

パターン認識

<http://www.cs.miyazaki-u.ac.jp/~date/lectures/pattern/>

伊達 章

宮崎大学 工学部 情報システム工学科

2014年10月22日

講義のスケジュール (案)

1. 講義の概要 10/1
2. 準備：確率・統計の基礎 10/22
3. 準備：octave の使い方 10/22
4. 教師あり学習. 識別関数 10/29
5. 最大事後確率則, 最小誤識別則, ベイズ決定則 11/5
6. 最尤推定法 1: ガウスモデル 11/12
7. 最尤推定法 2: 線形判別分析 11/19
8. 線形判別分析により手書き文字認識 1
9. 線形判別分析により手書き文字認識 2
10. 混合ガウスモデルの最尤推定 1
11. 混合ガウスモデルの最尤推定 2
12. ノンパラメトリックな手法 (1): カーネル密度推定法
13. ノンパラメトリックな手法 (2): k -最近傍則
14. ノンパラメトリックな手法 (3): パーセプトロン
15. 定期試験, 解説

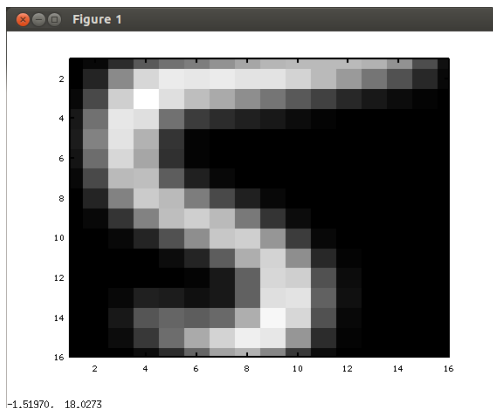
確率・統計の基礎

基本知識（確率・統計の復習）

- 確率変数, 確率密度関数
- 平均 μ , 分散 σ^2 , 標準偏差 σ
- 期待値, 分散共分散行列
- 独立, 相関係数, 無相関
- 確率分布: 一様分布, 正規分布, ガウス分布
- 同時確率, 条件付き確率, 周辺確率
- ベイズの公式, 事前確率・事後確率
- 擬似乱数の生成, 独立同一分布 (i.i.d.)
- 多次元正規分布

生成モデルに基づくパターン認識

'5' →
生成




→ y : '5'
認識

観測データ x , 推定対象 y

パターン認識の問題

識別関数 $f(x)$ を作ること

$$\mathbf{x} = (x_1, x_2, \dots, x_{256}) \rightarrow y = f(\mathbf{x})$$

	\mathbf{x}	y
0	00...00000000	$f(\mathbf{x}_0)$
1	00...00000001	$f(\mathbf{x}_1)$
2	00...00000010	$f(\mathbf{x}_2)$
3	00...00000011	$f(\mathbf{x}_3)$
	⋮	
k	00...11101011 	$f(\mathbf{x}_k) = 5$
	⋮	
$2^{256} - 1$	11...11111111	$f(\mathbf{x}_{2^{256}-1})$

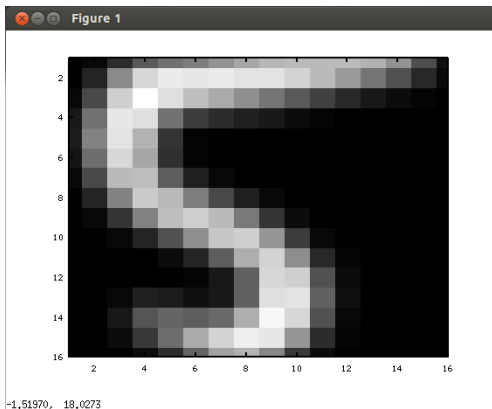
$x_i \in \{0, 1\}$ の場合. $2^{16 \times 16} = 2^{256} \approx 10^{75}$

モデル化

確率的生成モデル

確率的生成モデルに基づくパターン認識

'5' →
生成



→ y : '5'
認識

$y \sim p(y) \rightarrow$ データ $\mathbf{x} \sim p(\mathbf{x}|y) \rightarrow$ 認識 $\hat{y} = \underset{y}{\operatorname{argmax}} p(y|\mathbf{x})$

モデル化: $p(y)$ と $p(\mathbf{x}|y)$ を設計する!

