モデルベースクラスタリング

樋口さぶろお https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習　L08(2026-11-12 Thu)

最終更新: Time-stamp: "2026-11-12 Thu 18:21 JST hig"

今日の目標

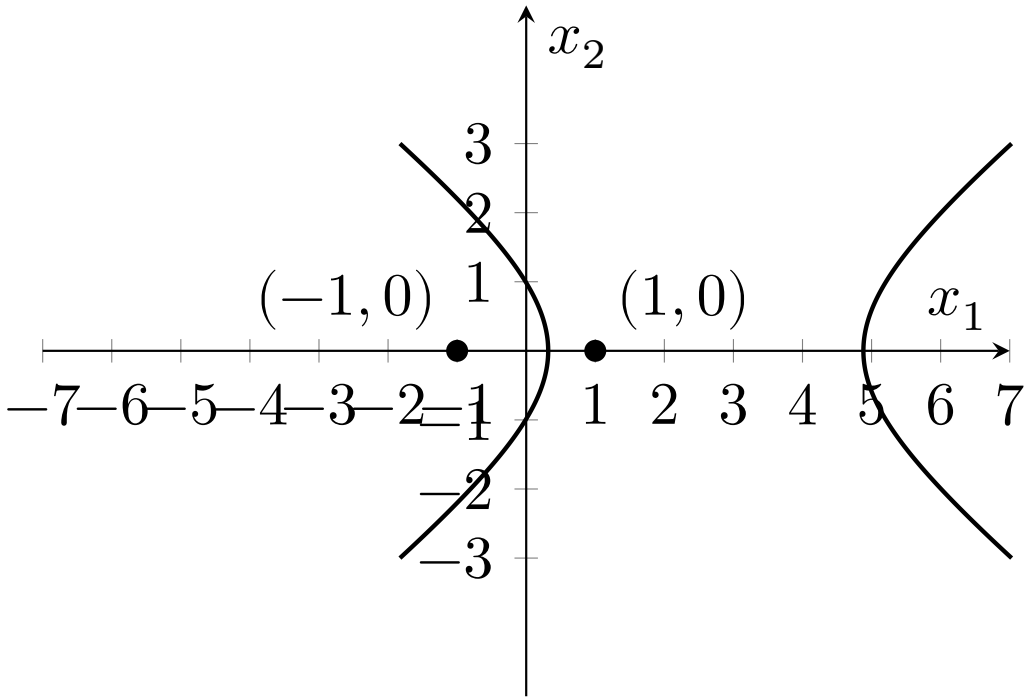
- 分類とクラスタ分析 (クラスタリング) の違いを説明できる
- ガウス混合モデルに対するモデルベースクラスタリングを説明できる

ここまで来たよ

8. モデルベースクラスタリング
   - 復習：ベイズ分類
   - クラスタリング
   - モデルベースクラスタリング

- 尤度比に基づくベイズ分類（どんな分布でも）
  - QDA=Quadratic Discriminant Analysis 2 次判別分析 (周辺分布がガウス混合分布 GMM のとき，境界線が 2 次曲線)
    - ~~QDA~~ マハラノビス距離のみによる 2 次判別分析 $\pi_y, \det \Sigma_y$ 無視.
    - LDA=Linear Discriminant Analysis 線形判別 ($\Sigma_0 = \Sigma_1$ のとき，判別境界が直線または平面)

QDA 判別境界は 2 次曲線



$\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ の間のマハラノビスの距離

$$D_\Sigma(\boldsymbol{x}, \boldsymbol{x}') = ((\boldsymbol{x} - \boldsymbol{x}')^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{x}'))^{1/2}$$

距離空間, 位相

簡単な例で理解 $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$ のとき, 'x の成分ごとに標準化'

$$D = \left[\left(\frac{x_1 - x_1'}{\sigma_1} \right)^2 + \left(\frac{x_2 - x_2'}{\sigma_2} \right)^2 \right]^{1/2}$$

簡単化：~~QDA~~ マラノビスの距離に基づく 2 次判別分析

QDA で一部の項を無視して，

$$-(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+(\boldsymbol{x}-\boldsymbol{\mu}_0)^\top \Sigma_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\geq 0$$

