ST3011 ASSIGNMENT

Q1. Single linkage between two clusters A and B defined as

$$d(A,B) = \min_{x \in A,\, y \in B} \{ d(x,y) \}$$

$d(A,B) \geq 0$

$d(A,B) = \min |B-a| = |B-A| = d(A,B) \geq 0$

If the two points are not the same the value will be >0.
If points are the same, $d(A,B) = 0$.

$d(A,B) = 0 \iff A = B$

If $A=B$ then A and B are on the same point.
So $d(A,B) = d(A,A) = d(B,B) = |A-A| = |B-B| = |0| = 0$

$d(A,B) = d(B,A)$

$d(A,B) = |B-A|$
$d(B,A) = |A-B|$
These values are equal $\Rightarrow$ property satisfied

For any other cluster $C = \{z_1, \ldots, z_m\}$ $d(A,C) \leq d(A,B) + d(B,C)$

Three examples.

1\.

B

A C

Clearly $d(A,C) \leq d(A,B) + d(B,C)$

2\.

A B C

Again, shortest distance from A to C is equal to shortest distance from A to B added to shortest distance from B to C.

3\.

A,B,C

(A,B,C) all the same point. All shortest distances are zero. Equality holds

$0 \leq 0+0$

Q2 Euclidean Dissimilarity and Average linkage

$$E.D. = \sqrt{(y_2-y_1)^2+(x_2-x_1)^2} \qquad A.L = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x,y).$$

$$d(A,A) = d\left\{ \binom{1}{1}\binom{1}{2},\ \binom{1}{1}\binom{1}{1},\ \binom{1}{2}\binom{1}{2} \right\}$$

$$\sqrt{(1-1)^2+(2-1)^2} = \sqrt{1}$$
$$\sqrt{(1-1)^2+(1-1)^2} = \sqrt{0} = 0$$
$$\sqrt{(2-2)^2+(1-1)^2} = \sqrt{0} = 0$$

$$A.L = \frac{\sqrt{1}+0+0}{(2)(2)} = \frac{1}{4}$$

$$d(B,B) = d\left\{ \binom{2}{3}\binom{3}{4},\ \binom{2}{3}\binom{2}{3},\ \binom{3}{4}\binom{3}{4} \right\}$$

$$\sqrt{(4-3)^2+(3-2)^2} = \sqrt{2}$$
$$\sqrt{(3-3)^2+(2-2)^2} = \sqrt{0} = 0$$
$$\sqrt{(4-4)^2+(4-4)^2} = \sqrt{0} = 0$$

$$A.L = \frac{\sqrt{2}+0+0}{(2)(2)} = \frac{\sqrt{2}}{4}$$

$$d(C,C) = d\left\{ \binom{4}{5}\binom{4}{5}\ \binom{5}{6}\binom{5}{6},\ \binom{1}{2}\binom{1}{2},\ \binom{4}{5}\binom{5}{6},\ \binom{4}{5}\binom{1}{2}\ \binom{5}{6}\binom{1}{2} \right\}$$

$$\sqrt{(5-5)^2+(4-4)^2} = \sqrt{0} = 0$$
$$\sqrt{(6-6)^2+(5-5)^2} = \sqrt{0} = 0$$
$$\sqrt{(2-2)^2+(1-1)^2} = \sqrt{0} = 0$$
$$\sqrt{(6-5)^2+(5-4)^2} = \sqrt{2}$$
$$\sqrt{(2-5)^2+(1-4)^2} = \sqrt{18}$$
$$\sqrt{(2-6)^2+(1-5)^2} = \sqrt{32}$$

$$= A.L \quad \frac{0+0+0+\sqrt{2}+\sqrt{18}+\sqrt{32}}{(3)(3)} = \cancel{4(+4\sqrt{2}}$$
$$= \frac{8\sqrt{2}}{9} = 1.257$$

$$d(A,B) = d\left\{ \binom{1}{1}\binom{2}{3},\ \binom{1}{1}\binom{3}{4},\ \binom{1}{2}\binom{2}{3},\ \binom{1}{2}\binom{3}{4} \right\}$$

$$\sqrt{(3-1)^2+(2-1)^2} = \sqrt{5}$$
$$\sqrt{(4-1)^2+(3-1)^2} = \sqrt{13}$$
$$\sqrt{(3-2)^2+(2-1)^2} = \sqrt{2}$$
$$\sqrt{(4-2)^2+(3-1)^2} = \sqrt{8}$$

$$A.L = \frac{\sqrt{5}+\sqrt{13}+\sqrt{2}+\sqrt{8}}{(2)(2)} = 2.52$$

DAVID WEITBRECHT
12300644

cont..

