1

08/11/2021

DWM
ASSIGNMENT 3
CLUSTERING

JUNAID GIRKAR
60004190057
TE COMPS A4

Q1

ANS.

CLUSTERING HOMOGENEITY :

Consider clustering c_1, where in a cluster $c \in c_1$, contains objects from two categories L_1, L_2 in the ground truth. Moreover consider clustering c_2, which is identical to c_1, except that c_1 is split into two clusters containing objects in L_1 and L_2 respectively. Let $c_1(o)$ and $c_2(o)$ be the cluster-id of o in c_1 and c_2 respectively.

B cubed precision of c_1 is :

$$\text{Precision}(c_1) = \frac{\sum_{o \notin c} \left(\sum_{o' \neq o,\ c_1(o)=c_1(o')} \text{correctness}(o,o') \Big/ \| \{o' \neq o \mid c_1(o) = c_1(o')\} \| \right) + \sum_{o \notin c} \left(\sum_{o' \neq o,\ c_1(o)=c_1(o')} \text{correctness}(o,o') \Big/ \| \{o' \neq o \mid c_1(o) = c_2(o')\} \| \right)}{\text{total number of objects in dataset.}}$$

The B cubed precision of c_2 is

$$\text{Rec}(c_2) = \frac{\sum_{o \notin c} \sum_{o' \neq o,\ c_2(o)=c_2(o')} \text{correctness}(o,o') \Big/ \| \{o' \neq o \mid c_2(o) = c_2(o')\} \| + \sum_{o \notin c} \sum_{o' \neq o,\ c_2(o)=c_2(o')} \text{correctness}(o,o') \Big/ \| \{o' \neq o \mid c_2(o) = c_2(o')\} \|}{\text{total number of objects in dataset.}}$$

Since $c_1 = c_2$, we have

$$\sum_{o \notin c} \frac{\sum_{o' \neq o,\ c_1(o)=c_1(o')} \text{correctness}(o,o')}{\| \{o' \neq o \mid c_1(o) = c_1(o')\} \|} = \sum_{o \notin c} \frac{\sum_{o' \neq o,\ c_2(o)=c_2(o')} \text{correctness}(o,o')}{\| \{o' \neq o \mid c_2(o) = c_2(o')\} \|}$$

In C_1, c is a mintuge of the objects from two categories L_1 and L_2. In C_2, c is divided into 2 clusters containing the objects in the 2 categories respectively. Therefore,

$$\sum_{o \in c} \frac{\sum_{o \neq o, C_1(o) = C_1(o'), \text{correctness}(o,o')}}{\| \{ o' \neq o \mid C_1(o) = C_1(o') \} \|} < \sum_{o \in c} \frac{\sum_{o' \neq o, C_2(o) = C_2(o'), \text{correctness}(o,o')}}{\| \{ o' \neq o \mid C_2(o) = C_2(o') \} \|}$$

Thus precision (C_1) < Precision (C_2)
Recall (C_1) < Recall (C_2)

CLUSTER COMPLETENESS : The proof of the clustel completeness is similar to that of the cluster homogenity.

RAG BAG : Consider a cluster C_1 and a clustel $c \in C_1$ such that all objects in c except for one, denoted by o, belong to the same category according to the ground truth. Consider a clustering from various categories according to the ground truth. In other words, $c' \in C_2$ is a rag bag. we show that B cubed precision of C_2 is greater than that of C_1.

In the B cubed precision calculation, precision (C_1) and precision (C_2) are the same except for those objects pairs involving o.

$$\text{Prec}(C_2) - \text{Prec}(C_1) = \frac{\left[\frac{(\|c'\| - 1)}{\|c'\| - 1} + \frac{\text{No of objects in } c' \text{ in the same category of } o}{\|c'\| - 1} - \frac{(\|c\| - 1)}{\|c\| - 1} \right]}{\text{total number of objects in the dataset}}$$

$$= \frac{\text{number of objects in } c' \text{ in the same category of } o}{(\|c'\| - 1) \times \text{total number of objects in the dataset}}$$

In the worst case, there is no object in c' that is in the same category of O, Prec (C_2) - Prec $(C_1) = 0$.
As long as there is atleast one object in c' that is the same category of O, Prec (C_2) – Prec $(C_1) > 0$.
The B cubed recall can be analyzed in a similar manner.

SMALL CLUSTER PRESERVATION: Suppose there are 2 categories, L_1 and L_2, in the ground truth such that $\|L_1\| > \|L_2\|$. Consider a clustering C_1 wherein the objects in L_1 are divided into 2 clusters C_{11} and C_{12}, and the objects in L_2 are contained in a cluster C_2. Consider a cluster C_2 identical to C_1 except that C_{11} and C_{12} are merged into one cluster C_1 and the objects in L_2 are divided into two clusters C_{21} and C_{22}. It is easy to verify that Recall (C_2) > Recall (C_1).

