6.混合分布モデル

6月10日　田中洸一

確率モデルによるクラスタリングとは（1）

※イメージ図

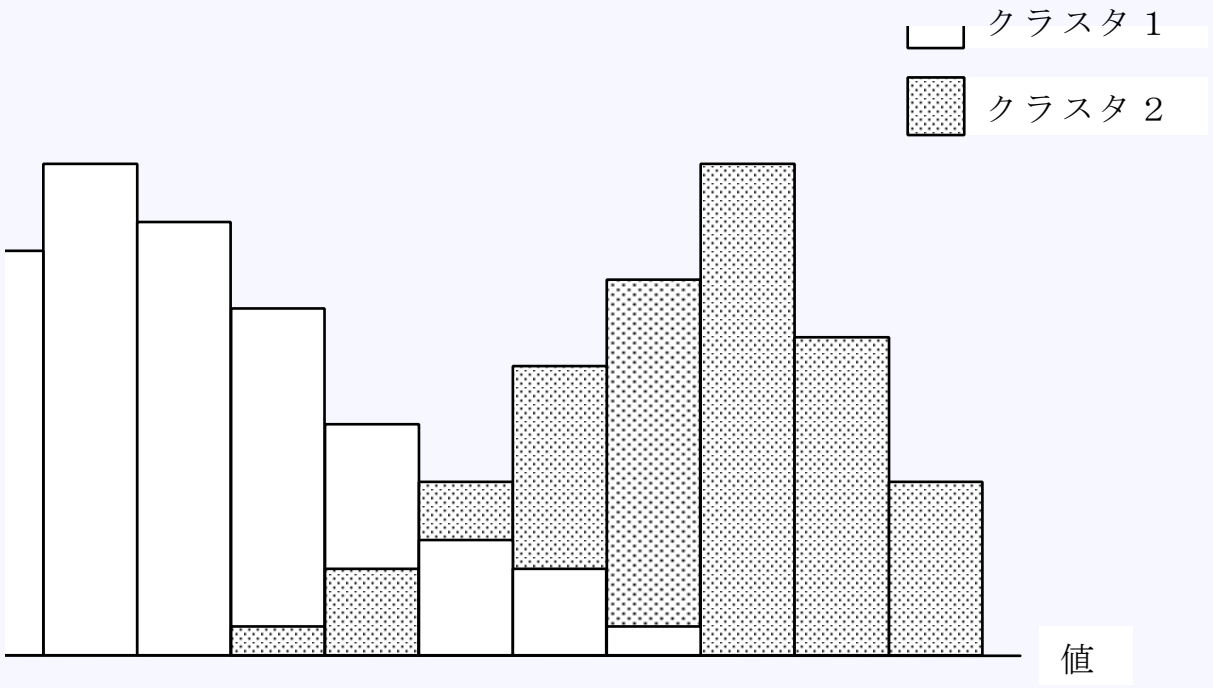

図1：クラスタ1とクラスタ2それぞれのデータからなるヒストグラム（独立）

確率モデルによるクラスタリングとは（2）

※イメージ図

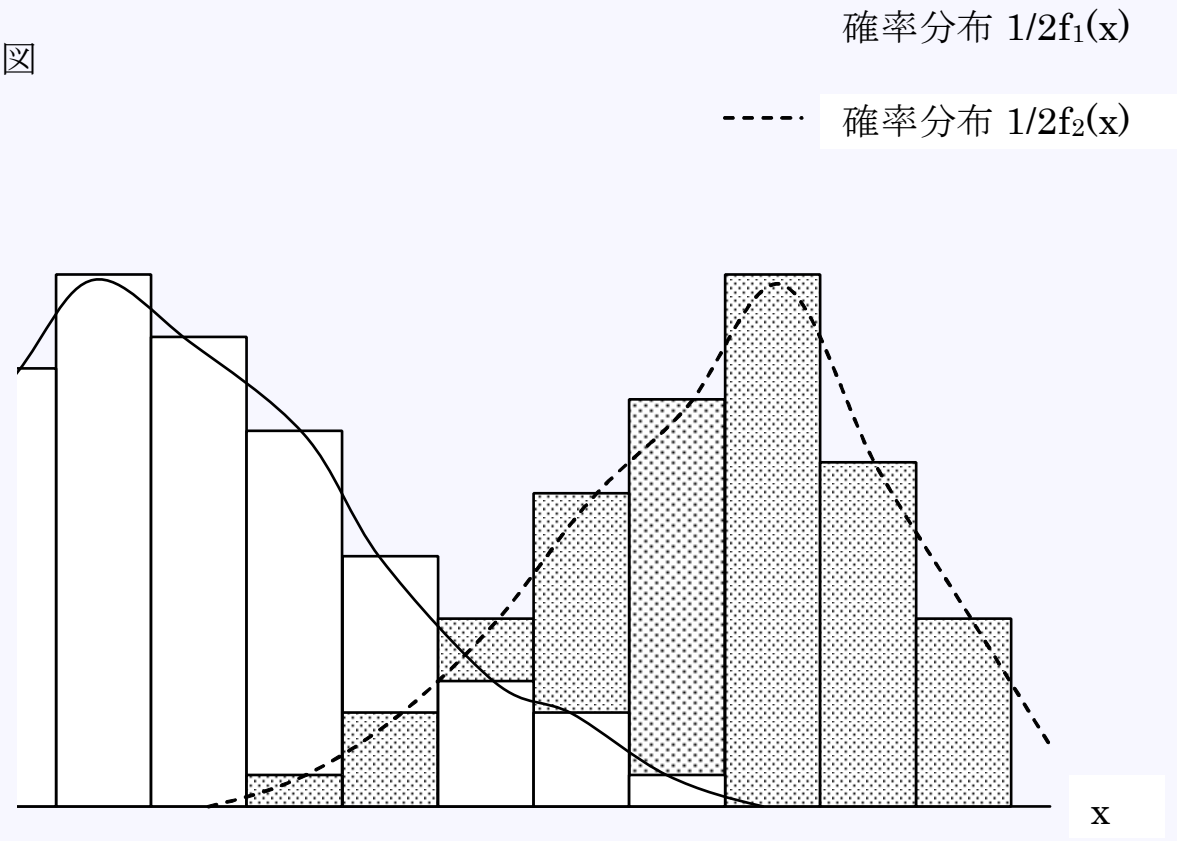

図２：図１を一様な確率分布とみなしたグラフ

確率モデルによるクラスタリングとは（３）

※イメージ図

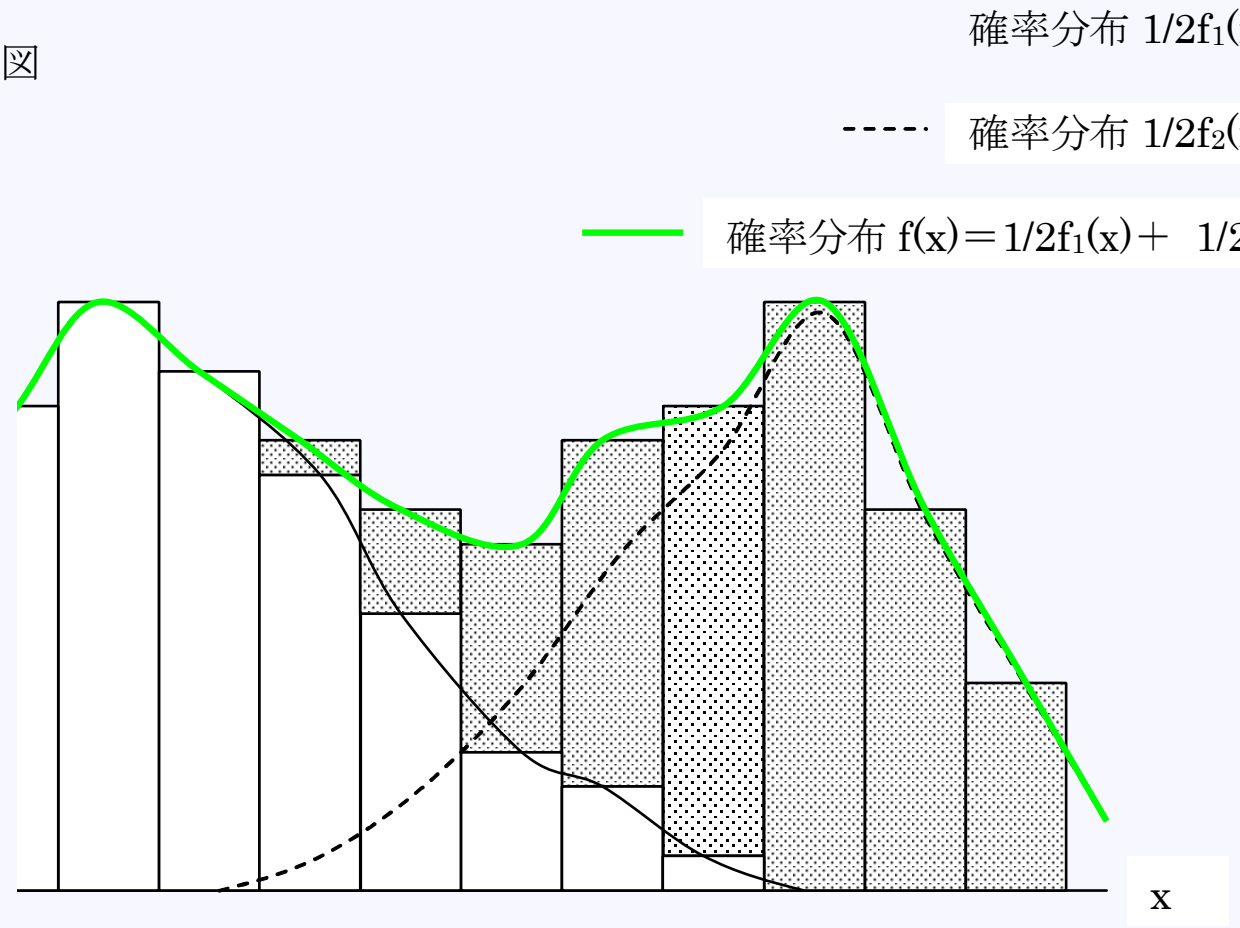

図３：確率分布 f(x)

確率モデルによるクラスタリングとは（4）

混合分布とは複数の分布を混合した分布

↓

混合分布モデルによるクラスタリングとは先ほどの$f(x)$を推定し、$f_c(x)$（実際は$\alpha_c f_c(x)$）が最大値となるc番目のクラスタに割り当てる

