CLUSTER ANALYSIS

Diamonds Hearts Spades Clubs

1 Introduction

Reference: Chapter 12, Johnson and Wichern

Clustering methods were developed to detect inherent structure of the data.

There number of clustering methods available to group multivariate data.

K
Q
J
A
1
2
⋮

Assumption :

The only assumption required is number of inherent clusters in the data. That is, assumption on distribution is not required to apply any clustering method. Therefore, it is a distribution free (non-parametric) procedure.

2 Minimum Distance Clustering

Let $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})^T (i = 1, 2, \ldots, n)$ and $\boldsymbol{x}_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,p})^T (j = 1, 2, \ldots, n)$ be two p-dimensional observations.

Definitions:

Straight line

(a) Euclidean Distance

$$\begin{aligned} D_{i,j} &= ||\boldsymbol{x}_i - \boldsymbol{x}_j|| = \left\{ (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \right\}^{1/2} \\ &= \left\{ \sum_{s=1}^{p} (x_{i,s} - x_{j,s})^2 \right\}^{1/2} \end{aligned}$$

$\sqrt{(x-y)'(x-y)}$

(b) L_l Distance

$$L_{i,j} = |\boldsymbol{x}_i - \boldsymbol{x}_j| = \sum_{s=1}^{p} |x_{i,s} - x_{j,s}|$$

Now group the observations so that the total distance (Euclidean or L_1) within the group is minimized.

L_1 distance is computationally faster to determine but it is less accurate than the Euclidean distance method.

3 Clustering Quality Indicator

mean of cluster (centre)

A common clustering criterion or quality indicator is the sum of squares of errors

$$\begin{aligned} \text{SSE} &= \textstyle\sum_{r=1}^{k} \sum_{\boldsymbol{x}_t \in C_r} ||\boldsymbol{x}_t - \boldsymbol{\mu}_r||^2 \\ &= \textstyle\sum_{r=1}^{k} \sum_{\boldsymbol{x}_t \in C_r} (\boldsymbol{x}_t - \boldsymbol{\mu}_r)^T (\boldsymbol{x}_t - \boldsymbol{\mu}_r) \\ &= \textstyle\sum_{r=1}^{k} \sum_{\boldsymbol{x}_t \in C_r} \sum_{s=1}^{p} (x_{t,s} - \mu_{r,s})^2 \end{aligned}$$

Where $\boldsymbol{\mu}_r = (\mu_{r,1}, \mu_{r,2}, \ldots, \mu_{r,p})$ is the mean of cluster $C_r (r = 1, \ldots, k)$ and $\boldsymbol{x}_t \in C_r$ implies that observation $\boldsymbol{x}_t$ is assigned to cluster C_r.

number of clusters

4 Migrating Means Algorithm (K Mean Method)

Step 1: Assume there are k clusters in the multivariate space. Select k different points arbitrary, say, $\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_k$ from the space to initialized the k means of the clusters. Or divide $\boldsymbol{x}_i$'s into k clusters arbitrary and compute the average of $\boldsymbol{x}_i$'s in C_r and let $\boldsymbol{m}_r$ be the average of $\boldsymbol{x}_i$'s in C_r.

clusters centre

Step 2: Compute EITHER Euclidean distance, $D_{i,j}$ or L_1 distance where

$$\begin{aligned} D_{i,j} &= ||\boldsymbol{x}_i - \boldsymbol{m}_j|| = \left\{ (\boldsymbol{x}_i - \boldsymbol{m}_j)^T (\boldsymbol{x}_i - \boldsymbol{m}_j) \right\}^{1/2} \\ &= \left\{ \textstyle\sum_{s=1}^{p} (x_{i,s} - m_{j,s})^2 \right\}^{1/2} \end{aligned}$$

and

$$L_{ij} = |\boldsymbol{x}_i - \boldsymbol{m}_j| = \sum_{s=1}^{p} |x_{i,s} - m_{j,s}|$$

where $\boldsymbol{m}_j = (m_{j,1}, m_{j,2}, \ldots, m_{j,p})^T$.

Step 3: Assign the observation, $\boldsymbol{x}_i$ to the nearest cluster for all i.

That is, assign $\boldsymbol{x}_i$ to C_r if $D_{i,r} = \min_j(D_{i,j})$ or for L_1 distance measure $L_{i,r} = \min_j(L_{i,j})$.

Step 4: Compute the new mean for clusters using the assigned observations. Let $\widehat{\boldsymbol{\mu}}_r = \frac{1}{n_r} \sum_{\boldsymbol{x}_t \in C_r} \boldsymbol{x}_t$ where n_r is the number of observations assigned to C_r.

