

機械学習

第2回 確率モデルと機械学習

兵庫県立大学 社会情報科学部

川嶋宏彰

kawashima@sis.u-hyogo.ac.jp

本日の講義内容

2

1. 確率モデルを使った分類

章番号は参考書（荒木）

- ベイズの定理で確率推論
- データを**ガウス分布（正規分布）**でモデル化

4章

5章前半 (p.82など)

2. 2次元ガウス分布による分類（教師あり）

- 2次元ガウス分布（2次元正規分布, 二変量…）
- 共分散は楕円のイメージ

p.201

3. クラスタリング（教師なし学習）

11章

- k-means 法
- 混合ガウス分布



確率を使って推論する (infer)

3

- 箱に以下の玉（カプセルトイ，不透明）が入っている
 - Aのおまけ入り赤玉が4個，白玉が2個（計6個） ●●●●○○
 - Bのおまけ入り赤玉が1個，白玉が3個（計4個） ●○○○
- 箱から玉が一つ無作為に取り出されるとする
- その玉はA, Bどちらのおまけが入っているか確率的に推論してみよう

事前確率 (prior)

- 取り出された玉の色を見る前： $P(A) = ?$ $\left[P(B) = 1 - P(A) \right]$

- 取り出された玉の色を見た後：

- 赤だったとき： $P(A | \bullet) = ?$
- 白だったとき： $P(A | \circ) = ?$

事後確率
(posterior)

玉の色：観測（入力）
から推論！

[復習] 確率の乗法定理

4

乗法定理

$$P(X, Y) = P(X | Y) P(Y) = P(Y | X) P(X)$$

最初の等号を
確認してみよう

| 色X 中身Y | ● | ○ | 計 |
|-----------|---|---|----|
| A | 4 | 2 | 6 |
| B | 1 | 3 | 4 |
| 計 | 5 | 5 | 10 |

玉の数

同時確率

$$P(X = \bullet, Y = A)$$

条件付き確率

$$P(X | Y = A)$$

| 色X 中身Y | ● | ○ | 計 $P(Y)$ |
|-----------|------|------|----------|
| A | 4/10 | 2/10 | 6/10 |
| B | 1/10 | 3/10 | 4/10 |
| 計 $P(X)$ | 5/10 | 5/10 | 1 |

$Y = A$ の確率

| 色X 中身Y | ● | ○ | 計 |
|-----------|-----|-----|---|
| A | 4/6 | 2/6 | 1 |
| B | 1/4 | 3/4 | 1 |

$Y = A$ の下での $X = \bullet$ の確率

[復習] ベイズの定理（重要！）

5

- 乗法定理

$$P(X, Y) = P(X | Y) P(Y) = P(Y | X) P(X)$$

これより

- ベイズの定理

$$P(Y | X) = \frac{P(X | Y) P(Y)}{P(X)}$$

ただし $P(X) \neq 0$ とする

まずは形を覚えよう

条件付き確率 $P(Y|X) = \frac{P(X,Y)}{P(X)}$ に
 $P(X, Y) = P(X|Y)P(Y)$ を代入する
と覚えてもよい

ベイズの定理で事後確率

6

- 箱に以下の玉（カプセルトイ，不透明）が入っている
 - Aのおまけ入り赤玉が4個，白玉が2個（計6個） ●●●●○○
 - Bのおまけ入り赤玉が1個，白玉が3個（計4個） ●○○○
- 箱から玉が一つ無作為に取り出されるとする
- その玉はA, Bどちらのおまけが入っているか確率的に推論してみよう

事前確率

- 取り出された玉の色を見る前： $P(A) = \frac{6}{10} = 0.6$ $P(B) = 0.4$

- 取り出された玉の色を見た後：

事後確率

- 赤だったとき： $P(A | \bullet) = \frac{P(\bullet | A) P(A)}{P(\bullet)} = \frac{\frac{4}{6} \times \frac{6}{10}}{\frac{5}{10}} = 0.8$ $P(B | \bullet) = 0.2$

事前確率 $P(A)$ と $P(\bullet|A)$ だけ知っていれば計算できる！

[レポート: 問1] ベイズの定理の復習

7

- 箱に以下の玉（カプセルトイ，不透明）が入っている
 - Aのおまけ入り赤玉が4個，白玉が2個（計6個） ●●●●○○
 - Bのおまけ入り赤玉が1個，白玉が3個（計4個） ●○○○
 - 箱から玉が一つ無作為に取り出されるとする
 - その玉はA, Bどちらのおまけが入っているか確率的に推論してみよう
- 取り出された玉の色を見る前： $P(A) = 0.6$ $P(B) = 0.4$
- 取り出された玉の色を見た後：
 - 白だったとき： $P(A | \bigcirc) =$ $P(B | \bigcirc) =$

[練習問題]ベイズの定理の復習

8

① ある病原体に感染しているか否かを調べるための検査を行う．検査対象となる人のうち，100人に2人の割合で感染があり (I)，100人に98人の割合で感染がない (U) とする．これを確率 $P(I) = 0.02$ や $P(U) = 0.98$ で表す．また，検査の精度は100%ではなく，病原体が検出される場合を pos ，検出されない場合を neg と表せば，実際に感染しているときに正しく検出する確率は $P(pos | I) = 0.99$ であり，感染していないときに検出しない確率は $P(neg | U) = 0.9$ である．さて，太郎君が1回の検査を受けたところ病原体が検出された．このとき，太郎君が実際に病原体に感染している確率（陽性適中率） $P(I | pos)$ を計算せよ．

• ヒント

- 確率の加法定理 $P(pos, I) + P(pos, U) = P(pos)$
- $P(pos | U)$ は $P(neg | U)$ から計算できる

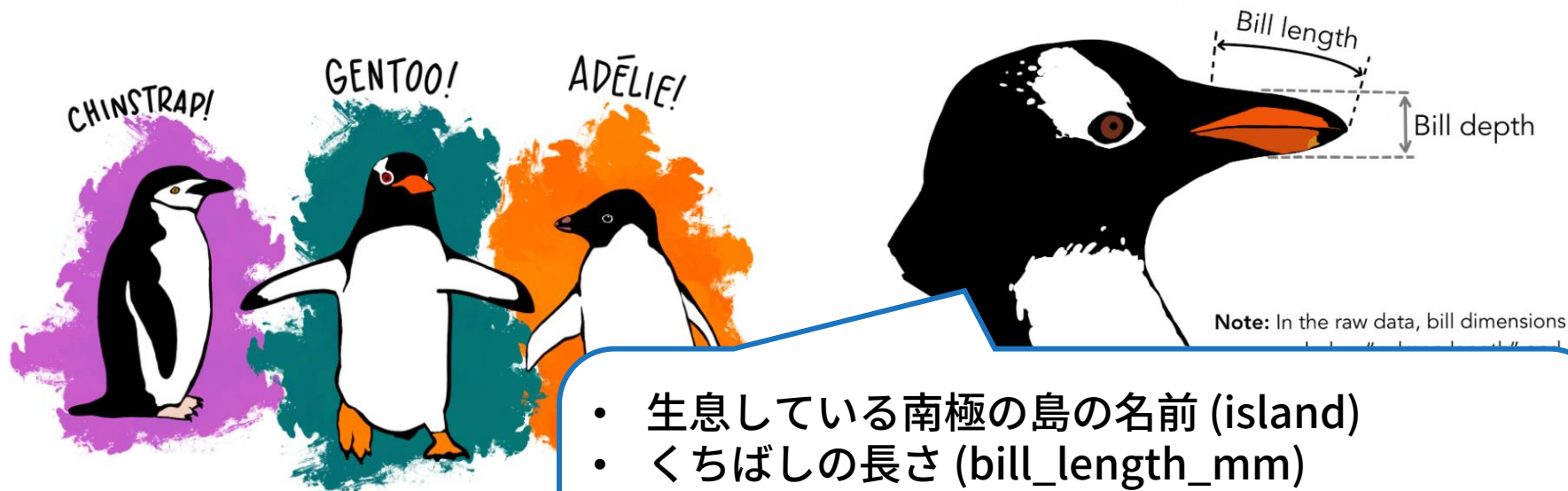
（注）求まる確率（陽性適中率）が結構小さいので意外に思うかもしれない．

② 実際の場合面では検査対象を絞る必要がある． $P(I) = 0.5$ の場合では陽性適中率はどうなるか？実際に確率を計算し，①の結果と比較して考察せよ．検査対象を絞らず全員を検査することの社会的な意味も考えてみよう（どのような場合に全員検査すべき？すべきでない？）

(続) Palmerpenguins データセット

9

- 3種のペンギン (今日もひとまず2種で)
 - アデリー (Adelie), ヒゲペンギン (Chinstrap), ジェンツー (Gentoo)



Artwork by @allison_host

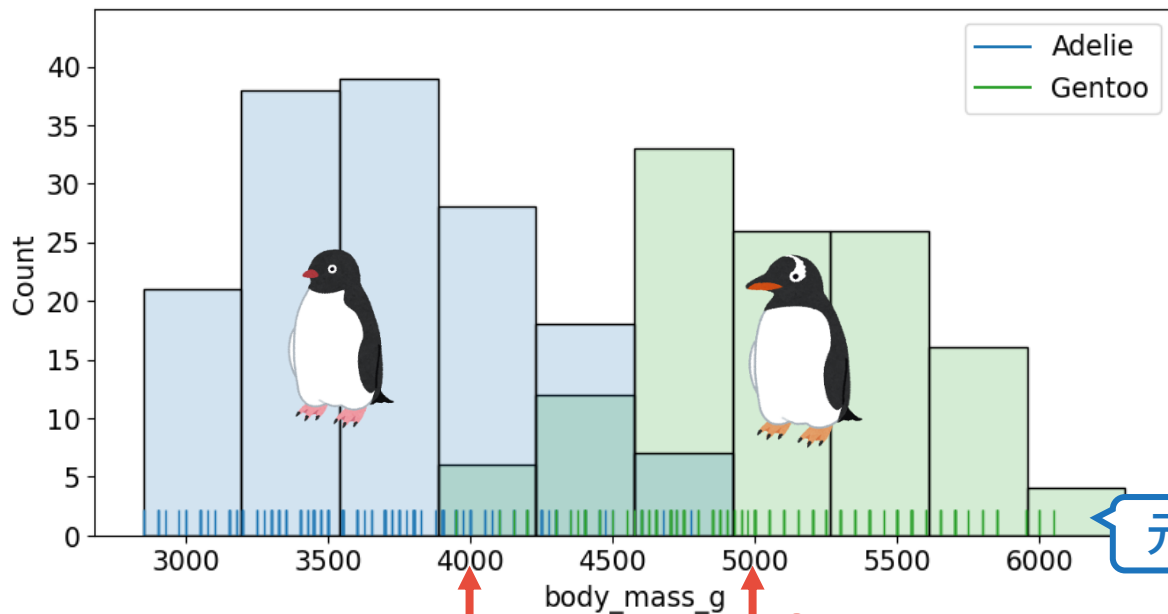
- 生息している南極の島の名前 (island)
- くちばしの長さ (bill_length_mm)
- くちばしの高さ (bill_depth_mm)
- フリッパー (翼) の長さ (flipper_length_mm)
- 体重 (body_mass_g), 性別 (sex) など

体重だけ分かった時にペンギンを分類したい

10

- 4 kg, 5kg のペンギンはそれぞれどちらと分類すればよい？
 - アデリーペンギン (Adelie)
 - ジェンツーペンギン (Gentoo)

