Cluster Analysis

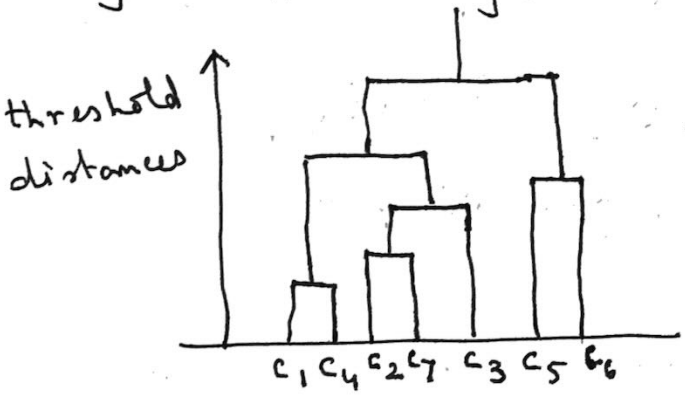
Grouping of objects (with multidimensional feature vectors) based on self similarities ("closeness").

Two approaches

- Hierarchical clustering
- non-hierarchical clustering

Hierarchical clustering

In HCA the observation vectors (cases) are grouped together on the basis of mutual distances in such a manner so as to ensure hierarchy in cluster formation. HCA output is usually visualized through a hierarchical tree, which is a nested set of partitions represented by a tree diagram called dendogram.



Dendogram tree

Characteristics

(i) Sectioning a tree at a particular level produces a partition into g ~~gr~~ disjoint groups

(ii) If two groups are chosen from different partitions (i.e. the results of partitioning at different levels) then either the groups are disjoint or one group contains the other

(iii) a numerical value is associated with each partition up/down the tree where branches join. This is a measure of distance between 2 merged clusters.

(iv) We get different trees corresponding to different distance measures (bet^n cases).

Two algorithms for HCA

- Agglomerative clustering algorithm
- Divisive clustering algorithm

Agglomerative clustering algorithm

- Operates with successive merger of objects
- Starts with n clusters, each containing a single data point.
- At each stage merges the 2 most similar groups to form a new cluster, thus reducing the number of clusters by 1.
- The process of merger of group of objects continue till all subgroups are fused together to form a single cluster

Divisive clustering algorithm

- Operates by successively splitting groups of objects
- starts with a single cluster having all the n objects
- Initial group split into 2 clusters ⇒ the distance bet^n the split groups ~~are~~ ~~now~~ is maximum, i.e the subgroups are as far as possible.

- Process of splitting continues till there are n clusters, each having a single object (leaf node)
- DCA is computationally inefficient.

Note: Result from agglomerative (or divisive) clustering algorithm displayed through dendogram.

Steps in agglomerative hierarchical algorithm

Step I : n clusters, each having a single object and an $n \times n$ symmetric matrix of distances ($D = (\!(d_{ij})\!)$)

Step II : Search the distance matrix for the nearest (most similar) pair of clusters Let the distance bet^n the most similar clusters, say U and V, be d_{UV}.

Step III : Merge U & V to form cluster (U, V), at a merger (fusion) level d_{UV}.
Update the distance matrix by
(i) deleting the rows & col^{m}s corresponding to clusters U & V
and (ii) adding a row & col^m giving the distance bet^n (U, V) and the remaining clusters (existing ones other than U & V).

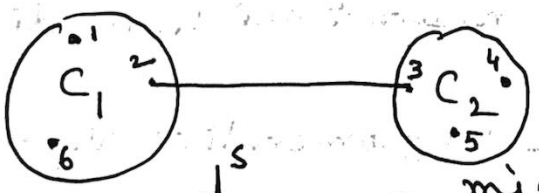
Step IV : Repeat (II) & (III) n-1 times, recording the identity of the clusters that are merged and the merger levels (i.e the

distances at which mergers take place).

Step 2 : Construct the dendogram tree from the information on mergers and level of mergers.

Remark: step iii requires calculation of distance bet^n clusters (for distance matrix updation). The following are commonly used distance measures bet^n clusters of objects

(i) Single link distance (minimum distance or nearest neighbor distance)



$$d^s_{C_1,C_2} = \min_{\substack{i \in C_1 \\ j \in C_2}} d_{ij}$$

(ii) Complete link distance (maximum distance or farthest distance)

$$d^c_{C_1,C_2} = \max_{\substack{i \in C_1 \\ j \in C_2}} d_{ij}$$

(iii) Average link distance

N_{C_1} : # of objects in C_1

N_{C_2} : # of objects in C_2

$$d^A_{C_1,C_2} = \frac{1}{N_{C_1} N_{C_2}} \sum_{\substack{i \in C_1 \\ j \in C_2}} d_{ij}$$