If $\widehat{\boldsymbol{\mu}}_r = \boldsymbol{m}_r$ for all r then the assignment of observations to clusters are completed and **STOP** the procedure. Otherwise, set $\boldsymbol{m}_r = \widehat{\boldsymbol{\mu}}_r$ and GOTO Step 2.

Example: Assign the following 4 observations into two groups:

k=2
n=4

k=2
n=4
p=2

Example: Assign the following 4 observations into two groups:

$$A=\begin{pmatrix} 5 \\ 3 \end{pmatrix},\ B=\begin{pmatrix} -1 \\ 1 \end{pmatrix},\ C=\begin{pmatrix} 1 \\ -2 \end{pmatrix} \text{ and } D=\begin{pmatrix} -3 \\ -2 \end{pmatrix}.$$

5 Hierarchical Clustering

This method does not require the user to specify the number of clusters. It produces an output that allows the user to decide the natural clusters in the population.

Initially, hierarchical clustering method assumes that all objects are individual clusters.

Next merges the neighbouring pair of clusters by checking the Distance Similarity matrix. Continue this until all objects appear in a single cluster.

Procedure

obs or variables

Step 1: Let $\boxed{N}$ be the number of objects in the multivariate population and $\mathcal{D} = (d_{i,j})_{\boxed{N\times N}}$ be the distance matrix (dissimilarity matrix), where $d_{i,j}$ is the distance (dissimilarity) between the objects (or observations) i and j.

distance matrix

Note that $d_{i,j}$ can be an Euclidean distance, L_1 distance or $\boxed{1 - r^2_{i,j}}$, where $r_{i,j}$ is the sample correlation between i and j, or any other similar quantity.

correlation

Step 2: Search the distance matrix for the most similar (nearest) pair of clusters. Let $\boxed{u}$ and $\boxed{v}$ be the most similar clusters(objects) and $\boxed{uv}$ be the new cluster formed by merging u and v. — new cluster

Step 3: Delete the rows and columns of the distance matrix corresponding to clusters u and v and add a row and column giving the distances between cluster $\boxed{uv}$ and the remaining clusters.

Repeat the Steps 2 and 3 a total of $N - 1$ times. Then all objects will be in a single cluster. Record the identity of clusters that are merged and the levels and present the results graphically in the form of a dendogram.

<u>**Linkage Methods**</u>

That is, methods of combining two or more objects to form a cluster. Commonly used linkage method are:

- Single Linkage — smallest distance
- Complete Linkage — largest distance
- Average Linkage — average distance

Single Linkage (Nearest Neighbour) Method

Find the smallest distance in $\boxed{\mathcal{D}}$ (matrix) and merge the corresponding objects, say u and v, to get the cluster $\boxed{uv}$. For the above Step 3, the distances between uv and another cluster(or object) w (say), is computed by

$$\boxed{d_{uv,w}} = \min(d_{u,w}, d_{v,w})$$

comparing distances

The other linkage methods proceeds in much the same way as single linkage, with one exception. That is, the computation of the distance between newly form cluster and other clusters(or objects).

Note that

- For **complete Linkage (Furthest neighbour) Method:**

$$d_{uv,w} = \boxed{\max}(d_{u,w}, d_{v,w})$$

- For **Average Linkage Method:**

$$d_{uv,w} = \frac{1}{\boxed{N_{uv}}\boxed{N_w}} \sum_i \sum_j d_{i,j}$$

where $d_{i,j}$ is the distance between object i in cluster uv and object j in cluster w and, $\boxed{N_{uv}}$ and $\boxed{N_w}$ are the number of objects in clusters uv and w respectively.

Example: The distances between pairs of five item are given below:

	1	2	3	4	5
1	0				
2	4	0			
3	6	9	0		
4	1	7	10	0	
5	6	3	5	8	0

Cluster the five items using the

(a) Single linkage

(b) Complete linkage and

(c) Average linkage

methods. Draw a dendograms.

Distance Measures for Binary Variables Present / Absent

Let $\boxed{x_{i,k}}$ be the score (1 or 0) of binary variable X_k on i^{th} object $\boldsymbol{x}_i$ and $\boxed{x_{j,k}}$ be the score (1 or 0) of X_k on j^{th} object for $k = 1, 2, \ldots, p$.

Let

$\boxed{a} = \boxed{(1-1)}$ matches between object i and j, both present

mis match $\boxed{b} = \boxed{(1-0)}$ mismatches between object i and j,

$\boxed{c} = \boxed{(0-1)}$ mismatches between object i and j, and

$\boxed{d} = \boxed{(0-0)}$ matches between object i and j. both absent

Then, $a + b + c + d = p$.