これを確率論で計算しよう



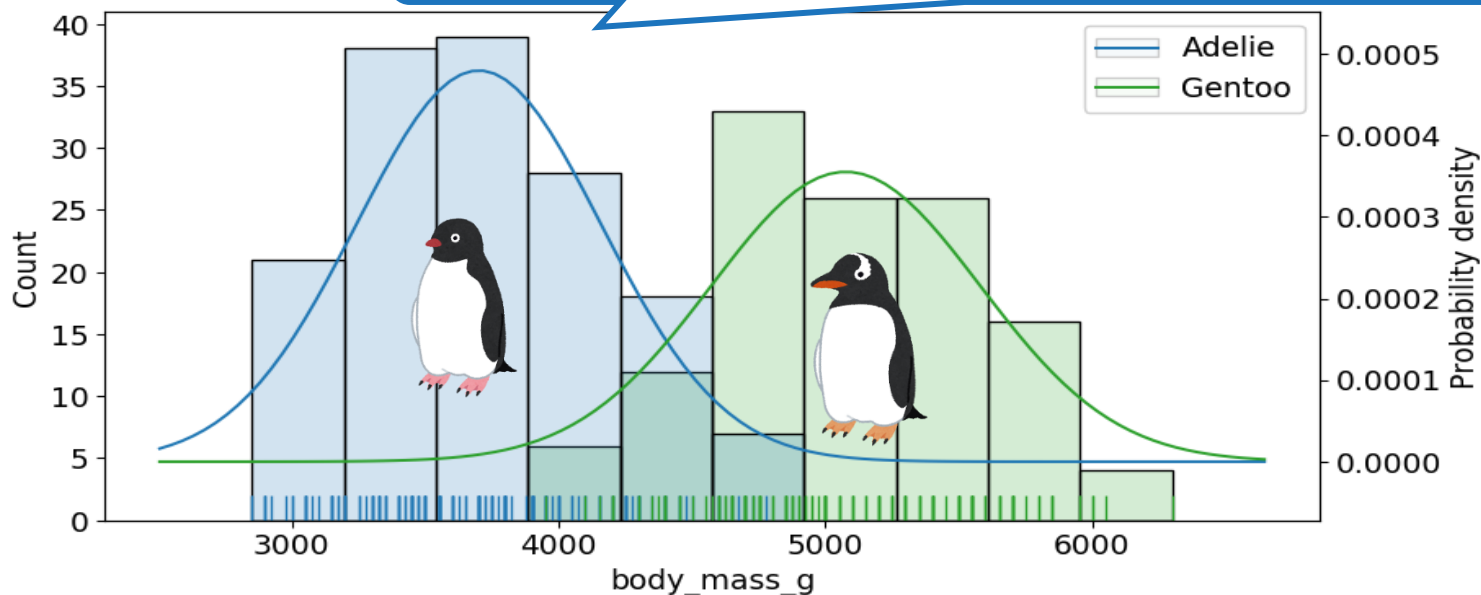
元データのrug plot

体重分布をガウス分布でモデル化

11

- 実数値データは「連続型の」確率分布でモデル化できる
 - ヒストグラム: ペンギン種ごとの度数分布を棒グラフで表示 (左軸)
 - 曲線: ペンギン種ごとのデータをガウス分布で近似 (右軸)

ひとまずヒストグラムが曲線で近似できそうなことに注目



[復習] 離散型の確率分布と確率質量関数

12

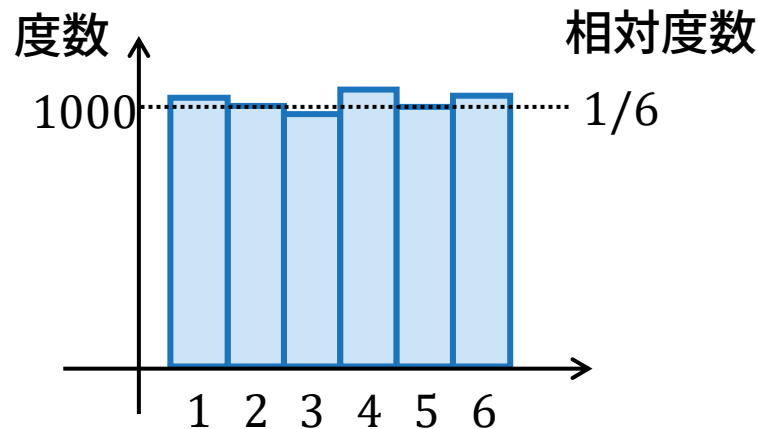
- 離散的な値を取るデータの分布は**確率質量関数**でモデル化
- 例: サイコロの出る目を確率変数 X で表す
 - 確率変数**: 確率的に値が決まる変数



確率質量関数: 確率変数 X の取りうる値とその確率の対応 (表)

| x | 1 | 2 | 3 | 4 | 5 | 6 |
|--------|---------------|---------------|---------------|---------------|---------------|---------------|
| $f(x)$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

公平なサイコロで出る目の確率モデル

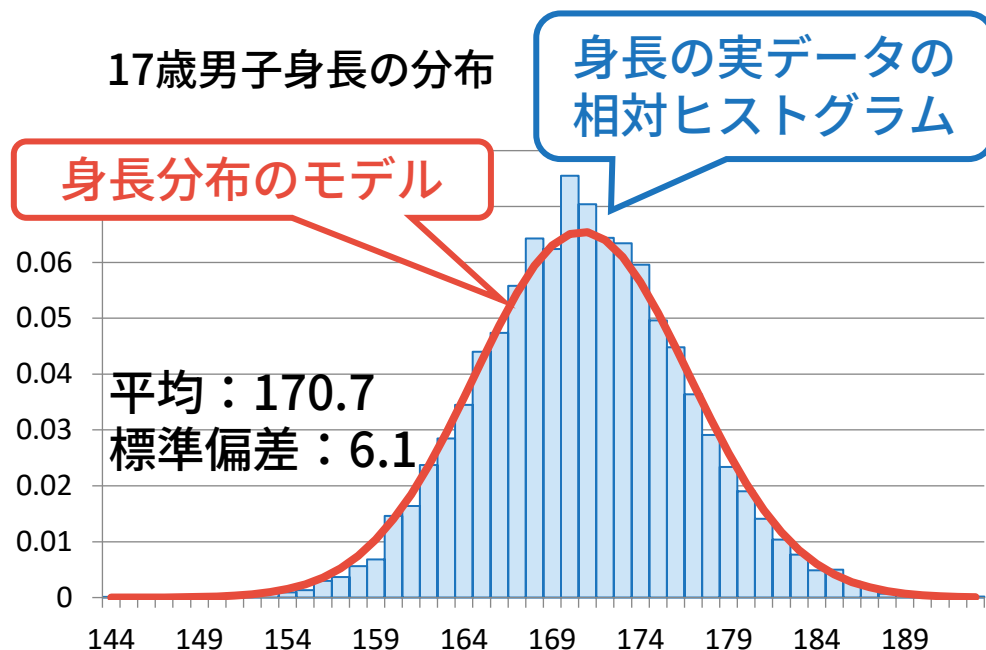


実際のデータ

[復習] 連続型の確率分布と確率密度関数

13

- 連続的な値を取るデータの分布は確率密度関数でモデル化
 - 確率密度 \neq 確率 であることに注意！



160cm以上, 161cm未満の確率は…

160.0cm以上, 160.1cm未満の確率は…

160.00cm以上, 160.01cm未満の確率は…

身長は元々実数…

どの幅で分布を考えればよい？

学校保健統計調査 平成24年度

十分細かな幅で確率を考える → 確率密度

[復習] 確率と確率密度

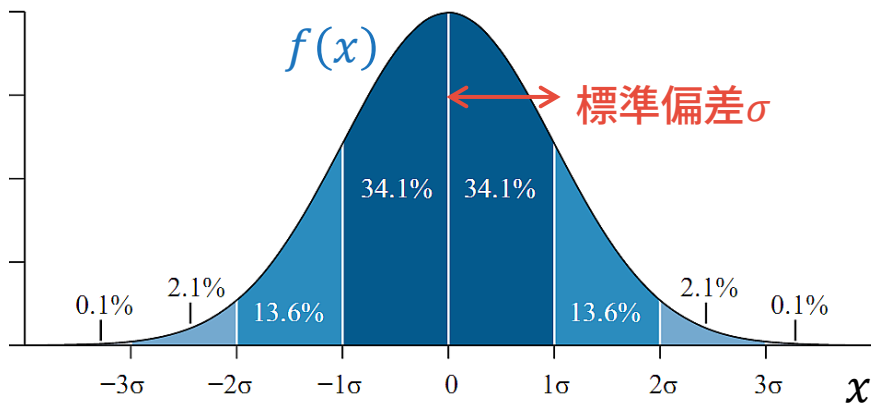
14

- 確率密度関数は積分するとその範囲の確率になる

- 連続変数 X が範囲 $[a, b]$ の値をとる確率： $P(a \leq X \leq b) = \int_a^b f(x) dx$
- X が取り得る全区間で積分すると1

例：平均0の正規分布
(別名ガウス分布)

この分布の密度関数は $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$



$$P(-\sigma \leq X \leq \sigma) = 0.6826$$

$$P(-2\sigma \leq X \leq 2\sigma) = 0.9544$$

$$P(-3\sigma \leq X \leq 3\sigma) = 0.9974$$

曲線 $f(x)$ の下面積

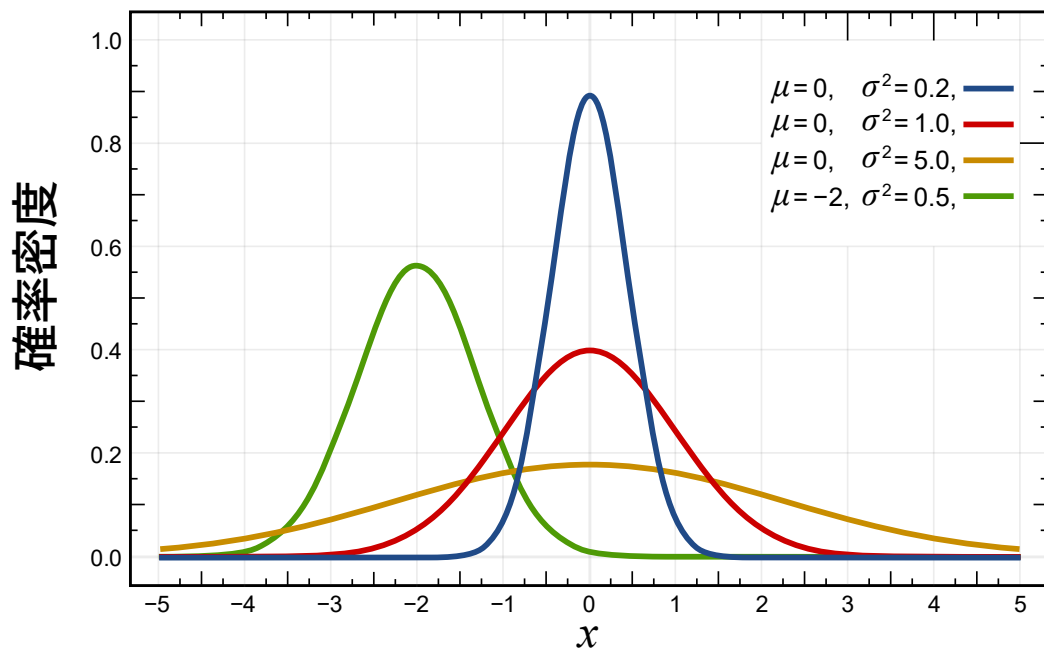
正規分布 (Normal distribution) (ガウス分布)

15

- 正規分布 $N(\mu, \sigma^2)$ の確率密度関数

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

平均値 μ : 中心
標準偏差 σ : ばらつき具合



補足: $f(x; \theta)$ の記法
「 θ をパラメタとして持つ
 x の関数 f 」 という意味

統計学では「正規分布 $N(\mu, \sigma^2)$ の密度関数は $f(x)$ 」と厳密に分けるが
機械学習では「正規分布 $N(x; \mu, \sigma^2)$ 」のように、分布と密度関数を区別せず呼ぶことが多い

1. 平均0のガウス分布で確率 $P(X = 0)$ の値は？(理由と共に)

確率 $P(X = 0)$ は0 (積分範囲が $[0, 0] \cdots P(0 \leq X \leq 0) = \int_0^0 f(x) dx = 0$)

ちなみに、確率密度 $p(0)$ は0ではない. $p(0) = \frac{1}{\sqrt{2\pi}\sigma}$

2. ガウス分布で平均の $\pm 3\sigma$ 以内の値を取る確率は99% 以上?

