100 2022 847
Sai Parthish. M.

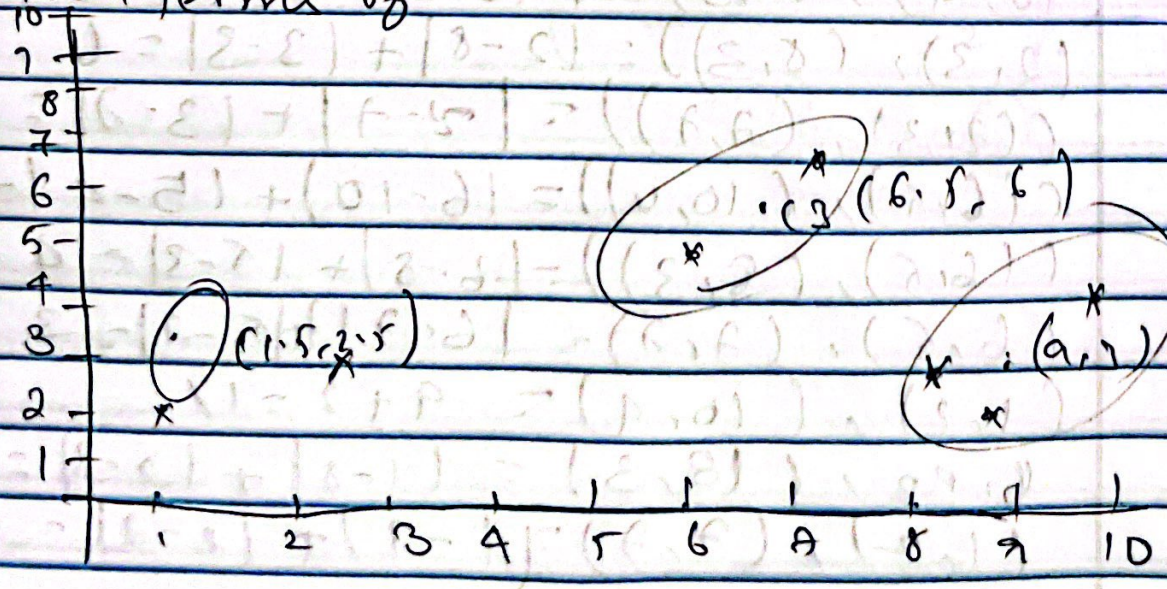
Data mining. 2

2// K-means clustering

Given 7 sensor nodes deployed in a network, The location of three sensor are describe in terms of x and y coordinate the given points are

	X	Y
1	10	4
2	8	3
3	7	4
4	2	3
5	6	5
6	1	2
7	9	2

7 sensor nodes deployed in a network locations of these sensors are described in terms of



Distance	m1(10,4)	m2(8,3)	m3(7,7)
1 (10,4)	0	3	6
2 (8,3)	3	0	5
3 (7,7)	6	5	0
4 (2,3)	9	6	9
5 (6,5)	5	4	3
6 (1,2)	11	8	11
7 (9,2)	3	9	7

By calculating manhatton distance from points to centroids

$$d(x,y) = \left(\sum_{i=1}^{n} (|x_i - y_i|)^p\right)^{1/p} \quad \text{where } p=1$$

