9. 階層的クラスター分析

概要: 一つのオブジェクトは多くの要素を持つ。 ここでは、各オブジェクトの近さを距離で定量的に定義し、その近似度を評価する。似ている者同士のことをクラスターと呼ぶ。

キーワード:クラスター分析;階層的クラスター分析;距離、ウォード法.

9.1. 序

試験の点数を考えよう。試験の点数は人によってさまざまである。しかし、理数系が得意な人とか、文系が得意な人とかが存在するであろう。ここでは、このようなメンバーのグループ化をしたい。

また、スーパーにおける客を考える。お客が似た買い物をする傾向があれば、そのグループに向けて個別に情報を配信できる。

以上のように、われわれはいくつかのデータからあるグループを構成したい。つまり、クラスターを構成したい。ここでは、その定量的な手順を示していく。

ここでは、まずクラスターの数を最初から決めつけない方法を紹介する。このことを階層的クラスター分析と呼ぶ。クラスターの数を最初から決めて解析する方法を非階層的クラスター分析とよび、次章で紹介する。

9.2. クラスター化の簡単な例

ここでは、10 人に対して行った英語と数学の成績について考える。

結果は表 1 のようなものであったとする。

メンバー A と B の距離を $d(A,B)$ とすると、それは以下となる。

$$\begin{aligned} d(A,B) &= \sqrt{(x_{A1}-x_{B1})^2+(x_{A2}-x_{B2})^2} \\ &= \sqrt{(67-43)^2+(64-76)^2} \\ &= 26.8 \end{aligned} \tag{1}$$

もしもメンバー A と B が全く同じ点数であれば、これは 0 になる。したがって、この距離は二人のメンバーの近さを表現している。

表 1　英語と数学の成績

Member ID	English x_1	Mathematics x_2
A	67	64
B	43	76
C	45	72
D	28	92
E	77	64
F	59	40
G	28	76
H	28	60
I	45	12
J	47	80

このような評価を他のメンバーの組み合わせに対しても行うと表 2 のような結果を得る。最小の距離はメンバー B と C である。したがって、我々はメンバーと C のクラスターを構成できる。

クラスターを構成後、そのクラスターの値をどうするのか、という問題がでてくる。一番簡単なものはその平均を取るということである。つまり、そのクラスターの英語と数学の点数はその平均値、つまり、

$$x_{BC1} = \frac{x_{B1} + x_{C1}}{2}, x_{BC2} = \frac{x_{B2} + x_{C2}}{2} \tag{2}$$

であるとする。このデータを使い、次のステップの評価をすると表 3 のようなものを得る。

表 2　各メンバー間の距離の第 1 ステップ

	A	B	C	D	E	F	G	H	I	J
A		26.8	23.4	48.0	10.0	25.3	40.8	39.2	56.5	25.6
B			4.5	21.9	36.1	39.4	15.0	21.9	64.0	5.7
C				26.2	33.0	34.9	17.5	20.8	60.0	8.2
D					56.4	60.5	16.0	32.0	81.8	22.5
E						30.0	50.4	49.2	61.1	34.0
F							47.5	36.9	31.3	41.8
G								16.0	66.2	19.4
H									50.9	27.6
I										68.0
J										

表 3　各メンバー間の距離の評価の第 2 ステップ

Member ID	English x_1	Mathematics x_2
A	67	64
B,C	44	74
D	28	92
E	77	64
F	59	40
G	28	76
H	28	60
I	45	12
J	47	80

	A	B,C	D	E	F	G	H	I	J
A		25.1	48.0	10.0	25.3	40.8	39.2	56.5	25.6
B,C			24.1	34.5	37.2	16.1	21.3	62.0	6.7
D				56.4	60.5	16.0	32.0	81.8	22.5
E					30.0	50.4	49.2	61.1	34.0
F						47.5	36.9	31.3	41.8
G							16.0	66.2	19.4
H								50.9	27.6
I									68.0
J									

この場合、クラスター BC と　J　の距離が最小となる。したがって、我々はクラスター BCJ を構成し、対応するデータを以下とする。

$$x_{BCJ1}=\frac{x_{B1}+x_{C1}+x_{J1}}{3}, x_{BC2}=\frac{x_{B2}+x_{C2}+x_{J2}}{3} \tag{3}$$

これらのデータを使い、我々は各メンバーの距離を評価し表 4 の結果を得る。

表 4　各メンバー間の距離の評価の第 3 ステップ

Member ID	English x_1	Mathematics x_2
A	67	64
B,C,J	45	76
D	28	92
E	77	64
F	59	40
G	28	76
H	28	60
I	45	12

	A	B,C,J	D	E	F	G	H	I
A		25.1	48.0	10.0	25.3	40.8	39.2	56.5
B,C,J			23.3	34.2	38.6	17.0	23.3	64.0
D				56.4	60.5	16.0	32.0	81.8
E					30.0	50.4	49.2	61.1
F						47.5	36.9	31.3
G							16.0	66.2
H								50.9
I								