モデル化：確率的生成モデル

- $p(y)$ ：事前分布，事前確率
どの数字が出現する確率が高い？
- $p(x|y)$ ：データモデル
例：数字「2」を書いてもらった。そのとき画像 x が描かれる確率
- x （画像データ，観測できる）
- y カテゴリ（認識の際は見えない，隠れ変数）

モデル化：確率的生成モデル

- $p(y)$ ：事前分布，事前確率
どの数字が出現する確率が高い？
- $p(x|y)$ ：データモデル
例：数字「2」を書いてもらった。そのとき画像 x が描かれる確率
- x （画像データ，観測できる）
- y カテゴリ（認識の際は見えない，隠れ変数）
↑これらはこの教科書での文字の使い方。
世の中では x と y の使い方が逆

ともかく確率が重要

確率，条件付き確率

	B_1 (風邪)	B_2 (風邪なし)	$p(A_i)$
A_1 (熱あり)	0.55	0.05	0.60
A_2 (熱なし)	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

例

同時確率 $p(A_1, B_1) = 0.55$

周辺確率 $p(A_1) = \sum_i p(A_1, B_i) = p(A_1) = 0.6$

条件付き確率

熱の有無を知る \Rightarrow 風邪であるかどうか検討がつく：

$$p(B_1|A_1) = \frac{p(B_1)p(A_1|B_1)}{p(A_1)} = \frac{p(A_1, B_1)}{p(A_1)} = \frac{0.55}{0.6} \approx 0.92$$

確率，条件付き確率

	B_1 (白)	B_2 (黒)	$p(A_i)$
A_1 (白)	0.55	0.05	0.60
A_2 (黒)	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

例： A と B はそれぞれ画像を構成するピクセル

同時確率 $p(A_1, B_1) = 0.55$

周辺確率 $p(A_1) = \sum_i p(A_1, B_i) = p(A_1) = 0.6$

条件付き確率

A の白黒を知る $\Rightarrow B$ が白黒どちらであるか検討がつく：

$$p(B_1|A_1) = \frac{p(B_1)p(A_1|B_1)}{p(A_1)} = \frac{p(A_1, B_1)}{p(A_1)} = \frac{0.55}{0.6} \approx 0.92$$

ベイズの公式

- ベイズの公式

熱があった (A_1) とする.

その時, 風邪のある (B_1), なし (B_2) の確率

$$p(B_1|A_1) = \frac{p(B_1)p(A_1|B_1)}{p(A_1)} = \frac{p(A_1, B_1)}{p(A_1)} = \frac{0.55}{0.6} \approx 0.92$$

$$p(B_2|A_1) = \frac{p(B_2)p(A_1|B_2)}{p(A_1)} = \frac{p(A_1, B_2)}{p(A_1)} = \frac{0.05}{0.6} \approx 0.08$$

- 事後確率最大化 (ベイズ推定) $\operatorname{argmax}_i p(B_i|A_1) = 1$

風邪であることの方が確率が大 \Rightarrow 風邪であると推定

入力 (観測値) : A 熱のあるなし

\Rightarrow 出力 (推定値) B 風邪かどうか

平均, 分散

- 平均 μ , 期待值 $E[x]$

$$\mu = E[x] = \sum_{i=1}^n x_i p(x_i), \quad \int_{-\infty}^{\infty} x p(x) dx$$

平均, 分散

- 平均 μ , 期待値 $E[x]$

$$\mu = E[x] = \sum_{i=1}^n x_i p(x_i), \quad \int_{-\infty}^{\infty} x p(x) dx$$

平均は分かった. 例: 数学のテストの平均 70 点
その周りにどの程度の幅でばらついているかも知りたい!

平均, 分散

- 平均 μ , 期待値 $E[x]$

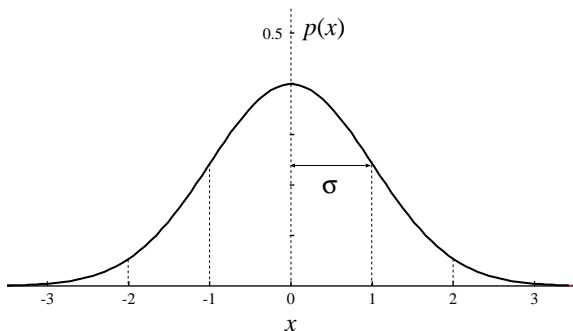
$$\mu = E[x] = \sum_{i=1}^n x_i p(x_i), \quad \int_{-\infty}^{\infty} x p(x) dx$$

平均は分かった. 例: 数学のテストの平均 70 点
その周りにどの程度の幅でばらついているかも知りたい!

