10. 非階層的クラスター分析（k-means 法）

概要: 前節では因子の似ている者同士を評価し、クラスターを構成した。これを階層的クラスター分析と呼ぶ。ここでは、最初からクラスターの数を指定し、データがどのクラスターに属するのかを評価する方法を示す。これをヒクラスター分析と呼ぶ。 この方法においては、各グループの平均を評価し、それをグループの特徴とする。各データがどのグループに近いかの距離を評価し、一番近いグループに属させる。その後、そのグループごとの平均を評価し直し、各データとの距離を再評価し、どのグループに属するかを決定する。この操作を繰り返し、データの移動、つまり各グループの平均の変化が無くなるまで続ける。

キーワード: クラスター分析;非階層的クラスター分析； k -means 法.

10.1. 序

前節では因子の似ている者同士を評価し、クラスターを構成した。ここでは、最初からクラスターの数を指定し、データがどのクラスターに属するのかを評価する方法を示す。

10.2. k-means 法

ここでは k -means 法を呼ばれる手法を扱う。この方法では、まずクラスターの数を設定する。このため、この方法のことを非階層的クラスター分析と呼ぶ。そして、平均値の初期値を設定する。ここでは、クラスターの数を k とする。

我々は p 種類の因子を考える。

初期値として、我々はそれぞれのクラスターのそれぞれの因子の平均値を以下のように設定する。

$$\mu_{1\_C_1}, \mu_{2\_C_1}, \cdots, \mu_{p\_C_1} \tag{1}$$

$$\mu_{1\_C_2}, \mu_{2\_C_2}, \cdots, \mu_{p\_C_2} \tag{2}$$

…

$$\mu_{1\_C_k}, \mu_{2\_C_k}, \cdots, \mu_{p\_C_k} \tag{3}$$

ここで、 $\mu_{i\_C_j}$ は i 因子のクラスター C_j における平均である。これらの平均値は仮のものであり、後に変更される。

上の平均値を使い、あるメンバーのデータ

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix} \tag{4}$$

を評価する。

このデータと $\mathbf{x}_i$ クラスター j の距離 $D^2_{C_j}[\mathbf{x}_i]$ として以下を定義する。

$$D^2_{C_1}[\boldsymbol{x}_i] = \left(x_{i1} - \mu_{1\_C_1}\right)^2 + \left(x_{i2} - \mu_{2\_C_1}\right)^2 + \cdots + \left(x_{ip} - \mu_{p\_C_1}\right)^2 \tag{5}$$

$$D^2_{C_2}[\boldsymbol{x}_i] = \left(x_{i1} - \mu_{1\_C_2}\right)^2 + \left(x_{i2} - \mu_{2\_C_2}\right)^2 + \cdots + \left(x_{ip} - \mu_{p\_C_2}\right)^2 \tag{6}$$

…

$$D^2_{C_k}[\boldsymbol{x}_i] = \left(x_{i1} - \mu_{1\_C_k}\right)^2 + \left(x_{i2} - \mu_{2\_C_k}\right)^2 + \cdots + \left(x_{ip} - \mu_{p\_C_k}\right)^2 \tag{7}$$

この中の最小のものを選択する。

$$Min\left[D^2_{C_1}[\boldsymbol{x}_i], D^2_{C_2}[\boldsymbol{x}_i], \cdots, D^2_{C_k}[\boldsymbol{x}_i]\right] = D^2_{C_\varsigma}[\boldsymbol{x}_i] \tag{8}$$

これにより、このデータは距離の二乗の最小値を与えるクラスターςに属すると判断する。この解析を全ての新しいデータに対して行う。すると、各データはどれかのクラスターに属することになる。