Q2

ANS (a) SINGLE LINK:

(i) At level 0, there are 6 clusters:

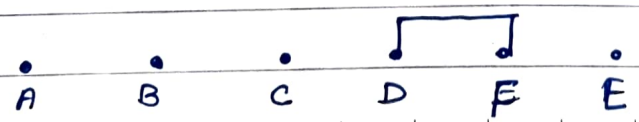
{A}, {B}, {C}, {D}, {E}, {F}

(ii) In the matrix items F and D are closest to each other i.e. minimum distance = 0.50

So we merge F and D into a single cluster.

clusters now are :-

{A}, {B}, {C}, {F, D}, {E}



computing distance from A to {F, D}

$\therefore$ dist({F, D}, {A}) = min (dist (F, A), dist (D, A))

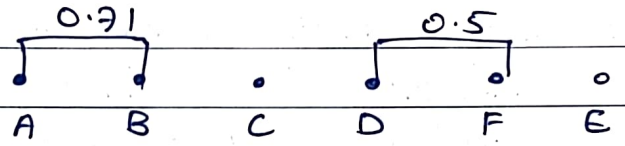
= min (3.20, 3.61)

= 3.20

(iii) In the matrix, cluster {A} and {B} are closest after {F, D} i.e. have minimum distance of 0.71 so we merge them into a single cluster.

$\therefore$ clusters now are :-

{A, B}, {C}, {F, D}, {E}



Compute new distance

dist ((A, B), (D, F)) = min (dist (A, D), dist (A, F), dist (B, D), dist (B, F))

= min (3.61, 3.20, 2.92, 2.50)

= 2.50

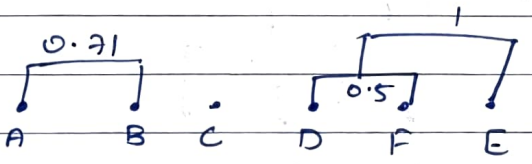
$\therefore$ Similarly we can calculate distance for all new distance matrix

ITEM	{A, B}	{C}	{D, F}	{E}
{A, B}	0			
{C}	4.95	0		
{D, F}	2.50	2.24	0	
{E}	3.54	1.41	1.00	0

(iv) In the above matrix, clusters {D, F} and {E} are closest. So, we merge these two.

clusters now are :-

{A, B}, {C}, {D, E, F}



From {A, B} to {D, E, F}

dist ({A, B}, {D, E, F}) = min (dist (A, D), dist (A, E), dist (A, F), dist (B, D), dist (B, E), dist (B, F))

= min (3.61, 2.92, 3.20, 2.50, 3.54, 4.24)

= 2.50

From {C} to {D, F, E}

∴ dist ({C}, {D, E, F}) = min (dist (C, D), dist (C, E), dist (C, F))

= min (2.24, 2.50, 1.41)

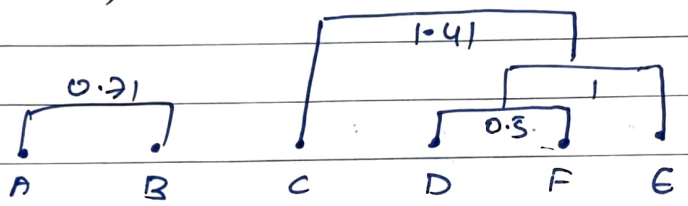
= 1.41

Now distance matrix

Item	{A, B}	{C}	{D, E, F}
{A, B}	0		
{C}	4.95	0	
{D, E, F}	2.50	1.41	0

(v) In the above distance matrix cluster {D, F, E} and {C} are closest. So we can merge them into single cluster. Clusters now are :-

{A, B}, {D, F, E, C}



From {A, B} to {C, D, F, E}

dist ((A, B), (C, D, F, E)) = min (dist(A, C), dist(A, D), dist(A, F), dist(A, E), dist(B, C), dist(B, D), dist(B, F), dist(B, E))

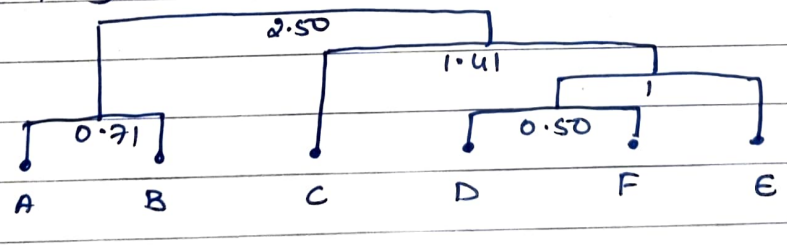
= min (5.66, 3.61, 4.24, 3.20, 4.95, 2.92, 3.54, 2.50)

= 2.50

∴ New distance matrix

Item	{A, B}	{C, D, F, E}
{A, B}	0	
{C, D, F, E}	2.50	0

(vi) So we can merge cluster {A, B} and {D, F, E, C} into one cluster.