$((10,4),(8,3)) = |10-8| + |4-3| = 2+1 = 3$

$((10,4),(7,7)) = |10-7| + |4-7| = 3+3 = 6$

$((8,3),(7,7)) = |8-7| + |3-7| = 1+4 = 5$

$((10,4),(2,3)) = |10-2| + |4-3| = 8+1 = 9$

$((2,3),(8,3)) = |2-8| + |3-3| = 6$

$((2,3),(7,7)) = |2-7| + |3-7| = 9$

$((6,5),(10,4)) = |6-10| + |5-4| = 5$

$((6,5),(8,3)) = |6-8| + |5-3| = 4$

$((6,5),(7,7)) = |6-7| + |5-7| = 3$

$((1,2),(10,4)) = 9+2 = 11$

$((1,2),(8,3)) = |1-8| + |2-3| = 8$

$((1,2),(7,7)) = |-7| + |2-7| = 11.$

$(9,12), (10,4) = |9-10| + |2-4| = 1+2 = 3$

$(9,2), (8,3), \quad |9-8| + |2-3| = 1+1 = 2$

$(9,2), (7,7) = |9-7| + |2-7| = 2+5 = 7$

$C_1 = \{(10,4)\} = \{2\}$

$C_2 = \{2, 4, 6, 7\}$

$C_3 = \{3, 5\}$

mean of C_1 is same it won't change as it only contains only one

$$\text{mean} = \left(\frac{8+2+1+9}{4}, \frac{3+3+2+2}{4}\right)$$

$$= (5, 2.5)$$

mean C_3 is where sensors

$$\left(\frac{7+6}{2}, \frac{7+5}{2}\right)$$

$$= (6.5, 6)$$

Distance	m1(10, 4)	m2(5, 2.5)	m3(6.5, 6)
1(10, 4)	0	6.5	5.5
2(8, 3)	3	3.5	4.5
3(2, 7)	6	6.5	1.5
4(2, 3)	9	3.5	9.5
5(6, 5)	5	3.5	1.5
6(1, 2)	11	4.5	9.5
7(9, 2)	3	4.5	6.5

New centroid

C1 = {1, 4, 7}

C2 = {2, 4, 6}

C3 = {3, 5}

As the sensors changed to different Centroids, the mean of C1 & C2 will change

C3 = remains as previous

$$C3 = \left(\frac{7+6}{2}, \frac{7+5}{2}\right) = (6.5, 6)$$

$$C1 = \left(\frac{10+8+9}{3}, \frac{4+3+2}{3}\right)$$

$$C1 = (9, 3)$$

$$C_2 = \left\{ \frac{2+1}{2}, \frac{3+2}{2} \right\}$$

$$C_2 : (1.5, 2.5)$$

Centroids are: $C_1 (9,3)$

$C_2 = (1.5, 2.5)$ $C_3 = (6.5, 6)$

Distance	$m_1(9,3)$	$m_2(1.5,2.5)$	$m_3(6.5,6$
1 (10, 4)	2	10	5.5
2 (8, 3)	1	7	4.5
3 (7, 7)	6	10	1.5
4 (2, 3)	4	1	7.5
5 (6, 5)	5	7	1.5
6 (1, 2)	9	1	9.5
7 (9, 2)	1	8	6.5

for the new centroid the sensors are remaining

C_1 has (1, 2, 7)

C_2 (4, 6)

$C_3 = (3, 5)$

So, we can lock the new centroid as main centroids

$C_1 (9,3)$ $C_2 (1.5, 2.5)$ $C_3 (6.5, 6)$

(3) Hierarchical Clustering

	A	B	C	D	E	F
A	0.00					
B	0.71	0.00				
C	5.66	4.95	0.00			
D	3.61	2.92	2.24	0.00		
E	4.24	3.54	1.41	1.00	0.00	
F	3.20	2.50	2.50	0.50	1.12	0.00

Step 1 (i) Single-link hierarchical clustering

As F.D having the shortest distance

	F.D	F.D	A	B	C	E
	F.D	0.00	3.20	2.50	2.24	1.00
	A		0.00	0.71	5.66	4.24
	B			0.00	4.95	3.54
	C				0.00	1.41
	E					0.00

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

Here we are combing A & B because they are having the shortest distance whole distance matrix.