(iv) Centroid distance : Distance bet^n cluster means

(v) Median distance : Distance bet^n cluster medians

Remark : Single linkage agglomerative algorithm

In step (iii) (distance matrix updation stage), we compute the distance bet^n (U, V) and any existing cluster W as

$$d_{(U,V),W} = \min(d_{UW}, d_{VW})$$

(d_{UW}, d_{VW} present in prev round)

d_{UW} & d_{VW} are nearest neighbor distances

Remark : Complete linkage agglomerative algorithm

In step (iii), compute distance bet^n (U, V) and any existing W as

$$d_{(U,V),W} = \max(d_{UW}, d_{VW})$$

Remark : Average linkage agglomerative algorithm

In step (iii), compute distance bet^n (U, V) and W computed as

$$d_{(U,V),W} = \frac{\sum_{i \in (U,V)} \sum_{K \in W} d_{iK}}{N_{(U,V)} \, N_W}$$

d_{iK} : distance bet^n object i in the cluster (U, V) and object K in cluster W

$N_{(U,V)}$: # of objects in (U, V) & N_W : # of objects in W

Example: Single linkage clustering

Five objects c_1, c_2, c_3, c_4, c_5

Distance matrix

$$D = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array} \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix} \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array}$$

$d_{c_3 c_5} = \min d_{ij} = 2$

$\Rightarrow$ Merge c_3 & c_5 to form cluster (c_3, c_5) at merger level 2 – total 4 clusters

Row col^m deletion containing c_3 & c_5

$$\begin{array}{c} \\ c_3 \\ \\ c_5 \end{array} \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix} \begin{array}{c} \\ \\ d_3 \\ \\ c_5 \end{array}$$

Computation of $d_{(c_3, c_5), c_i}$ for $i = 1, 2, 4$

$$d_{(c_3, c_5) c_1} = \min(d_{c_1 c_3}, d_{c_1 c_5}) = \min(3, 11) = 3$$

$$d_{(c_3, c_5) c_2} = \min(d_{c_2 c_3}, d_{c_2 c_5}) = \min(7, 10) = 7$$

$$d_{(c_3, c_5) c_4} = \min(d_{c_3 c_4}, d_{c_4 c_5}) = \min(9, 8) = 8.$$

Updated distance matrix

$$\begin{array}{c} \\ (C_3, C_5) \\ C_1 \\ C_2 \\ C_4 \end{array} \begin{pmatrix} 0 & & & \\ \boxed{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{pmatrix}$$

$d_{(C_3, C_5) C_1}$ is min

Merge (C_3, C_5) with C_1 to form cluster (C_1, C_3, C_5) at merger level 3 – total 3 clusters

Row column deletion step

$$\begin{array}{c} \\ (C_3, C_5) \\ C_1 \\ C_2 \\ C_4 \end{array} \begin{pmatrix} 0 & \not{7} & - & - \\ \not{3} & \not{9} & - & - \\ 7 & 9 & 0 & \\ \not{8} & 6 & 5 & 0 \end{pmatrix}$$

Calculate distance bet^n (C_3, C_5, C_1) & C_2 and (C_3, C_5, C_1) & C_4

$$d_{(C_1, C_3, C_5) C_2} = \min\left(d_{C_1 C_2}, d_{(C_3, C_5) C_2}\right)$$

$$= \min(9, 7) = 7$$

$$d_{(C_1, C_3, C_5) C_4} = \min\left(d_{C_1 C_4}, d_{(C_3, C_5)(C_4)}\right)$$

$$= \min(6, 8) = 6$$

Update distance matrix

$$\begin{matrix}(C_1, C_3, C_5) \\ C_2 \\ C_4\end{matrix}\begin{pmatrix} 0 & & \\ 7 & 0 & \\ 6 & \boxed{5} & 0 \end{pmatrix}$$

$d_{C_2 C_4}$ is min

Merge C_2 & C_4 to form cluster (C_2, C_4) at merger level 5 – total 2 clusters (C_1, C_3, C_5) & (C_2, C_4)

Row column deletion

$$\begin{matrix}(C_1, C_3, C_5) \\ C_2 \\ C_4\end{matrix}\begin{pmatrix} 0 & | & | \\ 7 & 0 & | \\ 6 & 5 & 0 \end{pmatrix}$$

Distance matrix updation

$$\begin{matrix}(C_1, C_3, C_5) \\ (C_2, C_4)\end{matrix}\begin{pmatrix} 0 & \\ ? & 0 \end{pmatrix}$$

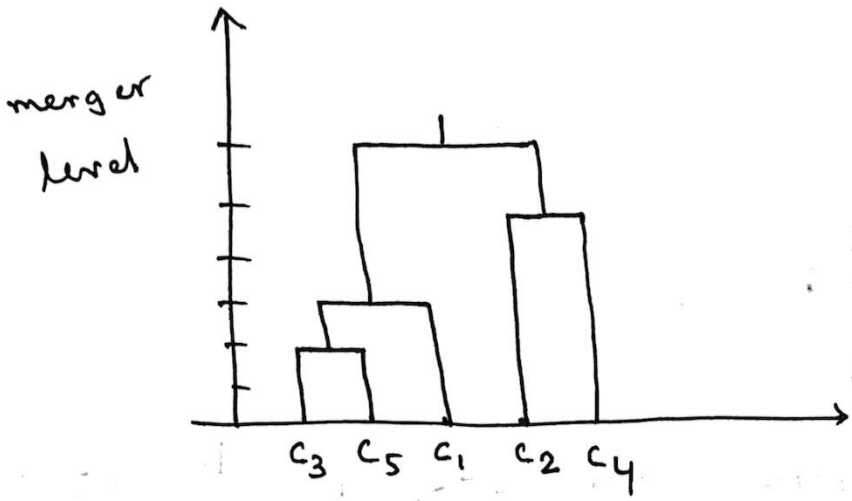
$$d_{(C_1, C_3, C_5)(C_2, C_4)} = \min\left(d_{(C_1, C_3, C_5) C_2}, d_{(C_1, C_3, C_5) C_4}\right)$$

$$= \min(7, 6) = 6$$

$$\begin{matrix}(C_1, C_3, C_5) \\ (C_2, C_4)\end{matrix}\begin{pmatrix} 0 & \\ 6 & 0 \end{pmatrix}$$

