混合ガウスモデルによる回帰分析および逆解析

Gaussian Mixture Regression GMR

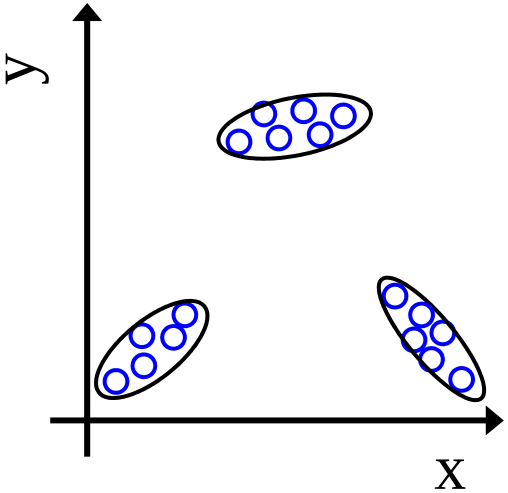
明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

GMR とは？

- ✓説明変数 X と目的変数 Y の関係を、複数の正規分布の重ね合わせで表現する手法
- ✓Y の変数が複数の場合でも、それらの関係を考慮しながら、他の手法と組み合わせることもなく、回帰分析および逆解析をすることが可能
- ✓逆解析においても、モデルの適用範囲を考慮した解が得られる

- ✓X と Y を合わせて混合ガウスモデルを構築
  - 混合ガウスモデル(Gaussian Mixture Models, GMM) についてはこちら: https://datachemeng.com/gaussianmixturemodel/
- ✓GMM は X と Y の同時確率密度分布 p(X, Y) に対応
- ✓確率の乗法定理とベイズの定理から、p(Y|X) を求めれば回帰分析、p(X|Y) を求めれば逆解析

GMR の概要



混合ガウスモデル
(Gaussian Mixture Models, GMM)

⇩

$p(\mathbf{x}, \mathbf{y})$ ＋ 確率の乗法定理
ベイズの定理

⇙ ⇘

回帰分析 $p(\mathbf{y} \mid \mathbf{x})$　　　$p(\mathbf{x} \mid \mathbf{y})$ 逆解析

混合正規分布 (混合ガウス分布)

✓説明変数 X と目的変数 Y とをつなげたもの (X の右に Y を追加したもの) を Z とする

✓X の変数の数を *k*, Y の変数の数を m, 正規分布の数を *n* とする

✓GMM を構築することは、以下の混合正規分布を得ることに対応する

(GMM の詳細はこちら: https://datachemeng.com/gaussianmixturemodel/)

$$p(\mathbf{z}) = \sum_{i=1}^{n} \pi_i N(\mathbf{z} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad \sum_{i=1}^{n} \pi_i = 1$$

$\mathbf{z}$: [$x_1, x_2, x_3, \ldots, x_k, y_1, y_2, y_3, \ldots, y_m$]
$\boldsymbol{\mu}_i$: i 番目の正規分布における 1 × (k+m) の平均ベクトル
$\boldsymbol{\Sigma}_i$: i 番目の正規分布における (k+m) × (k+m) の分散共分散行列
π_i : 混合係数 (各正規分布の重み)

X と Y を明示的に分けて書く

構築された混合正規分布は、X と Y の同時確率密度分布を意味する

$$p(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} \pi_i N\left(\begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}_{\mathrm{x},i} & \boldsymbol{\mu}_{\mathrm{y},i} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\mathrm{xx},i} & \boldsymbol{\Sigma}_{\mathrm{yx},i} \\ \boldsymbol{\Sigma}_{\mathrm{xy},i} & \boldsymbol{\Sigma}_{\mathrm{yy},i} \end{bmatrix} \right)$$

