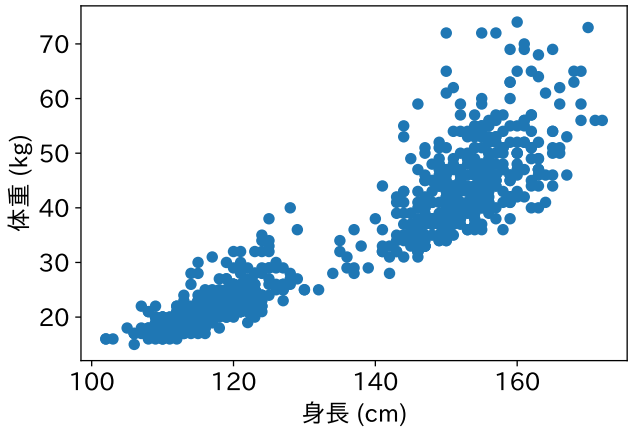
かたまりを見つける

松崎 拓也 *

2024 年 3 月 22 日

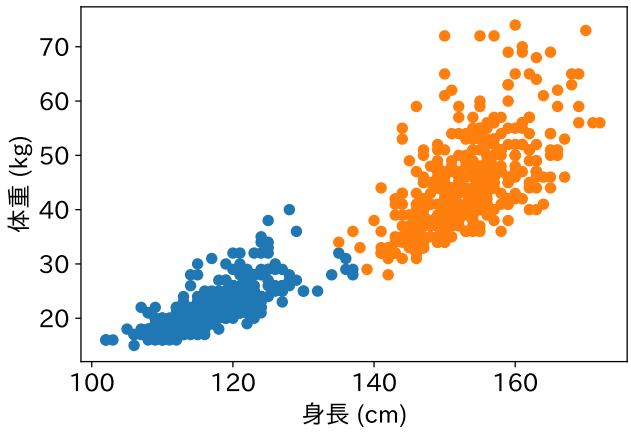
1　クラスタリングとは

日本全国の「1 年生」をたくさん集めて，身長と体重を調べたところ下の図のようになったとしましょう．ひとつの点が 1 人に対応し，横軸の値が身長，縦軸の値が体重を表します.[1]



なんだか変ですね．実は調べた人がうっかりしていて「小学校 1 年生」と「中学校 1 年生」が混じっているようです．ところで，この図から小学校 1 年生の平均身長と中学校 1 年生の平均身長は読み取れるでしょうか．

絶対確実とは言えませんが，おおむね図の左下のかたまりが小学校 1 年生，右上のかたまりが中学校 1 年生でしょうから，仮に下の図のオレンジと青のようにグループ分けしてみれば，それぞれの身長の平均値が計算できます．



*東京理科大学 理学部 応用数学科

[1] この図は，学校保険統計調査（令和 4 年）の「身長と体重の相関表及び身長別体重の平均値」の分布に従って 6 歳および 12 歳の男女それぞれ 200 人分（計 800 人分）のデータを生成したものです．

実際にグループごとに身長の平均値を計算してみると

- 青色の「小学校 1 年生らしき」グループの平均身長は 117.1cm
- オレンジ色の「中学校 1 年生らしき」グループの平均身長は 153.3cm

となりました.

学校保険統計調査（令和 4 年）によれば 2022 年度の 6 才の平均身長は女性が 116.0cm，男性が 117.0cm，12 才の平均身長は女性が 152.2cm，男性が 154.0cm ですから,「かたまりをもとにした推定」から，だいたい正しい値が分かったことになります．このように，人やものの特徴（この場合は身長と体重）をもとに，隠れたグループ（この場合は「小 1」と「中 1」）を見つけ出すことで，役に立つ情報をデータから引き出せることがあります.

上の例では，集めたデータは身長と体重の二つだけでしたので，それぞれを軸にとって図を描きグループを見つけるのは簡単でした．しかし，

> 足のサイズ，髪の毛の長さ，今日飲んだ水の量，先月読んだ本の冊数，昨日の睡眠時間，おとといの睡眠時間，...

といった具合に，人やものの特徴として非常にたくさんの種類のデータを集めることも，場合によっては可能です.

いま仮に，100 種類の特徴データをそれぞれの人から集めたとして，i 番目の人の j 番目の特徴（例えば「飼ってる猫の数」）を $x_j^{(i)}$ と書くことにします．そうすると，i 番目の人は

$$\begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_{100}^{(i)} \end{pmatrix}$$

という 100 次元のベクトルで表されることになります.