⇒ Merge (C_1, C_3, C_5) & (C_2, C_4) at merger level 6 – single cluster

Dendogram



Measures of dissimilarity

$\underset{p\times n}{\mathcal{X}} = (\underset{\sim}{x}_1 : \cdots : \underset{\sim}{x}_n)$

$\mathcal{X} \to \underset{p\times p}{D}$: dissimilarity matrix

$D = ((d_{rs}))$ ∋

(i) $d_{rs} \geq 0 \quad \forall\, r, s$

(ii) $d_{rr} = 0 \quad \forall\, r$

(iii) $d_{rs} = d_{sr} \quad \forall\, r, s$

Remark: (i) Symmetry cond^n may not always be satisfied

$D \to (D + D')/2$

(ii) If in addition to the cond^{n}s (i)–(iii), d_{rs} also satisfies triangle inequality

$d_{rs} \leq d_{rt} + d_{ts} \quad \forall\, r, s, t$

then the dissimilarity measure is a metric and we use the term "distance"

Some commonly used dissimilarity measures for numeric values

$\underset{\sim}{x}_i$, $\underset{\sim}{x}_j$: $p\times 1$ feature vector $\quad \underset{\sim}{x}_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{li} \\ \vdots \\ x_{pi} \end{pmatrix}$

(a) Euclidean distance

$$d_e = \left(\sum_{l=1}^{p} (x_{li} - x_{lj})^2 \right)^{1/2}$$

(b) Absolute value distance or City-block distance

$$d_{cb} = \sum_{l=1}^{p} |x_{li} - x_{lj}|$$

(c) Chebyshev distance

$$d_{ch} = \max_{l} |x_{li} - x_{lj}|$$

(d) Pearson distance

$$d_p = \left(\sum_{l=1}^{p} (x_{li} - x_{lj})^2 / s_l^2 \right)^{1/2}$$

s_l^2 : sample variance for l^{th} variable

(e) Quadratic distance

$$d_q = \left((\underset{\sim}{x}_i - \underset{\sim}{x}_j)' Q (\underset{\sim}{x}_i - \underset{\sim}{x}_j) \right)^{1/2}$$

Q is p.d.

e.g $Q = S^{-1}$: S = within group covariance matrix

Case of feature vector with ordinal values

Realizable values are ordered set.

e.g Academic grades: A, B, C, D, E, F

opinion ~~Opinion~~ : strongly disagree, disagree, neutral, agree, strongly agree

One approach is to transform M (say) ordinal values to

$$\frac{j - \frac{1}{2}}{M} \quad ; j = 1(1)M$$

in the prescribed order of their orginal values.

Treat the transformed values as ~~as~~ quantitative numeric values.

Only

Case of nominal (categorical, unordered) / ordinal values

Transform to binary variables set.

Nominal variable with K states

⟶ represent as K binary variables

nominal variable at m^{th} state

$$(0, \ldots \underset{\uparrow}{1}, \ldots 0)$$
m^{th}

Sly for ordinal

Similarity measures for binary variables

Let $\underset{\sim}{x}_{p\times 1}$ & $\underset{\sim}{y}_{p\times 1}$ be 2 vectors of binary variables.

Define

a : # of occurrences of $x_i = 1$ and $y_i = 1$

b : # of occurrences of $x_i = 0$ & $y_i = 1$

c : . . . . . $x_i = 1$ & $y_i = 0$

d : . . . . . $x_i = 0$ & $y_i = 0$

$a + b + c + d = p$, # of variables.

Some similarity measures:

(a) Simple matching coefficient:

$$s_{SM} = \frac{a+d}{a+b+c+d}$$

(b) Russell & Rao

$$s_{RR} = \frac{a}{a+b+c+d}$$

(c) Jaccard $\quad s_J = \frac{a}{a+b+c}$

(d) Czekanowski:

$$s_{cz} = \frac{2a}{2a+b+c}$$

Note:
dissimilarity measure
e.g. $d_{xy} = 1 - s_{xy}$ for s_{sm}

Mixed type : dimensions have both numerical / categorical

Transform all to binary.