$\mathbf{x}$: [x_1, x_2, x_3, …, x_k]
$\mathbf{y}$: [y_1, y_2, y_3, …, y_m]
$\boldsymbol{\mu}_{\mathrm{x},i}$: i 番目の正規分布における、X に対応する $1 \times k$ の平均ベクトル
$\boldsymbol{\mu}_{\mathrm{y},i}$: i 番目の正規分布における、Y に対応する $1 \times m$ の平均ベクトル
$\boldsymbol{\Sigma}_{\mathrm{xx},i}$: i 番目の正規分布における、X に対応する $k \times k$ の分散共分散行列
$\boldsymbol{\Sigma}_{\mathrm{yy},i}$: i 番目の正規分布における、Y に対応する $m \times m$ の分散共分散行列
$\boldsymbol{\Sigma}_{\mathrm{xy},i}$ ($\boldsymbol{\Sigma}_{\mathrm{yx},i}$): i 番目の正規分布における、X と Y の $m \times k$ ($k \times m$) の共分散行列

X から Y の推定 (回帰分析)

X から Y を推定することは、X が与えられたときの Y の事後分布を求めることに対応

確率の乗法定理とベイズの定理より、

$$p(\mathbf{y}\mid\mathbf{x})=\sum_{i=1}^{n}p\left(\mathbf{y}\mid\mathbf{x},\boldsymbol{\mu}_{\mathrm{x},i},\boldsymbol{\Sigma}_{\mathrm{xx},i}\right)p\left(\boldsymbol{\mu}_{\mathrm{x},i},\boldsymbol{\Sigma}_{\mathrm{xx},i}\mid\mathbf{x}\right)$$

$$=\sum_{i=1}^{n}p\left(\mathbf{y}\mid\mathbf{x},\boldsymbol{\mu}_{\mathrm{x},i},\boldsymbol{\Sigma}_{\mathrm{xx},i}\right)\frac{p\left(\mathbf{x}\mid\boldsymbol{\mu}_{\mathrm{x},i},\boldsymbol{\Sigma}_{\mathrm{xx},i}\right)p\left(\boldsymbol{\mu}_{\mathrm{x},i},\boldsymbol{\Sigma}_{\mathrm{xx},i}\right)}{\sum_{j=1}^{n}p\left(\mathbf{x}\mid\boldsymbol{\mu}_{\mathrm{x},j},\boldsymbol{\Sigma}_{\mathrm{xx},j}\right)p\left(\boldsymbol{\mu}_{\mathrm{x},j},\boldsymbol{\Sigma}_{\mathrm{xx},j}\right)}$$

$$=\sum_{i=1}^{n}p\left(\mathbf{y}\mid\mathbf{x},\boldsymbol{\mu}_{\mathrm{x},i},\boldsymbol{\Sigma}_{\mathrm{xx},i}\right)\frac{\pi_i\,p\left(\mathbf{x}\mid\boldsymbol{\mu}_{\mathrm{x},i},\boldsymbol{\Sigma}_{\mathrm{xx},i}\right)}{\sum_{j=1}^{n}\pi_j\,p\left(\mathbf{x}\mid\boldsymbol{\mu}_{\mathrm{x},j},\boldsymbol{\Sigma}_{\mathrm{xx},j}\right)}$$

Y の推定値の混合正規分布とその重み

$$p(\mathbf{y} \mid \mathbf{x}) = \sum_{i=1}^{n} p\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i}\right) \frac{\pi_i p\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i}\right)}{\sum_{j=1}^{n} \pi_j p\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathrm{x},j}, \boldsymbol{\Sigma}_{\mathrm{xx},j}\right)}$$

$$w_{\mathrm{x},i} = \frac{\pi_i p\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i}\right)}{\sum_{j=1}^{n} \pi_j p\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathrm{x},j}, \boldsymbol{\Sigma}_{\mathrm{xx},j}\right)}$$ とすると、

$$p(\mathbf{y} \mid \mathbf{x}) = \sum_{i=1}^{n} w_{\mathrm{x},i}\, p\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i}\right)$$

$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i})$: i 番目の正規分布における、Y の推定値の (多変量) 正規分布

