ASSIGNMENT- 3

1

Shaima Patel (V00949940)
Sannath Reddy Vemula (V00949217)

SENG 474, CSC 503: Assignment 3

1. (9 pt) Consider the dataset in Fig 1, with points belonging to two classes, blue squares and red circles.

(a) We will use Convex Hull method to draw SVM line separator as it is considered the best method

(b) We are given: $(1/2)w^2 = 2$

$$\text{Margin} = \frac{1}{\|w\|} = \frac{1}{\sqrt{w^2}} = \frac{1}{\sqrt{4}}$$

$\because w = 2$

$$\boxed{\text{margin} = 1/2}$$

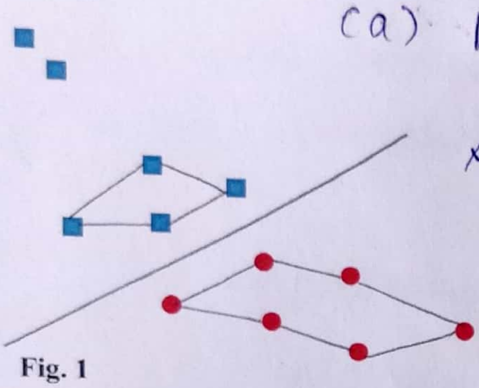


Fig. 1

(a) [1 pt] Draw (approximately) the SVM line separator.
(b) [1 pt] Suppose we find $(1/2)*\mathbf{w}^2$ to be 2 in the SVM optimization. What is the margin, i.e. the distance of closest points to the line?

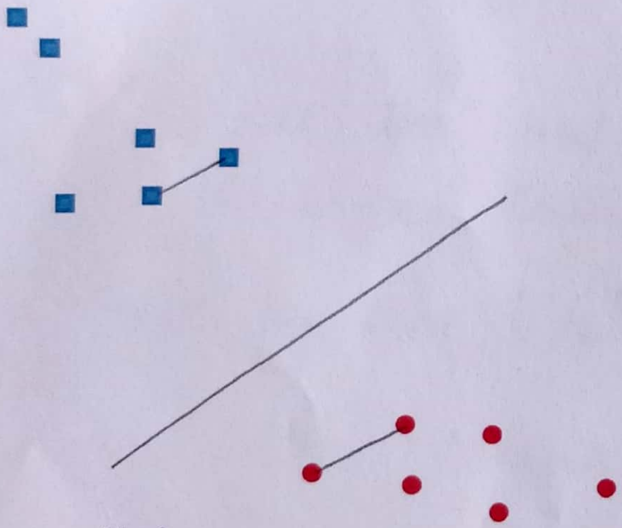


Fig. 2

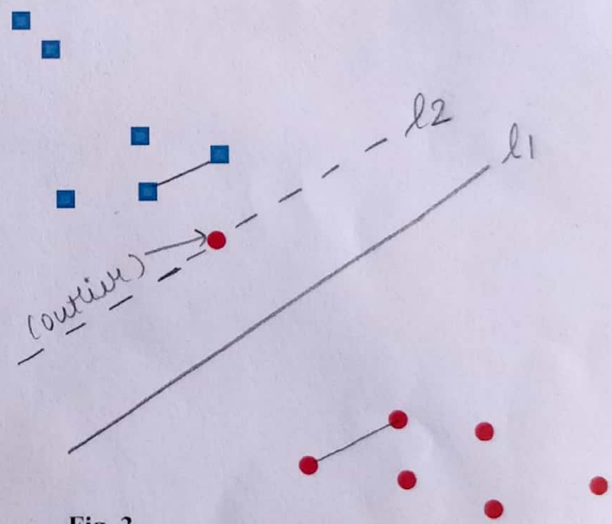


Fig. 3

(c) [1 pt] Now consider the dataset in Fig 2 (the red points are shifted below). Will $(1/2)*\mathbf{w}^2$ be smaller or greater than previously? Explain.
(d) [2 pt] Using a ruler, and the fact that $(1/2)*\mathbf{w}^2$ was 2 previously, find (approximately) the magnitude of the new line coefficient vector, $\mathbf{w}'$.
(e) [3 pt] Consider the dataset in Fig 3 (with one additional red circle quite close to the blue squares). Assuming optimization using slack variables and C=1, draw a line that does not perfectly separate the points, but which is nonetheless better than the line that perfectly separates the points. (Draw it in the figure, and explain why).
(f) [1 pt] Why would we rather prefer the line in (e) to the line that perfectly separates the points?

(c) Yes $(1/2)w^2$ will be smaller as when margin increases, value of $(1/2)w^2$ becomes smaller.