Example

	Height*	Weight⊛	Eye color	Hair color	Handedness	Gender
Ind 1	68	140	green	blond	R	F
Ind 2	73	185	brown	brown	R	M
Ind 3	67	165	blue	blond	R	M
Ind 4	64	120	brown	brown	R	F
Ind 5	76	210	brown	brown	L	M

* : in inch

⊛ : in lb

Define, $X_{1i} = \begin{cases} 1, & ht \geq 72 \text{ in} \\ 0, & o/w. \end{cases}$ $X_{2i} = \begin{cases} 1, & wt \geq 150 \text{ lb} \\ 0, & o/w \end{cases}$

$X_{3i} = \begin{cases} 1, & \text{brown eyes} \\ 0, & o/w \end{cases}$ $X_{4i} = \begin{cases} 1, & \text{blond hair} \\ 0, & o/w \end{cases}$

$X_{5i} = \begin{cases} 1, & R \\ 0, & L \end{cases}$ $X_{6i} = \begin{cases} 1, & \text{female} \\ 0, & \text{male} \end{cases}$

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

→

Ind 2 \ Ind 1	1	0
1	1 (a)	2 (b)
0	3 (c)	0 (d)

$$s_{sm}^{(1,2)} = \frac{1}{6}$$

Compute $\forall (i,j)$ $i \neq j$ pairs and get Similarity matrix

$$S = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & & & & & \\ 1/6 & 1 & & & & \\ - & - & 1 & & & \\ - & - & - & 1 & & \\ - & & - & - & 1 & \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

Distance bet^n groups of objects

(A) Methods based on feature vectors

Already discussed (single linkage, complete linkage, average linkage, centroid distance, median distance)

(B) Methods based on probabilistic distance

Distance measures use complete information about the structure of groups (classes) provided by the class conditional densities. The distance measure, say J, satisfies

(i) $J = 0$ iff $p(\underline{x}|\pi_1) = p(\underline{x}|\pi_2)$

(ii) $J \geq 0$

(iii) J attains the maximum value when the classes are disjoint

Commonly used probabilistic distances

(I) Bhattacharya distance

$$J_B = -\log\left(\int \left\{ p(\underline{x}|\pi_1)\, p(\underline{x}|\pi_2) \right\}^{1/2} d\underline{x} \right)$$

(II) Chernoff distance

$$J_C = -\log\left(\int \left\{ p(\underline{x}|\pi_1) \right\}^{s} \left\{ p(\underline{x}|\pi_2) \right\}^{1-s} d\underline{x} \right)$$

$s \in [0,1]$

(III) Divergence distance

$$J_D = \int \left(p(\underline{x}|\pi_1) - p(\underline{x}|\pi_2) \right) \log\left(\frac{p(\underline{x}|\pi_1)}{p(\underline{x}|\pi_2)} \right) d\underline{x}$$

(iv) Patrick - Fischer distance

$$J_{PF} = \left(\int \left\{ p_1 p(\underset{\sim}{x}|\pi_1) - p_2 p(\underset{\sim}{x}|\pi_2) \right\}^2 d\underset{\sim}{x} \right)^{1/2}$$

p_1 & p_2 are prior probabilities of the classes.

Remark: Probabilistic distances are computed under specific distributional setups.

Example:

Gr I objects $- N_p(\underset{\sim}{\mu}_1, \Sigma_1)$ $\quad \Sigma_i > 0$

Gr II objects $- N_p(\underset{\sim}{\mu}_2, \Sigma_2)$

Divergence distance calculations

$$J_D = \int \left((2\pi)^{-p/2} |\Sigma_1|^{-1/2} \exp\left(-\tfrac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)' \Sigma_1^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_1)\right) \right.$$

$$\left. -(2\pi)^{-p/2} |\Sigma_2|^{-1/2} \exp\left(-\tfrac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)' \Sigma_2^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_2)\right) \right)$$

$$\left(\log \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} + \log \left(\exp\left(-\tfrac{1}{2} \left\{ (\underset{\sim}{x} - \underset{\sim}{\mu}_1)' \Sigma_1^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_1) - (\underset{\sim}{x} - \underset{\sim}{\mu}_2)' \Sigma_2^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_2) \right\} \right) \right) \right)$$

$$= \log \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} \int \left(p(\underset{\sim}{x}|\pi_1) - p(\underset{\sim}{x}|\pi_2) \right) d\underset{\sim}{x}$$

$$+ \int \left(-\tfrac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)' \Sigma_1^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_1) \right) \left(p(\underset{\sim}{x}|\pi_1) - p(\underset{\sim}{x}|\pi_2) \right) d\underset{\sim}{x}$$

$$+ \int \left(\tfrac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)' \Sigma_2^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_2) \right) \left(p(\underset{\sim}{x}|\pi_1) - p(\underset{\sim}{x}|\pi_2) \right) d\underset{\sim}{x}$$

$$= \left(1 - 1\right) \log \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}}$$

$$+ \left(-\frac{1}{2}\right) \int \left(\underset{\sim}{x}' \Sigma_1^{-1} \underset{\sim}{x} + \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 - 2 \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{x}\right)$$
$$\left(p(\underset{\sim}{x}|\pi_1) - p(\underset{\sim}{x}|\pi_2)\right) d\underset{\sim}{x}$$

$$+ \left(\frac{1}{2}\right) \int \left(\underset{\sim}{x}' \Sigma_2^{-1} \underset{\sim}{x} + \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 - 2 \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{x}\right)$$
$$\left(p(\underset{\sim}{x}|\pi_1) - p(\underset{\sim}{x}|\pi_2)\right) d\underset{\sim}{x}$$