このようなベクトルが人数分集まったとして，さて，そこからどうやって隠れたグループを見つけたらよいでしょうか？

データをもとに隠れたグループを見つける方法は**クラスタリングアルゴリズム**と呼ばれ，色々な方法がこれまで提案されています．ここでは，そのうちの一つである **K-平均法**を紹介します.

2 K-平均法

全部で N 人の人について，D 種類の特徴を調べて，上のようにベクトルを作ったとします．i 番目の人のベクトルに

$$\vec{x}^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_D^{(i)} \end{pmatrix}$$

と名前をつけたとすると，以下の N 個のベクトルがデータとして与えられたということになります：

$$\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(N)}$$

これらを，なるべくまとまりのよいグループに分けたいわけです.

K-平均法では，データをいくつのグループに分けるかは，あらかじめ決めておきます．そのグループの数を K 個とします.

ところで「まとまりのよいグループ分け」とはどんなグループ分けでしょうか．色々な考え方がありえますが，K-平均法では以下の条件を満たすグループ分けを「まとまりがよい」と考えます：

どのデータも，他のどのグループの平均よりも自分が属するグループの平均に近い.

あるグループに属するデータが $\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_m$ であるとき，そのグループの平均を

$$\frac{1}{m}\left(\vec{z}_1 + \vec{z}_2 + \cdots + \vec{z}_m\right)$$

と定義します．例えばいま $D = 2$ として，以下の 3 つのベクトルからなるグループがあるとします：

$$\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}3\\4\end{pmatrix}.$$

このグループの平均は

$$\frac{1}{3}\left\{\begin{pmatrix}1\\2\end{pmatrix} + \begin{pmatrix}2\\0\end{pmatrix} + \begin{pmatrix}3\\10\end{pmatrix}\right\} = \begin{pmatrix}\frac{1+2+3}{3}\\\frac{2+0+10}{3}\end{pmatrix} = \begin{pmatrix}2\\4\end{pmatrix}$$

となります．つまり，あるグループの平均とは，そのグループに属するベクトルの各成分の平均値を並べたベクトルのことです．グループの平均は「大体どのあたりのベクトルが集まったグループか」を表しているので，そのグループの代表みたいなものだと考えられます．そうすると上の「まとまりのよいグループ」の条件は

どのデータも，各グループの代表のうち，自分のグループの代表に最も近い

と言い換えられます.

ここで，ふたつの D 次元ベクトル

$$\vec{x} = \begin{pmatrix}x_1\\x_2\\\vdots\\x_D\end{pmatrix}, \quad \vec{y} = \begin{pmatrix}y_1\\y_2\\\vdots\\y_D\end{pmatrix}$$

の間の距離を

$$||\vec{x} - \vec{y}|| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_D - y_D)^2}$$

と定義します.（ユークリッド距離と呼ばれます.）すると，あるグループの平均を

$$\vec{m} = \begin{pmatrix}m_1\\m_2\\\vdots\\m_D\end{pmatrix}$$

とするとき，データ $\vec{x}^{(i)}$ との距離は

$$||\vec{x}^{(i)} - \vec{m}|| = \sqrt{\left(x_1^{(i)} - m_1\right)^2 + \left(x_2^{(i)} - m_2\right)^2 + \cdots + \left(x_D^{(i)} - m_D\right)^2}$$

となります.

では「まとまりのよいグループ分け」を見つけるにはどうしたらよいでしょう．実は以下のように「まとまりがよくなりそうなグループにデータをどんどん移動させる」というやり方でいつかは「まとまりのよいグループ」が見つかることが分かります.

Step 0 : データ $\vec{x}^{(1)},\ldots,\vec{x}^{(N)}$ を適当に（ランダムに） K 個のグループに分ける.

Step 1 : 各グループの平均を計算し，それを $\vec{m}_1,\ldots,\vec{m}_K$ とする.

Step 2 : 各データ $\vec{x}^{(i)}$ $(i=1,\ldots,N)$ について，$\vec{m}_1,\ldots,\vec{m}_K$ との距離を計算し，最も近い平均が $\vec{m}_j$ ならば $\vec{x}^{(i)}$ を j 番目のグループに移動する.[2]

Step 3 : もしも Step 2 で移動するデータがひとつもなければ終了する．そうでなければ，新たなグループ分けに従って，各グループの平均 $\vec{m}_1,\ldots,\vec{m}_K$ を再計算し，Step 2 に戻る.