で考えると，マハラノビスの距離の大小で分類できる．

マラノビスの距離に基づく 2 次判別分析

$$D_{\Sigma_1}(\boldsymbol{x},\boldsymbol{\mu}_1)\leq D_{\Sigma_0}(\boldsymbol{x},\boldsymbol{\mu}_0)$$

ならば $\hat{y}=1$.
それぞれの正規分布の母平均値母分散で標準化．
判別境界は（QDA からずれた）2 次曲線．

超簡単化：LDA(線形判別分析)

QDA で一部の項を無視した上で，$\Sigma_0 = \Sigma_1 = \Sigma$ と仮定して，マハラノビスの距離だけで判定する

LDA

$$(\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^\top (\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)) \geq 0$$

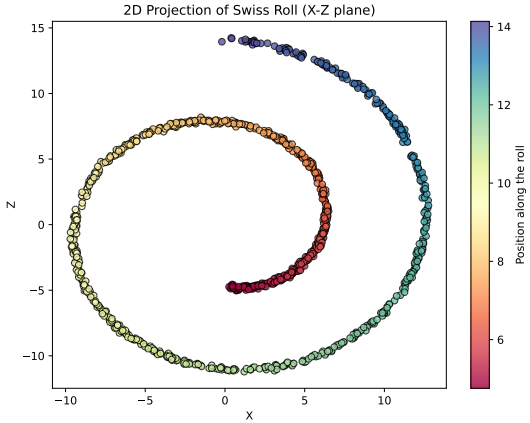
判別境界は，$\Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ に「直交」する直線.

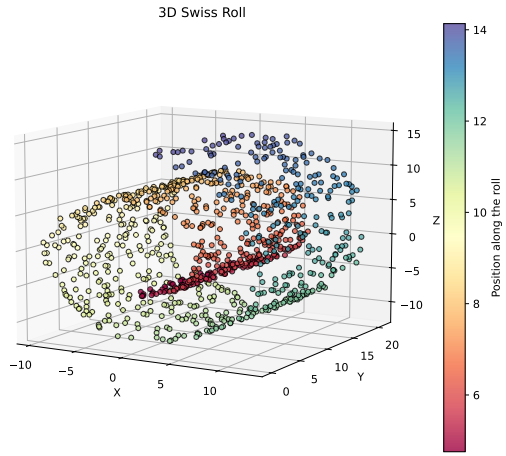
従来型分類手法の限界

- LDA 直線
- QDA 2 次曲線

空間内にスイスロール状に分布していたらどうする ⇝ 待て機械学習







ここまで来たよ

8. モデルベースクラスタリング
   - 復習：ベイズ分類
   - クラスタリング
   - モデルベースクラスタリング

多変量解析の手法の分類

教師あり

タイプ	教師 Y	手法・モデル
回帰	連続値	線形単回帰 L02 線形重回帰 L03,L04,L05
	非負整数	ポアソン回帰
分類	2 値（ラベル）	ロジスティック回帰 L06
	多値（ラベル・カテゴリ）	ベイズ分類 L07 QDA,LDA L07

時系列解析

L12 — とりあえず別世界

教師なし

タイプ	手法・モデル
クラスタリング	モデルベース L08 k-means 法 L09 ウォード法 L09
次元圧縮	主成分分析 (PCA) L10 因子分析 (Factor Analysis) L11

クラスター分析

ロジスティック回帰, ベイズ分類, QDA, LDA 分類. 教師あり学習 (supervised learning) $y_{\mathrm{train}i}$ は教師データ