- 分散 σ^2 , 標準偏差 σ : **【一つの指標】**

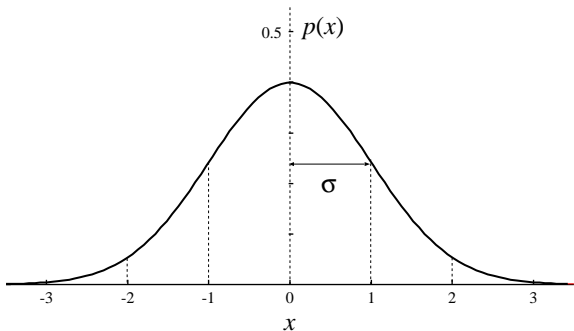
$$\sigma^2 = E[(x - \mu)^2]$$

正規分布, ガウス分布



「 x_1, x_2, \dots, x_{100} を平均 μ , 分散 σ^2 の互いに独立なガウス分布に従う確率変数とする」

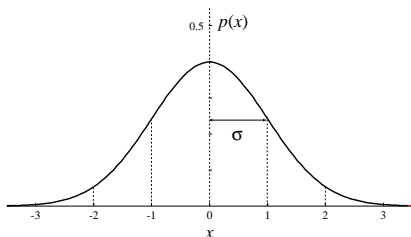
正規分布, ガウス分布 $\mathcal{N}(\mu, \sigma^2)$



$$p(x; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad p(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

正規分布, ガウス分布

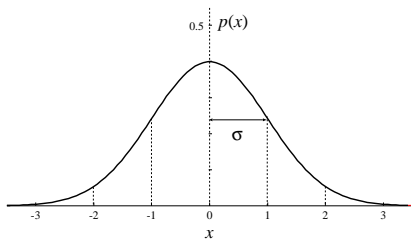


$$p(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$x_i, i = 1, \dots, 1000$ のうち約 68.26% が $-1 < x_i < 1$ に含まれている. その根拠:

$$\int_{-1}^1 p(x) dx = 0.6826$$

正規分布, ガウス分布



$$\int_{-2}^2 p(x) dx = 0.9544, \quad \int_{-3}^3 p(x) dx = 0.9974$$

基本知識（確率・統計の復習）

- 確率変数, 確率密度関数
- 平均 μ , 分散 σ^2 , 標準偏差 σ
- 期待値, 分散共分散行列
- 独立, 相関係数, 無相関
- 確率分布: 一様分布, 正規分布, ガウス分布
- 同時確率, 条件付き確率, 周辺確率
- ベイズの公式, 事前確率・事後確率
- 擬似乱数の生成, 独立同一分布 (i.i.d.)
- 多次元正規分布

擬似乱数

- 一様分布 [0:1]

```
octave:1> rand(5)
```

```
ans =
```

```
0.556212    0.518803    0.589602    0.645093    0.707168  
0.088337    0.307372    0.859300    0.790555    0.412982  
0.756140    0.217823    0.442209    0.815839    0.149388  
0.573751    0.336075    0.236351    0.863245    0.413433  
0.397294    0.884367    0.719179    0.476957    0.571799
```

```
octave:2> rand(5,1)
```

```
ans =
```

```
0.52111  
0.49983  
0.26851  
0.58936  
0.93169
```

擬似乱数

- 正規分布（ガウス分布）

平均 $\mu=72$, 標準偏差 $\sigma=8$ の正規分布 $\mathcal{N}(72, 8^2)$ に
したがうデータを 5 個生成

```
octave:14> 8*randn(1,5) + 72  
ans =
```

```
70.927    76.224    78.489    70.905    69.532
```

擬似乱数

- 正規分布（ガウス分布）

平均 $\mu=72$, 標準偏差 $\sigma=8$ の正規分布 $\mathcal{N}(72, 8^2)$ に
したがうデータを 5 個生成

```
octave:14> 8*randn(1,5) + 72  
ans =
```