$$= -\frac{1}{2} \left(\left\{ E_{\pi_1}\left(\underset{\sim}{X}' \Sigma_1^{-1} \underset{\sim}{X}\right) + \cancel{\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1} - 2 \underset{\sim}{\mu}_1' \Sigma_1^{-1} E_{\pi_1}(\underset{\sim}{X}) \right\} \right.$$
$$\left. - \left\{ E_{\pi_2}\left(\underset{\sim}{X}' \Sigma_1^{-1} \underset{\sim}{X}\right) + \cancel{\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1} - 2 \underset{\sim}{\mu}_1' \Sigma_1^{-1} E_{\pi_2}(\underset{\sim}{X}) \right\} \right)$$

$$+ \frac{1}{2} \left(\left\{ E_{\pi_1}\left(\underset{\sim}{X}' \Sigma_2^{-1} \underset{\sim}{X}\right) + \cancel{\underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2} - 2 \underset{\sim}{\mu}_2' \Sigma_2^{-1} E_{\pi_1}(\underset{\sim}{X}) \right\} \right.$$
$$\left. - \left\{ E_{\pi_2}\left(\underset{\sim}{X}' \Sigma_2^{-1} \underset{\sim}{X}\right) + \cancel{\underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2} - 2 \underset{\sim}{\mu}_2' \Sigma_2^{-1} E_{\pi_2}(\underset{\sim}{X}) \right\} \right)$$

— *

Note that

(i) $E_{\pi_i}(\underset{\sim}{X}) = \underset{\sim}{\mu}_i \quad ; i = 1, 2$

and we need to ~~do~~ calculate terms of type

$$E_{\pi_i}\left(\underset{\sim}{X}' \Sigma_j^{-1} \underset{\sim}{X}\right) \qquad i, j = 1, 2$$

$= ?$

If $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$, then

$$E(\underset{\sim}{X}' \Sigma^{-1} \underset{\sim}{X}) = E\left(tr\left(\underset{\sim}{X}' \Sigma^{-1} \underset{\sim}{X}\right)\right)$$

$$= E\left(tr\left(\Sigma^{-1} \underset{\sim}{X}\, \underset{\sim}{X}'\right)\right)$$

$$= tr\left(E\left(\Sigma^{-1} \underset{\sim}{X}\, \underset{\sim}{X}'\right)\right)$$

$$= tr\left(\Sigma^{-1} E\left(\underset{\sim}{X}\, \underset{\sim}{X}'\right)\right)$$

$$= tr\left(\Sigma^{-1}\left(\Sigma + \underset{\sim}{\mu}\, \underset{\sim}{\mu}'\right)\right)$$

$$= tr\, \Sigma^{-1}\Sigma + tr\left(\Sigma^{-1} \underset{\sim}{\mu}\, \underset{\sim}{\mu}'\right)$$

$$= tr\, I_p + tr\left(\underset{\sim}{\mu}' \Sigma^{-1} \underset{\sim}{\mu}\right)$$

$$= tr\, I_p + \underset{\sim}{\mu}' \Sigma^{-1} \underset{\sim}{\mu} \quad \text{— (**)}$$

In general,

$$E(\underset{\sim}{X}' A\, \underset{\sim}{X}) = tr\, A\Sigma + \underset{\sim}{\mu}' A\, \underset{\sim}{\mu}$$

From (*),

$$J_D = \left(\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2\right)$$

$$- \frac{1}{2}\left(E_{\pi_1}(\underset{\sim}{X}' \Sigma_1^{-1} \underset{\sim}{X}) - E_{\pi_2}(\underset{\sim}{X}' \Sigma_1^{-1} \underset{\sim}{X}) - E_{\pi_1}(\underset{\sim}{X}' \Sigma_2^{-1} \underset{\sim}{X}) + E_{\pi_2}(\underset{\sim}{X}' \Sigma_2^{-1} \underset{\sim}{X})\right)$$

$$= \left(\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_1 \right)$$

$$- \frac{1}{2} \left(\left(\text{tr}\, \Sigma_1^{-1} \Sigma_1 + \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 \right) - \left(\text{tr}\, \Sigma_1^{-1} \Sigma_2 + \underset{\sim}{\mu}_2' \Sigma_1^{-1} \underset{\sim}{\mu}_2 \right) \right.$$

$$\left. - \left(\text{tr}\, \Sigma_2^{-1} \Sigma_1 + \underset{\sim}{\mu}_1' \Sigma_2^{-1} \underset{\sim}{\mu}_1 \right) + \left(\text{tr}\, \Sigma_2^{-1} \Sigma_2 + \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 \right) \right)$$

$$= \left(\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_1 \right.$$

$$\left. - \frac{1}{2} \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 + \frac{1}{2} \underset{\sim}{\mu}_2' \Sigma_1^{-1} \underset{\sim}{\mu}_2 + \frac{1}{2} \underset{\sim}{\mu}_1' \Sigma_2^{-1} \underset{\sim}{\mu}_1 - \frac{1}{2} \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 \right)$$

$$+ \left(- \frac{1}{2} \text{tr}\, I_p + \frac{1}{2} \text{tr}\, \Sigma_1^{-1} \Sigma_2 + \frac{1}{2} \text{tr}\, \Sigma_2^{-1} \Sigma_1 - \frac{1}{2} \text{tr}\, I_p \right)$$