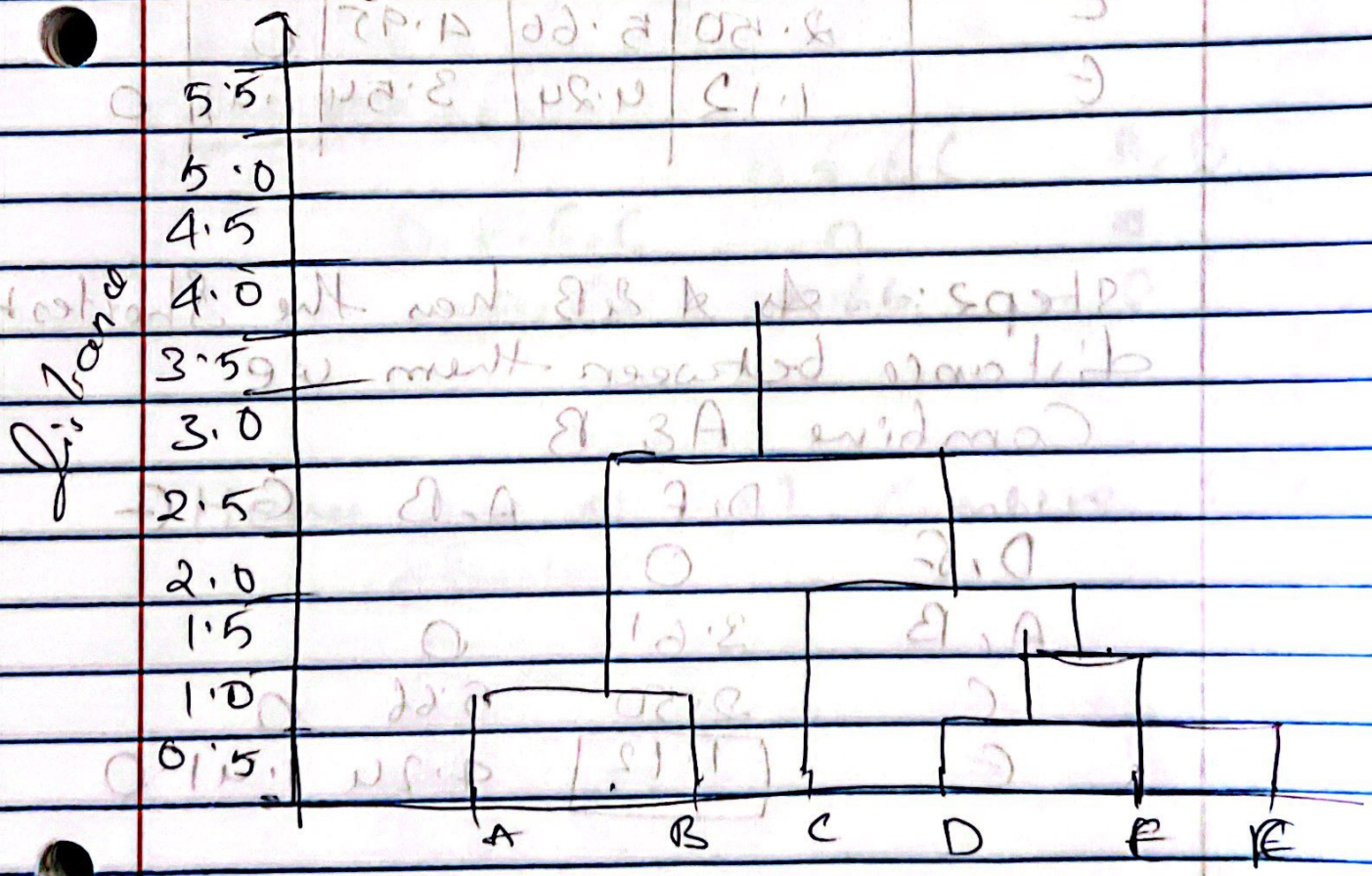
Steps:- from the distance matrix as (F,D), E) are having the shortest distance Hence we combine (F,D,E)

	F,D,E	A,B	C
F,D,E	0	2.50	1.41
(A,B)		0	4.95
C			0

Step 4:- Here we combine (F, D, C) & (A, B) which are having shortest distance from the above distance matrix

	(F, D, E, C)	(A, B)
F, D, E, C)	0	2.50
(A, B)		0

Step 3:- At last we combine ((A, D), E) & (A, B)



(b) Complete link hierarchical clustering

~~Step2~~

As from the table

Step 1: As D&F having shortest distance in whole data matrix we combine D&F.