$d(A,C) = d\left\{ \binom{1}{1}\binom{4}{5} \; \binom{1}{1}\binom{5}{6} \; \binom{1}{1}\binom{1}{2} \; \binom{1}{2}\binom{4}{5} \; \binom{1}{2}\binom{5}{6} \; \binom{1}{2}\binom{1}{2} \right\}$

$\sqrt{(5-1)^2+(4-1)^2} = \sqrt{25} = 5$

$\sqrt{(6-1)^2+(5-1)^2} = \sqrt{41}$

$\sqrt{(2-1)^2+(1-1)^2} = \sqrt{1} = 1$

$\sqrt{(5-2)^2+(4-1)^2} = \sqrt{18}$

$\sqrt{(6-2)^2+(5-1)^2} = \sqrt{32}$

$\sqrt{(2-2)^2+(1-1)^2} = \sqrt{0} = 0$

$$A.L = \frac{5+\sqrt{41}+1+\sqrt{18}+\sqrt{32}+0}{(2)(3)} = 3.717$$

$d(B,C) = d\left\{ \binom{2}{3}\binom{4}{5} \; \binom{2}{3}\binom{5}{6} \; \binom{2}{3}\binom{1}{2} \; \binom{3}{4}\binom{4}{5} \; \binom{3}{4}\binom{5}{6} \; \binom{3}{4}\binom{1}{2} \right\}$

$\sqrt{(5-3)^2+(4-2)^2} = \sqrt{8}$

$\sqrt{(6-3)^2+(5-2)^2} = \sqrt{18}$

$\sqrt{(2-3)^2+(1-2)^2} = \sqrt{2}$

$\sqrt{(5-4)^2+(4-3)^2} = \sqrt{2}$

$\sqrt{(6-4)^2+(5-3)^2} = \sqrt{8}$

$\sqrt{(2-4)^2+(1-3)^2} = \sqrt{8}$

$$A.L = \frac{\sqrt{8}+\sqrt{18}+\sqrt{2}+\sqrt{2}+\sqrt{8}+\sqrt{8}}{(2)(3)} = \frac{11\sqrt{2}}{6} = 2.59$$

Resulting Dissimilarity matrix

	A	B	C
A	0.25	2.52	3.717
B	2.52	0.35	2.54
C	3.717	2.59	1.257

DAVID WEMBRECHT
12300644

Q3 Rand Index = $\dfrac{\binom{n}{2} + 2\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{c_1}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{c_2}\binom{n_{\cdot j}}{2}\right]}{\binom{n}{2}}$

		Cluster A		
		Group1	Group2	
Cluster B	Group1	25	3	28
	Group 2	4	36	40
	Group3	6	7	13
		35	46	81

$$\frac{\binom{81}{2} + 2\left[\binom{25}{2}+\binom{3}{2}+\binom{4}{2}+\binom{36}{2}+\binom{6}{2}+\binom{7}{2}\right] - \left[\binom{28}{2}\binom{40}{2}\binom{13}{2}\binom{35}{2}\binom{46}{2}\right]}{\binom{81}{2}}$$

$$\frac{3240+2[300+3+6+630+15+21] - [378+780+78+595+1035]}{3240}$$

$$\frac{3240+1950-2866}{3240} = \frac{2324}{3240} = 0.71728$$

Rand Index $= 0.71728$

DAVID WETTBRECHT
12300644

Q2 Euclidean Dissimilarity and Average Linkage

$E.D = \sqrt{(y_2-y_1)^2+(x_2-x_1)^2}$ $\qquad$ $A.L = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x,y)$

$d(A,A) = d\left\{\binom{1}{1}\binom{1}{1},\ \binom{1}{1}\binom{1}{2},\ \binom{1}{2}\binom{1}{1},\ \binom{1}{2}\binom{1}{2}\right\}$

$\sqrt{(1-1)^2+(1-1)^2} = \sqrt{0} = 0$

$\sqrt{(1-1)^2+(2-1)^2} = \sqrt{1} = 1$

$\sqrt{(1-2)^2+(1-1)^2} = \sqrt{1} = 1$

$\sqrt{(2-2)^2+(1-1)^2} = \sqrt{0} = 0$

$A.L = \frac{0+1+1+0}{4} = \frac{1}{2}$

$d(B,B) = d\left\{\binom{2}{3}\binom{2}{3},\ \binom{2}{3}\binom{3}{4},\ \binom{3}{4}\binom{2}{3},\ \binom{3}{4}\binom{3}{4}\right\}$