つまり，この方法は

グループ分け $\to$ 平均を計算 $\to$ 再グループ分け $\to$ 平均を再計算 $\to\cdots$

という風にグループ分けと平均の計算を繰り返すわけです.

もしもどこかの時点で再グループ分け（Step 2）の際にひとつもデータが移動しなかったとすると，次の Step 3 で手続きは終わります．それは各データが，全てのグループのうちで，自分がいま属するグループの平均に最も近いということを意味します．ですので，もしも上の手続きが終了したとすれば，それはこの節の最初の方で説明した「まとまりのよいグループ分け」がうまく見つかった，ということになります．最初のページのオレンジと青のグループ分けは，実際に 800 個のデータについて K-平均法を行なって得られたものです.

ここで気になるのは，この手続きで，どんなデータでも，いつでも「まとまりのよいグループ分け」が見つかるのかということです．手続きが終われば条件を満たすグループ分けが見つかっていることは分かりましたので，問題は「この手続きは必ずいつか終わるのか」ということになります．どう思いますか？次の節の説明を見る前に，少し考えてみてください.

3　K-平均法のもう一つの見方

K-平均法が必ずいつか終了して,「まとまりのよいグループ分け」が見つかることを示すために，次のようなことを考えます．まず，K-平均法の実行途中のある時点で，あるグループに属するデータが n 個あったとし，それらを $\vec{x}^{(\ell_1)},\ldots,\vec{x}^{(\ell_n)}$ とします．ここで，ある D 次元ベクトル $\vec{p}$ から，このグループの各データまでの距離の 2 乗の和，すなわち以下の量 $d(\vec{p})$ を考えます

$$d(\vec{p}) = ||\vec{x}^{(\ell_1)}-\vec{p}||^2+\cdots+||\vec{x}^{(\ell_n)}-\vec{p}||^2. \tag{1}$$

これを最小にするようなベクトル $\vec{p}$ はどんなベクトルでしょう？ベクトル $\vec{p}$ の成分を

$$\vec{p}=\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_D \end{pmatrix}$$

[2] より正確には，最も近い平均を持つグループが，現在属するグループを含めて二つ以上ある場合は（つまり現在のグループと他のグループが同点 1 位のときは）移動しない，というルールです．これは後で手続きが必ず止まることを示すときに重要になります.

として $d(\vec{p})$ を成分で表せば

$$d(\vec{p}) = \sum_{i=1}^{n} \left\{ (x_1^{(\ell_i)} - p_1)^2 + \cdots + (x_D^{(\ell_i)} - p_D)^2 \right\} \tag{2}$$

$$= \sum_{j=1}^{D} \left\{ (x_j^{(\ell_1)} - p_j)^2 + \cdots + (x_j^{(\ell_n)} - p_j)^2 \right\} \tag{3}$$

となります（添字に気をつけて確かめてください）．よって，$d(\vec{p})$ を最小にするベクトル $\vec{p}$ を見つけるには，各成分ごとに

$$(x_j^{(\ell_1)} - p_j)^2 + \cdots + (x_j^{(\ell_n)} - p_j)^2 \tag{4}$$

を最小にする p_j を見つけて，それを並べたベクトルを作ればよいことになります．

ところで式 (4) は

$$\begin{aligned}
&(x_j^{(\ell_1)} - p_j)^2 + \cdots + (x_j^{(\ell_n)} - p_j)^2 \\
=&\left(x_j^{(\ell_1)}\right)^2 + \cdots + \left(x_j^{(\ell_n)}\right)^2 - 2\left(x_j^{(\ell_1)} + \cdots + x_j^{(\ell_n)}\right) p_j + n p_j^2 \\
=&n\left(p_j - \frac{1}{n}\left(x_j^{(\ell_1)} + \cdots x_j^{(\ell_n)}\right)\right)^2 + (p_j \text{ に関係ない部分})
\end{aligned}$$

と p_j について平方完成できますから，式 (4) を最小にする p_j は

$$p_j = \frac{1}{n}\left(x_j^{(\ell_1)} + \cdots + x_j^{(\ell_n)}\right)$$

つまり $\vec{x}^{(\ell_1)}, \ldots, \vec{x}^{(\ell_n)}$ の第 j 成分の平均値です．よって，式 (1) の $d(\vec{p})$ を最小にするベクトル $\vec{p}$ は，各次元の成分の平均値を並べたベクトルということになり，つまりそれは（予想通り？）このグループの平均

$$\vec{m} = \frac{1}{n}\left(\vec{x}^{(\ell_1)} + \cdots + \vec{x}^{(\ell_n)}\right)$$