この場合、メンバー A　と　E の距離が最小となる。したがって、クラスター AE　を構成し、そのデータを以下とする。

$$x_{AE1}=\frac{x_{A1}+x_{E1}}{2}, x_{AE2}=\frac{x_{A2}+x_{E2}}{2} \tag{4}$$

これらのデータをつかり、各クラスターとメンバーの距離を比較し、表 5 **エラー! 参照元が見つかりません。**の結果を得る。

表 5　各メンバー間の距離の評価の第 4 ステップ

Member ID	English x_1	Mathematics x_2
A,E	72	64
B,C,J	45	76
D	28	92
F	59	40
G	28	76
H	28	60
I	45	12

	A,E	B,C,J	D	F	G	H	I
A,E		29.5	52.2	27.3	45.6	44.2	58.6
B,C,J			23.3	38.6	17.0	23.3	64.0
D				60.5	16.0	32.0	81.8
F					47.5	36.9	31.3
G						16.0	66.2
H							50.9
I							

この場合、D，G ,および H 間の距離が最小となる。したがって、クラスター DGH を構成し、クラスターの値を以下とする。

$$x_{DGH1}=\frac{x_{D1}+x_{G1}+x_{H1}}{3}, x_{DGH2}=\frac{x_{D2}+x_{G2}+x_{H2}}{3} \tag{5}$$

これらの値を使い、各クラスターとメンバーの間の距離を評価し、表 6 の結果を得る。

表 6　各メンバー間の距離の評価の第 5 ステップ

Member ID	English x_1	Mathematics x_2
A,E	72	64
B,C,J	45	76
D,G,H	28	76
F	59	40
I	45	12

	A,E	B,C,J	D,GH	F	I
A,E		29.55	45.61	27.29	58.59
B,C,J			17.00	38.63	64.00
D,GH				47.51	66.22
F					31.30
I					

この場合、クアスター BCJ　とクラスター　DGH の距離が最小になる。したがって、これらを合わせたクラスター $BCJDGH$ を構成し、そのクラスターの値を以下とする。

$$x_{BCJDGH1}=\frac{x_{B1}+x_{C1}+x_{J1}+x_{D1}+x_{G1}+x_{H1}}{6}, x_{BCJDGH2}=\frac{x_{B2}+x_{C2}+x_{J2}+x_{D2}+x_{G2}+x_{H2}}{6} \tag{6}$$

これらの値を使い、クラスターとメンバーの距離を評価し、表 7 の結果を得る。

表 7　各メンバー間の距離の評価の第 6 ステップ

Member ID	English x_1	Mathematics x_2
A,E	72	64
B,C,J,D,G,H	36.5	76
F	59	40
I	45	12

	A,E	B,C,J,D,G,H	F	I
A,E		37.47	27.29	58.59
B,C,J,D,G,H			42.45	64.56
F				31.30
I				

この場合、クラスター AE　と F　の間の距離が最小であるから、クラスター AEF　を構成し、そのクラスターに対応する値を以下とする。

$$x_{AEF1}=\frac{x_{A1}+x_{E1}+x_{F1}}{3}, x_{AEF2}=\frac{x_{A2}+x_{E2}+x_{F2}}{3} \tag{7}$$

表 8　各メンバー間の距離の評価の第 7 ステップ

Member ID	English x_1	Mathematics x_2
A,E,F	67.67	56.00
B,C,J,D,G,H	36.50	76.00
I	45.00	12.00

	A,E,F	B,C,J,D,G,H	I
A,E,F		37.03	49.50
B,C,J,D,G,H			64.56
I			

これらのデータをつかり、クラスターとメンバーの距離を評価し、表 8 の結果を得る。この場合、クラスター *AEF* とクラスター *BCJDGH* の間の距離が最小になる。したがって、我々はクラスター *AEFBCJDGH* を構成し、対応するクラスターの値を以下とする。

最後に、我々は Table 9 の結果を得る。

Table 1 各メンバー間の距離の評価の第 8 ステップ。

Member ID	English x_1	Mathematics x_2
A,E,F,B,C,J,D,G,H	46.89	69.33
I	45.00	12.00

	A,E,F,B,C,J,D,G,H	I
A,E,F,B,C,J,D,G,H		57.36
I		

この一連のプロセスの後、図 1 のような樹形図を作成できる。これから、臨界の距離を設定することにより、我々はいくつかのグループを構成できる。もし、臨界の距離を 25 とすると以下の 4 つのグループを得ることができる。