$\sqrt{(3-3)^2+(2-2)^2} = \sqrt{0} = 0$

$\sqrt{(4-3)^2+(3-2)^2} = \sqrt{2} = \sqrt{2}$

$\sqrt{(3-4)^2+(2-3)^2} = \sqrt{2} = \sqrt{2}$

$\sqrt{(4-4)^2+(3-3)^2} = \sqrt{0} = 0$

$A.L = \frac{0+\sqrt{2}+\sqrt{2}+0}{(2)(2)} = \frac{\sqrt{2}}{2} = 0.71$

$d(C,C) = d\left\{\binom{4}{5}\binom{4}{5},\ \binom{4}{5}\binom{5}{6},\ \binom{4}{5}\binom{1}{2},\ \binom{5}{6}\binom{4}{5},\ \binom{5}{6}\binom{5}{6},\ \binom{5}{6}\binom{1}{2},\ \binom{1}{2}\binom{4}{5},\ \binom{1}{2}\binom{5}{6},\ \binom{1}{2}\binom{1}{2}\right\}$

$\sqrt{(5-5)^2+(4-4)^2} = \sqrt{0} = 0$

$\sqrt{(6-5)^2+(5-4)^2} = \sqrt{2}$

$\sqrt{(2-5)^2+(1-4)^2} = \sqrt{18}$

$\sqrt{(5-6)^2+(4-5)^2} = \sqrt{2}$

$\sqrt{(6-6)^2+(5-5)^2} = \sqrt{0} = 0$

$\sqrt{(2-6)^2+(1-5)^2} = \sqrt{32}$

$\sqrt{(5-2)^2+(4-1)^2} = \sqrt{18}$

$\sqrt{(6-2)^2+(5-1)^2} = \sqrt{32}$

$\sqrt{(2-2)^2+(1-1)^2} = \sqrt{0} = 0$

$A.L = \frac{0+\sqrt{2}+\sqrt{18}+\sqrt{2}+0+\sqrt{32}+\sqrt{18}+\sqrt{32}+0}{(3)(3)} = \frac{16\sqrt{2}}{9} = 2.51$

$d(A,B) = d\left\{\binom{1}{1}\binom{2}{3},\ \binom{1}{1}\binom{3}{4},\ \binom{1}{2}\binom{2}{3},\ \binom{1}{2}\binom{3}{4}\right\}$

$\sqrt{(3-1)^2+(2-1)^2} = \sqrt{5}$

$\sqrt{(4-1)^2+(3-1)^2} = \sqrt{13}$

$\sqrt{(3-2)^2+(2-1)^2} = \sqrt{2}$

$\sqrt{(4-2)^2+(3-1)^2} = \sqrt{8}$

$A.L = \frac{\sqrt{5}+\sqrt{13}+\sqrt{2}+\sqrt{8}}{(2)(2)} = 2.52$

DAVID WEITBRECHT

Q2 cont.

$d(A,C) = d\left\{ \begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}4\\5\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}5\\6\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}4\\5\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}5\\6\end{pmatrix} \begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} \right\}$

$\sqrt{(5-1)^2+(4-1)^2} = \sqrt{25} = 5$

$\sqrt{(6-1)^2+(5-1)^2} = \sqrt{41}$

$\sqrt{(2-1)^2+(1-1)^2} = \sqrt{1} = 1$

$\sqrt{(5-2)^2+(4-1)^2} = \sqrt{18}$

$\sqrt{(6-2)^2+(5-1)^2} = \sqrt{32}$

$\sqrt{(2-2)^2+(1-1)^2} = \sqrt{0} = 0$

$A.L = \dfrac{5+\sqrt{41}+1+\sqrt{18}+\sqrt{32}+0}{(2)(3)} = 3.717$

$d(B,C) = d\left\{ \begin{pmatrix}2\\3\end{pmatrix}\begin{pmatrix}4\\5\end{pmatrix}, \begin{pmatrix}2\\3\end{pmatrix}\begin{pmatrix}5\\6\end{pmatrix}, \begin{pmatrix}2\\3\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}3\\4\end{pmatrix}\begin{pmatrix}4\\5\end{pmatrix}, \begin{pmatrix}3\\4\end{pmatrix}\begin{pmatrix}5\\6\end{pmatrix}, \begin{pmatrix}3\\4\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} \right\}$