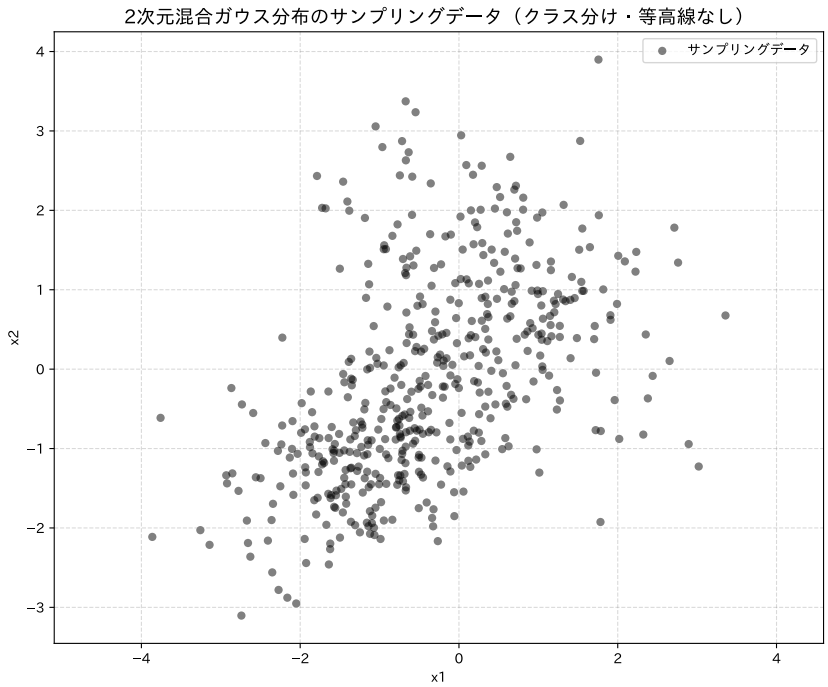
- 主に $y = 0, 1$ の 2 クラス. 固定.
- 訓練データ $(x_{\mathrm{train}i}, y_{\mathrm{train}i})$ $(i = 1, \dots, n)$ で学習 (=母数を推定). 教師=正解ラベル y_{train}.
- モデルと母数を使ってテストデータ x_{test} に対して, 予測値 $\hat{y}_{\mathrm{test}}$ を出力.

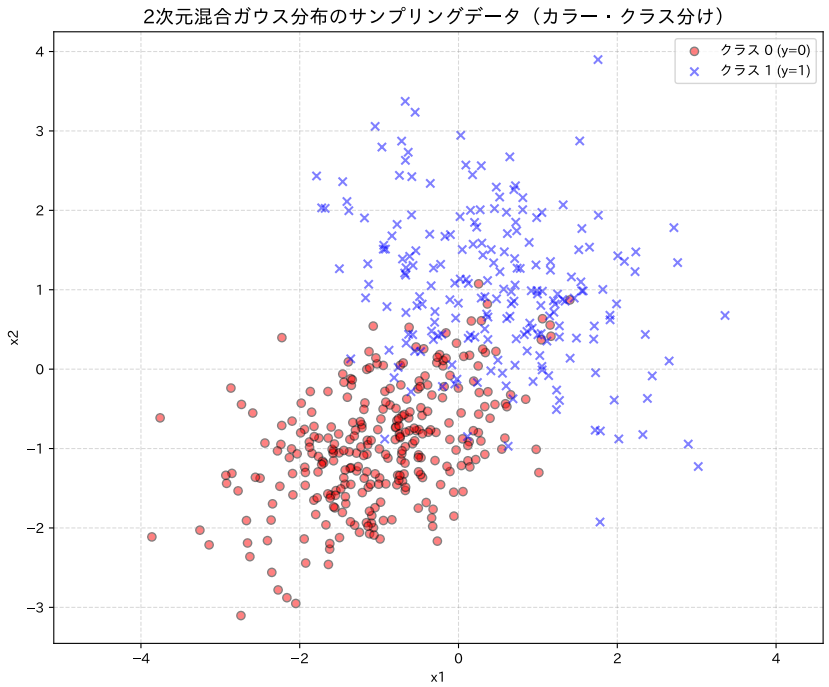
クラスター分析, クラスタリング 分類のようなものだが，正解ラベルはなく，いくつかの近くまとまった塊「クラスタ」に分ける. 教師なし学習 (unsupervised learning)

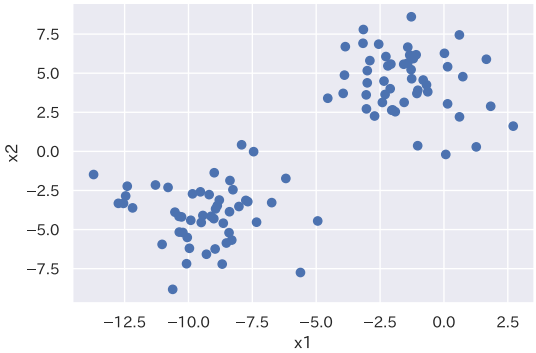
- クラスタの個数 K は人為的に指定, または後出しで決める. 正解はない.
- テストデータ $(x_{\mathrm{test}i})$ $(i = 1, \dots, n)$ を,「近い」点をまとめてクラスター (データ点の集合) に類別.
  - 一発の式はなく, アルゴリズム的，反復法的な感じで. ニュートン法数値計算法