[B,C,J,D,G,H], [A,E], F,I

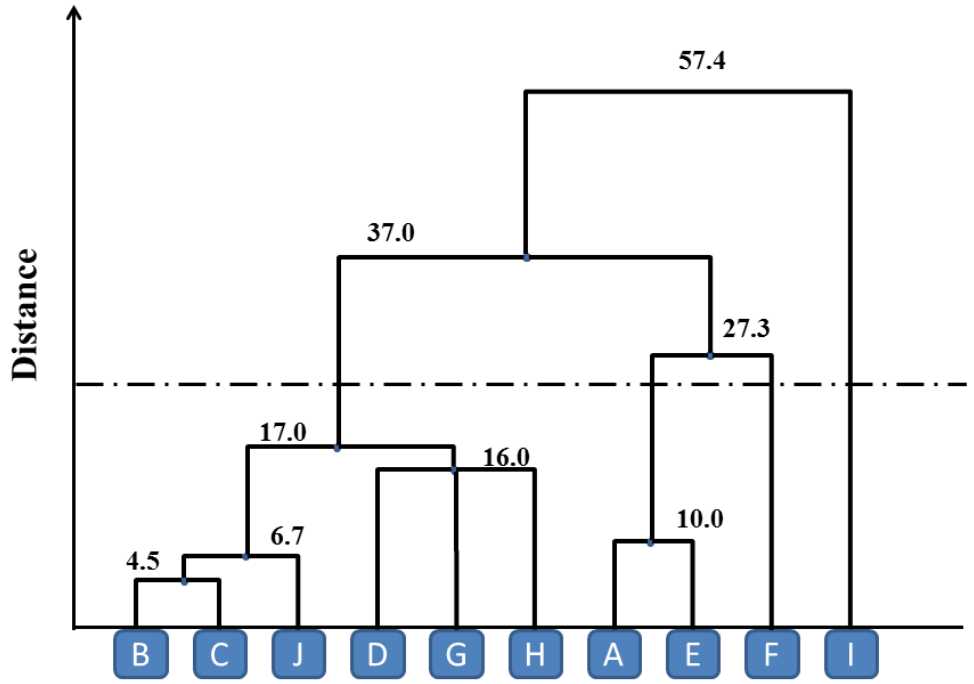

図 1　樹形図

9.3. データの規格化

データ間の距離の定義にはいろいろある。この場合は、成績であったため、基本的には同じ評価尺度である。しかし、まったく違う量からなる構成要素の場合もある。このような場合は規格化値を用いる。

それぞれの因子の平均と分散が以下のように分かっているとする。

$$\mu_{x_1} = \frac{1}{n}\sum_i x_{i1}, \sigma_{x_1} = \frac{1}{n}\sum_i \left(x_{i1} - \mu_{x_1}\right)^2 \tag{8}$$

$$\mu_{x_2} = \frac{1}{n}\sum_i x_{i2}, \sigma_{x_2} = \frac{1}{n}\sum_i \left(x_{i2} - \mu_{x_2}\right)^2 \tag{9}$$

ここで、n はデータの数である。

その場合、規格化値は以下となる。

$$z_{i1} = \frac{x_{i1} - \mu_{x_1}}{\sigma_{x_1}}, z_{i2} = \frac{x_{i2} - \mu_{x_2}}{\sigma_{x2}} \tag{10}$$

この規格化値を使って上の解析をしていく。

9.4. 距離と分散の関係

これまでの解析で二つのデータの距離を定義してきた。その二乗は以下である。

$$\begin{aligned} d^2 &= \left(x_{ax_1} - x_{bx_1}\right)^2 + \left(x_{ax_2} - x_{bx_2}\right)^2 \\ &= \left[\left(x_{ax_1} - \bar{x}_{x_1}\right) - \left(x_{bx_1} - \bar{x}_{x_1}\right)\right]^2 + \left[\left(x_{ax_2} - \bar{x}_{x_2}\right) - \left(x_{bx_2} - \bar{x}_{x_2}\right)\right]^2 \\ &= \left(x_{ax_1} - \bar{x}_{x_1}\right)^2 + \left(x_{bx_1} - \bar{x}_{x_1}\right)^2 - 2\left(x_{ax_1} - \bar{x}_{x_1}\right)\left(x_{bx_1} - \bar{x}_{x_1}\right) \\ &\quad + \left(x_{ax_2} - \bar{x}_{x_2}\right)^2 + \left(x_{bx_2} - \bar{x}_{x_2}\right)^2 - 2\left(x_{ax_2} - \bar{x}_{x_2}\right)\left(x_{bx_2} - \bar{x}_{x_2}\right) \\ &= S_{ab}^{(2)} - 2\left[\left(x_{ax_1} - \bar{x}_{x_1}\right)\left(x_{bx_1} - \bar{x}_{x_1}\right) + \left(x_{ax_2} - \bar{x}_{x_2}\right)\left(x_{bx_2} - \bar{x}_{x_2}\right)\right] \end{aligned} \tag{11}$$

ここで、

$$\begin{aligned} \left(x_{ax_1} - \bar{x}_{x_1}\right)\left(x_{bx_1} - \bar{x}_{x_1}\right) &= \left(x_{ax_1} - \frac{x_{ax_1} + x_{bx_1}}{2}\right)\left(x_{bx_1} - \frac{x_{ax_1} + x_{bx_1}}{2}\right) \\ &= \frac{x_{ax_1} - x_{bx_1}}{2}\frac{x_{bx_1} - x_{ax_1}}{2} \\ &= -\frac{1}{4}\left(x_{ax_1} - x_{bx_1}\right)^2 \end{aligned} \tag{12}$$

同じような解析で以下を得る。

$$\left(x_{ax_2} - \bar{x}_{x_2}\right)\left(x_{bx_2} - \bar{x}_{x_2}\right) = -\frac{1}{4}\left(x_{ax_2} - x_{bx_2}\right)^2 \tag{13}$$