Yes

3. 確率は1までだが、密度関数の値は1を超えることがある？
(理由と共に. 1を超える場合があるなら具体例を)

Yes 例：ガウス分布で σ が小さい場合

例：矩形をした密度関数で横幅0.1 → 高さ10

4. $P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x) dx = 1$

体重分布をガウス分布でモデル化

17

- 種ごとの体重分布をガウス分布で近似

$$p(x | \text{Adelie}) = N(x; \mu_{\text{Adelie}}, \sigma_{\text{Adelie}}^2)$$

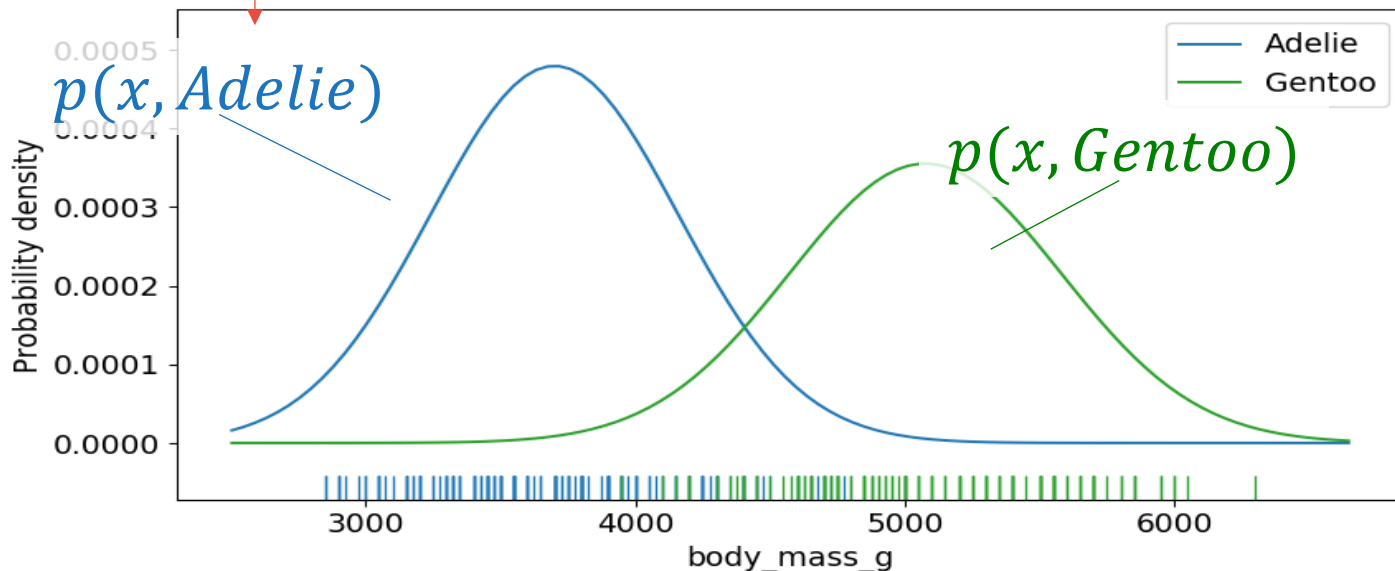
$$p(x | \text{Gentoo}) = N(x; \mu_{\text{Gentoo}}, \sigma_{\text{Gentoo}}^2)$$

Adelieならば体重 x
はこの分布に従う

Gentooならばこの分布

これより $p(x, \text{Adelie}) = p(x | \text{Adelie})P(\text{Adelie})$

出現確率も考慮



事後確率による分類

18

- 観測を得ると「事前確率」が「事後確率」に変わる

$$\bullet \quad P(\text{Adelie} \mid x) = \frac{p(x, \text{Adelie})}{p(x)} = \frac{p(x \mid \text{Adelie})}{p(x)} P(\text{Adelie})$$

事後確率

事前確率

- 事後確率の大小で分類できる

$$\bullet \quad P(\text{Adelie} \mid x) = \frac{p(x, \text{Adelie})}{p(x)} = \frac{p(x \mid \text{Adelie}) P(\text{Adelie})}{p(x)}$$

$$\bullet \quad P(\text{Gentoo} \mid x) = \frac{p(x, \text{Gentoo})}{p(x)} = \frac{p(x \mid \text{Gentoo}) P(\text{Gentoo})}{p(x)}$$

この授業での記法

$P(\cdot)$: 確率

$p(\cdot)$: 確率密度関数

分母 $p(x)$ は共通なので分子だけ比較すればよい

事後確率の大小による決定境界

19

$$p(x | \text{Adelie})P(\text{Adelie}) \text{ vs } p(x | \text{Gentoo})P(\text{Gentoo})$$

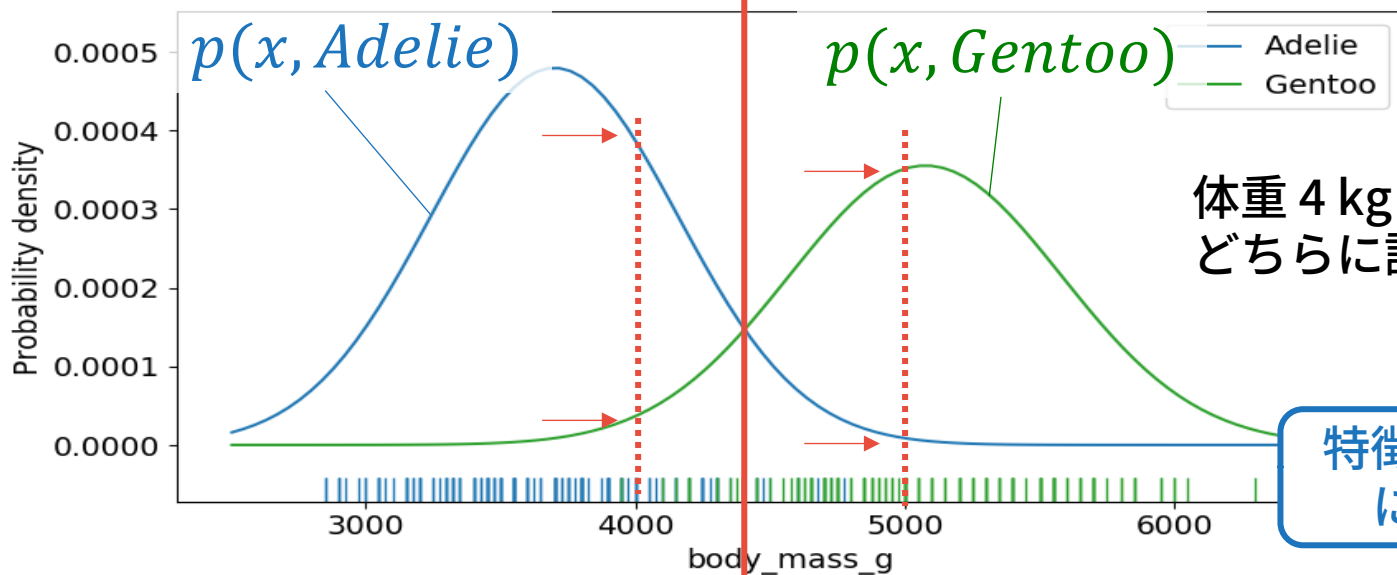
$$\parallel N(x; \mu_{\text{Adelie}}, \sigma_{\text{Adelie}}^2)$$

Adelieならば体重 x
はこの分布に従う

$$\parallel N(x; \mu_{\text{Gentoo}}, \sigma_{\text{Gentoo}}^2)$$

Gentooならば体重 x
はこの分布に従う

決定境界: 4400gあたり



体重 4 kg や 5 kg は
どちらに認識される?

特徴量が2次元
になると?

2次元ガウス分布は山のイメージ

20

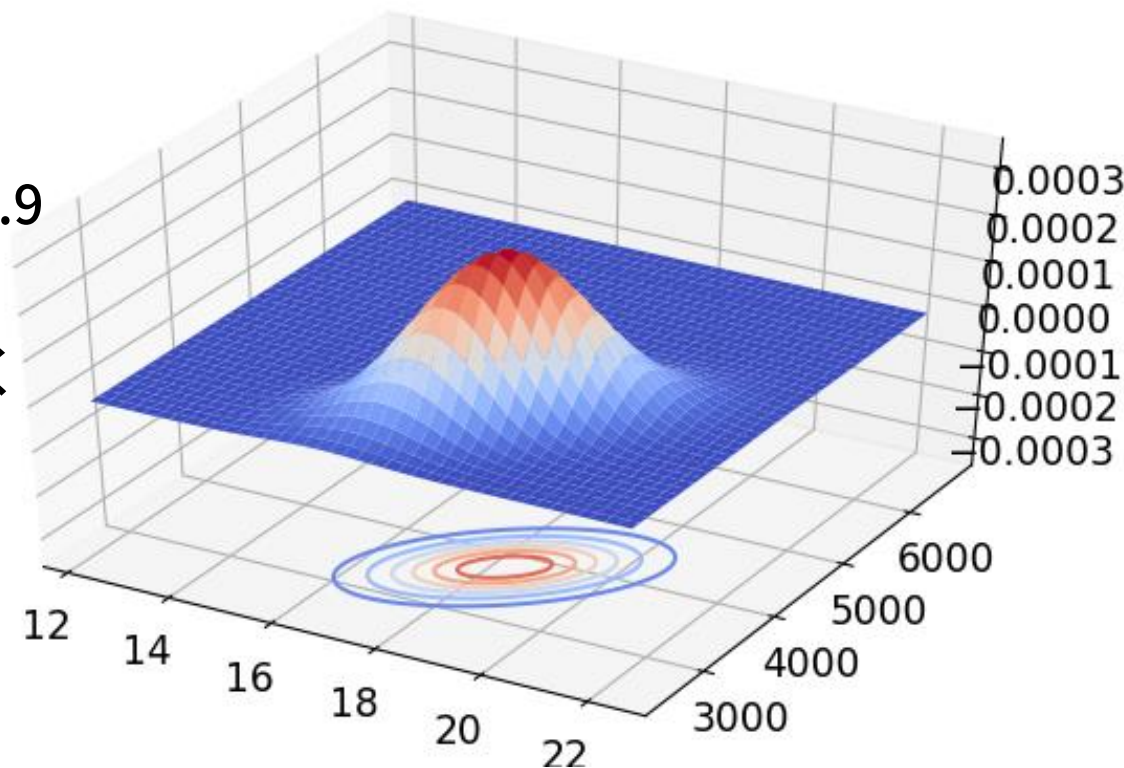
- ・ ガウス分布では確率密度の等しい点を結ぶと楕円になる
(山の等高線のイメージ)

- ・ 確率楕円とも呼ばれる

(例) 90%確率楕円:

楕円内を積分すると確率0.9

(注意) 等確率線ではなく
確率密度の等しい線



d次元ガウス分布の式

21

• 1次元ガウス分布（正規分布）

「次元」でなく「変量」と呼ぶ場合も

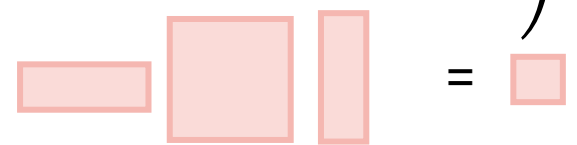
$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

パラメタ
は2つ

指数部が小さくて見えにく
いのでよくこのように書く

exp の中は
スカラー

• d次元ガウス分布（正規分布）

$$N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$


行列式 (det Σ とも書く)

平均・共分散とデータからの推定

22

分散

共分散

定義
• 確率変数 $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ の平均 $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ と共分散行列 $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$

- 平均: $\mu_1 = E[X_1]$, $\mu_2 = E[X_2]$
- 分散: $\sigma_{11} = E[(X_1 - \mu_1)^2]$, $\sigma_{22} = E[(X_2 - \mu_2)^2]$
- 共分散: $\sigma_{12} = \sigma_{21} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$

 \parallel
 σ_{12}

分散

期待値で定義されるが
データから推定できる

• N 個の2次元データ $x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \end{pmatrix}$ $i = 1, \dots, N$ からの推定 (estimate)

推定方法

データから近似的に
推定するための式
(点推定)

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x^{(i)}$$

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^N (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} (x_1 \ x_2) = \begin{pmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{pmatrix}$$

推定したパラメタは hat などつけて区別

(分母 N で推定する場合も:
最尤推定など)

データの広がり方と共分散行列の関係

23

- 共分散行列と散布図との対応はイメージできるようにしよう

平均は全て $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$\Sigma = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

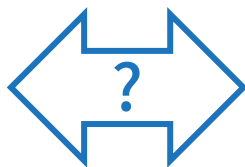
$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