$w_{\mathrm{x},i}$: i 番目の正規分布における、Y の推定値の分布の重み

Y の推定値の(多変量)正規分布の平均

i 番目の正規分布における、Y の推定値の多変量正規分布 $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i})$ について、平均ベクトル $\mathbf{m}_i(\mathbf{x})$ は、

$$\mathbf{m}_i\left(\mathbf{x}\right) = \boldsymbol{\mu}_{\mathrm{y},i} + \left(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{x},i}\right)\boldsymbol{\Sigma}_{\mathrm{xx},i}^{-1}\boldsymbol{\Sigma}_{\mathrm{xy},i}$$

となる
式変形はビショップの本
[C.M. Bishop, パターン認識と機械学習 上下, 丸善出版, 2006]
を参照のこと

$\mathbf{x}$ に、X のサンプルを入力することで、i 番目の正規分布における Y の値を推定できる

Y の推定値の正規分布の重み

i 番目の正規分布における、Y の推定値の分布の重み $w_{\mathrm{x},i}$ について、

$$w_{\mathrm{x},i} = \frac{\pi_i p\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i}\right)}{\sum_{j=1}^{n} \pi_j p\left(\mathbf{x} \mid \boldsymbol{\mu}_{\mathrm{x},j}, \boldsymbol{\Sigma}_{\mathrm{xx},j}\right)}$$

$p(\mathbf{x} \mid \boldsymbol{\mu}_{\mathrm{x},i}, \boldsymbol{\Sigma}_{\mathrm{xx},i})$ は、平均ベクトル $\boldsymbol{\mu}_{\mathrm{x},i}$, 分散共分散行列 $\boldsymbol{\Sigma}_{\mathrm{xx},i}$ の

多変量正規分布における $\mathbf{x}$ での確率密度を計算することに対応する

最終的な Y の推定値をどうするか？

ある一つのサンプルについて、それぞれの Y の推定値は、
n 個の推定値とそれらの重みとして与えられる

それらから最終的な推定値を計算するには、

- ✓ mode (最頻値) : 重みが最も大きい推定値を選ぶ
- ✓ mean (平均値) : n 個の推定値の重み付け平均とする

の二通りある

X と Y との間の関係が非線形のとき、mean では上手くいかない場合があることが確認されているが、どちらがよいかはまだ定かではない

Y から X の推定 (逆解析)

p. 5-9 について、X と Y を入れ替えて同じことをすれば、

Y から X を推定でき、逆解析ができる

正規分布の数をどう決めるか？

✓正規分布の 1, 2, 3, … と振って GMM を行い、それぞれベイズ情報量規準 (Bayesian Information Criterion, BIC) を計算する

$$BIC = -2\log L + M \log N$$

- L : 尤度 (http://datachemeng.com/maximumlikelihoodestimation/)
- M : 推定するパラメータの数
  - 今回は詳細を記載しないが、分散共分散行列 $\boldsymbol{\Sigma}_k$ に制限を与えることで、M が変化する (制限しないときは考えなくてよい)
- N : サンプル数

✓BIC の値が最小となる正規分布の数とする

✓詳細は記載しなかったが、分散共分散行列の種類も一緒に決めることができる

正規分布の数をどう決めるか？ 補足

✓他の回帰分析手法と同じように、クロスバリデーションで正規分布の数や分散共分散行列の種類を最適化することも可能

✓回帰分析の性能をできるだけ上げたいのならクロスバリデーションによる最適化のほうがよく、逆解析の性能も考慮したいのなら BIC による最適化のほうがよいかもしれない (確証なし)

その他

- ✓今回は、X と Y を分けて、GMR の説明をしたが、基本的にすべての変数は同等に扱われる

- ✓そのため、X や Y の一部の変数を固定して、回帰分析や逆解析を行うことなど、柔軟な解析もできる

どうやって実際に GMR を実行するか？

✓GMR をするための Python のプログラムを作りました！

✓GMR のデモと、BIC で正規分布の数や分散共分散行列を最適化するデモも付いています！

https://github.com/hkaneko1985/sgmm

参考文献

- ✓T. Miyao, H. Kaneko, K. Funatsu, J. Chem. Inf. Model., 56, 286-299, 2016
- ✓C.M. Bishop, パターン認識と機械学習 上下, 丸善出版, 2006