したがって、以下となる。

$$\begin{aligned} d^2 &= S_{ab}^{(2)} + \frac{1}{2}\left[\left(x_{ax_1} - x_{bx_1}\right)^2 + \left(x_{ax_2} - x_{bx_2}\right)^2\right] \\ &= S_{ab}^{(2)} + \frac{1}{2}d^2 \end{aligned} \tag{14}$$

これは以下のように変形できる。

$$d^2 = 2S_{ab}^{(2)} \tag{15}$$

したがって、この距離は分散に 2 倍ということになる。

9.5. 様々な距離

クラスター分析においては距離が重要な指標となっている。その最も単純なものは、前節で利用した、以下のものである。

$$d_{ab} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ia} - x_{ib}\right)^2} \tag{16}$$

しかしながら、この距離の定義には様々なものがある。ここでは、それについて簡単に触れる。

まず、この各距離に重みをつける、という方法がある。その場合は以下となる。

$$d_{ab} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i \left(x_{ia} - x_{ib}\right)^2} \tag{17}$$

距離の一般化に対しては、以下がある。

$$d_{ab} = \left[\frac{1}{n}\sum_{i=1}^{n} \left|x_{ia} - x_{ib}\right|^{\alpha}\right]^{\frac{1}{\alpha}} \tag{18}$$

このモデルにおいては、α が大きくなるほど、どれから突出して違うとそれで違うと判断することになる。

9.6. 評価項目間の相互作用

ここでは、距離に対して評価する項目間の相互作用を考慮していなかった。その相互作用は判別分析の場合と同様に簡単に取り入れることができる。それは以下となる。

$$d_{ab} = \sqrt{\frac{1}{n}\left(\boldsymbol{\mu}_{z_a} - \boldsymbol{\mu}_{z_b}\right)^T \Sigma^{-1} \left(\boldsymbol{\mu}_{z_a} - \boldsymbol{\mu}_{z_b}\right)} \tag{19}$$

ここで、項目に対する評価は規格化されているものとする。

9.7. 鎖効果

鎖効果とは、一つの因子が全体を支配してしまうことである。この特殊な場合では、そのループは常に二つになってしまう。この状況は我々の欲するものではない。ここで紹介した単純な距離の評価よりも、これから紹介するウォード法のほうが、この鎖効果を低減できる、といわれている。

9.8. 二つの因子に対するウォード法

ウォード法は、データの分散を最小にするようにクラスターを構成する、という手法である。この方法は上で指摘した鎖効果に強いと言われており、頻繁に利用される。

表 9　日本語と英語の成績

Member ID	Japanese x_1	English x_2
1	5	1
2	4	2
3	1	5
4	5	4
5	5	5

表 10　メンバー1 と 2 の絵日本語と英語の成績

Member ID	Japanese x_1	English x_2
1	5	1
2	4	2
Average	4.5	1.5

表 9 に示すように日本語と英語の 5 人の成績を考える。

この中でメンバー1 の 2 の距離を考える。

メンバー1 と 2 の国語と英語の成績の平均を表 10 のように考え、それを基準にした二乗和を以下のように評価する。

$$\begin{aligned} K_{12} &= \sum_{i=1}^{2}\sum_{k=1}^{2}\left(x_{ik}-\bar{x}_k\right)^2 \\ &= \left(5-4.5\right)^2+\left(4-4.5\right)^2+\left(1-1.5\right)^2+\left(2-1.5\right)^2 \\ &= 1.00 \end{aligned} \tag{20}$$

表 11 に示すように他の組み合わせに対しても同様な評価をする。最小の分散が一番近いと判断する。この場合は、メンバー4 と 5 が近いのでこの二人でクラスター C_1 を構成する。

表 11　クラスター間の二乗平方和

ID	1	2	3	4	5
1					
2	1				
3	16	9			
4	4.5	2.5	8.5		
5	8	5	8	0.5	

クラスター C_1 とメンバー 1 の間の評価をする。

対応するデータを表 12 に示す。

表 12　C_1 とメンバ 1 の成績表と平均

Cluster ID	Member ID	Japanese x_1	English x_2
1	1	5	1
C1	4	5	4
	5	5	5
Average		5.00	3.33

この場合の国語の平均は 5 であり、英語の平均は.33 である。対応する二乗和 $K_{C_1 1}$ を以下のように評価する。

$$\begin{aligned} K_{C_1 1} &= \sum_{i=1}^{3}\sum_{k=1}^{2}\left(x_{ik}-\bar{x}_k\right)^2 \\ &= \left(5-5.00\right)^2+\left(5-5.00\right)^2+\left(5-5.00\right)^2 \\ &\quad +\left(1-3.33\right)^2+\left(4-3.33\right)^2+\left(5-3.33\right)^2 \\ &= 8.67 \end{aligned} \tag{21}$$

この二乗和をそのまま利用しない。しかし、その二乗和の増加 $\Delta K_{C_1 1}$ を以下のように評価する。

$$\begin{aligned} \Delta K_{C_1 1} &= K_{C_1 1}-\left(K_{C_1}+K_1\right) \\ &= 8.67-\left(0.5+0\right) \\ &= 8.17 \end{aligned} \tag{22}$$