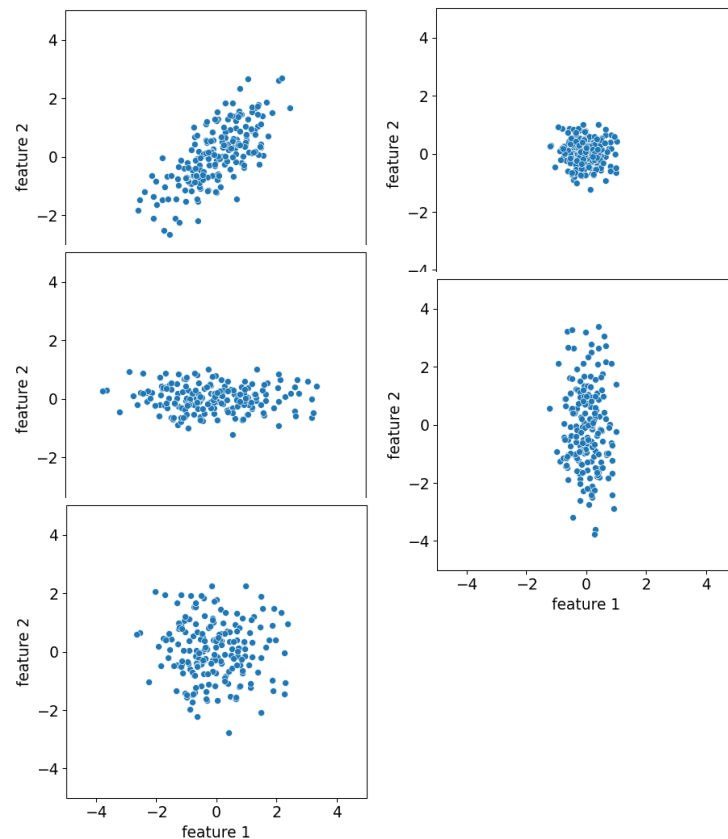
$$\Sigma = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0.2 & 0 \\ 0 & 2 \end{bmatrix}$$

どの共分散行列が
どのデータに対応？



$N(x; \mu, \Sigma)$ より正規乱数で生成したデータ (各200点)



2次元ガウス分布の等高線

24

分散

共分散

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \quad \text{ただし } \sigma_{12} = \sigma_{21}$$

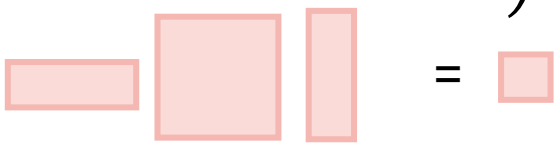
\parallel
 σ_{12}

$\sigma_{jj} = \sigma_j^2 \quad (j = 1, 2)$

分散

分散 = 標準偏差の2乗

等高線は次式を満たす \mathbf{x} の集合:

$$N(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) = C \text{ (const.)}$$


整理すると

$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = C' \text{ (const.)}$$

これが楕円の式に相当!

Σ は正定値対称行列

共分散行列と等高線（楕円）の形

25

- $\Sigma = \text{一般}$

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = \text{const.}$$

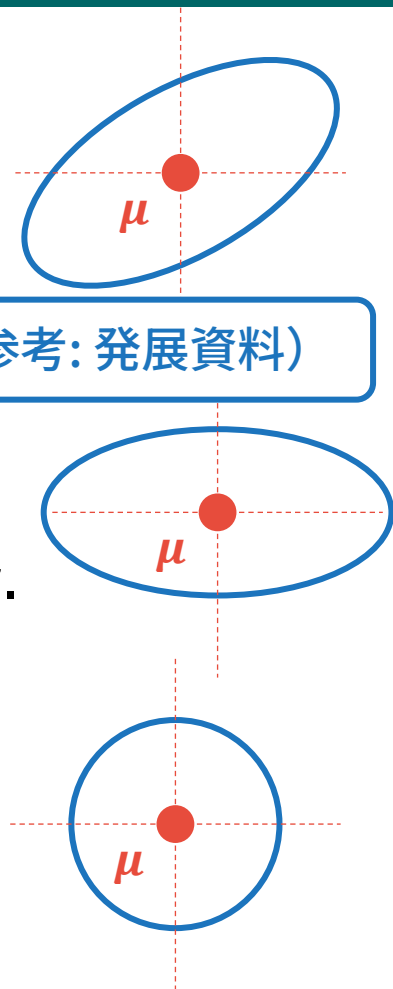
回転含む（参考：発展資料）

- $\Sigma = \text{対角行列（共分散がゼロ）}$

$$(x - \mu)^T \begin{pmatrix} \sigma_{11}^{-1} & 0 \\ 0 & \sigma_{22}^{-1} \end{pmatrix} (x - \mu) = \text{const.}$$

- $\Sigma = \text{単位行列の定数倍（各軸の分散が等しい）}$

$$(x - \mu)^T (x - \mu) = \text{const.}$$

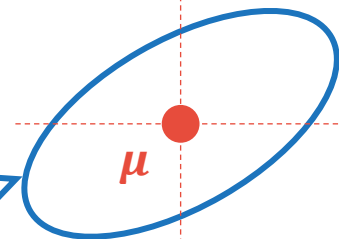


マハラノビス距離

26

- $d_M = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$ を μ からのマハラノビス距離と呼ぶ

楕円上が等距離となる
→ データの分布を踏まえた距離の定義

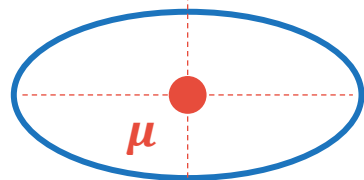


例

Σ = 対角行列

$$d_M^2 = (x - \mu)^T \begin{pmatrix} \sigma_{11}^{-1} & 0 \\ 0 & \sigma_{22}^{-1} \end{pmatrix} (x - \mu) = \frac{(x_1 - \mu_1)^2}{\sigma_{11}} + \frac{(x_2 - \mu_2)^2}{\sigma_{22}}$$

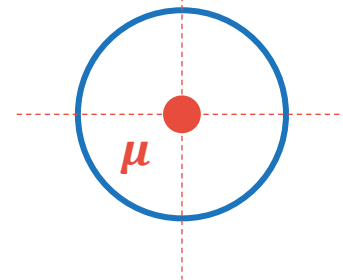
各軸が標準化される



Σ = 単位行列

$$d_M^2 = (x - \mu)^T (x - \mu) = (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2$$

いつもの距離！



[レポート: 問2] 共分散行列

27

- ① スライド22にある $\hat{\mu}$ と $\hat{\Sigma}$ の式は、平均や分散・共分散を、データから推定するための式でありベクトル・行列形式でまとめて書かれている。
(ベクトルや行列の) 各要素はどのような計算式になるか？
 - 平均 μ_1 , 分散 σ_{11} , 共分散 σ_{12} について示すこと
- ② スライド26の $\Sigma =$ 対角行列 のときの式変形を確認せよ。
 - 途中式を示すこと

各クラスを2次元ガウス分布でモデル化

28

$$N(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \frac{1}{2\pi|\boldsymbol{\Sigma}_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

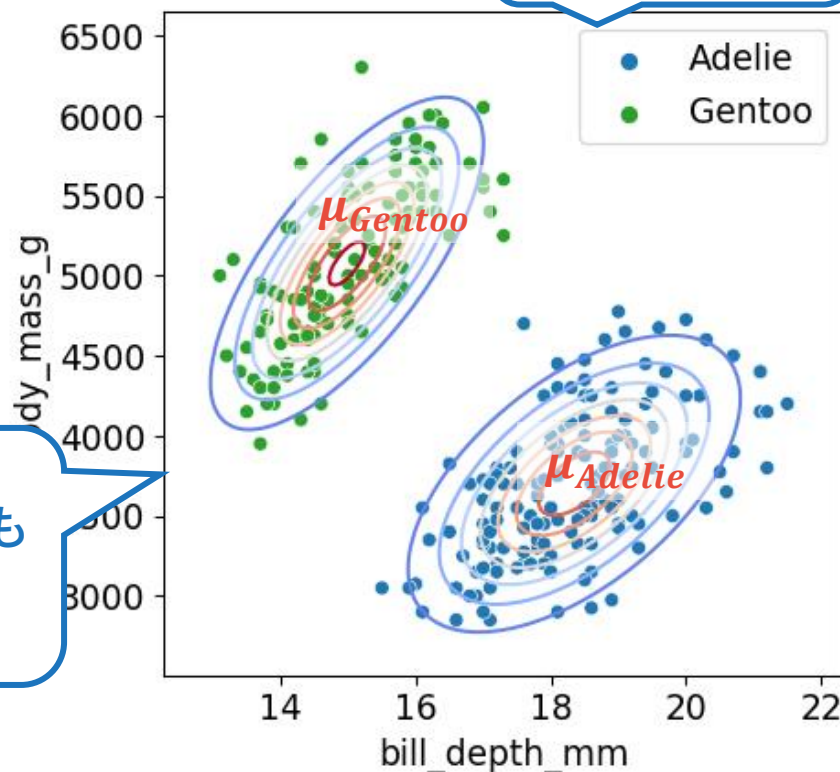
($c = \text{Adelie}, \text{Gentoo}$)

- モデルパラメタ (計10個)

- $(\boldsymbol{\mu}_{\text{Adelie}}, \boldsymbol{\Sigma}_{\text{Adelie}}) : 2 + 3$
- $(\boldsymbol{\mu}_{\text{Gentoo}}, \boldsymbol{\Sigma}_{\text{Gentoo}}) : 2 + 3$

データの広がる方向 (相関) も
うまくモデル化できている

くちばし高さ
と体重



各クラスを共分散0のガウス分布でモデル化

29

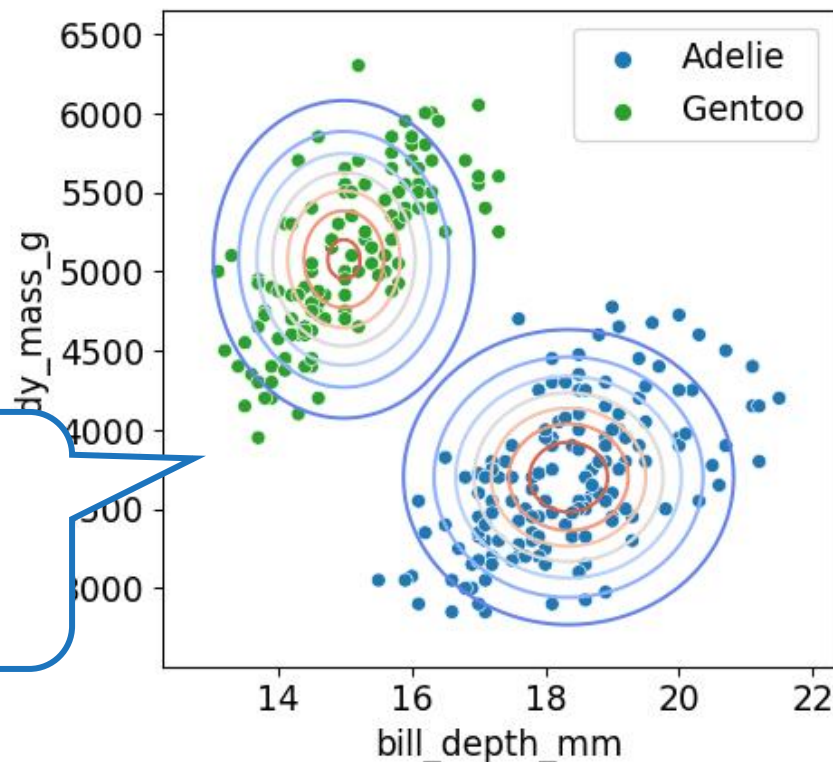
$$N(\mathbf{x}; \boldsymbol{\mu}_c, \Sigma_c) = \frac{1}{2\pi\sqrt{\sigma_{c,11}\sigma_{c,22}}} \exp\left(-\frac{1}{2}\left(\frac{(x_1 - \mu_{c,1})^2}{\sigma_{c,11}} + \frac{(x_2 - \mu_{c,2})^2}{\sigma_{c,22}}\right)\right)$$

($c = \text{Adelie}, \text{Gentoo}$)

- モデルパラメタ (計8個)

- $(\boldsymbol{\mu}_{\text{Adelie}}, \Sigma_{\text{Adelie}}) : 2 + 2$
- $(\boldsymbol{\mu}_{\text{Gentoo}}, \Sigma_{\text{Gentoo}}) : 2 + 2$