教師データなし.「正解に近いか?」とは別の基準. クラスター分析=クラスタリング=まずはクラスターを作成すること

クラスタリングのタスク







ハード/ソフトクラスタリング

- ハードクラスタリング: データ点 x_i が，クラスタ k に所属，と出力 (or 途中で考慮)
- ソフトクラスタリング: データ点 x_i が，クラスタ $k = 0, 1, \ldots, K-1$ に所属する確率 r_{ik} をすべて出力 (or 途中で考慮)

ここまで来たよ

8. モデルベースクラスタリング
   - 復習：ベイズ分類
   - クラスタリング
   - モデルベースクラスタリング

混合分布

- 確率密度関数 $f_1(x), f_2(x), \dots, f_K(x)$
- 重み $\pi_1, \dots, \pi_K$ ただし，$\pi_k \geq 0, \sum_k \pi_k = 1$

があるとき，和 $f(x) = \sum_k \pi_k f_k(x)$ は確率密度関数になっており，確率変数 X の分布を定める．これを，$f_1, \dots, f_K$ の混合分布という．
$f_1, \dots, f_K$ が (多変量) ガウス分布であるとき，$f(x)$ をガウス混合分布 (Gaussian mixture) という

一般の（多変量）正規分布を（多変量）ガウス分布ともいう

これまで，混合ガウス分布とも書いていたが，同じもの

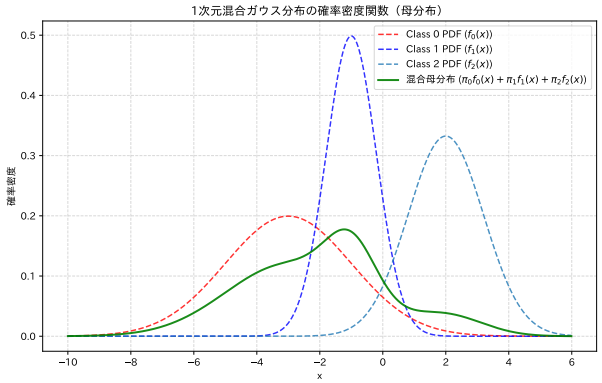
周辺分布 $f_{\mathrm{X}}(x)$ がガウス混合分布であるような，同時分布

$$f(x, y) = \pi_y f(x|\theta_y) = \pi_y \frac{1}{(2\pi\sigma_y^2)^{1/2}} \mathrm{e}^{-\frac{1}{2\sigma_y^2}(x-\mu_y)^2}$$

で，y を潜在変数（値が隠されることがある変数）とみなしたものを ガウス混合モデル (GMM) という

ガウス混合モデルベースクラスタリング

前提 $\boldsymbol{x}_i, y_i$ はガウス混合モデルにしたがう．同時分布は
$f(\boldsymbol{x}, y) = \pi_y f(\boldsymbol{x}|\boldsymbol{\mu}_y, \Sigma_y)$. y はクラスタを表す潜在変数 (値が隠されている)
クラスタは K 個． $y = 0, 1, 2, \ldots, K-1$
データ点 i がクラスタ y に属するかを表す確率変数 $Z_{iy} = 0, 1$.
入力 $\{\boldsymbol{x}_i\}(i = 1, \ldots, nZ)$
出力 責任度 $r_{iy} = P(Y_i = y) = \mathrm{E}[Z_{iy}]$: データ点 i がクラスタ y に属する確率
出力 データ点 i の属するクラスタ y(r_{iy} が最大となるクラスタ y)
$K = 3$ で，データ点 i が $y = 0$ のとき，$Z_{i0} = 1, Z_{i1} = 0, Z_{i2} = 0$.
$y = 2$ のとき，$Z_{i0} = 0, Z_{i1} = 0, Z_{i2} = 1$.