ここで K_1 は 0 である。何故ならば、それは一つのデータであるからである。

この二乗和の増加を他の組み合わせに対してもおこなってまとめたものが表 13 である。最小のものがメンバー 1 と 2 の組み合わせである。したがって、我々はメンバー1 と 2 でクラスター C_2 を構成する。

表 13　二乗和の評価の第 2 ステップ

ID	1	2	3	C_1
1				
2	1			
3	16	9		
C_1	8.17	4.83	10.83	

次にクラスター C_1，C_2，およびメンバー 3 の間で、二乗和の増加を考え、その結果を表 14 に示す。その差が最小なのはクラスター C_1 とクラスター C_2 を組み合わせた場合であ

る。したがって、対応する樹系図は図 2 のようになる。

表 14　二乗和の評価の第 3 ステップ

ID	3	C_1	C_2
3			
C_1	10.83		
C_2	16.33	9.25	

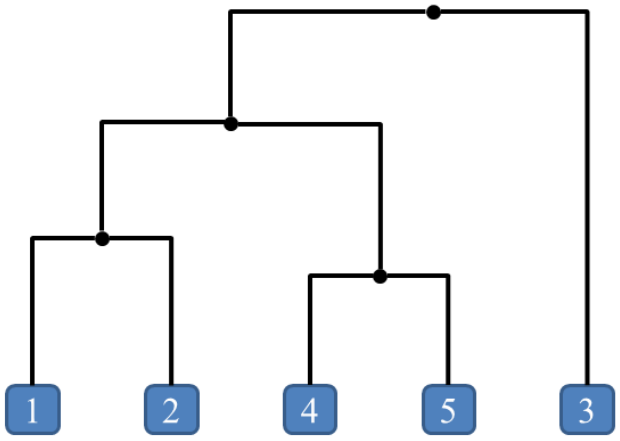

図 2　ウォード法によって構成した樹形図

9.9. ウォード法の多変数への拡張

前節では二つの因子に対するウォード法を示した。このウォード法は 2 因子以上のものに容易に拡張することができる。

ここでは一般に n 個の因子を考える。

クラスター l とクラスター m をマージしてクラスタ lm を構成することを考える。クラスター l とクラスター m はデータ x_{ilk}, x_{imk} ,を持っているとする。そして、それぞれのクラスターのデータ数を n_l および n_m とする。対応する二乗和は以下となる。

$$K_l = \sum_{i=1}^{n}\sum_{k=1}^{n_l}\left(x_{ilk} - \bar{x}_{il}\right)^2 \tag{23}$$

$$K_m = \sum_{i=1}^{n}\sum_{k=1}^{n_m}\left(x_{imk} - \bar{x}_{im}\right)^2 \tag{24}$$

マージした二乗和は以下となる。

$$K_{lm} = K_l + K_m + \Delta K_{lm} \tag{25}$$

二乗和の差は以下となる。

$$\Delta K_{lm} = \frac{n_l n_m}{n_l + n_m} \sum_{i=1}^{n} \left(\bar{x}_{il} - \bar{x}_{im} \right)^2 \tag{26}$$

これはウォード法を運用するうえで大変重要な定理であり、以下のように証明される。

二乗和 K_{lm} は以下のように与えられる。

$$K_{lm} = \sum_{i=1}^{n} \left[\sum_{k=1}^{n_l} \left(x_{ilk} - \bar{x}_{ilm} \right)^2 + \sum_{k=1}^{n_m} \left(x_{imk} - \bar{x}_{ilm} \right)^2 \right] \tag{27}$$

ここで、マージしたクラスタ lm の平均値は以下となる。

$$\bar{x}_{ilm} = \frac{\sum_{k=1}^{n_l} x_{ilk} + \sum_{k=1}^{n_m} x_{imk}}{n_l + n_m} \tag{28}$$

この方程式を K_l と K_m に以下のように関連付ける。

$$\begin{aligned}
K_{lm} &= \sum_{i=1}^{n} \left[\sum_{k=1}^{n_l} \left(x_{ilk} - \bar{x}_{il} + \bar{x}_{il} - \bar{x}_{ilm} \right)^2 + \sum_{k=1}^{n_m} \left(x_{imk} - \bar{x}_{im} + \bar{x}_{im} - \bar{x}_{ilm} \right)^2 \right] \\
&= \sum_{i=1}^{n} \left[\begin{array}{l} \sum_{k=1}^{n_l} \left(x_{ilk} - \bar{x}_{il} \right)^2 + 2 \sum_{k=1}^{n_l} \left(x_{ilk} - \bar{x}_{il} \right) \left(\bar{x}_{il} - \bar{x}_{ilm} \right) + n_l \left(\bar{x}_{il} - \bar{x}_{ilm} \right)^2 \\ + \sum_{k=1}^{n_m} \left(x_{imk} - \bar{x}_{im} \right)^2 + 2 \sum_{k=1}^{n_m} \left(x_{imk} - \bar{x}_{im} \right) \left(\bar{x}_{im} - \bar{x}_{ilm} \right) + n_m \left(\bar{x}_{im} - \bar{x}_{ilm} \right)^2 \end{array} \right] \\
&= \sum_{i=1}^{n} \left[\sum_{k=1}^{n_l} \left(x_{ilk} - \bar{x}_{il} \right)^2 + \sum_{k=1}^{n_m} \left(x_{imk} - \bar{x}_{im} \right)^2 \right] \\
&\quad + \sum_{i=1}^{n} \left[n_l \left(\bar{x}_{il} - \bar{x}_{ilm} \right)^2 + n_m \left(\bar{x}_{im} - \bar{x}_{ilm} \right)^2 \right] \\
&= K_l + K_m + \sum_{i=1}^{n} \left[n_l \left(\bar{x}_{il} - \bar{x}_{ilm} \right)^2 + n_m \left(\bar{x}_{im} - \bar{x}_{ilm} \right)^2 \right]
\end{aligned} \tag{29}$$