では、どのような確率モデルの表し方をすればよいか？

確率モデルの表し方

データを発生する確立モデルがK個あるとする。

c番目のモデルの下でデータxが発生する確率を$p_c(x)$とし、データxが発生する確率を線形和で示す。

$$p(x) = \sum_{c=1}^{K} \alpha_c p_c(x)$$

※$\alpha_c \geq 0$かつ$\sum_{c=1}^{K} \alpha_c = 1$

データが連続値の場合は確率密度関数f(x)

$$f(x) = \sum_{c=1}^{K} \alpha_c f_c(x)$$

$f_c(x)$は正規分布を扱う。データxはn次元ベクトルより、
n変数の多変量正規分布。

$$f_c(x) = \frac{1}{(2\pi)^{n/2} \left| \sum_c \right|^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu_c)^t \sum_c^{-1} (x-\mu_c) \right\}$$

$$\mu_c = (\mu_1, \mu_2, \cdots, \mu_n)^t$$

μ_cはn次元の平均ベクトル

Σ_cはn×nの分散共分散行列

$$\sum_c = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix}$$

※cの添え字は省略

パラメータ

- f(x)を求めるには、パラメータ$\mu_c, \Sigma_c, \alpha_c$を求める必要がある。

$$\theta = (\alpha_1, \alpha_2, \cdots, \alpha_K, \mu_1, \mu_2 \cdots, \mu_K, \sigma_1^2, \sigma_2^2 \cdots, \sigma_K^2)$$