Dendrogram for single linkage.

(2) AVERAGE UNKAGE

Distance matrix:

	A	B	C	D	E	F
A	0					
B	0.71	0				
C	5.66	4.95	0			
D	3.61	2.92	2.24	0		
E	4.24	3.54	1.41	1.00	0	
F	3.2	2.5	2.5	0.5	1.12	0

At level 0:

clusters : {A}, {B}, {C}, {D}, {E}, {F}

At level 1 : avg = 1

∵ dist (D, F) ≤ 1, Merge DF (0.5)

∴ Clusters : {A}, {B}, {C}, {DF}, {E}

Average distance matrix

	A	B	C	DF	E
A	0				
B	0.71	0			
C	5.66	4.95	0		
DF	3.41	2.71	2.37	0	
E	4.24	3.54	1.41	1.06	0

∵ dist (A, B) ≤ 1 (0.71)

Merge A, B

∴ Clusters :- {AB}, {C}, {DF}, {E}

Avg distance matrix

	AB	C	DF	E
AB	0			
C	5.31	0		
DF	3.058	2.37	0	
E	3.89	1.41	(1.06)	0

At avg = 2

∵ avg dist (DF, E) ≤ 2 (1.06)

∴ Merge DF and E

clusters : {A, B}, {C}, {D, F, E}

Avg dist matrix

	AB	C	DFE
AB	0		
C	5.31	0	
DFE	3.335	2.05	0

At avg = 3

∵ avg dist (DFE, C) ≤ 3 (2.05)

∴ Merge (DFE, C)

clusters : {A, B}, {C, D, F, E}

Avg dist matrix

	AB	CDEF
AB	0	
CDEF	(3.8275)	0

5

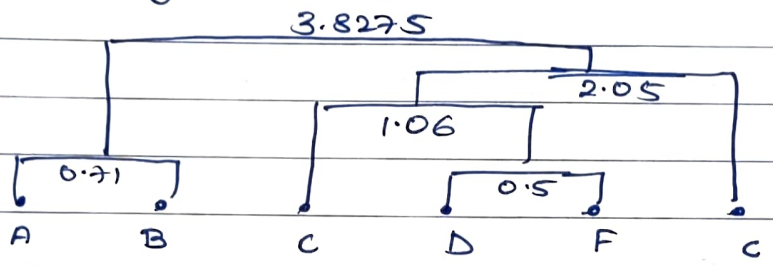
At avg = 4

∵ avg dist (AB, CDEF) ≤ 4 (3.8275)

Merge AB, CDEF

Clusters : {ABCDEF}

Dendrogram for Average link

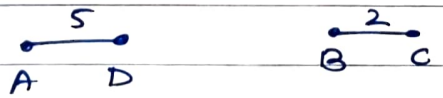


Q3

ANS

	A	B	C	D
A	0			
B	1	0		
C	4	2	0	
D	5	6	3	0

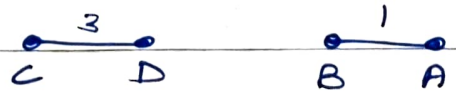
let A & B be initial mediods



Mediods : {A, B}

Non-Mediods : {C, D}

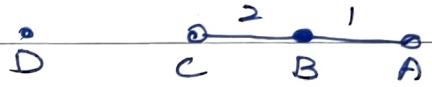
Replace A → C



$TC_{AC} = 1 + 0 + (-2) + (-2) = -3$

Reduced by +3

Replace $A \rightarrow D$



$TC_{AD} = 1 + 0 + 0 + (-5) = -4$

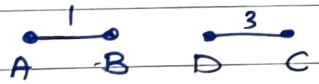
Reduced by +4

Replace $B \rightarrow C$

$TC_{BC} = 0 + 1 + (-2) + (-2) = -3$

Reduced by +3

Replace $B \rightarrow D$



$TC_{BD} = 0 + 1 + 1 + (-5) = -3$

Reduced by -3

∴ Cost is reduced by replacing D with A.