ここで以下を利用している。

$$\sum_{k=1}^{n_l} \left(x_{ilk} - \bar{x}_{il} \right) \left(\bar{x}_{il} - \bar{x}_{ilm} \right) = \left(\bar{x}_{il} - \bar{x}_{ilm} \right) \sum_{k=1}^{n_l} \left(x_{ilk} - \bar{x}_{il} \right) = 0 \tag{30}$$

$$\sum_{k=1}^{n_m} \left(x_{imk} - \bar{x}_{im} \right) \left(\bar{x}_{im} - \bar{x}_{ilm} \right) = \left(\bar{x}_{im} - \bar{x}_{ilm} \right) \sum_{k=1}^{n_l} \left(x_{imk} - \bar{x}_{im} \right) = 0 \tag{31}$$

$$
\begin{aligned}
\bar{x}_{il} - \bar{x}_{ilm} &= \frac{\sum_{k=1}^{n_l} x_{ilk}}{n_l} - \frac{\sum_{k=1}^{n_l} x_{ilk} + \sum_{k=1}^{n_m} x_{imk}}{n_l + n_m} \\
&= \frac{\left(n_l + n_m\right)\sum_{k=1}^{n_l} x_{ilk} - n_l\left[\sum_{k=1}^{n_l} x_{ilk} + \sum_{k=1}^{n_m} x_{imk}\right]}{n_l\left(n_l + n_m\right)} \\
&= \frac{n_m \sum_{k=1}^{n_l} x_{ilk} - n_l \sum_{k=1}^{n_m} x_{imk}}{n_l\left(n_l + n_m\right)} \\
&= \frac{n_m n_l \dfrac{\sum_{k=1}^{n_l} x_{ilk}}{n_l} - n_l n_m \dfrac{\sum_{k=1}^{n_m} x_{imk}}{n_m}}{n_l\left(n_l + n_m\right)} \\
&= \frac{n_m n_l \left(\bar{x}_{il} - \bar{x}_{im}\right)}{n_l\left(n_l + n_m\right)} \\
&= \frac{n_m \left(\bar{x}_{il} - \bar{x}_{im}\right)}{n_l + n_m}
\end{aligned}
\tag{32}
$$

同様に以下が成り立つ。

$$
\bar{x}_{im} - \bar{x}_{ilm} = \frac{n_l \left(\bar{x}_{im} - \bar{x}_{in}\right)}{n_l + n_m} \tag{33}
$$

したがって、以下を得る。

$$
\begin{aligned}
n_l\left(\bar{x}_{il} - \bar{x}_{ilm}\right)^2 + n_m\left(\bar{x}_{im} - \bar{x}_{ilm}\right)^2 &= n_l\left[\frac{n_m\left(\bar{x}_{il} - \bar{x}_{im}\right)}{n_l + n_m}\right]^2 + n_m\left[\frac{n_l\left(\bar{x}_{im} - \bar{x}_{in}\right)}{n_l + n_m}\right]^2 \\
&= \frac{\left(n_l n_m^2 + n_m n_n^2\right)\left(\bar{x}_{il} - \bar{x}_{im}\right)^2}{\left(n_l + n_m\right)^2} \\
&= \frac{n_l n_m}{n_l + n_m}\left(\bar{x}_{il} - \bar{x}_{im}\right)^2
\end{aligned}
\tag{34}
$$

したがって、 Eq. (26) は証明された。

距離は以下となる。

$$
d_{lm} = \sqrt{\Delta K_{lm}} \tag{35}
$$

9.10. ウォード法摘要の例

ここでは、実際にウォード法を具体的な例題に適用する。

メンバーの各項目で評価する場合、クラスター化は二通りある。

すなわち、メンバーをクラスター化する場合と、項目をクラスター化する場合である。

メンバーをクラスター化する場合は、似たもの同士のグループを選択することになる。

項目をクラスター化する場合は似た評価項目のグループを選択することになる。

ここでは、項目の類似度を評価する。

表 15 に CS データを示す。成績は. 5 点満点である。メンバー数を n と表記し、ここでは 20 である。ここでは、以下の 5 つの因子が顧客の満足度を左右するものとする。

