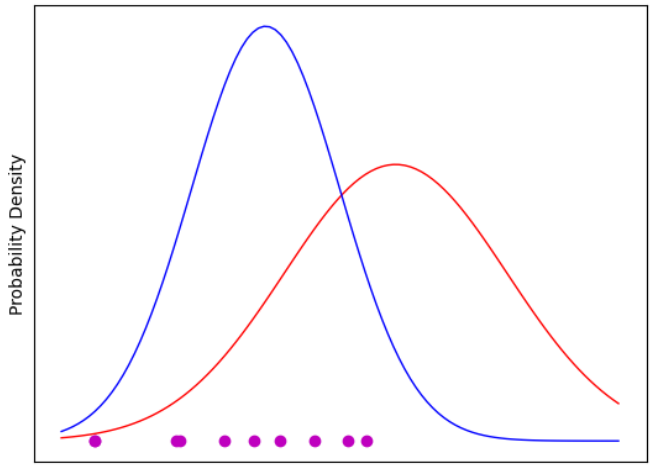
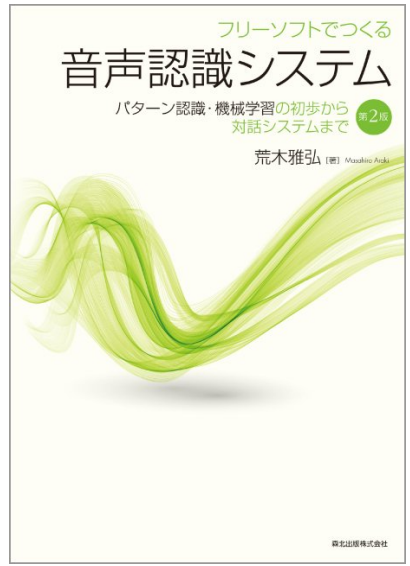
8. 未知データを推定しよう – 統計的方法 –



- 8.1 間違う確率を最小にしたい
- 8.2 データの広がりを推定する
- 8.3 実践的な統計的識別



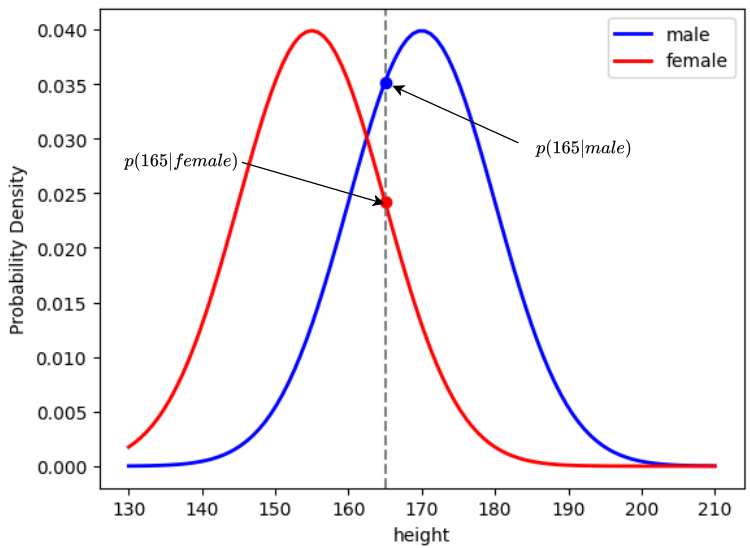
- 荒木雅弘:『フリーソフトでつくる音声認識システム(第2版)』(森北出版, 2017年)
- スライドとJupyter notebook
- サポートページ

8.1 間違う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
  - 誤差最小のシステム（ニューラルネットワークなど）
    - 学習データに最適化してしまい、未知データでは機能しないかもしれない
  - 誤り確率最小のシステム
    - 未知データの確率分布を推定し、誤り確率を最小にしたい
  - （期待損失最小のシステム）
    - 誤りの種類に依存したドメイン依存の損失を推定する必要があるので、ここでは扱わない

8.1.1 誤り確率最小の判定法 (1/2)

- 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
  - 統計で与えられるのは、性別がわかったときの身長の分布