各軸のばらつきは表現できている
特徴量間の相関はモデル化できない



2次元ガウス分布で共分散ゼロ

30

- 共分散ゼロの2次元ガウス分布は1次元ガウス分布の積になる

$$\begin{aligned} N(\mathbf{x}; \boldsymbol{\mu}, \Sigma) &= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}}} \exp\left(-\frac{1}{2}\left(\frac{(x_1 - \mu_1)^2}{\sigma_{11}} + \frac{(x_2 - \mu_2)^2}{\sigma_{22}}\right)\right) \\ &= \frac{1}{\sqrt{2\pi\sigma_{11}}} \frac{1}{\sqrt{2\pi\sigma_{22}}} \exp\left(-\frac{1}{2}\frac{(x_1 - \mu_1)^2}{\sigma_{11}}\right) \exp\left(-\frac{1}{2}\frac{(x_2 - \mu_2)^2}{\sigma_{22}}\right) \\ &= N(x_1; \mu_1, \sigma_{11}) N(x_2; \mu_2, \sigma_{22}) \end{aligned}$$

$\sigma_{jj} = \sigma_j^2$ は分散

- ガウス分布で共分散ゼロ（無相関）とすると

$p(\mathbf{x}) = p(x_1, x_2) = p(x_1)p(x_2)$ となり各特徴量の独立性が仮定される

無相関 = 独立は
ガウス分布特有！

ちなみに各特徴量の独立性を仮定した事後確率
(ベイズの定理) による分類法を **Naïve Bayes**
(**ナイーブベイズ**) と呼ぶ

2次元ガウス分布による決定境界

31

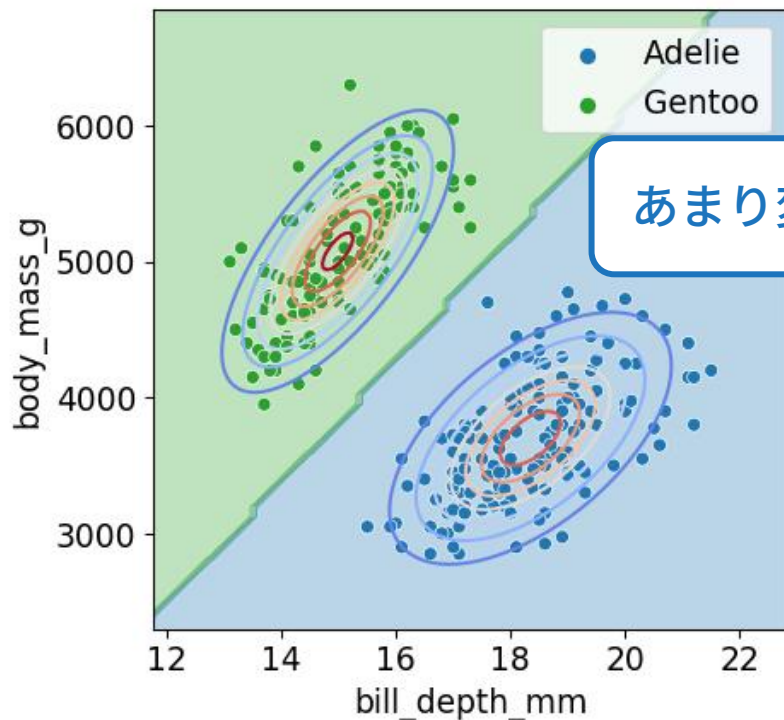
事後確率

共分散 $\neq 0$

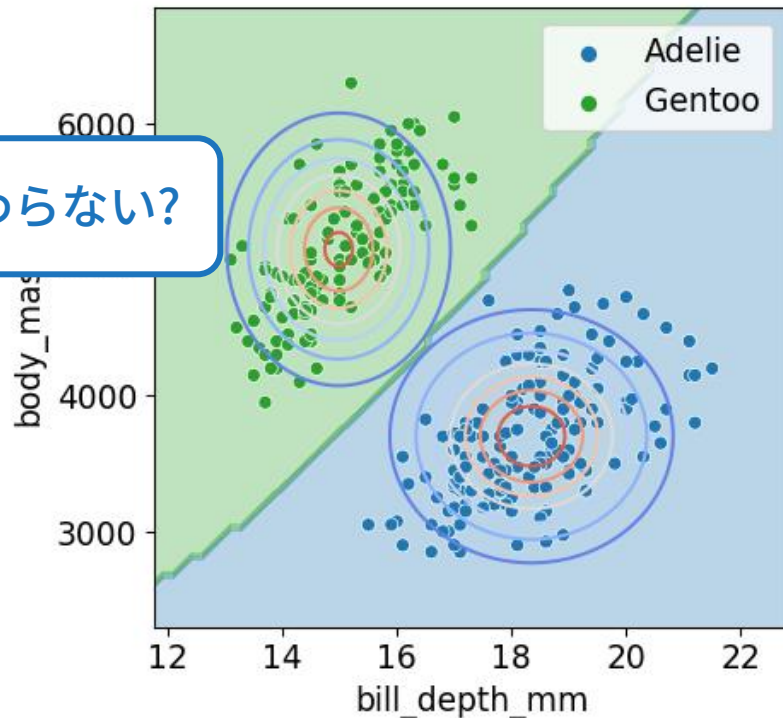
共分散 $= 0$

$$\begin{aligned} P(\text{Adelie} | \mathbf{x}) &\propto p(\mathbf{x} | \text{Adelie}) P(\text{Adelie}) \\ P(\text{Gentoo} | \mathbf{x}) &\propto p(\mathbf{x} | \text{Gentoo}) P(\text{Gentoo}) \end{aligned}$$

$$\begin{aligned} P(\text{Adelie} | \mathbf{x}) &\propto p(x_1 | \text{Adelie}) p(x_2 | \text{Adelie}) P(\text{Adelie}) \\ P(\text{Gentoo} | \mathbf{x}) &\propto p(x_1 | \text{Gentoo}) p(x_2 | \text{Gentoo}) P(\text{Gentoo}) \end{aligned}$$



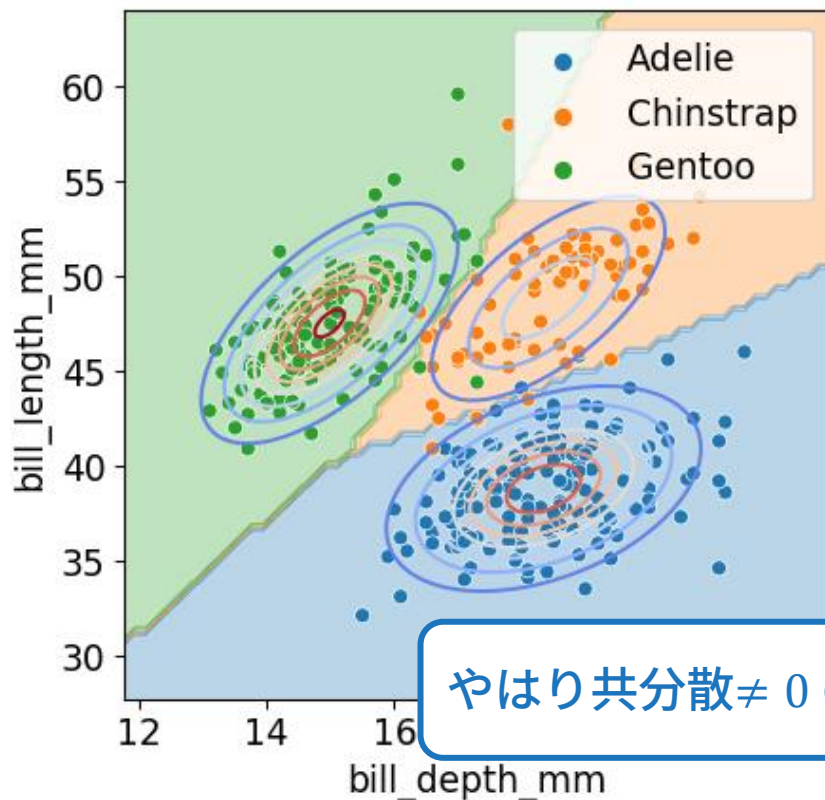
あまり変わらない?



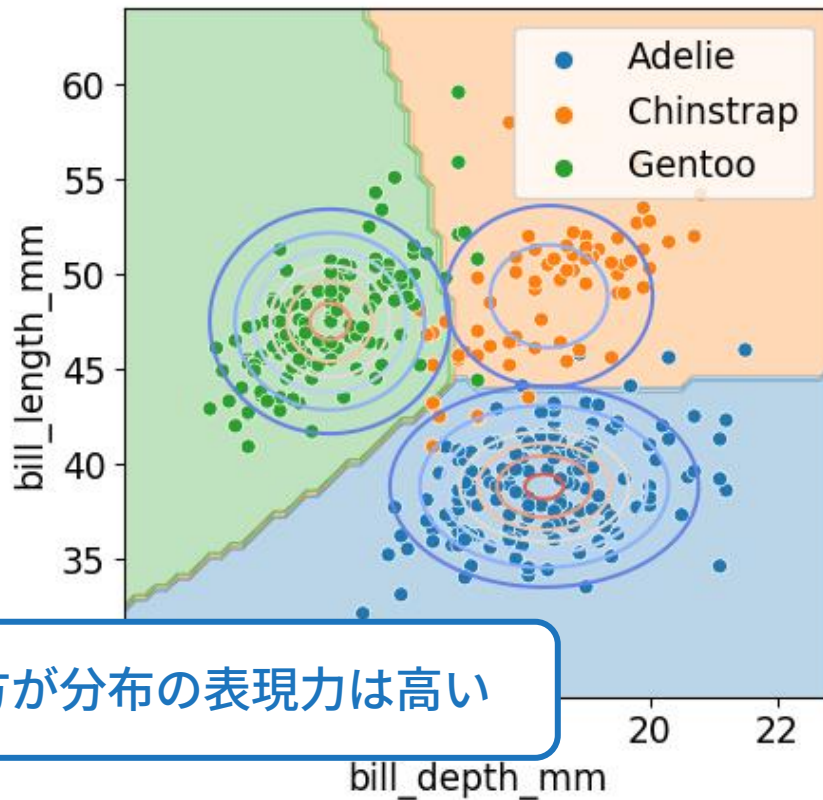
3クラス分類での決定境界の比較

32

共分散 $\neq 0$ (相関もモデル化)



共分散 = 0 (無相関を仮定)



やはり共分散 $\neq 0$ の方が分布の表現力が高い

[レポート: 問3] 決定境界が直線になる場合

33

- どのクラスの共分散行列も同一であると仮定したモデルでは、決定境界が直線になる。これを、スライド29のモデルで式変形とともに示せ。

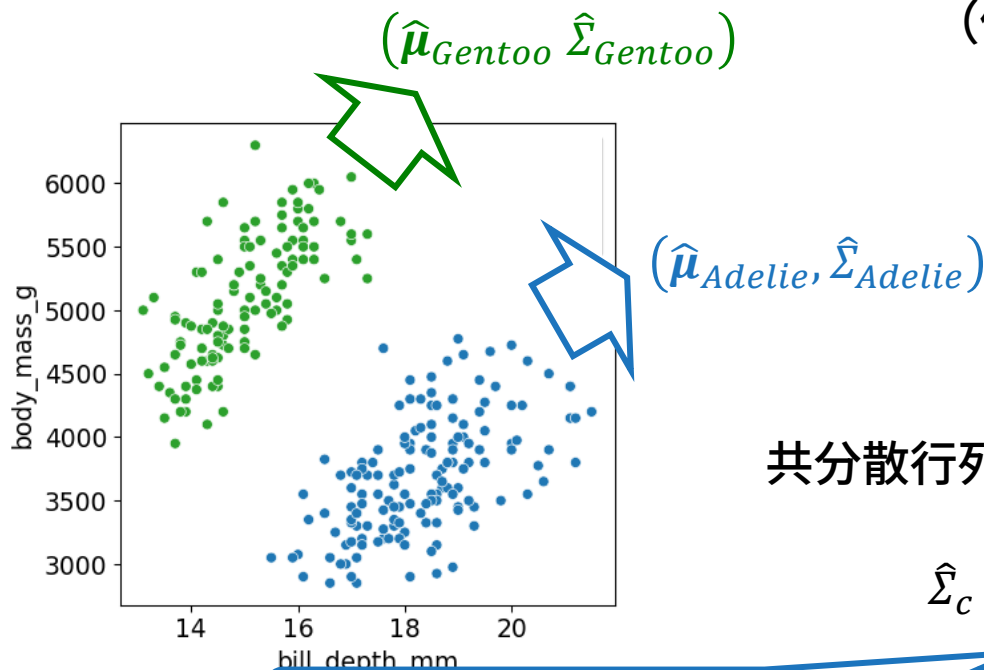
(ヒント)