GMM のモデルベースクラスタリング

反復ループカウンタ $t = 0, 1, 2 \ldots,$. $t \to +\infty$ での値が出力.
母数 $\boldsymbol{\theta} = (\pi_0, \ldots, \pi_{K-1}, \boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_{K-1}, \Sigma_0, \ldots, \Sigma_{K-1})$
入力: $\boldsymbol{x}_i, i = 1, \ldots, n.$

- 初期化 $\boldsymbol{\theta}^{(t=0)}$, を適当に設定.
- 反復ループ $t = 0, 1, 2, \ldots, T$ で以下を繰り返す.
  - E ステップ: 与えられたパラメタ $\boldsymbol{\theta}^{(t)}$ に対して, ベイズ分類的に 責任度 $P(Y_i = y) = r_{iy}$ を計算.

$$r_{iy}^{(t)} = \frac{\pi_y^{(t)} f(\boldsymbol{x}_i | \boldsymbol{\mu}_y^{(t)}, \Sigma_y^{(t)})}{\sum_{y=0}^{K-1} \pi_y^{(t)} f(\boldsymbol{x}_i | \boldsymbol{\mu}_y^{(t)}, \Sigma_y^{(t)})}$$

  - M ステップ: 責任度 r_{iy} を重みとして, パラメタ $\boldsymbol{\theta}^{(t)}$ を更新.

$$\begin{aligned}
\boldsymbol{\mu}_y^{(t+1)} &= \frac{\sum_{i=1}^n r_{iy}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^n r_{iy}^{(t)}}, \\
\Sigma_y^{(t+1)} &= \frac{\sum_{i=1}^n r_{iy}^{(t)} (\boldsymbol{x}_i - \boldsymbol{\mu}_y^{(t+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_y^{(t+1)})^\top}{\sum_{i=1}^n r_{iy}^{(t)}}, \\
\pi_y^{(t+1)} &= \frac{\sum_{i=1}^n r_{iy}^{(t)}}{n}
\end{aligned}$$

ソフトクラスタリングの出力: 確率 $r_{iy} = P(Y_i = y)$
ハードクラスタリングの出力: 確率 $r_{iy} = P(Y_i = y)$ が最大のクラスタ y.

EM アルゴリズム

潜在変数（値の見えない変数）$Z_{ik} = \begin{cases} 1 & (i\text{が}k\text{に属する}) \\ 0 & (\text{他}) \end{cases}$ を考える．

目的 パラメタ $\boldsymbol{\theta}$ を推定する．副産物として，Z_{ik} も求めたい．

$r_{ik} = \mathrm{E}[Z_{ik}]$ はその母期待値．

手順

- Expectation ステップ: 値が不明な Z_{ik} に対して，x の条件付き確率で，その期待値 $r_{ik} = \mathrm{E}[Z_{ik}]$ を計算する．
- Maximization ステップ: 計算した事後分布を使って，パラメタ $\boldsymbol{\theta}$ を更新する．尤度を最大化するような母数を選ぶことから．

EM アルゴリズムは，欠測があるときに，欠測値を推定しながら最大化を行うアルゴリズムで，様々なところに現れる．

L08-Q1

ミニチュアモデルベースクラスタリング

データ $\{\boldsymbol{x}_i\}_{i=1,\dots,n}$ が次の通り与えられる．

$$\boldsymbol{x}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

潜在変数 $y = 0, 1$ を考え，データが母数未知の同時分布

$$f(\boldsymbol{x}, y) = \pi_y \frac{1}{(2\pi)^{2/2} (\det \Sigma_y)^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^\top \Sigma_y^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_y)} \quad (y = 0, 1, \quad \boldsymbol{x} \in \mathbb{R}^2)$$

の周辺分布 $f_{\mathrm{x}}(\boldsymbol{x})$ にしたがうと仮定する．
ただし，$\Sigma_0 = \Sigma_1 = \Sigma = \begin{pmatrix} 1^2 & 0 \\ 0 & 1^2 \end{pmatrix}$ は既知である．
EM アルゴリズムを 1 ステップのみ実行する．その際，母数 $\boldsymbol{\theta} = (\pi_y, \boldsymbol{\mu}_y)$ のみを更新する．Σ は固定する．母数の初期値 $\boldsymbol{\theta}^{(0)}$ は次の通りとする．

$$\pi_0^{(0)} = 1 - 1/3, \quad \pi_1^{(0)} = 1/3, \quad \boldsymbol{\mu}_0^{(0)} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_1^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$