8.1.1 誤り確率最小の判定法（2/2）

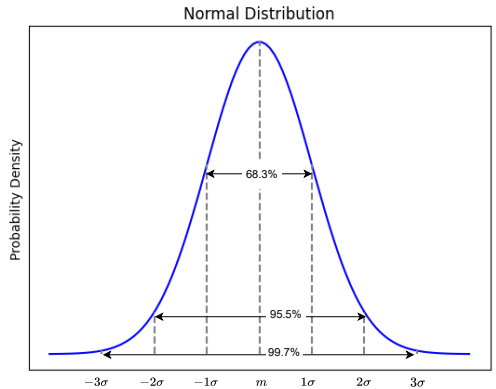
- 確率を用いたパターン認識
  - 事後確率最大化識別（ベイズ決定則）
    - $P(\omega_i|\boldsymbol{x})$ を最大にするクラス ω_i を識別結果とする

$$\underset{i=1,\ldots,c}{\arg\max}\, P(\omega_i|\boldsymbol{x}) = k \quad \Rightarrow \quad \boldsymbol{x} \in \omega_k$$

  - 身長による成人男女の判別システムの場合
    - 例）入力が 185.0 cmの時、$P(\omega_{male}|185.0)$ と $P(\omega_{female}|185.0)$ 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 $P(\omega|\boldsymbol{x})$ 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 $P(\omega_i|185.0)$ 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): $\delta = 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$
    - 二項分布に基づく比率の推定における95%信頼区間の計算



p=0.5のときの95%信頼区間

- n=100: 50.0%±10.0
- n=2,000: 50.0%±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（1/2）

- 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質量関数
  - 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（2/2）

- 事後確率 $P(\omega_i|\boldsymbol{x})$
  - $\boldsymbol{x}$ が生起したとき、そのクラスが ω_i である確率
- 事前確率 $P(\omega_i)$
  - クラス ω_i の生起確率
- クラスによらない $\boldsymbol{x}$ の生起確率 $p(\boldsymbol{x})$
- 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

$$p(\boldsymbol{x}|\omega_i) \quad (i = 1, \ldots, c)$$

    - クラス ω_i の分布から $\boldsymbol{x}$ が出現する確率

8.1.4 厄介者 $p(\boldsymbol{x})$ を消そう

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

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

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 ω_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「尤度 $p(\boldsymbol{x}|\omega_i)$」

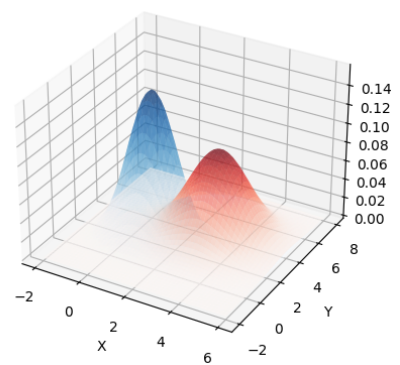
- 尤度 $p(\boldsymbol{x}|\omega_i)$ の求め方
  - 尤度とは
    - あるクラスのデータ集合から、ある特徴ベクトル $\boldsymbol{x}$ が観測される確率をあらわす確率密度関数
    - $\boldsymbol{x}$ の各要素が識別に役立つ特徴として選ばれているものならば、いくつかの値は観測されやすく、それらから遠くなるに従って観測されにくくなるような性質を持つはず
  - 確率分布の形を仮定して、そのパラメータを学習データから推定
    - 例）正規分布の場合：パラメータは平均ベクトルと共分散行列

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（1/2）

- 確率密度関数の例
  - 正規分布（d次元）
    - m_i：平均ベクトル，Σ_i：共分散行列

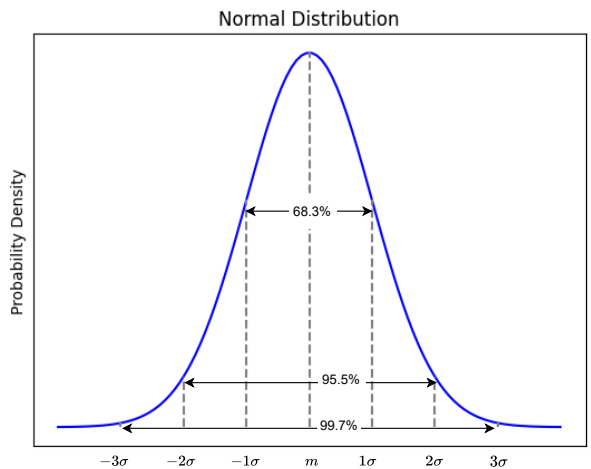
$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i)\}$$