- $c = \text{Adelie}, \text{Gentoo}$ だが適当にAとかGという名前にした方が楽かも
- クラスAの事後確率とクラスGの事後確率が等しいときの (x_1, x_2) の式が線形（直線の式）になるか
- 事前確率 $P(A) = p$ とかで置いておこう ($P(G) = 1 - p$)
- まず両辺の自然対数 (e を底とする対数) を取ろう

確率モデルの教師あり学習

34

- 学習はクラスごとにデータを分けてパラメタ推定するだけ
 - ガウス分布のパラメタ = 平均と共分散行列



(例) Adelie クラスの平均ベクトル

$$\hat{\mu}_{Adelie} = \frac{1}{N_{Adelie}} \sum_{i: c(i) = Adelie} \mathbf{x}^{(i)}$$

Adelie のデータ (個体) だけ取り出して平均を取る

共分散行列も各クラス $c = Adelie, Gentoo$ ごとに

$$\hat{\Sigma}_c = \frac{1}{N_c - 1} \sum_{i: c(i) = c} (\mathbf{x}^{(i)} - \hat{\mu}_c)(\mathbf{x}^{(i)} - \hat{\mu}_c)^T$$

scikit-learn の実装は不偏分散で推定

scikit-learn で学習する場合

35

- 一般の共分散行列: QuadraticDiscriminantAnalysis クラス

```
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from mlutils import plot_decision_boundary # mlutils.py で定義

clf_qda = QuadraticDiscriminantAnalysis()
clf_qda.fit(X, y) # 学習
```

- 共分散 = 0 (対角行列) : GaussianNB クラス

```
from sklearn.naive_bayes import GaussianNB

clf_gnb = GaussianNB()
clf_gnb.fit(X, y) # 学習
```

ガウス分布での
ナイーブベイズ (NB)

第1回レポート／今日のコード

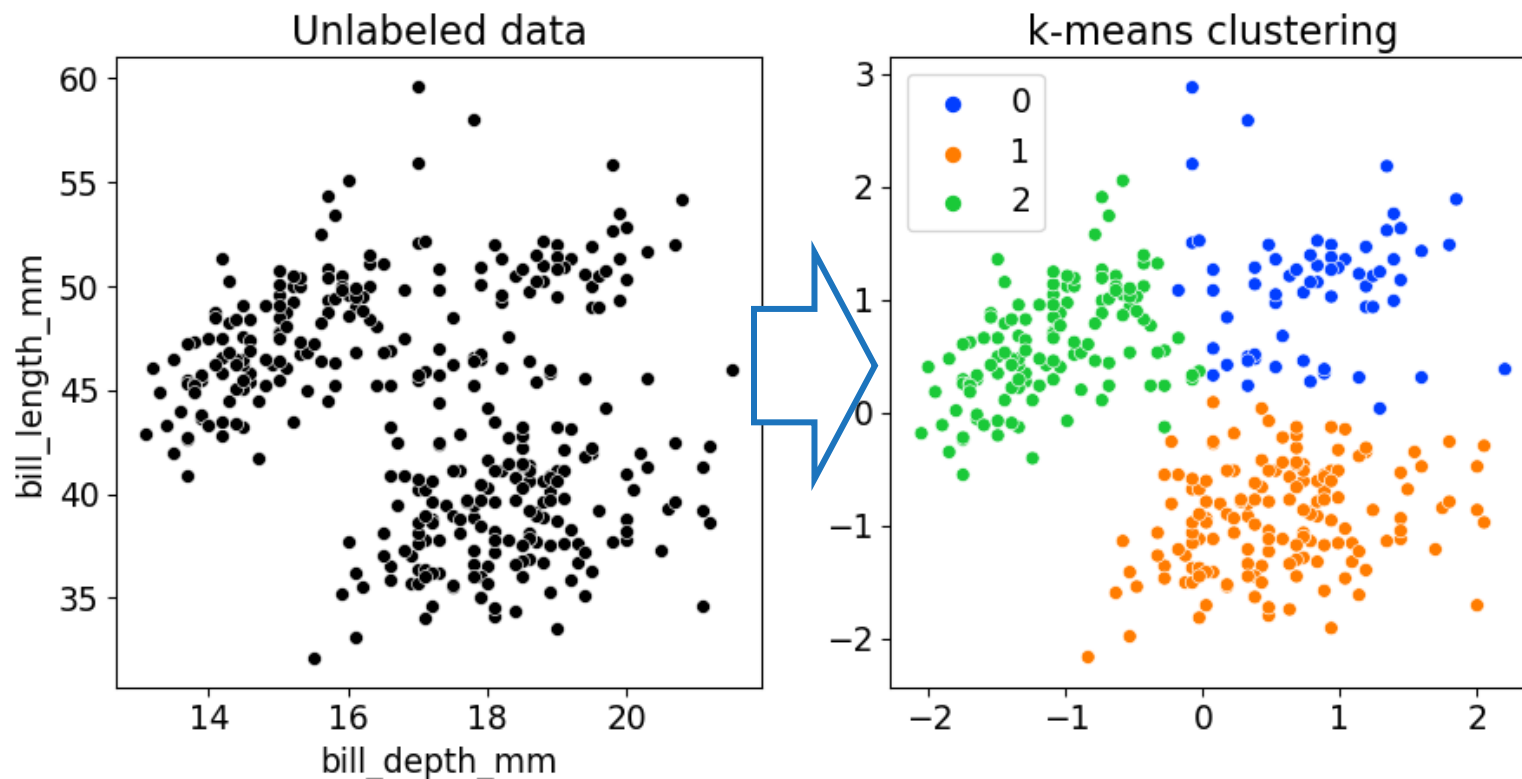
36

- レポート
 - 今週と来週でレポートを出題
 - レポートと書かれたスライドが対象
 - ✕切: 5/9（金）今週分は今週末までには取り組もう
（注意）
 - GWで質問対応できる日が限られる
 - 5/19に小テストを予定 → GWで復習を入れておこう
- 今日のコード（参考）
 - <https://colab.research.google.com/drive/1NXzGwnX2bSG9ets5CI35Nkxqply36eaJ?usp=sharing>

クラスタリング（教師なし学習）

37

- 教師なし学習ではラベルが振られていない
 - 似たパターンをグループ化したい？ → クラスタリング！

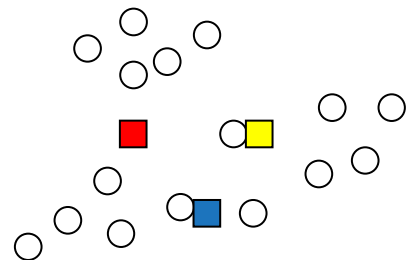


k -means法（定番のクラスタリング手法）

38

- 暫定的なラベル付与とパラメタ推定を繰り返す

- 各データにラベル付与 = 帰属クラスタを決める

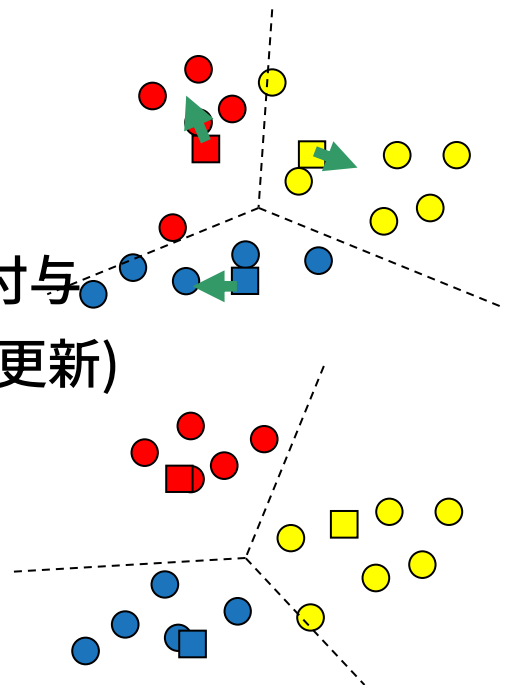


- アルゴリズム

- k 個のクラスタ中心を配置

- ループ（収束するまで）：

- 各データへ最も近いクラスタ中心のラベルを付与
- 同じラベルのデータからクラスタ中心を計算(更新)



平均: k -means 法
中央値: k -medoids 法

混合ガウス分布の推定

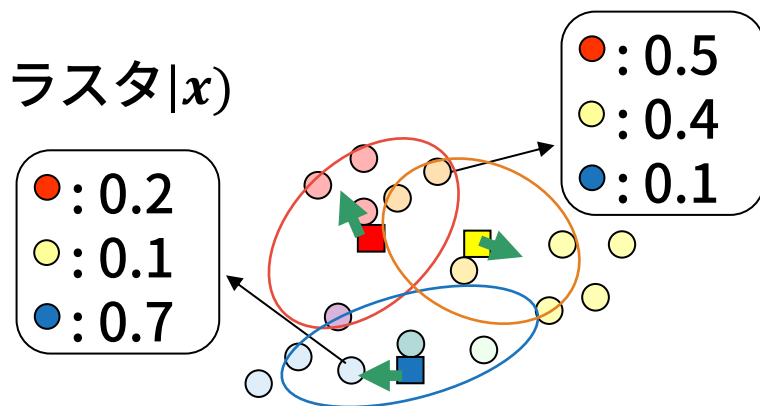
39

- 各クラスタ = ガウス分布としたソフトクラスタリング

- k 個のクラスタ = k 個のガウス分布
 - 各クラスタへの帰属度 = 事後確率 $P(\text{クラスタ} | x)$

- アルゴリズム

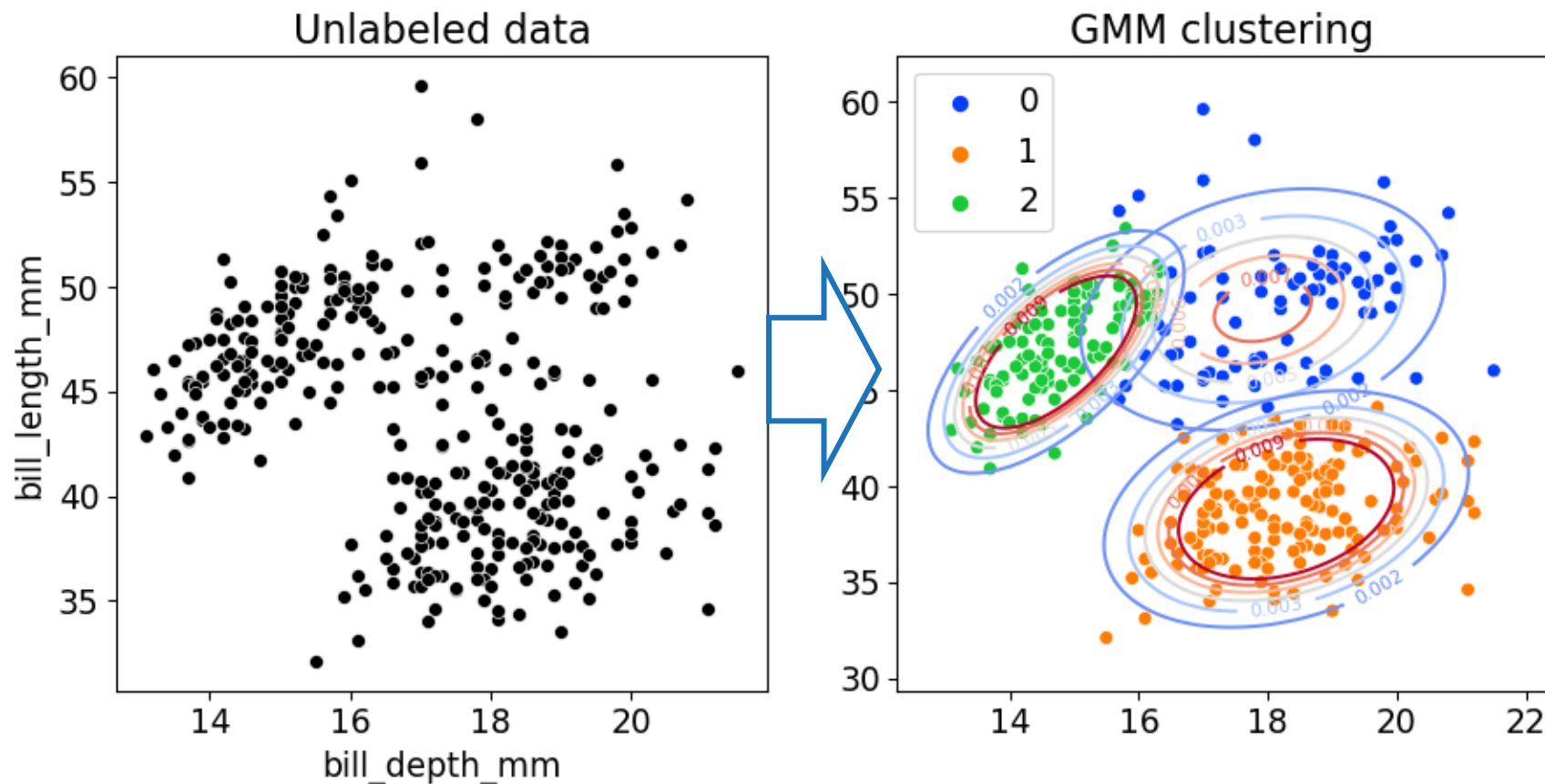
- k 個のガウス分布を配置
 - ループ（収束するまで）：
 - 各データに対して各ガウス分布への帰属度（事後確率）を計算
 - 帰属度で重みづけした全データで各ガウス分布の平均と共分散行列を計算（更新）



