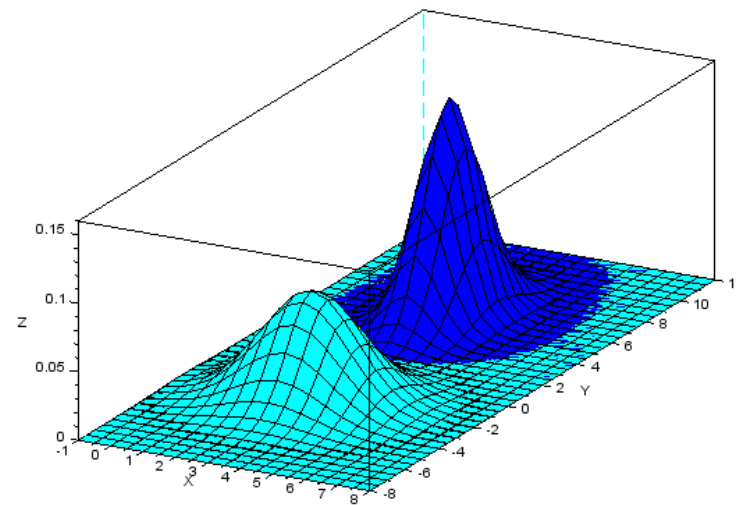
### 8.2.1 未知データの統計的性質を予測する

- 確率密度関数の例

- ◆ 正規分布 ( $d$ 次元)

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

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \Sigma_i^{-1} (\mathbf{x} - \mathbf{m}_i)\right\}$$

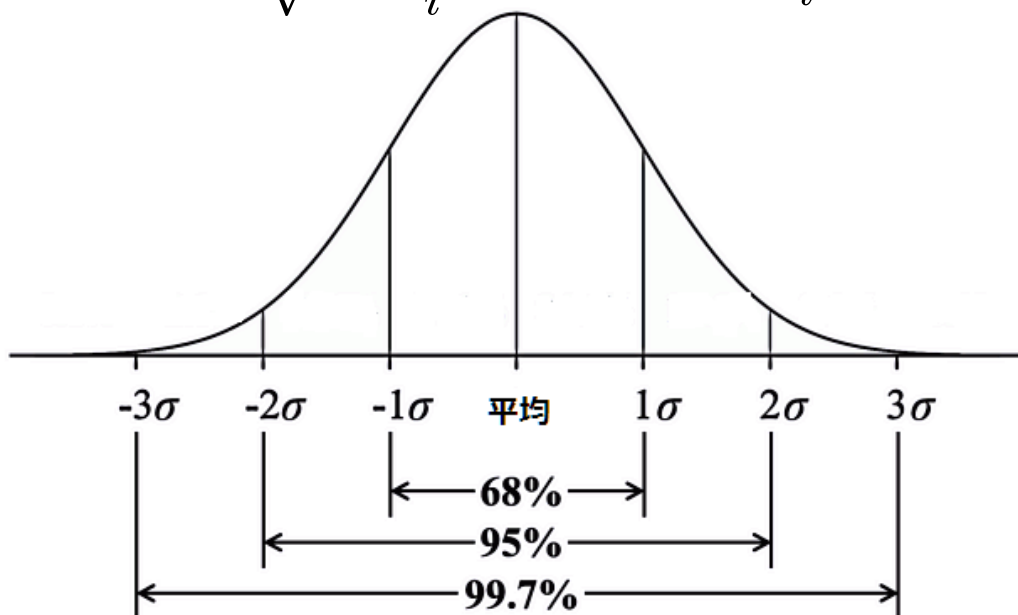


## 8.2.1 未知データの統計的性質を予測する

- 確率密度関数の例

- ◆ 正規分布 (1次元)

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(x - m_i)^2}{2\sigma_i^2}\right\}$$

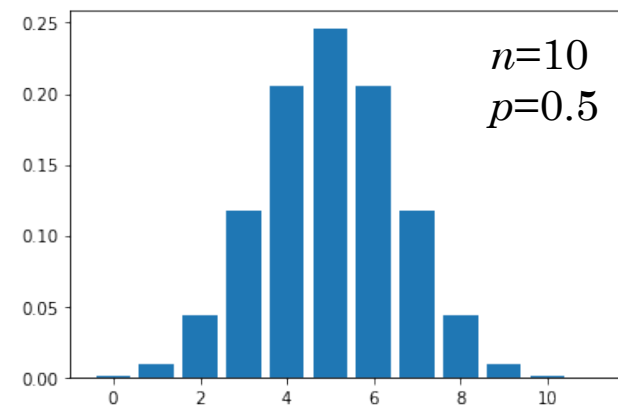


# 正規分布とは

- 離散型二項分布の例  $P(X = k) = {}_n C_k p^k (1 - p)^{n-k}$

◆  $n$ 枚のコインを投げた時の、表の枚数 $k$ の度数

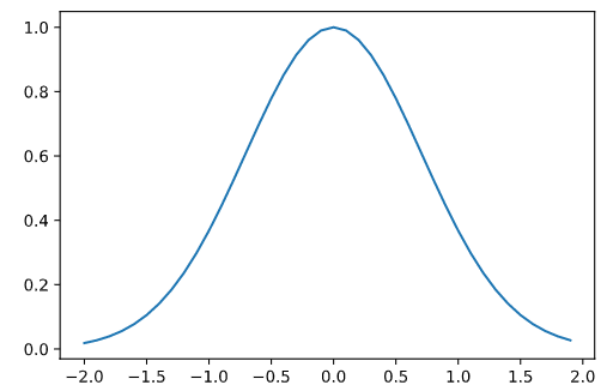
$n=1$		1	1				
$n=2$		1	2	1			
$n=3$		1	3	3	1		
$n=4$		1	4	6	4	1	
$n=5$		1	5	10	10	5	1
...							