です．

いま，K-平均法のある時点の Step 2 の直前，つまり，グループへの再配置を始める前に，

$\vec{x}^{(1)}$ はグループ i_1 に属し，
$\vec{x}^{(2)}$ はグループ i_2 に属し，
$\ldots$,
$\vec{x}^{(N)}$ はグループ i_N に属していた

とします．ここで $i_1, i_2, \ldots, i_N \in \{1, 2, \ldots, K\}$ です．そして，Step 2 を実行したところ

$\vec{x}^{(1)}$ はグループ j_1 に属し，
$\vec{x}^{(2)}$ はグループ j_2 に属し，
$\ldots$,
$\vec{x}^{(N)}$ はグループ j_N に属す

という新たなグループ分けになったとします．（データ $\vec{x}^{(\ell)}$ がグループを移動しなかったときは $i_\ell = j_\ell$ とします．）さらに，Step 2 の直前の，$i_1, \ldots, i_N$ で表されるグループ分けに対する，各グループ k ($k = 1, \ldots, K$) の平均を $\vec{m}_k^{\text{old}}$ とし，Step 2 の実行後の，$j_1, \ldots, j_N$ で表されるグループ分けに対する，各グループ k の平均を $\vec{m}_k^{\text{new}}$ とします．

ここで，Step 2 を実行する前後で，各データと，それが属するグループの平均の距離の 2 乗の和がどう変化するか考えます．まず，Step 2 を実行したときにデータ $\vec{x}^{(\ell)}$ がグループ i_ℓ から j_ℓ に移動した（$i_\ell \neq j_\ell$）とすると，$\vec{x}^{(\ell)}$ は，移動前のグループ i_ℓ の平均 $\vec{m}_{i_\ell}^{\rm old}$ よりも移動先のグループ j_ℓ の（再計算前の）平均 $\vec{m}_{j_\ell}^{\rm old}$ に近いはずですから

$$||\vec{x}^{(\ell)} - \vec{m}_{i_\ell}^{\rm old}|| > ||\vec{x}^{(\ell)} - \vec{m}_{j_\ell}^{\rm old}||$$

が成り立ちます．両辺ともゼロ以上ですから，辺々 2 乗して

$$||\vec{x}^{(\ell)} - \vec{m}_{i_\ell}^{\rm old}||^2 > ||\vec{x}^{(\ell)} - \vec{m}_{j_\ell}^{\rm old}||^2 \tag{5}$$

でもあります．また，Step 2 でデータ $\vec{x}^{(\ell)}$ がグループを移動しなかったとすれば $i_\ell = j_\ell$ ですから当然

$$||\vec{x}^{(\ell)} - \vec{m}_{i_\ell}^{\rm old}||^2 = ||\vec{x}^{(\ell)} - \vec{m}_{j_\ell}^{\rm old}||^2 \tag{6}$$

となります．従って，Step 2 で少なくともひとつのデータがグループを移動したとすれば，全てのデータ $\vec{x}^{(1)}, \ldots, \vec{x}^{(N)}$ について，グループを移動した場合は式 (5)，しなかった場合は式 (6) の両辺の和をとることで

$$\sum_{\ell=1}^{N} ||\vec{x}^{(\ell)} - \vec{m}_{i_\ell}^{\rm old}||^2 > \sum_{\ell=1}^{N} ||\vec{x}^{(\ell)} - \vec{m}_{j_\ell}^{\rm old}||^2 \tag{7}$$

となります．次にこの式の右辺を，移動後のグループが同じデータごとの和，つまり，同じ j_ℓ を持つデータ $\vec{x}^{(\ell)}$ の和にまとめると

$$\sum_{\ell=1}^{N} ||\vec{x}^{(\ell)} - \vec{m}_{j_\ell}^{\rm old}||^2 = \sum_{k=1}^{K} \sum_{\ell:j_\ell=k} ||\vec{x}^{(\ell)} - \vec{m}_{k}^{\rm old}||^2 \tag{8}$$

