7. Cluster Analysis

樋口さぶろお

龍谷大学大学院理工学研究科数理情報学専攻

理論物理学特論 L09(2020-11-23 Mon)

最終更新: Time-stamp: "2020-11-21 Sat 15:10 JST hig"

今日の目標

- k-means clustering を行列で表現できる

クラスター分析とは KA MDA §7

$n \times p$ データ行列 X. KA MDA Table 6.2(p.87)

- 行数: n 学生数=職業数
- 列数: p 科目数=特性の個数
- 行列の要素: X_{ij} 学生 i の科目 j の点数, 職業 i の特性 j の値

学生・職業を点数・特性値が「近い」k 個のグループ (cluster) に分けたい.

Membership Matrix G

$G = (g_{ik})$ KA MDA (7.1) の 1 個前 KA MDA Table 7.1

- 行数: n 学生数=職業数
- 列数: K クラスターの個数
- 行列の要素: g_{ik} 属すれば 1, 属さなければ 0.

各行に成分 1 はちょうど 1 個だけ. 学生, 職業はちょうど 1 個のクラスターに属する.

$\rightsquigarrow \mathrm{G}\mathbf{1}_K = \mathbf{1}_n$

Cluster Feature Matrix C KA MDA §7.2

$C = (c_{kj})$ KA MDA (7.1) の 1 個前 KA MDA Table 7.2

- 行数: K クラスターの個数
- 列数: p 特性の個数
- 行列の要素: c_{ik} クラスターを代表する特性の値

Formulation KA MDA §7.2

KA MDA §7.2

$$\mathrm{X} = \mathrm{GC} + \mathrm{E} \tag{7.3}$$

Minimize

$$f(\mathrm{G}, \mathrm{C}) = \|\mathrm{E}\|^2 = \|\mathrm{X} - \mathrm{GC}\|^2 \tag{7.4}$$

K 次元散布図 KA MDA Fig.7.2
点: X の各行 (青四角) $\tilde{\boldsymbol{x}}'_i$, C の各行 (赤丸) $\tilde{\boldsymbol{c}}'_k$,
K 個の座標: K 個の各特性の値

Iterative Algorithm KA MDA 7.4

- explicit solution (陽解法) 解=既知の量の式 PCA
- implicit solution (陰解法) 解の満たす方程式がわかるだけ f(解)=0
  - iterative solution (反復解法) だんだん真の解に収束していく (かも)

Algorithm (KMC=K-means Clustering) KA MDA p.98 下

Step 1. $\mathrm{G}_{[t]}, \mathrm{C}_{[t]}$ を $t = 0$, 適当な可能解 $\mathrm{G}_{[0]}, \mathrm{C}_{[0]}$ にセット.
Step 2. $f(\mathrm{G}_{[t]}, \mathrm{C})$ を C について最小化, $\mathrm{C}_{[t+1]}$ とおく.
Step 3. $f(\mathrm{G}, \mathrm{C}_{[t+1]})$ を G について最小化, $\mathrm{G}_{[t+1]}$ とおく.
Step 4. 「収束したら」終了, そうでなければ Step 2. に戻る.

収束条件

収束条件の例「あまり変化しなくなった」

$$f(\mathrm{G}_{[0]}, \mathrm{C}_{[0]}) - f(\mathrm{G}_{[1]}, \mathrm{C}_{[1]}) \leq \epsilon$$

振動しないの? $\rightsquigarrow$

$$f(\mathrm{G}_{[0]}, \mathrm{C}_{[0]}) \geq f(\mathrm{G}_{[0]}, \mathrm{C}_{[1]}) \geq f(\mathrm{G}_{[1]}, \mathrm{C}_{[1]}) \geq f(\mathrm{G}_{[1]}, \mathrm{C}_{[2]}) \geq f(\mathrm{G}_{[2]}, \mathrm{C}_{[2]}) \geq \cdots$$

真の最小値に到達する前に等号になっちゃわないの? $\rightsquigarrow$ 収束先が最小値でない心配
global minimum 対 local minimum

Step 2. の具体的手続き KA MDA §7.5

$$f(\mathrm{G}, \mathrm{C}) = \|\mathrm{X} - \mathrm{GC}\|^2$$

を C に関して最小化したい.

KA MDA A.2.2 によれば

$$F(\mathrm{B}) = \|\mathrm{Y} - \mathrm{XB}\|^2 \tag{A.2.11}$$

を最小化するには, X への射影 $\mathrm{P_{XY}} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$ を使って,

$$\mathrm{XB} = \mathrm{P_X Y} \tag{A.2.12}$$

となる, つまり $\mathrm{B} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$ とせよ, とのことだった.
$\mathrm{Y} = \mathrm{X}, \mathrm{X} = \mathrm{G}, \mathrm{B} = \mathrm{C}$ として C を決めると,

$$\mathrm{C} = (\mathrm{G}'\mathrm{G})^{-1}\mathrm{G}'\mathrm{X} = \mathrm{D}^{-1}\mathrm{G}'\mathrm{X} \tag{7.13}$$

Step 2. の直観的説明

$$\mathrm{C} = (\mathrm{G}'\mathrm{G})^{-1}\mathrm{G}'\mathrm{X} = \mathrm{D}^{-1}\mathrm{G}'\mathrm{X} \tag{7.13}$$

D は対角行列で, 対角成分は, クラスタに所属する学生数.

$$\begin{aligned} c_{kj} =& \overline{x}_{kj} = \text{変数 } j \text{ の値 } x_{ij} \text{ の, クラスター } k \text{ 内の平均値} \\ =& \frac{1}{\sum_i g_{ik}} \sum_{i \in \text{cluster } k} x_{ij} \end{aligned} \tag{7.15}$$

Step 3. の具体的手続き KA MDA §7.6

$$f(\mathrm{G}, \mathrm{C}) = \|\mathrm{X} - \mathrm{GC}\|^2$$

を G に関して最小化したい.
行列にしたからきれいに書けるわけではない. K 択が n 個あるだけだから大したことない.
いちばん小さくなる成分を 1 にしろ $\cdots$ (7.19)

練習問題

KA MDA (7.5) の 10×2 のデータ行列 X を考える.

1. $G_{[t]} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ に対して, KA MDA §7.4 の Step 2 により, $C_{[t+1]}$ を求めよう.
2. $C_{[t']} = \begin{pmatrix} 7 & 3 \\ 4 & 1 \\ 6 & 7 \end{pmatrix}$ に対して, KA MDA §7.4 の Step 3 により, $G_{[t'+1]}$ を求めよう (3 行目までででいい).

Excel にコピーして計算することをおすすめ.