$\sqrt{(5-3)^2+(4-2)^2} = \sqrt{8}$

$\sqrt{(6-3)^2+(5-2)^2} = \sqrt{18}$

$\sqrt{(2-3)^2+(1-2)^2} = \sqrt{2}$

$\sqrt{(5-4)^2+(4-3)^2} = \sqrt{2}$

$\sqrt{(6-4)^2+(5-3)^2} = \sqrt{8}$

$\sqrt{(2-4)^2+(1-3)^2} = \sqrt{8}$

$A.L. = \dfrac{\sqrt{8}+\sqrt{18}+\sqrt{2}+\sqrt{2}+\sqrt{8}+\sqrt{18}}{(2)(3)} = \dfrac{11\sqrt{2}}{6} = 2.59$

Resulting Dissimilarity Matrix:

	A	B	C
A	0.5	2.52	3.717
B	2.52	0.71	2.59
C	3.717	2.59	2.51

Single Linkage

$d(A,B) = \min_{x \in A,\, y \in B} \underbrace{d(x,y)}_{\text{distance}}$

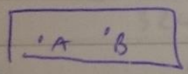
takes shortest distance

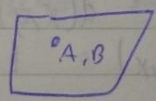
- In begining each element is in cluster of its own
- clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster
- At each step, the two clusters seperated by the shortest distance are combined

$d(A,B) \geq 0$

A and B are any two points. Then distance will always be ≥ 0 as shown in graph

if $A=B$ distance $=0$





Satisfies property

$d(A,B) = 0$ if and only if $x=y$

If $x \neq y$ then distance will be >0, thus if there is a distance of 0, A and B must be at the same point.

$d(x,y) = d(y,x)$

This will always be the case with single linkage, the distance will not change depending on which point comes first/second

12/12/14 Assignment Correction

1 i. $d(A,B) = \min_{x \in A\, y \in B} (d(x,y))$ ← minimum of non negative ⇒ so non negative itself.

$d(x,y) \geq 0$

$A \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} \neq B \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right)$

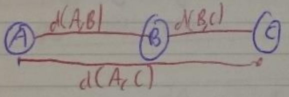
But $d(A,B) = d\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) = 0$

$x=y$ part of property is NOT true

ii. $D(\cancel{A,B})$ $D(x,y) = d(y,z)$ $\min_{x \in A\, y \in B} (d(y,x)) = d(B,A)$

iii. Third property is false for single linkage

Counter example.



$d(A,C) \geq d(A,B) + d(B,C)$

2 $d(A,A)$ $A = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}$

$\dfrac{d\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} + d\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\} + d\left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} + d\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right)}{4} = \dfrac{2}{4}$

$D(A,B)$ $A = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}$
$B = \left\{ \begin{pmatrix} 3 \\ 3 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right\}$

$\dfrac{d\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \end{pmatrix} \right) + d\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right) + d\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \end{pmatrix} \right) + d\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} \right)}{4} = \dfrac{\sqrt{5} + \sqrt{13} + \sqrt{2} + \sqrt{8}}{4}$

$d(B,B) = 2\sqrt{2}/4$

$d(C,C) = (2\sqrt{2} + 2\sqrt{18} + 2\sqrt{32})/9$

$d(A,C) = d(C,A) = (\sqrt{25} + \sqrt{41} + 1 + \sqrt{18} + \sqrt{32})/6$

$d(B,C) = d(C,B) = (\sqrt{8} + \sqrt{18} + \sqrt{2} + \sqrt{8} + \sqrt{2} + \sqrt{2})/6$

2

3. $$\frac{\binom{81}{2} + 2\left[\binom{25}{2}+\binom{4}{2}+\binom{6}{2}+\binom{3}{2}+\binom{36}{2}+\binom{7}{2}\right] - \left[\binom{35}{2}+\binom{46}{2}+\binom{38}{2}+\binom{40}{2}+\binom{3}{2}\right]}{\binom{81}{2}} \approx 0.72$$

4. 1m Follow report guidline
1m Standardise the data
1m Was a PCA preformed?
1.5m First PC intrepreted
1.5m Second PC intrepreted
1 How many PC's to include
1.5 Contrast results with non standardised data
1 Was clustering performed?
1.5 Choice of dissimilarity/effect of dissimilarity
1.5 Choice/effect of linkage, alter conclusion
1 How many clusters should be used?
1.5 For chosen clusters, what are they like? What makes it different?
3 Consider cluster analysis on the lower dimensional PCA results
1 Looking at Rand Index
1 Relating back to motivating story