8.2.1 未知データの統計的性質を予測する（2/2）

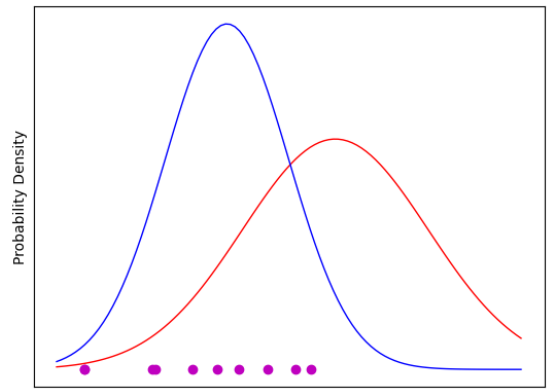
- 確率密度関数の例
  - 正規分布(1次元)

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



8.2.2 最尤推定 (1/3)

- データ χ 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
    - $\boldsymbol{\theta}$: 分布のパラメータ

$$\log L(\chi|\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in\chi} \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

8.2.2 最尤推定 (2/3)

- 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 (3/3)

- 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よぶ
  → 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的に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よぶ
  → 4～7章の学習アルゴリズム

8.2.3 統計的な識別 (1/2)

- 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

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

- クラス分布の部分に正規分布の式を適用(→ 識別関数は $\boldsymbol{x}$ の2次関数)

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

8.2.3 統計的な識別 (2/2)

- 共分散行列が全クラスで等しい場合
  - $\boldsymbol{x}$ 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

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

- 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)とし、かつ事前確率が全クラスで等しいとする
  - $\rightarrow$ 識別関数は最近傍決定則と同じになる

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

8.3 実践的な統計的識別

8.3.1 単純ベイズ法

- 特徴空間各次元の独立性を仮定
  - 推定対象の分布が1次元正規分布に単純化される
  - 少ないデータで確率密度関数を推定できる

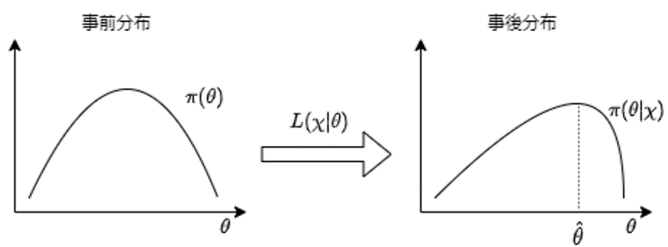
$$
\begin{aligned}
g_i(\boldsymbol{x}) &= p(\boldsymbol{x}|\omega_i)P(\omega_i) \\
&\sim \prod_{j=1}^{d} p(x_j|\omega_i)P(\omega_i)
\end{aligned}
$$

8.3.2 ベイズ推定

- 分布のパラメータを確率変数と考える
  - パラメータの事前分布 $\pi(\theta)$ を仮定し、データ観測後の事後分布 $\pi(\theta|\chi)$ をベイズの定理を用いて求める

$$\pi(\theta|\chi) \propto L(\chi|\theta)\pi(\theta)$$

    - 事前分布と事後分布が同じ分布のとき共役分布とよぶ
      - 事前分布が正規分布、尤度関数が正規分布であれば、事後分布も正規分布
      - 事前分布がベータ分布、尤度関数が二項分布であれば、事後確率もベータ分布
    - 求めた事後分布の平均値や最大値でパラメータ $\hat{\theta}$ を推定する

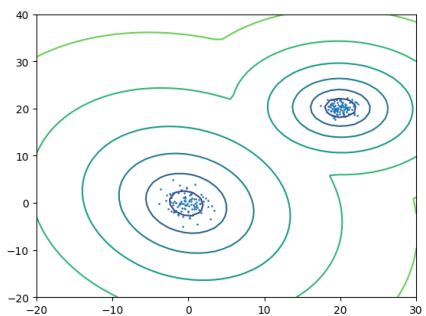


8.3.3 複雑な確率密度関数の推定

- 複数の正規分布の重み付き和(混合分布)を用いる

$$p(\boldsymbol{x}|\omega_i) = \sum_{j=1}^{k} w_{ij}\mathcal{N}(\boldsymbol{\theta}_{ij})$$

  - 各正規分布のパラメータと重みをEMアルゴリズムで学習
    - 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、その確率を重みとして分布のパラメータを最尤推定することを繰り返す



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が最大となるクラスを決める問題に変換
  - 事前確率と尤度を最尤推定
- 実践的な統計的識別
  - ナイーブベイズ法、ベイズ推定、混合分布
- Jupyter notebook