a: Understanding customer

b: Prompt response

c: Flexible response

d: Producing ability

e:Providing useful information

f:Active proposal

20 人のデータからクラスターを構成する。

因子 a-f を k とし、メンバー1-20 を i で表す。

表 15　様々な項目の満足度

Member ID	a: Understanding customer	b:Prompt response	c:Flexible response	d: Producing ability	e:Providing uselul Informatio	f:Active proposal
1	4	4	3	4	5	4
2	4	3	4	3	3	2
3	3	1	2	3	2	1
4	5	5	5	5	5	5
5	4	4	4	4	4	4
6	2	2	2	2	2	2
7	5	5	5	5	5	5
8	4	3	3	3	3	3
9	2	1	1	1	1	2
10	2	1	1	1	1	3
11	4	5	5	5	5	5
12	3	3	3	2	2	2
13	3	3	3	3	3	3
14	3	4	4	3	3	3
15	4	3	3	3	3	3
16	1	1	1	1	1	1
17	5	5	5	5	5	5
18	4	1	1	1	3	3
19	4	4	3	3	4	3
20	4	4	4	3	3	3

Step1: 二乗和を評価する。

二乗和は以下のように評価される。

$$K_l = \sum_{i=1}^{n}\sum_{k=1}^{n_l}\left(x_{ilk} - \bar{x}_{il}\right)^2 \tag{36}$$

まず因子 a を考え、その二乗を評価する。

因子 a を考えているから、l は a である。ここで、またクラスターは構成していないから $k = a$ である。要素は一つであるから、平均もそのデータと同じである。つまり、以下である。

$$\bar{x}_{ia} = x_{ia} \tag{37}$$

したがって、その二乗和は以下である。

$$K_a = 0 \tag{38}$$

他の因子でも同じであるから以下が成り立つ。

$$K_a = K_b = \cdots = K_f = 0 \tag{39}$$

これはK_lの初期化になっている。

Step2:全ての組み合わせの二乗和を評価する。

全ての組み合わせは以下である。

a-b, a-b, a-d, a-e, a-f,a-e,a-f

b-c,b-d,b-e,b-f

c-d,c-e,c-f

d-e,d-f

e-f

この中でa と bの組み合わせを考える。この場合の二乗和の増加は以下となる

$$\Delta K_{ab} = \frac{n_a n_b}{n_a + n_b} \sum_{i=1}^{20} \left(\bar{x}_{ia} - \bar{x}_{ib} \right)^2 \tag{40}$$

ここで、 a と b は単因子であり、データ数は$n_a = n_b = 1$である。つまり、 $\bar{x}_{ia} = x_{ia}$ および $\bar{x}_{ib} = x_{ib}$である。その二乗和は以下である。

$$\sum_{i=1}^{20} \left(\bar{x}_{ia} - \bar{x}_{ib} \right)^2 = 20 \tag{41}$$

表 16　項目 a と b の二乗和の評価

Member ID	a:Understanding customer	b:Prompt response	Square of deviation
1	4	4	0
2	4	3	1
3	3	1	4
4	5	5	0
5	4	4	0
6	2	2	0
7	5	5	0
8	4	3	1
9	2	1	1
10	2	1	1
11	4	5	1
12	3	3	0
13	3	3	0
14	3	4	1
15	4	3	1
16	1	1	0
17	5	5	0
18	4	1	9
19	4	4	0
20	4	4	0
		Sum	20
		ΔK	10

したがって、このクラスター構成による二乗和の増分は以下となる。

$$\begin{aligned}\Delta K_{ab} &= \frac{n_a n_b}{n_a + n_b}\sum_{i=1}^{20}\left(\bar{x}_{ia} - \bar{x}_{ib}\right)^2 \\ &= \frac{1\times 1}{1+1}\times 20 \\ &= 10\end{aligned} \tag{42}$$

この結果を表 16 に示す。

他の組み合わせでも同様な解析をし、まとめたものが表 17 である。

表 17　二乗和の評価のまとめ

Item	a	b	c	d	e	f
a		10	9	9	5.5	8
b			2	4	4.5	7
c				3	6.5	9
d					3.5	7
e						4.5
f						

Step3: 最初のクラスターの構成

最小の 2 乗和の増加 ΔK_{bc} はクラスター b と c である。

対応する距離 d_{bc} は以下である。

$$\begin{aligned} d_{bc} &= \sqrt{\Delta K_{bc}} \\ &= \sqrt{2} \end{aligned} \tag{43}$$

クラスター bc と関連する二乗和は以下である。

$$\begin{aligned} K_{bc} &= K_b + K_c + \Delta K_{bc} \\ &= \Delta K_{bc} \\ &= 2 \end{aligned} \tag{44}$$