となります．ここで右辺の内側の総和 $\sum_{\ell:j_\ell=k}$ は，$j_\ell = k$ となるような ℓ についての和を表します．式 (2) で定義した $d(\vec{p})$ を思い出すと，上の式の右辺の内側の総和は，新たなグループ k についての $d(\vec{m}_k^{\rm old})$ と同じになっています．上で示したように，この d は新たなグループ k の平均つまり $\vec{m}_k^{\rm new}$ によって最小になりますから

$$\sum_{k=1}^{K} \sum_{\ell:j_\ell=k} ||\vec{x}^{(\ell)} - \vec{m}_{k}^{\rm old}||^2 \geq \sum_{k=1}^{K} \sum_{\ell:j_\ell=k} ||\vec{x}^{(\ell)} - \vec{m}_{k}^{\rm new}||^2 \tag{9}$$

が成り立ちます．この右辺について，式 (8) と同様に

$$\sum_{k=1}^{K} \sum_{\ell:j_\ell=k} ||\vec{x}^{(\ell)} - \vec{m}_{k}^{\rm new}||^2 = \sum_{\ell=1}^{N} ||\vec{x}^{(\ell)} - \vec{m}_{j_\ell}^{\rm new}||^2 \tag{10}$$

が成り立ちますので，式 (7) から式 (10) を合わせると

$$\sum_{\ell=1}^{N} ||\vec{x}^{(\ell)} - \vec{m}_{i_\ell}^{\rm old}||^2 > \sum_{\ell=1}^{N} ||\vec{x}^{(\ell)} - \vec{m}_{j_\ell}^{\rm new}||^2$$

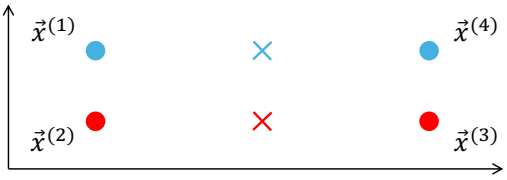
が分かります．これは，Step 2 でデータを再グループ化した際，少なくともひとつのデータがグループを移動したのであれば「各グループのメンバーとグループの平均の距離の 2 乗の和」（以下，この量を「グループ内の散らばり具合」と呼ぶことにします）が Step 2 の前後で必ず減少することを示しています．

従って，K-平均法の手順が繰り返される限り「グループ内の散らばり具合」は減少していくことになります．しかし，N 個のデータを K 個のグループに分けるやり方は有限通りしかありませんから，いくらでも減少し続けるということはありません．従って，K-平均法の手順はいつか必ず停止することが分かりました．

4　クラスタリングは運任せ？

ひとつ注意が必要なのは，K-平均法で必ず「グループ内の散らばり具合」が最も小さいグループ分けが見つかるとは限らないことです．例えば，いま以下の図に示す 4 つの 2 次元ベクトル $\vec{x}^{(1)}, \ldots, \vec{x}^{(4)}$ を 2 つのグループにクラスタリングするとします．

もしも K-平均法の Step 0 で最初のグループ分けが下図のようになったとすると，$\vec{x}^{(1)}$ と $\vec{x}^{(4)}$ は現在自分が属する水色のグループの平均 × により近く，$\vec{x}^{(2)}$ と $\vec{x}^{(3)}$ は現在自分が属する赤のグループの平均 × により近いため，次の Step 2 で移動するデータはありません．



よって，この場合，K-平均法の手順はただちに終了します．しかし，下図のグループ分けの方が，より「グループの散らばり具合」が小さいことは明らかです．

このように，Step 0 のランダムなグループ分けの結果によって，K-平均法で得られる最終的なグループ分けは異なる場合があり，それは必ずしも「グループの散らばり具合」が最も小さいグループ分けとは限りません．その意味で，K-平均法で得られる結果のよさ（「グループの散らばり具合」の小ささ）には多少「運任せ」なところがあります．

5　おわりに

このコラムでは，自動的にデータをグループ分けする方法のひとつである K-平均法について紹介しました．「ものごとをグループに分ける」ということは「おなじ種類のものをまとめて名前をつける」ということの第一歩とも言えます．例えば昔の人は，そこらを飛んだり這ったりしている 6 本足の小さな生き物を，似た特徴をもつものごとにまとめて「蝶（ちょう）」とか「蟻（アリ）」とか名前をつけました．ところでフランス語では「パピヨン（papillon）」というと日本語でいう「蝶」と「蛾」がどちらも含まれるそうです．日本語の感覚からすると「蝶」と「蛾」は大違いな気がしますが，フランス語では区別しないということです．同じデータの集まりから始めても K-平均法では違うグループ分けが得られることがある，ということと何か少し関係があるかもしれませんね．