混合ガウス分布の推定

混合ガウス分布の推定

40



実は「混合ガウス分布推定の特殊な場合が k-means 法」

発展

(確率分布を用いた機械学習を扱う
場合には知っておくと便利な知識)

共分散行列の分解

42

- 共分散行列を固有値分解すると回転とスケールを取り出せる

固有値分解 $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = E \Lambda E^T$

対称行列

E : 回転
(+ 鏡映)

Λ : 各軸のスケール情報

$E^{-1} = E^T$: 回転 (+ 鏡映)
(E と逆の回転)

$|E| = -1$ だと一方
の符号が反転

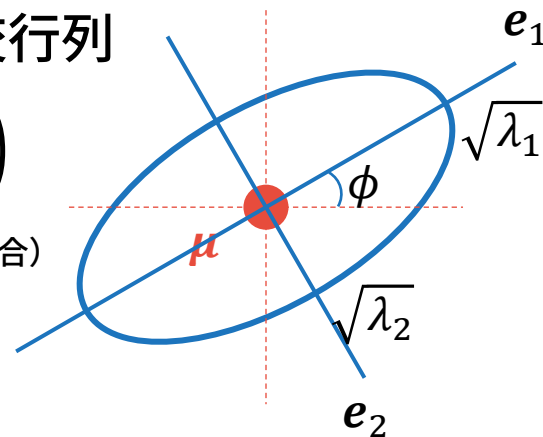
- E : 固有ベクトル e_1, e_2 を列ベクトルとする直交行列

$$E = (e_1 \ e_2) = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}$$

(行列式1となるよう E を定めた場合)

- Λ : 固有値 λ_1, λ_2 を対角成分とする対角行列

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$



[発展問題] 共分散行列

43

スライド24の固有値分解を利用すれば、スライド21の最後の式は

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = y^T \Lambda^{-1} y$$

となる．ここで $\Sigma^{-1} = E \Lambda^{-1} E^T$ であることを利用し $y = E^T (x - \mu)$ とした．

この変換では，原点を μ とし，軸が楕円の長軸・短軸になるよう座標系を考えたことになる（変換後は共分散行列が Λ ）．

では， $(x - \mu)^T \Sigma^{-1} (x - \mu) = z^T z$ とするにはどのように変換すればよい
か？また，変換後の共分散行列はどうなるか？（ヒント: $\Lambda^{-1} = (\Lambda^{-\frac{1}{2}})^2$ ）

$$\Lambda^{-\frac{1}{2}} = \begin{pmatrix} \lambda_1^{-1/2} & 0 \\ 0 & \lambda_2^{-1/2} \end{pmatrix}$$

白色化(whitening)と
呼ばれる重要な変換

対角行列の逆行列

$$\bullet \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_{11}^{-1} & 0 \\ 0 & \sigma_{22}^{-1} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_{11}} & 0 \\ 0 & \frac{1}{\sigma_{22}} \end{pmatrix}$$

$$\bullet \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_{11}} & 0 \\ 0 & \frac{1}{\sigma_{22}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

回転行列

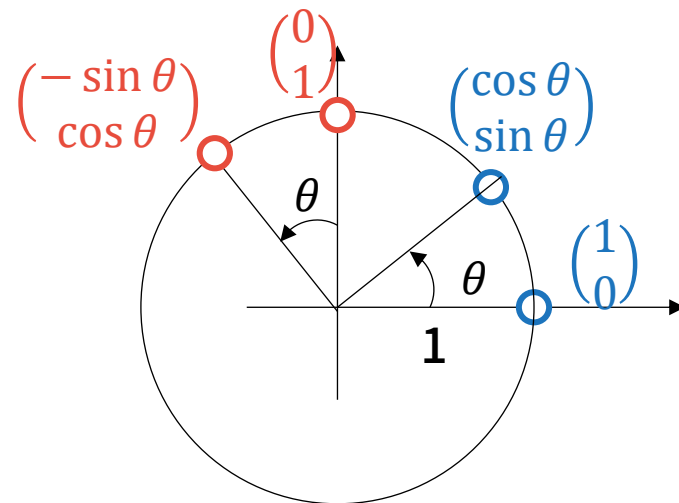
45

- $R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ とおく

$$R \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

$$R \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$$

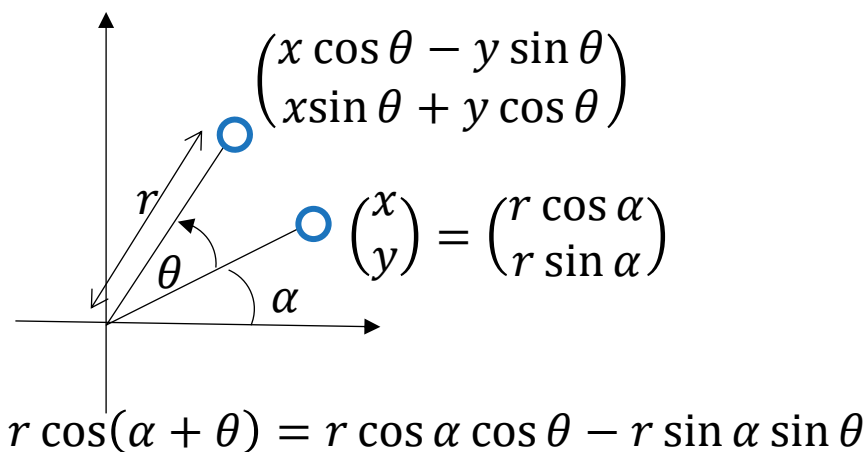
R でベクトル
が回転する



- R は一般のベクトル (x, y) も回転させる

$\begin{pmatrix} x \\ y \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ に R をかけると

$$\begin{aligned} R \begin{pmatrix} x \\ y \end{pmatrix} &= x R \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y R \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= x \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} + y \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} \\ &= \begin{pmatrix} x \cos\theta - y \sin\theta \\ x \sin\theta + y \cos\theta \end{pmatrix} \end{aligned}$$



座標変換

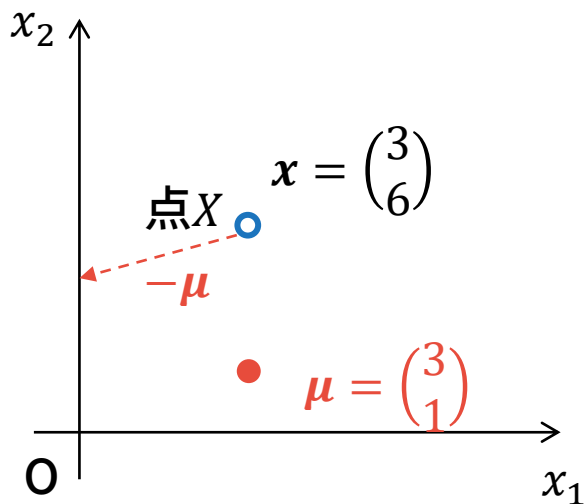
46

- $x' = x - \mu$ と置くと「平行移動」→ 座標変換として解釈すると？

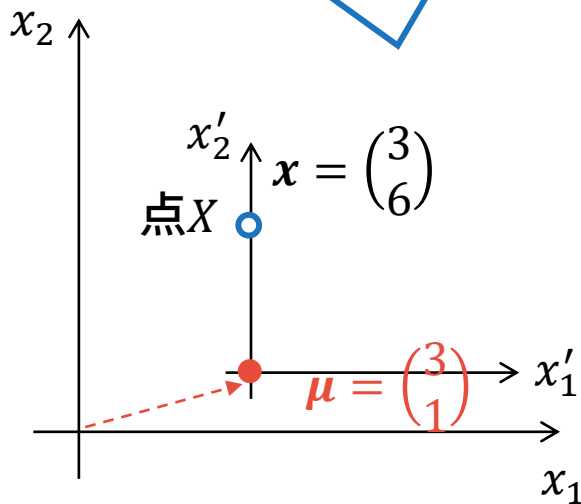
$$\bullet \begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \end{pmatrix}$$

μ が原点になるよう xy 座標系
を平行移動した $x'_1x'_2$ 座標系を
考える

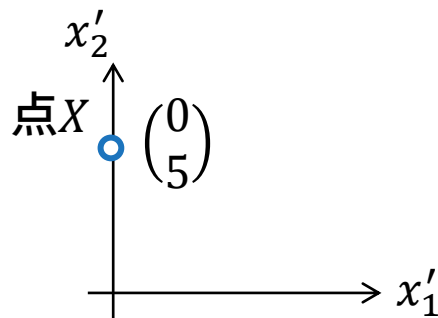
$x'_1x'_2$ 座標系の世界で
点 X をみると・・・



平行移動としての解釈



座標変換としての解釈



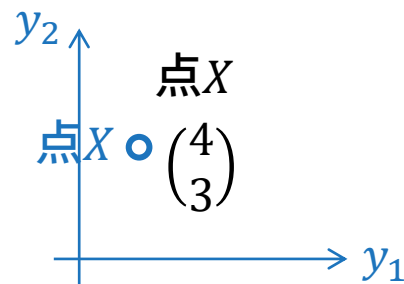
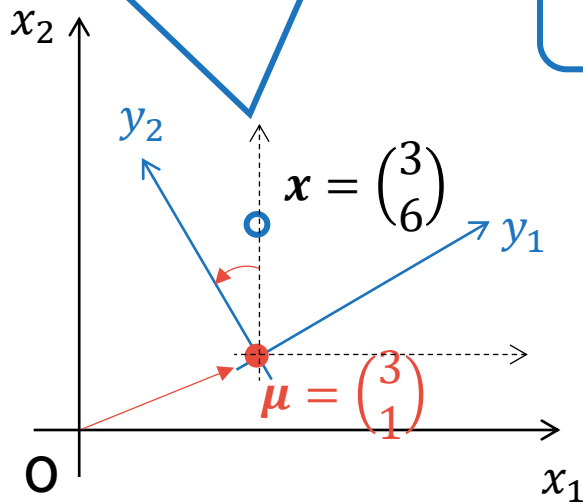
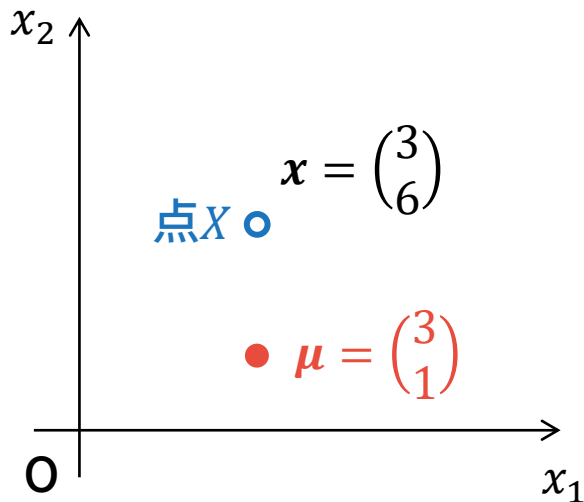
座標変換