対応する因子数は以下である。

$$n_{bc} = 2 \tag{45}$$

クラスターの平均を評価し、表 18 を得る。

表 18　最初に構成したクラスター

Cluster 1	a	[bc]			d	e	f
Member ID	a: Understanding customer	b:Prompt response	c:Flexible response	[bc] mean	d: Producing ability	e:Providing uselul Informati	f:Active proposal
1	4	4	3	3.5	4	5	4
2	4	3	4	3.5	3	3	2
3	3	1	2	1.5	3	2	1
4	5	5	5	5	5	5	5
5	4	4	4	4	4	4	4
6	2	2	2	2	2	2	2
7	5	5	5	5	5	5	5
8	4	3	3	3	3	3	3
9	2	1	1	1	1	1	2
10	2	1	1	1	1	1	3
11	4	5	5	5	5	5	5
12	3	3	3	3	2	2	2
13	3	3	3	3	3	3	3
14	3	4	4	4	3	3	3
15	4	3	3	3	3	3	3
16	1	1	1	1	1	1	1
17	5	5	5	5	5	5	5
18	4	1	1	1	1	3	3
19	4	4	3	3.5	3	4	3
20	4	4	4	4	3	3	3

Step4: 次のステップの二乗和を評価する。

以下の組み合わせを考える。

a-[bc], a-d, a-e, a-f,a-e,a-f

[bc]-d,[bc]-e,[bc]-f

d-e,d-f

e-f

a とクラスター [bc] の二乗和の増分は以下となる。

$$\begin{aligned}\Delta K_{a[bc]} &= \frac{n_a n_{bc}}{n_a + n_{bc}} \sum_{i=1}^{20} \left(\bar{x}_{ia} - \bar{x}_{ibc} \right)^2 \\ &= \frac{1 \times 2}{1+2} \times 18 \\ &= 12\end{aligned} \tag{46}$$

対応するものを表 19 に示す。

表 19　項目 a とクラスターの二乗和の評価

Member ID	a: Understanding customer	[bc] mean	Square of deviation
1	4	3.5	0.25
2	4	3.5	0.25
3	3	1.5	2.25
4	5	5	0
5	4	4	0
6	2	2	0
7	5	5	0
8	4	3	1
9	2	1	1
10	2	1	1
11	4	5	1
12	3	3	0
13	3	3	0
14	3	4	1
15	4	3	1
16	1	1	0
17	5	5	0
18	4	1	9
19	4	3.5	0.25
20	4	4	0
		Sum	18
		ΔK	12

他の組み合わせに対しても同様な解析をし、表 20 に示す結果を得る。

表 20　分散の評価の第 2 ステップ

ID	a	bc	d	e	f
a		12	9	5.5	8
bc			4	6.67	10
d				3.5	7
e					4.5
f					

Step5: 2 回目のクラスター構成

最小の二乗和の増分は ΔK_{de} である。したがって、d と e でクラスアターを構成する。

対応する距離 d_{bc} 以下である。

$$\begin{aligned} d_{de} &= \sqrt{\Delta K_{de}} \\ &= \sqrt{3.5} \end{aligned} \tag{47}$$

クラスター de と関連する二乗和は以下となる。

$$\begin{aligned} K_{de} &= K_d + K_e + \Delta K_{de} \\ &= \Delta K_{de} \\ &= 3.5 \end{aligned} \tag{48}$$

対応する因子数は以下である。

$$n_{de} = 2 \tag{49}$$

9.11. まとめ

この章を纏める。

二つの因子間の距離は以下で与えられる。

$$d(A,B) = \sqrt{(x_{A1} - x_{B1})^2 + (x_{A2} - x_{B2})^2}$$

この距離を全ての因子間でおこない、最小のものを選択し、クラスターを構成する。そ

のクラスターの平均値を以下のように定義する。

$$x_{AB1}=\frac{x_{A1}+x_{B1}}{2}, x_{AB2}=\frac{x_{A2}+x_{B2}}{2}$$

上のデータを使い、同じプロセスを繰り返していく。

クラスター分析においては距離の評価が重要である。その距離の一般的な評価として以下がある。

$$d=\left[\frac{1}{n}\sum_{i=1}^{n}\left|x_{iA}-x_{iB}\right|^{\alpha}\right]^{\frac{1}{\alpha}}$$

もっと別の観点からクラスターを構成する方法がウォード法である。こちらのほうが鎖効果に強いと言われ、よく採用される。

ウォード法においては、二乗和の増分が最小になるようにクラスターを構成していく。

クラスター l とクラスター m からクラスター lm を構成することを考える。クラスター l とクラスター m はそれぞれデータ x_{ilk}，x_{imk} を持つと考える。データ数をそれぞれ n_l および n_m とする。すると、それぞれに二乗和は以下となる。

$$K_l=\sum_{i=1}^{n}\sum_{k=1}^{n_l}\left(x_{ilk}-\bar{x}_{il}\right)^2$$

$$K_m=\sum_{i=1}^{n}\sum_{k=1}^{n_m}\left(x_{imk}-\bar{x}_{im}\right)^2$$

ここで、二乗和の増分は以下となる。

$$\Delta K_{lm}=\frac{n_l n_m}{n_l+n_m}\sum_{i=1}^{n}\left(\bar{x}_{il}-\bar{x}_{im}\right)^2$$

これを全ての組み合わせに対して行い、その最小値を与えるものでクラスターを構成する。マージされた二乗和は以下で表現される。

$$K_{lm}=K_l+K_m+\Delta K_{lm}$$

上のプロセスを繰り返していく。