Example: Certain characteristics associated with six workers in a factory are listed below:

Person	Married	School	Sex
1	yes	Private	Female
2	no	state	Female
3	yes	state	Male
4	yes	state	Male
5	no	Private	Female
6	no	Private	Male

Draw a dendograms and identify similar groups (clusters) among these persons.

The following tables (Table 12.1, Johnson and Wichern, p.675) lists commonly use similarity coefficients for bivariate data.

Table: Similarity Coefficients For Clustering

Binary data

Coefficient	Rationale
1. $\boxed{\frac{a+d}{p}}$	Equal weights for 1-1 matches and 0-0 matches.
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1-1 matches and 0-0 matches.
3. $\frac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\frac{a}{p}$	No 0-0 matches in numerator.
5. $\frac{a}{a+b+c}$	No 0-0 matches in numerator or denominator. The 0-0 matches are treated as irrelevant.
6. $\frac{2a}{2a+b+c}$	No 0-0 matches in numerator or denominator. Double weight for 1-1 matches.
7. $\frac{a}{a+2(b+c)}$	No 0-0 matches in numerator or denominator. Double weight for unmatched pairs.
8. $\frac{a}{b+c}$	Ratio of matches to mismatches with 0-0 matches excluded.

no about match

Note that the similarity coefficients 1, 2 and 3 are monotonic.

Example 1

$k=2$ (two groups)

arbitrary assign obs to groups

cluster 1 $A=\begin{pmatrix}5\\3\end{pmatrix}$ $B=\begin{pmatrix}-1\\1\end{pmatrix}$

cluster 2 $C=\begin{pmatrix}1\\-2\end{pmatrix}$ $D=\begin{pmatrix}-3\\-2\end{pmatrix}$

Find mean of cluster (centre)

	$\bar{x}_1$	$\bar{x}_2$	
cluster 1	$\frac{5+-1}{2}=2$	$\frac{3+1}{2}=2$	$(2,2)$
cluster 2	$\frac{1+-3}{2}=-1$	$\frac{-2+-2}{2}=-2$	$(-1,-2)$

Calculate distances for A

$d^2(A,(A'B)) = (5-2)^2 + (3-2)^2 = \boxed{10}$

$d^2(A,(CD)) = (5--1)^2 + (3--2)^2 = 61$

If we move A to cluster 2

cluster$_{adj}$ 1 mean $(-1,1)$ (B only)

cluster$_{adj}$ 2 mean $\frac{2\times -1+5}{3}=1$ $\frac{2\times -2+3}{3}=-0.33$

$(1,-0.33)$

$d^2(A,(B)) = (5--1)^2 + (3-1)^2 = 40$

$d^2(A,(ACD)) = (5-1)^2 + (3--0.33)^2 = \boxed{27.1}$

A is closer to AB than to ACD

consider B $\rightarrow$ cluster 2

originally

$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = \boxed{10}$

$d^2(B,(CD)) = (-1--1)^2 + (1+2)^2 = 9$

with B in cluster 2

new cluster 2 $(-1,-1)$

$d^2(B,(A)) = 40$

$d^2(B,(BCD)) = (-1 - -1)^2 + (1 - -1)^2 = \boxed{4}$

move B to cluster 2

check C

$d^2(C,(A)) = (1-5)^2 + (-2-3)^2 = 41$

$d^2(C,(BCD)) = (1+1)^2 + (-2+1)^2 = \boxed{5}$

if C was in cluster 1

center $(3, 0.5)$

$d^2(C,(AC)) = (1-3)^2 + (-2-0.5)^2 = \boxed{10.25}$

center $(-2, -0.5)$

$d^2(C,(BD)) = (1 - -2)^2 + (-2 - -0.5)^2 = 11.25$

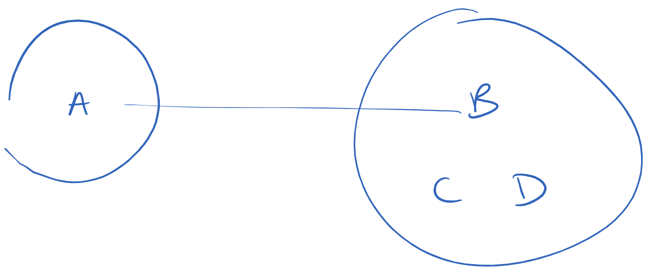
C is closer to CBD cluster 2 than AC

$\rightarrow$ leave C in cluster 2

eventually

		A	B	C	D
cluster 1	A	0	40	41	89
cluster 2	BCD	52	4	5	5

squared distance to centre



Repeat with different allocation to clusters at the start

Example 12.3

Distance matrix for 5 objects (variables or observations)

$$D = d_{ik} = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 9 & 0 & & & \\ 3 & 3 & 7 & 0 & & \\ 4 & 6 & 5 & 9 & 0 & \\ 5 & 11 & 10 & 2 & 8 & 0 \end{array}$$

symmetrical

Use single linkage (smallest distance)