	D,F	A	B	C	E
D,F	0				
A	3.61	0			
B	2.92	0.91	0		
C	2.50	5.66	4.95	0	
E	1.12	4.24	3.54	1.41	0

Step2: As A & B has the shortest distance between them we Combine A & B

	D,F	A,B	C	E
D,F	0			
A,B	3.61	0		
C	2.50	5.66	0	
E	1.12	4.24	1.41	0

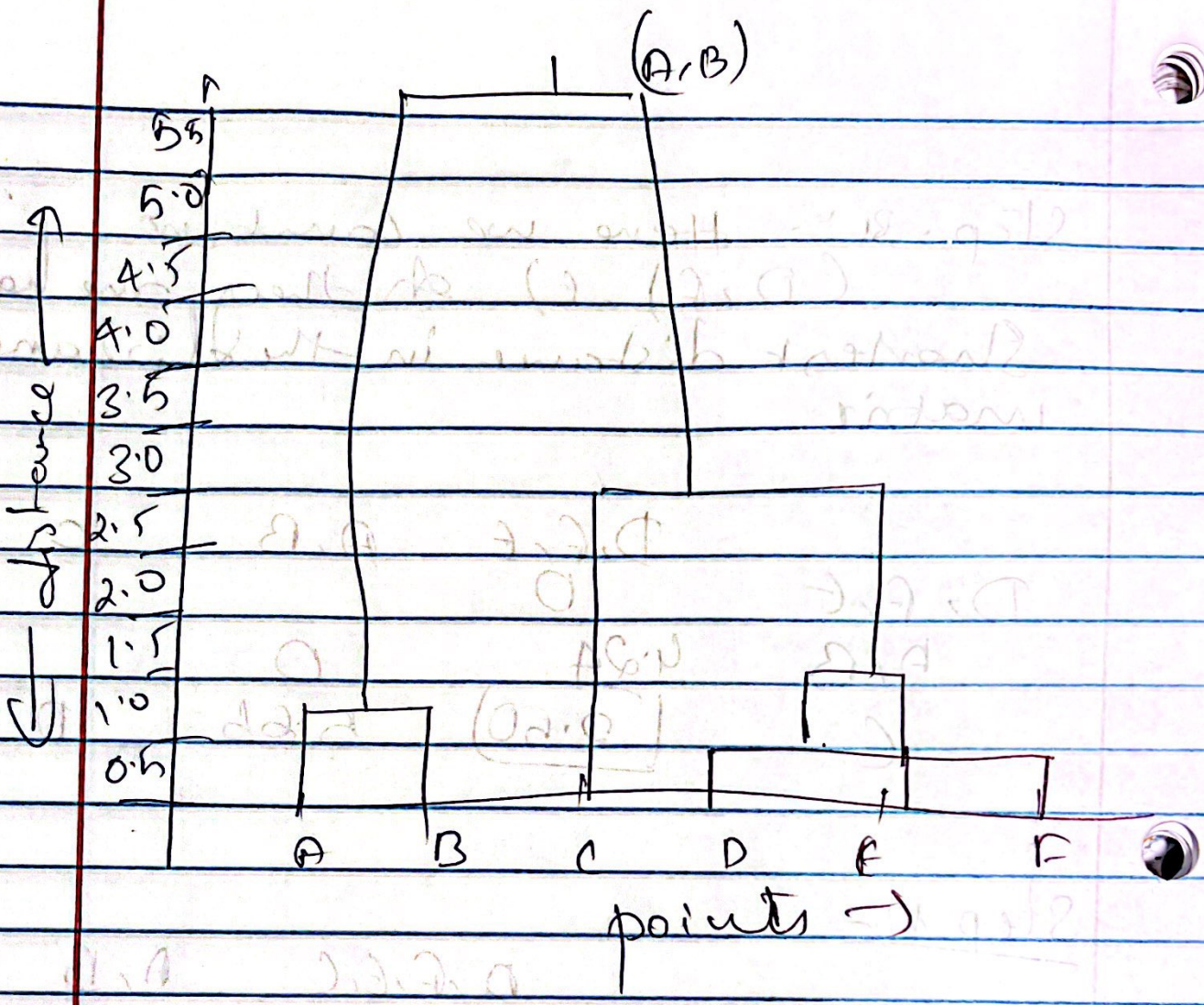
Step-3:- Here we combine (D,F),E) as they are having shortest distance in the distance matrix