$$= \left(\frac{1}{2} \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 + \frac{1}{2} \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 + \frac{1}{2} \underset{\sim}{\mu}_2' \Sigma_1^{-1} \underset{\sim}{\mu}_2 + \frac{1}{2} \underset{\sim}{\mu}_1' \Sigma_2^{-1} \underset{\sim}{\mu}_1 \right.$$

$$\left. - \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_2 \right)$$

$$+ \frac{1}{2} \left(\text{tr}\, \Sigma_1^{-1} \Sigma_2 + \text{tr}\, \Sigma_2^{-1} \Sigma_1 - 2 \text{tr}\, I_p \right)$$

$$= \frac{1}{2} \left(\underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_2 + \underset{\sim}{\mu}_2' \Sigma_1^{-1} \underset{\sim}{\mu}_2 + \underset{\sim}{\mu}_1' \Sigma_2^{-1} \underset{\sim}{\mu}_1 \right.$$

$$\left. - 2 \underset{\sim}{\mu}_1' \Sigma_1^{-1} \underset{\sim}{\mu}_2 - 2 \underset{\sim}{\mu}_2' \Sigma_2^{-1} \underset{\sim}{\mu}_1 \right)$$

$$+ \frac{1}{2} \left(\text{tr}\, \Sigma_1^{-1} \Sigma_2 + \text{tr}\, \Sigma_2^{-1} \Sigma_1 - 2 \text{tr}\, I_p \right)$$

$$= \frac{1}{2} \left(\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1 \right)' \left(\Sigma_1^{-1} + \Sigma_2^{-1} \right) \left(\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1 \right)$$

$$+ \frac{1}{2} \text{tr} \left(\Sigma_1^{-1} \Sigma_2 + \Sigma_2^{-1} \Sigma_1 - 2 I_p \right).$$

Note: If $\Sigma_1 = \Sigma_2$, then

$$J_D = (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1)' \Sigma^{-1} (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1)$$

→ square of Mahalanobis distance

Note: For $N_p(\underset{\sim}{\mu}_1, \Sigma_1)$ & $N_p(\underset{\sim}{\mu}_2, \Sigma_2)$ setup.

Chernoff distance

$$J_c = \frac{1}{2} s(1-s) (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1)' \Sigma_s^{-1} (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1) + \frac{1}{2} \log \left(\frac{|\Sigma_s|}{|\Sigma_1|^{1-s} |\Sigma_2|^s} \right)$$

$$\Sigma_s = (1-s)\Sigma_1 + s\Sigma_2 \quad ; \quad s \in [0, 1]$$

Note: If $\Sigma_1 = \Sigma_2$ i.e. $N_p(\underset{\sim}{\mu}_1, \Sigma)$ & $N_p(\underset{\sim}{\mu}_2, \Sigma)$

$$J_D = 8 J_B = (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1)' \Sigma^{-1} (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1).$$

Note: For $N_p(\underset{\sim}{\mu}_1, \Sigma_1)$ & $N_p(\underset{\sim}{\mu}_2, \Sigma_2)$ setup

$$J_{PF} = (2\pi)^{-p/2} \left(|2\Sigma_1|^{-1/2} + |2\Sigma_2|^{-1/2} - 2 |\Sigma_1 + \Sigma_2|^{-1/2} \right.$$

Patrick-Fischer $\left. \exp\left(-\frac{1}{2} (\mu_2 - \mu_1)' (\Sigma_1 + \Sigma_2)^{-1} (\underset{\sim}{\mu}_2 - \underset{\sim}{\mu}_1)' \right) \right)$

Non-hierarchical clustering methods

Features

- no distance/dissimilarity matrix calculations
- no hierarchy of clusters
- starts from either
  (i) a random initial partition of items into groups
  or (ii) an initial set of randomly selected seed points, which will form the nuclei of clusters.

K-means clustering or iterative relocation method

K-means method is an iterative algorithm that assigns each item to the cluster having the nearest centroid (mean)

Steps for K-means algorithm

(I) Partition the items into K initial clusters.

(II) Reassign items to the cluster whose centroid is nearest (in Euclidean sense mostly). Recalculate the centroids for the cluster receiving the new item and for the cluster losing that item.

(III) Repeat step (II) till no more reassignment can take place.

Note: Rather than starting with a partition of items into K initial groups in step (I), we can also specify

K initial seed points (centroids) and then proceed to step (II) after a walk through the data.

Remark: Let's try to understand the logic behind K-means method!!

Let there be N objects to be put into K clusters.

Suppose $c(i) = k$ denote that object i in cluster k

$k = 1(1)K \; ; \; i = 1(1)N$

A natural "loss function" (criterion) for clustering:

$$W(c) = \frac{1}{2} \sum_{k=1}^{K} \sum_{c(i)=k} \sum_{c(i')=k} d_{ii'} \quad — (1)$$

where, $d_{ii'} = d(\underset{\sim}{x}_i, \underset{\sim}{x}_{i'})$

$W(c)$ is within cluster point scatter, a measure of compactness of clusters. It characterizes the extent to which observations assigned, under a partition, to same clusters tend to be close to one another.