- 観測されたデータ$x_1, x_2, \cdots, x_N$確率モデルのθを求めるには対数尤度関数を最大にするθを求めればよい

$$L(\theta) = \sum_{i=1}^{N} \log \left\{ \sum_{c=1}^{K} \alpha_c f_c(x_i) \right\} \qquad (6.2)$$

EMアルゴリズム

(6.2)の最大値を与えるパラメータθを求めることは困難

↓

EMアルゴリズム

隠れ変数が存在 → パラメータを推定

推定されたパラメータ → 隠れ変数を推定

c番目のクラスタから発生したという情報を隠れ変数とし、それぞれが収束するまで繰り返す

隠れ変数推定（1）

$$\mathcal{Q} = E\left[\log q(y \mid x, \theta^{(t)})\right]$$ ※$q(y): y = (x, c)$の分布

θを固定して、尤度を最大にする隠れ変数を求める

y=(x,c)の確率密度関数　$\alpha_c f_c(x)$

xとθ(t)が与えられたときのcの分布　$\dfrac{\alpha_c^{(t)} f_c^{(t)}(x)}{\sum_{k=1}^{K} \alpha_k^{(t)} f_k^{(t)}(x)}$

$$\mathcal{Q} = \sum_{c=1}^{K} \frac{\alpha_c^{(t)} f_c^{(t)}(x)}{\sum_{k=1}^{K} \alpha_k^{(t)} f_k^{(t)}(x)} \log(\alpha_c f_c(x))$$

隠れ変数推定（2）

観測データは複数個の$x_1, x_2, \cdots, x_N$だが、
サンプルの独立性から各Qの和が全体のQ

$$Q = \sum_{i=1}^{N} \sum_{c=1}^{K} \frac{\alpha_c^{(t)} f_c^{(t)}(x_i)}{\sum_{k=1}^{K} \alpha_k^{(t)} f_k^{(t)}(x_i)} \log(\alpha_c f_c(x_i))$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{K} g_{ic}^{(t)} \log(\alpha_c f_c(x_i))$$

$$g_{ic}^{(t)} = \frac{\alpha_c^{(t)} f_c^{(t)}(x_i)}{\sum_{k=1}^{K} \alpha_k^{(t)} f_k^{(t)}(x_i)}$$

最終的に得られるg_{ic}が
クラスタリング結果

パラメータ推定

Qを最大にするθを求める

$$\alpha_c^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} g_{ic}^{(t)}$$

$$\mu_c^{(t+1)} = \frac{\sum_{i=1}^{N} g_{ic}^{(t)} x_i}{\sum_{i=1}^{N} g_{ic}^{(t)}}$$

$$\sum_c^{(t+1)} = \frac{\sum_{i=1}^{N} g_{ic}^{(t)} \left(x_i - \mu_c^{(t+1)}\right)^t \left(x_i - \mu_c^{(t+1)}\right)}{\sum_{i=1}^{N} g_{ic}^{(t)}}$$

初期値

EMアルゴリズムにおいて初期値 $\alpha_c^{(0)}, \mu_c^{(0)}, \Sigma_c^{(0)}$ が必要。

- $\alpha_c^{(0)} = 1/K$
- 全体のデータの分散から $\Sigma_c^{(0)}$
- K個のデータから $\mu_c^{(0)}$

しかし、初期値次第で結果が異なる

↓

事前にk-meanなどで初期値を設定することもある

Σ_cのモデル

Σ_cのすべての要素を求める

↓

・計算が大変　・モデルが複雑になる

・Σ_cのすべての要素を可変にすると尤度はいくらでも大きくなる

↓

簡略化されたΣ_cのモデルを提案

EII

$$\sum_c = \sigma^2 \mathrm{I} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

$$f_c(x_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} \frac{\left(x_j^i - \mu_{cj}\right)^2}{\sigma^2} \right\}$$