47

- $y = E^T(x - \mu)$ とおくことの意味は「平行移動」 + 「回転」

μ を原点とし、軸を回転させた
 $y_1 y_2$ 座標系を考える

(y_1, y_2) の世界で点 X
をみると・・・



発展問題 ②

48

$(x - \mu)^T \Sigma^{-1} (x - \mu)$ に $\Sigma^{-1} = E \Lambda^{-1} E^T$ を代入すると

変換

$$(x - \mu)^T E \Lambda^{-1} E^T (x - \mu)$$

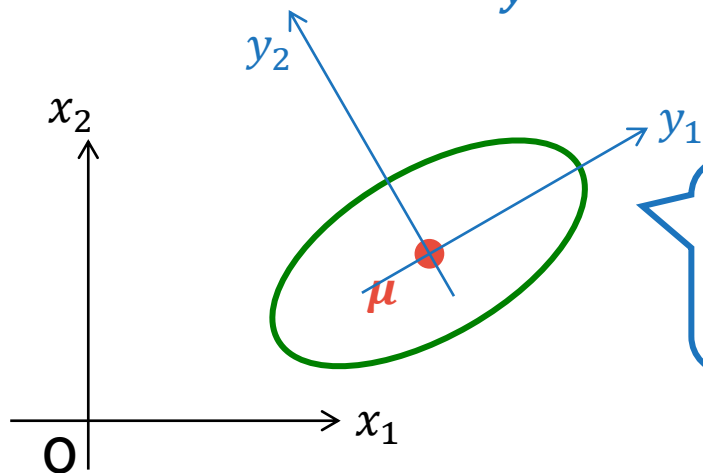
ここで $y = E^T (x - \mu)$ とおくと

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, E = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

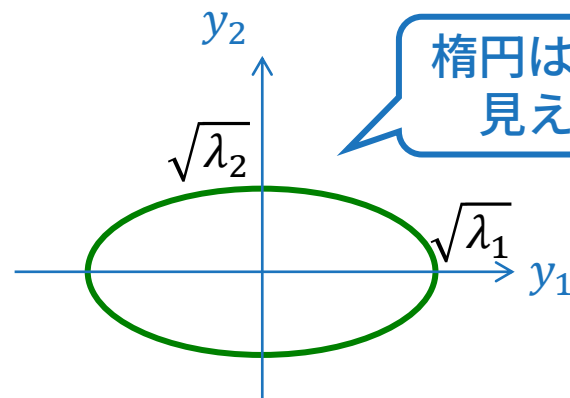
公式 $(Ax)^T = x^T A^T$

$$\underbrace{(x - \mu)^T E \Lambda^{-1}}_{y^T} \underbrace{E^T (x - \mu)}_y = y^T \Lambda^{-1} y$$

今は $\det E = 1$
と仮定



変換後の
 $y_1 y_2$ 座標系
では



楕円はこう
見える

発展問題 ②

49

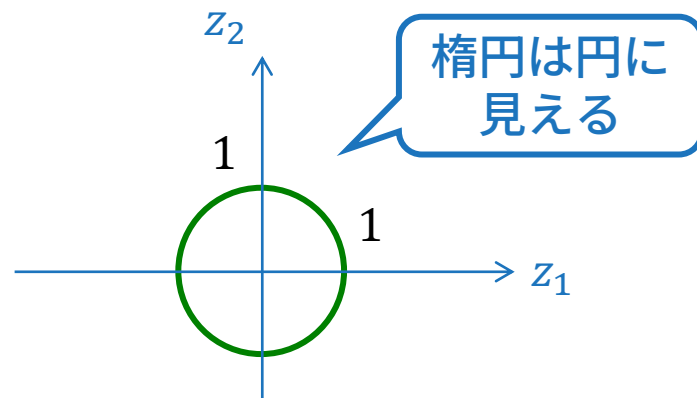
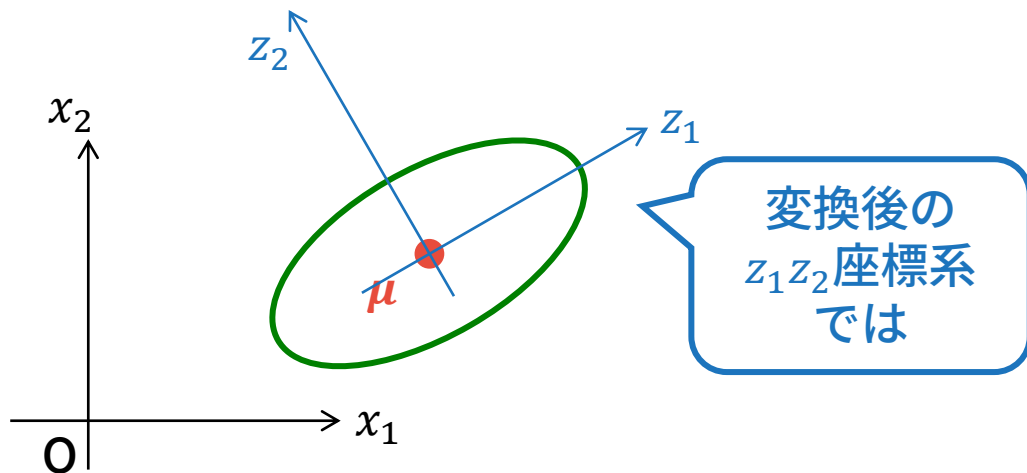
- $(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{E} \boldsymbol{\Lambda}^{-1} \mathbf{E}^T (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{z}^T \mathbf{z}$ したい

- $(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{E} \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{E}^T (\mathbf{x} - \boldsymbol{\mu})$

$$\begin{aligned}\boldsymbol{\Lambda}^{-1} &= \begin{pmatrix} \lambda_1^{-1} & 0 \\ 0 & \lambda_2^{-1} \end{pmatrix} = \begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{pmatrix} \\ &= \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{-\frac{1}{2}}\end{aligned}$$

ここで $\mathbf{z} = \dots$ の変換を考える

対角行列を掛けると軸
がスケールされる



発展問題 ②

50

- $(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{E} \boldsymbol{\Lambda}^{-1} \mathbf{E}^T (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{z}^T \mathbf{z}$ としたい

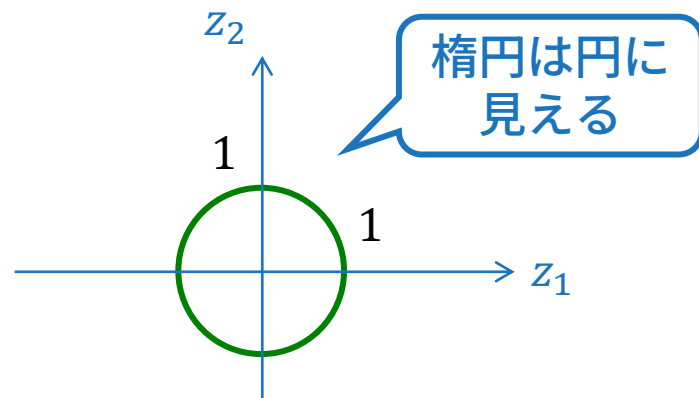
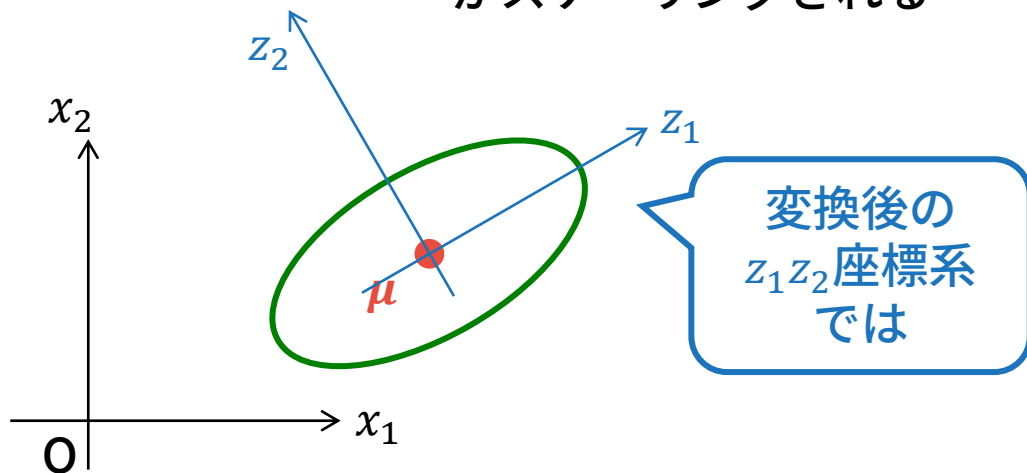
- $(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{E} \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{E}^T (\mathbf{x} - \boldsymbol{\mu})$

\mathbf{z}^T

\mathbf{z}

$$\begin{aligned} \boldsymbol{\Lambda}^{-1} &= \begin{pmatrix} \lambda_1^{-1} & 0 \\ 0 & \lambda_2^{-1} \end{pmatrix} = \begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{pmatrix} \\ &= \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{-\frac{1}{2}} \end{aligned}$$

対角行列を掛けると軸
がスケールされる



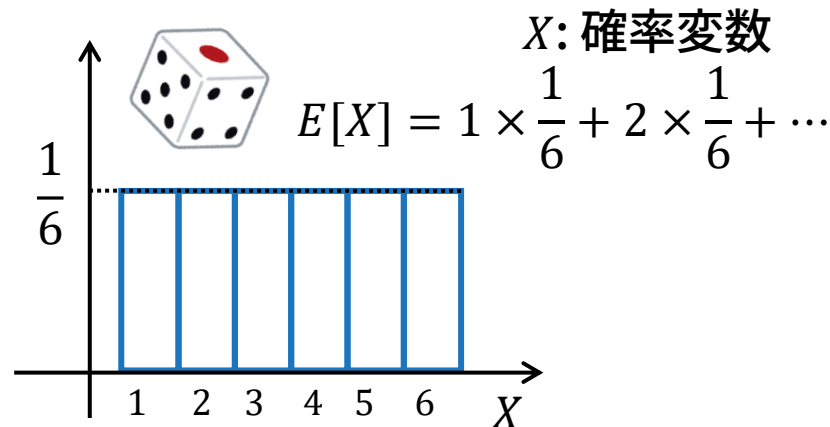
付録

期待値の計算方法

52

- 期待値は「確率変数」に対して定義される
- 離散分布の期待値

$$E[X] = \sum_k x_k P(X = x_k)$$



- 連続分布の期待値

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

$f_X(x)$ は確率密度関数 $\int_{-\infty}^{\infty} f_X(x) dx = 1$

教科書によっては $E[X]$
を $E(X)$ と書くことも

記述統計の平均値との比較

53

- 記述統計におけるデータの平均値

$$\bar{x} = \frac{3 + 1 + 2 + 3 + 2 + 2 + 2 + 1 + 4 + 1}{10}$$

1が3回, 2が4回,
3が2回, 4が1回

- 度数を使って書き直してみる

$$\bar{x} = \frac{1 \times 3 + 2 \times 4 + 3 \times 2 + 4 \times 1}{3 + 4 + 2 + 1} = 1 \cdot 0.3 + 2 \cdot 0.4 + 3 \cdot 0.2 + 4 \cdot 0.1$$

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_K f_K}{N} = x_1 \frac{f_1}{N} + x_2 \frac{f_2}{N} + \cdots + x_K \frac{f_K}{N} = \sum_{k=1}^K x_k p_k$$

- ただし $N = f_1 + f_2 + \cdots + f_K$, $p_k = \frac{f_k}{N}$ とおいた

- 一方, 確率変数 X に対しては具体的な頻度 f_k でなく分布

$P(X = x_k)$ が与えられる: 期待値 $E[X] = \sum_k x_k P(X = x_k)$

- 期待値 $E[X]$ を確率変数 X の平均と呼ぶことも多い

共分散 (covariance)

54

- 二変数 x, y データに対する共分散 c_{xy}

- 偏差積の平均 (データのバラツキを表現)
 - 偏差 $(x_i - \bar{x})$ と偏差 $(y_i - \bar{y})$ の符号が一致 (緑領域) なら+
 - 偏差 $(x_i - \bar{x})$ と偏差 $(y_i - \bar{y})$ の符号が不一致 (青領域) なら-
- 共分散の絶対的な大きさのみでは相関の強さを評価できない
 - x, y の単位やスケールに影響されるため

$$c_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

記述統計なら n で割る