Further, consider the total point scatter

$$T = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{ii'}$$

i.e. $$T = \frac{1}{2} \sum_{k=1}^{K} \sum_{c(i)=k} \left(\sum_{c(i')=k} d_{ii'} + \sum_{c(i') \neq k} d_{ii'} \right)$$

i.e. $$T = \frac{1}{2} \sum_{k=1}^{K} \sum_{c(i)=k} \sum_{c(i')=k} d_{ii'} + \frac{1}{2} \sum_{k=1}^{K} \sum_{c(i)=k} \sum_{c(i') \neq k} d_{ii'}$$

i.e. $T = W(C) + B(C)$

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i') \neq k} d_{ii'}$$ is between-cluster point scatter. A measure of separation of clusters; large when different clusters are far apart.

Note:

$W(C)$ & $B(C)$ depends on cluster assignment

T is indep of cluster assignment

Note: It is natural to maximize $B(C)$ or equivalently minimize $W(C)$ over all possible cluster assignments

Note: The combinatorial optimization is not feasible even for moderate N & K.

The number of distinct cases (assignments) is

$$S(N,K) = \frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^N$$

e.g $S(10,4) = 34,105$

$S(19,4) \approx 10^{10}$

Feasible strategies are based on iterative greedy descent - K-means algorithm is one such algorithm.

K-means algorithm

Iterative descent clustering algorithm

Let
$$d(\underset{\sim}{x}_i, \underset{\sim}{x}_{i'}) = d_{ii'} = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \| \underset{\sim}{x}_i - \underset{\sim}{x}_{i'} \|^2$$

Within-cluster point scatter
$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} \| \underset{\sim}{x}_i - \underset{\sim}{x}_{i'} \|^2$$
$$= \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| \underset{\sim}{x}_i - \bar{\underset{\sim}{x}}_k \|^2$$

where $N_k = \sum_{i=1}^{N} I_{(C(i)=k)}$

$\bar{\underset{\sim}{x}}_k$: mean of of k^{th} cluster

Objective is thus to get C^* ;
$$C^* = \underset{C}{\operatorname{argmin}} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| \underset{\sim}{x}_i - \bar{\underset{\sim}{x}}_k \|^2$$

and we can obtain C^* by enlarged optimization problem
$$\min_{C, \underset{\sim}{m}_1, \ldots \underset{\sim}{m}_K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| \underset{\sim}{x}_i - \underset{\sim}{m}_k \|^2 \quad \text{— (2)}$$

as for any set of S observations
$$\bar{\underset{\sim}{x}}_S = \underset{m}{\operatorname{argmin}} \sum_{i=1}^{S} \| \underset{\sim}{x}_i - \underset{\sim}{m} \|^2$$

K-mean algorithm is an alternating optimization procedure for minimization of (2)

K-means clustering algorithm

Step I. For a given cluster assignment C, the total within cluster variability.

$$\sum_{k=1}^{K} N_k \sum_{C(i)=k} \| \underset{\sim}{x}_i - \underset{\sim}{m}_k \|^2$$

is minimized w.r.t. $\underset{\sim}{m}_1, \ldots \underset{\sim}{m}_k$ by taking means of the currently assigned clusters

Step II: Given a current set of cluster means, i.e. $m_k = \bar{\underset{\sim}{x}}_k$

$$\sum_{k=1}^{N} N_k \sum_{C(i)=k} \| \underset{\sim}{x}_i - \bar{\underset{\sim}{x}}_k \|^2$$

is minimized by assigning each obsn to the closest cluster mean

i.e. $$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} \| \underset{\sim}{x}_i - \underset{\sim}{m}_k \|^2$$

Step I & II are iterated until the assignments do not change.

Remark: The result may represent a suboptimal local minimum. One should start the algorithm with many initial random choices and choose the soln having minimum objective function value.

Example : K-means

Cases	Variables x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

$K = 2$

Step I : Arbitrary partition (A, B) (C, D)

Cluster centroids

Cluster	$\bar{x}_1$	$\bar{x}_2$
(A, B)	2	2
(C, D)	-1	-2

Step II : Compute distances from cluster centroids and reassign each item to the nearest group

(If an item is moved from the initial configuration, the cluster centroid must be updated before we can proceed further)

A

$$d^2(A, (A,B)) = 3^2 + 1^2 = 10$$

$$d^2(A, (C,D)) = 6^2 + 5^2 = 61$$

$\Rightarrow$ A is closer to (A, B) than (C, D)

$\Rightarrow$ No reassignment required.

B

$$d^2(B, (A,B)) = 3^2 + 1^2 = 10$$

$$d^2(B, (C,D)) = 0^2 + 3^2 = 9$$

⇒ Reassign B to (C, D) to form (B, C, D)

Cluster centroid updation

Cluster	Coordinates of centroid $\bar{x}_1$	$\bar{x}_2$
A	5	3
(B, C, D)	-1	-1

$d^2(A, A) = 0$; $d^2(A, (B, C, D)) = 52$

$d^2(B, A) = 40$; $d^2(B, (B, C, D)) = 4$

$d^2(C, A) = 41$, $d^2(C, (B, C, D)) = 5$

$d^2(D, A) = 89$, $d^2(D, (B, C, D)) = 5$

⇒ no reassignment is necessary and relocation iteration stops.

(A), (B, C, D) is the final clustering.

Comparing different cluster partitions

Objective is to have a criterion function for comparing different cluster partitions of the data and hence to choose the best cluster partition.

Let the N data points be given by $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_N$.