$d_{35} = 2$ smallest distance $\therefore$ first cluster

create cluster (35)

$$d_{(35),1} = \min(d_{3,1}, d_{5,1}) = \min(3, 11) = 3$$

$$d_{(35),2} = \min(d_{3,2}, d_{5,2}) = \min(7, 10) = 7$$

$$d_{(35),4} = \min(d_{3,4}, d_{5,4}) = \min(9, 8) = 8$$

construct new D matrix $(5-1 \times 5-1)$
(4×4)

$$D = \begin{array}{c|cccc} & (35) & 1 & 2 & 4 \\ \hline (35) & 0 & & & \\ 1 & 3 & 0 & & \\ 2 & 7 & 9 & 0 & \\ 4 & 8 & 6 & 5 & 0 \end{array}$$

new cluster (135)

$$d_{(135),2} = \min(7, 9) = 7$$

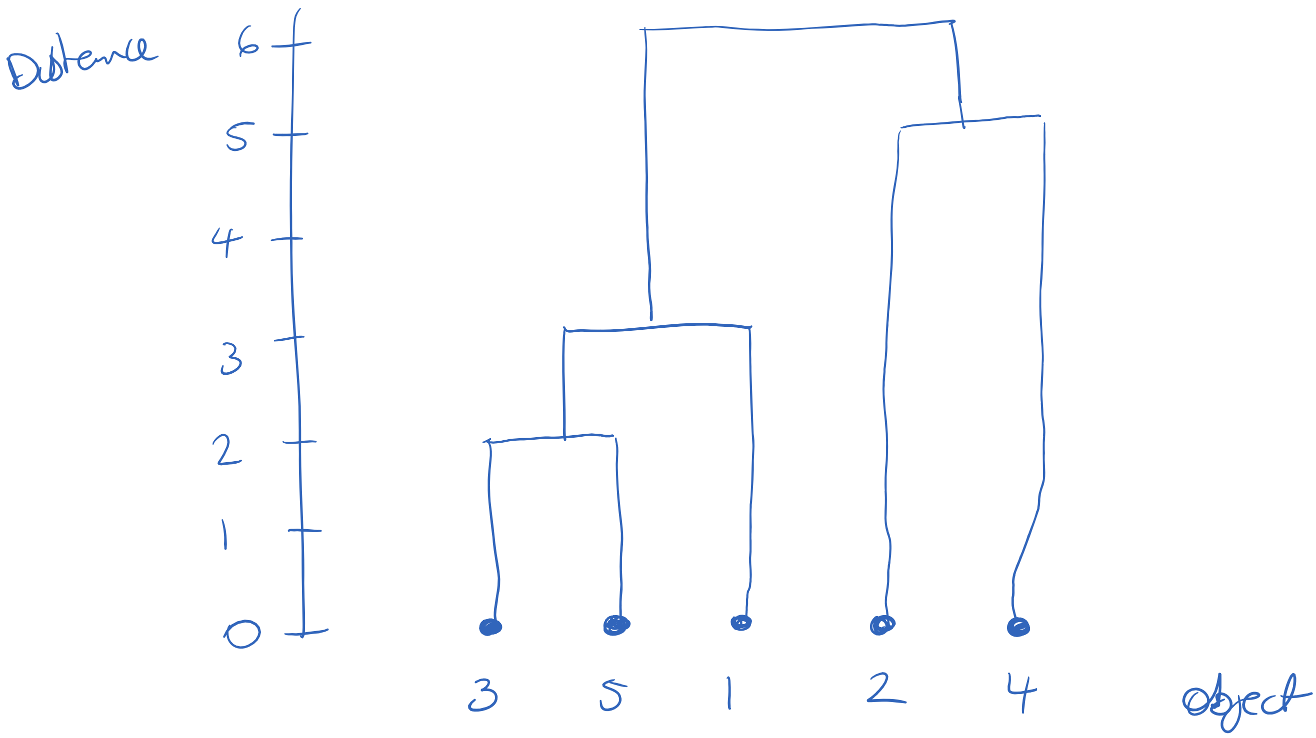
$$d_{(135),4} = \min(8, 6) = 6$$

$$D_{3\times3} = \begin{array}{c|ccc} & (135) & 2 & 4 \\ \hline (135) & 0 & & \\ 2 & 7 & 0 & \\ 4 & 6 & 5 & 0 \end{array}$$