ここで、我々は各クラスターの平均値を再度評価する。

$$\mu_{1\_C_1} = \frac{\sum_{i=1}^{n_{C_1}} x_{i1}}{n_{C_1}}, \mu_{2\_C_1} = \frac{\sum_{i=1}^{n_{C_1}} x_{i2}}{n_{C_1}}, \cdots, \mu_{p\_C_1} = \frac{\sum_{i=1}^{n_{C_1}} x_{ip}}{n_{C_1}} \quad for\ \boldsymbol{x}_i \in C_1 \tag{9}$$

$$\mu_{1\_C_2} = \frac{\sum_{i=1}^{n_{C_2}} x_{i1}}{n_{C_2}}, \mu_{2\_C_2} = \frac{\sum_{i=1}^{n_{C_2}} x_{i2}}{n_{C_2}}, \cdots, \mu_{p\_C_2} = \frac{\sum_{i=1}^{n_{C_2}} x_{ip}}{n_{C_2}} \quad for\ \boldsymbol{x}_i \in C_2 \tag{10}$$

…

$$\mu_{1\_C_k} = \frac{\sum_{i=1}^{n_{C_k}} x_{i1}}{n_{C_k}}, \mu_{2\_C_k} = \frac{\sum_{i=1}^{n_{C_k}} x_{i2}}{n_{C_k}}, \cdots, \mu_{p\_C_k} = \frac{\sum_{i=1}^{n_{C_k}} x_{ip}}{n_{C_k}} \quad for\ \boldsymbol{x}_i \in C_k \tag{11}$$

ここで、n_{C_ς} はクラスターςのデータ数である。

この新たな平均値を使い、我々は全てのデータを評価し直す。このプロセスをすべての平均値が変動しなくなるまで繰り返す。 つまり、所属を変更するデータがなくなるまで繰り返す。

このプロセスを終えると、我々は最終的な k 個のクラスターを得る。

10.3. k-means 法の初期設定

この方法には、初期のグループ数を決定することと、そのグループの平均の設定に任意性がある。

一つは、データをランダムに振り分ける方法である。これは、あるデータに対して乱数を発生させてどれかのグループに属させることをする。その後にクラスターを構成をする。

一つは、項目に対してある一定の値を k 個設定し、それに対して各データを振り分ける方法である。この場合は、一定の値は最後までその値のままである。

もう一つは、初期の比較的少ないデータで階層的クラスター分析をおこない、クラスター数、平均値を設定する。その後は、これまで示してきた k-means 法を適用する。

10.4. まとめ

この章のまとめを行う。

p 個の因子があるとする。ここで、k 個のクラスターがあると設定する。そして、クラスターςの仮の平均を以下のように設定する。

$$\mu_{1\_C_\varsigma}, \mu_{2\_C_\varsigma}, \cdots, \mu_{p\_C_\varsigma}$$

ここで $\varsigma = 1,2,\cdots,k$ である。

これらを使い、各データとクラスターの距離の二乗を以下のように評価する。

$$D^2_{C_\varsigma}[\boldsymbol{x}_i] = \left(x_{i1} - \mu_{1\_C_\varsigma}\right)^2 + \left(x_{i2} - \mu_{2\_C_\varsigma}\right)^2 + \cdots + \left(x_{ip} - \mu_{p\_C_\varsigma}\right)^2$$

この距離が最小のものをそのデータが属するクラスターとする。これを全てのデータで繰り返す。その後、平均値を以下のように再評価する。

$$\mu_{1\_C_\varsigma} = \frac{\sum_{i=1}^{n_{C_\varsigma}} x_{i1}}{n_{C_\varsigma}}, \mu_{2\_C_1} = \frac{\sum_{i=1}^{n_{C_\varsigma}} x_{i2}}{n_{C_\varsigma}}, \cdots, \mu_{p\_C_\varsigma} = \frac{\sum_{i=1}^{n_{C_\varsigma}} x_{ip}}{n_{C_\varsigma}} \quad for\ \boldsymbol{x}_i \in C_\varsigma$$

ここで、n_{C_ς} はクラスターςのデータ数である。

このプロセスを平均値が変動しなくなるまで繰り返す。最終的には、我々はk 個のクラスターを得る。

この方法の初期設定に、3 種類の方法を示した。