The sample variance covariance matrix $\hat{\Sigma}$ is given by

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\underset{\sim}{x}_i - \underset{\sim}{m})(\underset{\sim}{x}_i - \underset{\sim}{m})'$$

where $\underset{\sim}{m} = \frac{1}{N} \sum_{i=1}^{N} \underset{\sim}{x}_i = \underset{\sim}{\bar{x}}$

Let there be K clusters and define

$$z_{ji} = \begin{cases} 1, & \text{if } \underset{\sim}{x}_i \in \text{cluster } j \\ 0, & \text{o/w} \end{cases}$$

$$\underset{\sim}{m}_j = \frac{1}{N_j} \sum_{i=1}^{N} z_{ji} \underset{\sim}{x}_i \left(= \underset{\sim}{\bar{x}}_j\right) : \text{mean of cluster } j$$

$$N_j = \sum_{i=1}^{N} z_{ji} : \#\text{ of } \underset{\sim}{x}_i\text{'s in cluster } j$$

The within-cluster sum of squares and cross product (SSCP) scatter matrix is

$$S_W = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} z_{ji} (\underset{\sim}{x}_i - \underset{\sim}{m}_j)(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'$$

(the pooled within cluster scatter matrix over K clusters)

Note that

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\underset{\sim}{x}_i - \underset{\sim}{m})(\underset{\sim}{x}_i - \underset{\sim}{m})'$$

$$= \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}(\underset{\sim}{x}_i - \underset{\sim}{m})(\underset{\sim}{x}_i - \underset{\sim}{m})'$$

$$= \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}(\overline{\underset{\sim}{x}_i - \underset{\sim}{m}_j} + \overline{\underset{\sim}{m}_j - \underset{\sim}{m}})(\underset{\sim}{x}_i - \underset{\sim}{m}_j + \underset{\sim}{m}_j - \underset{\sim}{m})'$$

$$= \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}\Big((\underset{\sim}{x}_i - \underset{\sim}{m}_j)(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'$$

$$+ (\underset{\sim}{m}_j - \underset{\sim}{m})(\underset{\sim}{m}_j - \underset{\sim}{m})'$$

$$+ \cancel{(\underset{\sim}{x}_i - \underset{\sim}{m}_j)}^{0}(\underset{\sim}{m}_j - \underset{\sim}{m})'$$

$$+ \cancel{(\underset{\sim}{m}_j - \underset{\sim}{m})}^{0}(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'\Big)$$

i.e. $$\hat{\Sigma} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}(\underset{\sim}{x}_i - \underset{\sim}{m}_j)(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'$$

$$+ \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}(\underset{\sim}{m}_j - \underset{\sim}{m})(\underset{\sim}{m}_j - \underset{\sim}{m})'$$

(e.g. $$\frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}(\underset{\sim}{m}_j - \underset{\sim}{m})(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'$$

$$= \frac{1}{N}\sum_{j=1}^{K}(\underset{\sim}{m}_j - \underset{\sim}{m})\sum_{i=1}^{N} z_{ji}(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'$$

$$= \frac{1}{N}\sum_{j=1}^{K}(\underset{\sim}{m}_j - \underset{\sim}{m})\left(N_j\,\underset{\sim}{m}_j - N_j\,\underset{\sim}{m}_j\right)' = 0 \Big)$$

Sly $$\frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}(\underset{\sim}{x}_i - \underset{\sim}{m}_j)(\underset{\sim}{m}_j - \underset{\sim}{m})' = 0$$

i.e.

$$\hat{\Sigma} = S_W + \frac{1}{N}\sum_{j=1}^{K} N_j (\underset{\sim}{m}_j - \underset{\sim}{m})(\underset{\sim}{m}_j - \underset{\sim}{m})'$$

$$= S_W + S_B$$

S_B : between-cluster sum of squares and cross product (SSCP) scatter matrix

(indicates the scatter of the cluster means about total grand mean)

Popular optimum clustering criteria are based on univariate (scalar) functions of above matrices, $S_W, S_B, \hat{\Sigma}$

e.g. (a) Minimization of $tr(S_W)$

$$tr(S_W) = tr\left(\frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}(\underset{\sim}{x}_i - \underset{\sim}{m}_j)(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'\right)$$

$$= \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}\, tr\left((\underset{\sim}{x}_i - \underset{\sim}{m}_j)(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'\right)$$

$$= \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji}\, tr\left((\underset{\sim}{x}_i - \underset{\sim}{m}_j)'(\underset{\sim}{x}_i - \underset{\sim}{m}_j)\right)$$

$$= \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N} z_{ji} \| \underset{\sim}{x}_i - \underset{\sim}{m}_j \|^2$$

i.e. $tr(S_W) = \frac{1}{N}\sum_{j=1}^{K} S_j$

$$S_j = \sum_{i=1}^{n} z_{ji} \| \underset{\sim}{x}_i - \underset{\sim}{m}_j \|^2$$

the within-cluster sum of squares for j^{th} cluster

Thus minimization of $tr(S_W)$ is same as minimization of total within cluster sum of squares about the K centroids. (equiv to max $tr\, S_B$)

(b) Min $\frac{|S_W|}{|\hat{\Sigma}|}$

(c) Min $tr(\hat{\Sigma}^{-1} S_W)$

(d) Max $tr(S_W^{-1} S_B)$