New mediods {D, B}

clusters:

Mediods {D, B} Non-mediods {A, C}

Replace D → A

$TC_{DA} = -1 + 0 + 0 + 5 = 4$

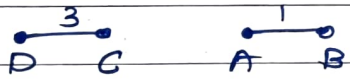
Increase by 4

Replace D → C

$TC_{DC} = 0 + 0 + (-2) + 3 = 1$

Increase by 1

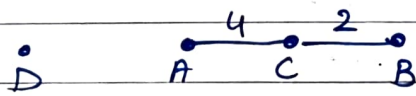
Replace B with A



$TC_{BA} = (-1) + 1 + 1 + 0 = 1$

Increased by 1

Replace B with C

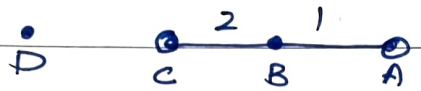


$TC_{BC} = 3 + 2 + (-2) + 0 = 3$

Increased by 3

∵ No one reduced the cost, original mediods {B, D} remain same

Final clusters :



Mediods : {D}, {B}
Non-Mediods: {C}, {A}

Qs $A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $B_1(5,8)$, $B_2(7,5)$,
ANS $B_3(6,4)$, $C_1(1,2)$, $C_2(4,9)$

Initial centroids: $A_1(2,10)$, $B_1(5,8)$, $C_1(1,2)$

Euclidean distance : $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$
* → represents min distance from centroid

Iteration 1	C_1	C_2	C_3
A_1	0 *	3.61	8.06
A_2	5	4.24	3.16 *
A_3	8.49	5 *	7.28
B_1	3.61	0 *	7.21
B_2	7.07	3.61 *	6.71
B_3	7.22	4.12 *	5.39
C_1	8.06	7.21	0 *
C_2	2.24	1.41 *	7.62

Clusters after 1st round of execution:

$\{A_1\}$ $\{A_3, B_1, B_2, B_3, C_2\}$ $\{A_2, C_1\}$

New centroids = (2,10), (6,6), (1.5, 3.5)

Iteration 2	(2,10)	(6,6)	(1.5, 3.5)
A_1	0 *	5.65	6.5
A_2	5	4.12	1.58 *
A_3	8.48	2.82 *	6.51
B_1	8.61	2.23 *	5.7
B_2	7.07	1.91 *	5.7
B_3	7.21	2	4.5
C_1	8.06	6.4	1.58 *
C_2	2.24 *	3.6	6.04

Clusters after 2nd round of execution:

$\{A_1, C_2\}$, $\{A_3, B_1, B_2, B_3\}$, $\{A_2, C_1\}$

New centroids :- (3, 9.5), (6.5, 5.25), (1.5, 3.5)

Iteration 3	(3, 9.5)	(6.5, 5.25)	(1.5, 3.5)
A_1	1.12 *	6.54	6.51
A_2	4.61	4.51	1.58 *
A_3	7.43	1.95 *	6.51
B_1	2.5 *	3.13	5.7
B_2	6.02	0.56 *	5.7
B_3	6.26	1.35 *	4.5
C_1	7.76	6.39	1.58 *
C_2	1.12 *	4.51	6.04

clusters after 3rd round of execution:
$\{A_1, B_1, C_2\}$, $\{A_3, B_2, B_3\}$, $\{A_2, C_1\}$

New centroids: (3.67, 9), (7, 4.33), (1.5, 3.5)

Iteration 4

Iteration 4	(3.67, 9)	(7, 4.33)	(1.5, 3.5)
A_1	1.95 *	6.01	6.51
A_2	4.33	5.04	1.58 *
A_3	6.61	1.05 *	6.51
B_1	1.66 *	4.14	5.7
B_2	5.2	0.67 *	5.7
B_3	5.52	1.05 *	4.5
C_1	7.49	6.44	1.58 *
C_2	0.33 *	5.55	6.04

clusters after 4th round of execution:
$\{A_1, B_1, C_2\}$, $\{B_2, B_3, A_3\}$, $\{A_2, C_1\}$
∵ There are no new assignments to the clusters. The algorithm terminates and the above clusters are the final three clusters

	$\{A1, B1, C2\}$	$\{A3, B1, B2\}$	$\{A2, C1\}$
centroids:	(3.67, 9)	(7, 4.33)	(1.5, 3.5)

a) clusters after first round of execution

→ $\{A1\}$ (2, 10); $\{A3, B1, B2, B3, C2\}$ (5, 8); $\{A2, C1\}$ (1, 2)

b) Final three clusters

→ $\{A1, B1, C2\}$ (3.67, 9); $\{A3, B1, B2\}$ (7, 4.33); $\{A2, C1\}$ (1.5, 3.5)