	D,F,E	A,B	C
D,F,E	0		
A,B	4.24	0	
C	2.50	5.66	0

Step 4:-

	D,F,E,C	A,B
D,F,E,C	0	5
A,B	5.66	0

Hence we at last Combine to one cluster.



4) Association Rule.

Transaction id	Items Purchased
1	A, B, D, E
2	B, C, D
3	A, D, F
4	A, B, C, D, F
5	A, B
6	C, E, F

@ the Apriori algorithm on the above table with min Support = 0.25

Items	Support	min-Support
A	4	0.6
B	4	0.6
C	3	0.5
D	4	0.6
E	2	0.3
F	3	0.5

i ϕ = (All items are generated)
ii No items are pruned - ϕ
iii ϕ - No items are pruned after scan
iv) {A}, {B}, {C}, {D}, {E}, {F}

2nd Iteration

Here k = 2, So take all the item set of length '2'

Items	Support	min. Support
{A, B}	3	0.5
{A, C}	1	0.16 x
{A, D}	3	0.5
{A, E}	1	0.16 x
{A, F}	2	0.3
{B, C}	2	0.3
{B, D}	3	0.5
{B, E}	1	0.16
{B, F}	1	0.16
{C, D}	2	0.3
{C, E}	1	0.16
{C, F}	2	0.3
{D, E}	1	0.16

{D, F}	2	2/6 = 0.3
{E, F}	1	1/6 = 0.16 ✗

(i) a) - All items are generated
ii) b) - No items are pruned before Scan
iii) {A, C}, {A, E}, {B, E}, {B, F} {C, E}, {D, E}, {E, F} are pruned after Scan, as they don't meet the min support of '0.25'.

(iv) {A, B}, {A, D}, {A, F}, {B, C} {B, D}, {C, D} {C, F}, {D, F}

After Pruning the item set present

items	min- Support
{A, B}	0.5
{A, D}	0.5
{A, F}	0.3
{B, C}	0.3
{B, D}	0.5
{C, D}	0.3

{C, F} = 0.3
{D, F} = 0.3

(ii) rd iteration

Hence K=3, so take all the items set of length l=3 from the pruned set

Items	Pruned	min support
{A, B, D}	No	0.3
{A, B, F}	Pruned	
{A, D, F}	No	0.3
{A, B, C}	pruned	
{B, C, D}	No	0.3
{C, D, F}	No	0.16
{A, C, D}	pruned	
{A, C, F}	pruned	
{B, C, F}	pruned	
{B, D, F}	pruned	

Items not even generated are

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

{C, D, E}, {C, E, F}, {D, E, F}

ii) Items pruned without Scan are
{A, B, F}, {A, B, C} {A, C, D}
{A, C, F}, {B, C, F} {B, D, F}

iii) Items pruned after Scan
{C, D, F}

iv) frequent items sets are
{A, B, D}, {A, D, F} {B, C, D}

{A, B, D}, {A, D, F}, {B, C, D}
are find frequent sets are
we can't find frequent
sets further.

The frequent items are {A, B}
{A, D}, {A, F}, {B, C}, {B, D}
{C, D}, {C, F}, {D, F}

$A \rightarrow B$ = Confidence = $\frac{\text{Support}(A \cup B)}{\text{Support}(A)}$