$$d_{(24),(135)} = \min(7, 6) = 6$$

$$D_{2\times2} \begin{array}{c|cc} (135) & 0 & \\ (24) & 6 & 0 \end{array}$$

Construct a dendrogram



Example 12.5

Using complete linkage (largest distance)

$$D = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

first cluster (35)

$$d_{(35),1} = \max(d_{3,1}, d_{5,1}) = \max(3, 11) = 11$$

$$d_{(35),2} = \max(d_{3,2}, d_{5,2}) = \max(7, 10) = 10$$

$$d_{(35),4} = \max(d_{3,4}\ d_{5,4}) = \max(8, 9) = 9$$

construct $D_{4\times 4}$

$$D = \begin{array}{c} (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

$$d_{(24)(35)} = \max(10, 9) = 10$$

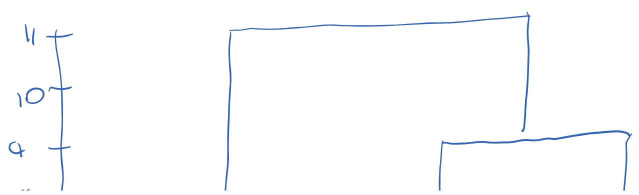
$$d_{(24),1} = \max(9, 6) = 9$$

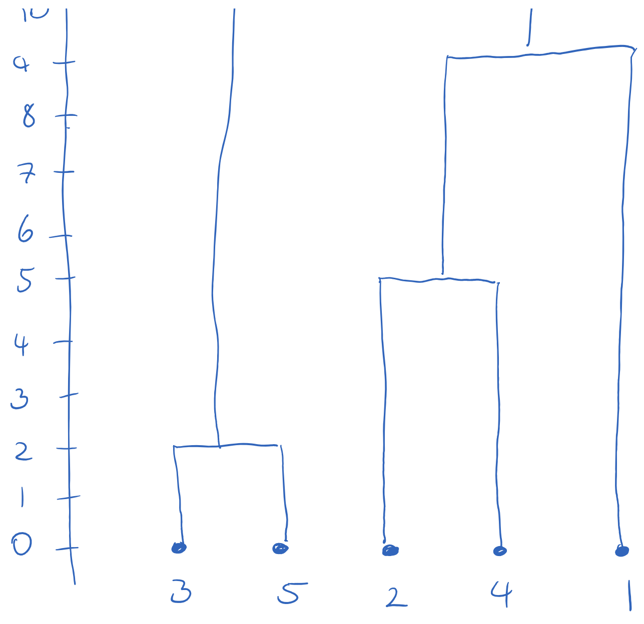
$$D_{3\times 3} = \begin{array}{c} (35) \\ (24) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & 9 & 0 \end{bmatrix} \end{array}$$

$$d_{(124)(35)} = \max(11, 10) = 11$$

$$D_{2\times 2} = \begin{array}{c} (35) \\ (124) \end{array} \begin{bmatrix} 0 & \\ 11 & 0 \end{bmatrix}$$

Construct dendrogram





objects by complete linkage

Example 12.1

Binary variables 6

X_1 height 1 > 180cm

X_2 weight 1 > 67.5kg

X_3 eye colour 1 = brown

X_4 hair colour 1 = blonde

X_5 handed 1 = right

X_6 gender 1 = female

Use for similarity of individuals

	X_1	X_2	X_3	X_4	X_5	X_6
Individual 1	0	0	0	1	1	1
Individual 2	1	1	1	0	1	0

summarize matches / mismatches

Ind 1 \ Ind 2	1	0	
1	a = 1	b = 2	3
0	c = 3	d = 0	3
	4	2	6

6 variables

$a+b+c+d=p$ total no. of binary variables

equal weighting of presence & absence

X_3 = eye colour

1 = brown

0 = { blue, grey, green, hazel ... }

similarity $\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}$

construct a similarity matrix (say 5 individuals)

$$S = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 1 & & & & \\ 2 & 1/6 & 1 & & & \\ 3 & 4/6 & 3/6 & 1 & & \\ 4 & 4/6 & 3/6 & 2/6 & 1 & \\ 5 & 0 & 5/6 & 2/6 & 2/6 & 1 \end{array}$$

symmetrical

most distinct

most similar

$$d_{ik} = \sqrt{2(1-\tilde{s}_{ik})}$$