◆  $n \rightarrow \infty$  のときが正規分布

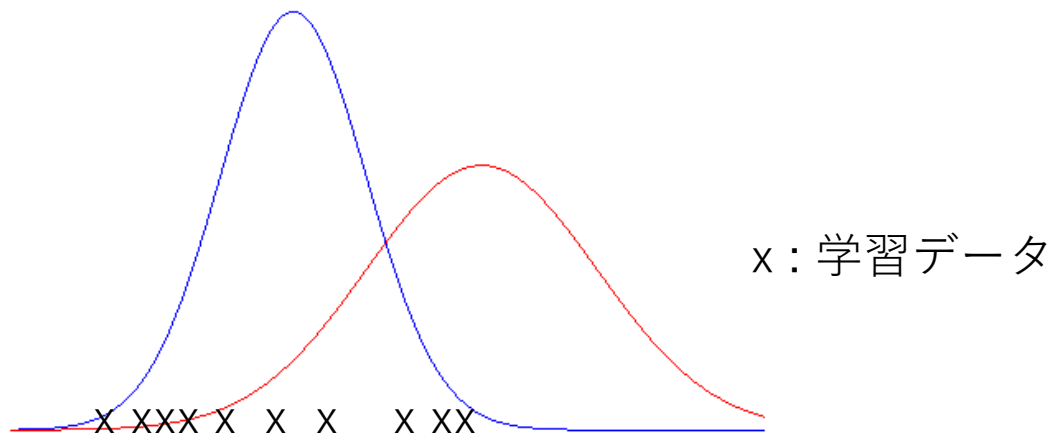
- 独立な要因の合計によって生じた数値がこの分布にあてはまる

例)  $n$ 問からなるテストの点数



## 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

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

## 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

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

## 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

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

- ◆ アンダーフローを避けるため対数をとる

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

- ◆ 尤度の部分に正規分布の式を適用 (→ 識別関数は  $\mathbf{x}$  の2次関数)

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



## 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
  - ◆  $\mathbf{x}$  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

$$g_i(\mathbf{x}) = \mathbf{x}^T \Sigma_0^{-1} \mathbf{m}_i - \frac{1}{2} \mathbf{m}_i^T \Sigma_0^{-1} \mathbf{m}_i + \log P(\omega_i)$$

- 共分散行列を単位行列(特徴間の相関がなく、分散も等しい)かつ事前確率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

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

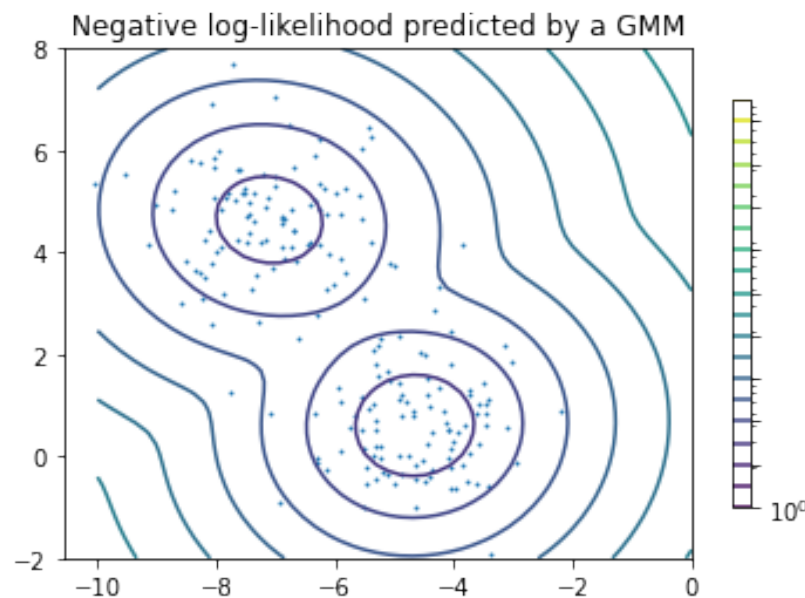
- 例題8.1

## 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
- 演習問題8-3

## 8.3 実践的な統計的識別

- EMアルゴリズム
  - ◆ 与えられた個数の正規分布をランダムに生成
  - ◆ それぞれの分布からデータが生成される確率を計算し、それを重みとして分布のパラメータを調整することを繰り返す



## 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

# まとめ

- 統計的パターン認識
  - ◆ 未知データに対する誤り確率を最小とするモデルを学習データから推定する
  - ◆ 事後確率最大化識別
    - ベイズの定理で尤度と事前確率の積最大となるクラスを決める問題に変換
    - それぞれを最尤推定
- 実践的な統計的識別
  - ◆ ナイーブベイズ法: 特徴ベクトルの各次元の独立性を仮定
  - ◆ 混合分布: 複数の正規分布で尤度関数をモデル化