The Association rules are :-

$\{A\} \rightarrow \{B\}$ ⇒ Confidence = $\frac{\sigma(A,B)}{\sigma(A)} = 0.75$

$\{B\} \rightarrow \{A\} = \frac{\sigma(A,B)}{\sigma(B)} = 3/4 = 0.75$

ii: $\{A\} - \{D\}$

Confidence = $\frac{\sigma(A,D)}{\sigma(A)} = 3/4 = 0.75$

iv) $\{D\} - \{A\}$

confidence = $\frac{\sigma(D,A)}{\sigma(D)} = 0.75$

v) $\{F\} \rightarrow \{A\}$

Confidence = $\frac{\sigma(F,A)}{\sigma(F)} = 0.66$

vi) $\{A\} \rightarrow \{F\} = \frac{\sigma(A,F)}{\sigma(A)} = 2/4 = 0.5$

vii) $\{B\} \rightarrow \{C\}$ $C = \frac{\sigma(B,C)}{\sigma(B)} = 0.5$

(viii) {C} → {B}

Confidence = $\frac{\sigma(C,B)}{\sigma(C)}$ = 0.6

{B} → {D}

Confidence = $\frac{\sigma(B,D)}{\sigma(B)}$ = 3/4 = 0.75

(x) {D} → {B}

Confidence = $\frac{\sigma(D,B)}{\sigma(D)}$ = 3/4 = 0.75

(xi) {C} → {D}

Confidence = $\frac{\sigma(C,D)}{\sigma(C)}$ = 0.66

(xii) {D} → {C}

Confidence = $\frac{\sigma(D,C)}{\sigma(D)}$ = 0.5

(xiii) {C} → {F}

Confidence = $\frac{\sigma(C,F)}{\sigma(C)}$ = 0.66,

{F} → {D}

Confidence: $\frac{\sigma(F \cup D)}{\sigma(F)} = 2/3 = 0.66$

1002022847

Data mining Assignment

(1) Classification Evaluation

we have developed two classifiers (1 and C2 which determine the posterior probability of movies nominated for the academy Awards. The given we can take nominated as (+) and not nominated as (-) classes

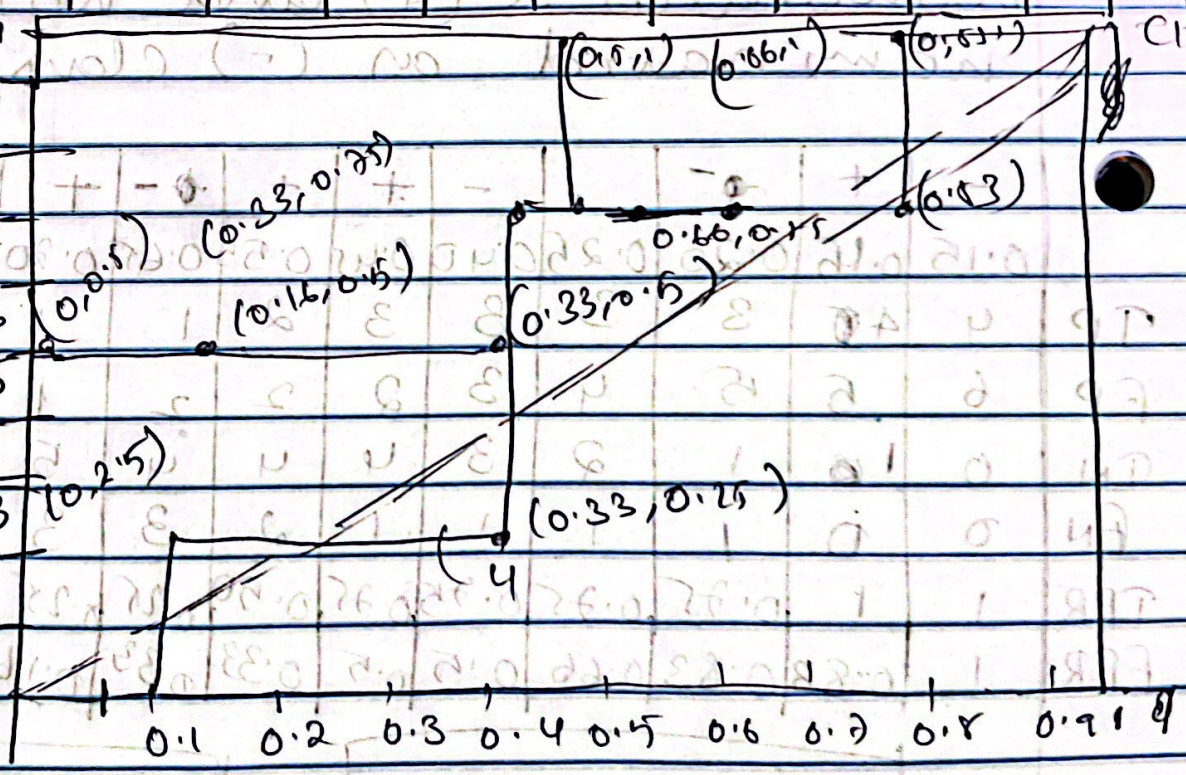
	-	+	-	-	-	+	+	-	+	-	-
	0.15	0.15	0.20	0.25	0.40	0.45	0.55	0.65	0.70	0.75	1.00
TP	4	4	3	3	3	3	2	1	1	0	0
FP	6	5	5	4	3	2	2	2	1	1	0
TN	0	1	1	2	3	4	4	4	5	5	6
FN	0	0	1	1	1	1	2	3	3	4	4
TPR	1	1	0.75	0.75	0.75	0.75	0.50	0.25	0.25	0	0
FPR	1	0.83	0.83	0.66	0.5	0.5	0.33	0.33	0.16	0.16	0