70.927 76.224 78.489 70.905 69.532

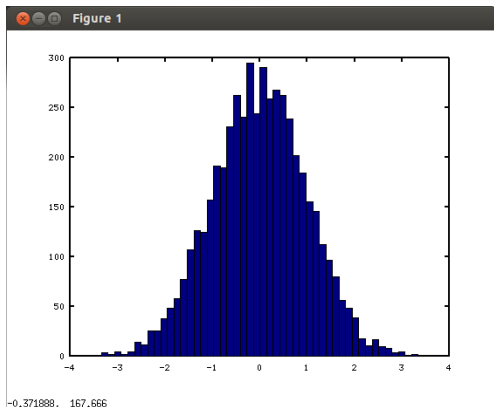
- 標準偏差 σ の意味？！

octave で正規分布にしたがうデータを生成

```
1 octave:1> x = 0.7*randn(200,1);
2 octave:2> mean(x)
3 ans = 0.076972
4 octave:3> sqrt(var(x))
5 ans = 0.68542
```

- 平均 $\mu = 0$, 標準偏差 $\sigma = 0.7$ の正規分布にしたがうデータを 200 個生成
- 正規分布 (= Gauss 分布) とは？
- seed の設定: `randn("seed", 20141022)`

正規分布（ガウス分布）



```
1 octave:19> x=randn(5000,1);  
2 octave:21> hist(x,50)
```

- $\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(0, 1)$ にしたがう 5000 のデータ
- $[-1 : 1]$ にあるデータは何%？ $[-3 : 3]$ は？

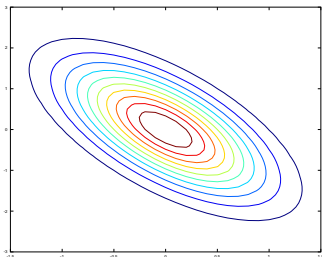
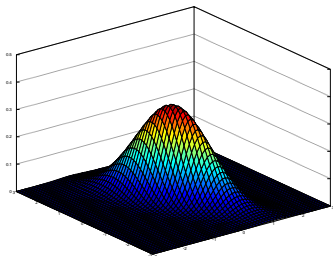
確かめてみる

```
1 n=5000 % 生成するデータの個数
2 s=2.0; % 1,2,3 と値を変えてみる
3 x=randn(n,1); % 正規分布の乱数を生成
4
5 c=0;
6 for i=1:n
7     if ( x(i) > -s && x(i) < s )
8         c=c+1;
9     end
10 end
11 c, c/n
```

※ for 文を使っているので、参考にしすぎないこと！

2次元正規分布

2次元正規分布

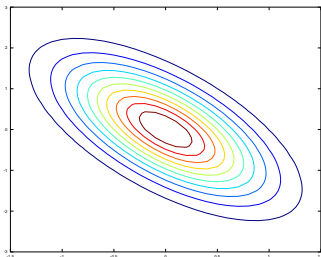
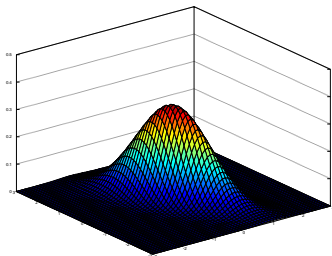


$$p(\mathbf{x}; \boldsymbol{\mu}, V) = c e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T V^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(\mathbf{x}; \boldsymbol{\mu}, V) d\mathbf{x} = 1$$

$$c = \frac{1}{2\pi\sqrt{|V|}}, \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, V = \begin{pmatrix} \sigma_{x^2} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{y^2} \end{pmatrix}$$

多次元正規分布



$$p(\mathbf{x}; \boldsymbol{\mu}, V) = c e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T V^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(\mathbf{x}; \boldsymbol{\mu}, V) d\mathbf{x} = 1$$

$$c = \frac{1}{(\sqrt{2\pi})^n \sqrt{|V|}}$$

分散共分散行列 V

$$\begin{aligned} V &= \begin{pmatrix} \sigma_{x^2} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{y^2} \end{pmatrix} \\ \sigma_x^2 &= \text{E}[(x - \mu_x)^2] \\ \sigma_{xy} &= \text{E}[(x - \mu_x)(y - \mu_y)] \\ V &= \text{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] \\ &\approx \frac{1}{n} \sum_{\alpha=1}^n (\mathbf{x}^\alpha - \boldsymbol{\mu})(\mathbf{x}^\alpha - \boldsymbol{\mu})^T \end{aligned} \quad (1)$$

V は対称行列, 正定値 (すべての固有値が正)

終