※ $x_i = \left(x_1^i, x_2^i, \cdots, x_n^i\right)^t$

※ $\mu_c = \left(\mu_{c1}, \mu_{c2}, \cdots, \mu_{cn}\right)^t$

・EMアルゴリズムによる更新式（g_{ic}, α_c, μ_cの更新式は前と同じ）

$$\sigma^{2(t+1)} = \frac{1}{nN} \sum_{i=1}^{N} \sum_{c=1}^{K} g_{ic}^{(t)} \sum_{j=1}^{n} \left(x_j^i - \mu_{cj}^{(t+1)}\right)^2$$

VII

$$\sum_c = \sigma_c^2 \mathrm{I} = \begin{pmatrix} \sigma_c^2 & 0 & \cdots & 0 \\ 0 & \sigma_c^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_c^2 \end{pmatrix}$$

$$f_c(x_i) = \frac{1}{(2\pi\sigma_c^2)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} \frac{\left(x_j^i - \mu_{cj}\right)^2}{\sigma_c^2} \right\}$$

※ $x_i = \left(x_1^i, x_2^i, \cdots, x_n^i\right)^t$

※ $\mu_c = \left(\mu_{c1}, \mu_{c2}, \cdots, \mu_{cn}\right)^t$

・EMアルゴリズムによる更新式（g_{ic}, α_c, μ_cの更新式は前と同じ）

$$\sigma_c^{2(t+1)} = \frac{\sum_{i=1}^{N} g_{ic}^{(t)} \sum_{j=1}^{n} \left(x_j^i - \mu_{cj}^{(t+1)}\right)^2}{\sum_{i=1}^{N} g_{ic}^{(t)}}$$

EEI

$$\sum_c = \sum = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

$$f_c(x_i) = \frac{1}{(2\pi)^{\eta/2}\sqrt{\prod_{j=1}^{n}\sigma_j^2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\frac{\left(x_j^i - \mu_{cj}\right)^2}{\sigma_j^2}\right\}$$

※ $x_i = \left(x_1^i, x_2^i, \cdots, x_n^i\right)^t$

※ $\mu_c = \left(\mu_{c1}, \mu_{c2}, \cdots, \mu_{cn}\right)^t$

・EMアルゴリズムによる更新式（g_{ic}, α_c, μ_cの更新式は前と同じ）

$$\sigma_j^{2(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} g_{ic}^{(t)}\left(x_j^i - \mu_{cj}^{(t+1)}\right)^2$$

VEI

$$\sum_c = \begin{pmatrix} \sigma_{c1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{c2}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{cn}^2 \end{pmatrix}$$

$$f_c(x_i) = \frac{1}{(2\pi)^{n/2}\sqrt{\prod_{j=1}^{n}\sigma_{cj}^2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\frac{\left(x_j^i - \mu_{cj}\right)^2}{\sigma_{cj}^2}\right\}$$

※ $x_i = \left(x_1^i, x_2^i, \cdots, x_n^i\right)^t$

※ $\mu_c = \left(\mu_{c1}, \mu_{c2}, \cdots, \mu_{cn}\right)^t$

・EMアルゴリズムによる更新式（g_{ic}, α_c, μ_cの更新式は前と同じ）

$$\sigma_{cj}^{2(t+1)} = \frac{\sum_{i=1}^{N} g_{ic}^{(t)}\left(x_j^i - \mu_{cj}^{(t+1)}\right)^2}{\sum_{i=1}^{N} g_{ic}^{(t)}}$$