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Table for C2

	−	−	−	+	−	+	−	−	+	+
Φ	0.25	0.30	0.35	0.55	0.60	0.65	0.75	0.80	0.85	0.90	1.00
TP	4	4	4	4	3	3	2	2	2	1	0
FP	6	5	4	3	3	2	2	1	0	0	0
TN	0	1	2	3	3	4	4	5	6	6	6
FN	0	0	0	0	1	1	2	2	2	3	4
TPR	1	1	1	1	0.75	0.75	0.5	0.5	0.5	0.25	0
FPR	1	0.83	0.66	0.5	0.5	0.33	0.33	0.16	0	0	0



(b) Calculate the F-measure for all the threshold values in C

As we know that

$$\text{f-measure} = \frac{2(TP)}{2(TP)+FN+FP}$$

At θ is 0.15

$$\text{f-measure} = \frac{2(4)}{2(4)+5+0} = 8/14 = 0.57$$

At θ is 0.45

$$\text{f-measure} = \frac{2(3)}{6+2+1} = 6/9 = 0.66$$

Similarly calculate for other values 'θ' value at (0.45) given as the highest f-measure

Confusion matrix at $\theta = 0.45$

Actual class	Predicted class: Nominated	Predicted class: not-nominated
class: nominated (+)	TP(a) 3	FN(b) 1
class: NOT-nominated (−)	FP(c) 2	TN(d) 4

Precision:

$$(p) = \frac{a}{a+c} = 3/3+2 = 3/5$$

recall

$$(r) = \frac{a}{a+b} = \frac{3}{3+1} = 3/4$$

f-measure (F):

$$F = \frac{2a}{2a+b+c} = \frac{2(3)}{2(3)+1+2}$$

$$= 6/9$$

① At $\theta = 0.65$ in C1

The matrix values are

FP(a) = FN(b) = 3, FP(c) = 2

Given the weight matrix with

w1 = 1, w2 = 10, w3 = 4, w4 = 1

We know that

weighted accuracy = $\frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$

$$\frac{1(1)+4(1)}{1(1)+3(10)+2(4)+4(1)}$$

$$\frac{5}{43}$$

$$= 0.116$$

At $\theta = 0.65$ in C2

The confusion matrix values are

TP(a) = 3, FN(b) = 1 FP(c) = 1

TN(d) = 4

Given

$w1 = 1, w2 = 10, w3 = 4, w4 = 1$

Now,

$$WA = \frac{1(3)+4(1)}{1(3)+10(1)+4(1)+4(1)}$$

$$\Rightarrow 7/25 \Rightarrow 0.28$$

C2 has provided the better accuracy at $\theta = 0.65$

we can conclude 'C2' Classifer is better.