As we shifted red points in Fig 2. margin increased and $(1/2)w^2$ decreased

(d) We approximate the value of $(1/2)w^2 = 2$

$$\text{Margin} = \frac{1}{\|w'\|} = 2 \quad \text{(Applying value)}$$

$$\|w'\| = 1/2 \qquad \left\{ \because \text{From hint } 1/\|w'\| = 4 \times \frac{1}{\|w\|} \right.$$

(e) To see the best amongst 2 lines, let's calculate cost

Cost of line : $(1/2)w^2 + C\varepsilon$

Here Margin $= 2$

$C = 1$

$\varepsilon = 1.5$

Cost of L1	Cost of L2
$= 1/2 * (1/2)^2 + 1 * 1.5$	$= 1/2 * (2)^2$ $\{ \because$ As point not is outlier $\varepsilon = 0$.
$\boxed{L1 = 1.6}$	$\boxed{L2 = 2}$

As we can see that cost of L1 < cost of L2,

$\therefore$ we select $\boxed{L1}$

(f) We will choose line L1 and not line L2 because L1 gives less cost and better margin. It also can easily classify new data if entered.

Question - 2

Let's assume:

Total samples in data = 2n
" " Test data = n
" " Train data = n
" " Positive samples = n

(a) Our dataset have equal (ie n) +ve & -ve records, so our confusion matrix will have n/2 = +ve = -ve

∴ Testing dataset ⇒
Confusion Matrix :

		Predicted -ve	Predicted +ve
Actual	-ve	0	n/2
	+ve	0	n/2

$$\text{Accuracy} = \frac{n/2 + 0}{0+0+n/2+n/2} = \frac{n/2}{n}$$

$$\therefore \text{Accuracy} = 1/2 = 50\%$$

Error Rate = 1 - Accuracy

$= 1 - 1/2$

$= 1/2 = 50\%$

(b) now given ;

- Probability of +ve test sample = 80% = 0.8
  " " -ve " " = 20% = 0.2

- +ve samples will predict ⇒ $0.8 \times n/2 = \boxed{0.4n}$ +ve ; $\frac{n}{2} - 0.4n = \boxed{0.1n}$ -ve

- -ve samples will predict ⇒ $0.2 \times n/2 = \boxed{0.1n}$ -ve ; $n/2 - 0.1n = \boxed{0.4n}$ +ve

- Confusion Matrix :

		Predicted -ve	Predicted +ve
Actual	-ve	0.1n	0.4n
	+ve	0.1n	0.4n

- Accuracy $= \frac{0.1n + 0.4n}{n} = \frac{0.5n}{n} = \boxed{50\%}$

- Error Rate = 1 - Accuracy

$= 1 - 0.5$

$= 0.5$

$= \boxed{50\%}$

(c) Now give ;

No of +ve samples $= 2/3\,n$

" " -ve " " $= 1/3\,n$

Probability of +ve samples $= 1$

- Prediction of +ve samples $= 2/3\,n * 1$

$= \boxed{2/3\,n} \Rightarrow$ +ve

$= 2/3\,n * 0$

$= \boxed{0} \Rightarrow$ -ve

- Prediction of -ve samples $= 1/3\,n * 1$

$= \boxed{1/3\,n} \Rightarrow$ +ve

$= 1/3\,n * 0$

$= \boxed{0} \Rightarrow$ -ve

- Confusion Matrix:

		Predicted -ve	Predicted +ve
Actual	-ve	0	$1/3\,n$
	+ve	0	$2/3\,n$

- Accuracy $= \dfrac{2/3\,n}{n} = 2/3$

- Error rate $= 1 - 2/3 = 1/3$

$= \boxed{33.3\%}$

(d) • Probability of +ve classified = 2/3 ⑥

• Probability of -ve classified = 1/3

• Prediction for +ve samples:

$2/3 * 2/3\, n = \boxed{4/9\, n} \Rightarrow +ve$

$2/3\, n - 4/9\, n = \boxed{2/9\, n} \Rightarrow -ve$

• Prediction for -ve samples:

$1/3\, n * 2/3\, n = \boxed{2/9\, n} \Rightarrow +ve$

$1/3\, n - 2/9\, n = \boxed{1/9\, n} \Rightarrow -ve$

• Confusion Matrix:

		Predicted -ve	Predicted +ve
Actual	-ve	1/9n	2/9n
	+ve	2/9n	4/9n

• Accuracy $= \dfrac{4/9\, n + 1/9\, n}{n}$

$= 5/9$

• Error rate $= 1 - 5/9$

$= 4/9$

$= \boxed{44.4\%}$

Question- 3

(a) ROC Graph for A & B
(∵ Descending Order)

Instances	True Class	Classifier A
1	+	0.73
2	+	0.69
5	+	0.67
4	−	0.55
6	+	0.47
9	+	0.45
3	−	0.44
10	−	0.35
8	−	0.15
7	−	0.08

Instances	True Class	Classifier B
3	−	0.68
1	+	0.61
5	+	0.45
7	−	0.38
4	−	0.31
6	+	0.09
8	−	0.05
10	−	0.04
2	+	0.03
9	+	0.01

(b) Threshold ; t = 0.5

So the values > 0.5 ⇒ +ve
values < 0.5 ⇒ −ve

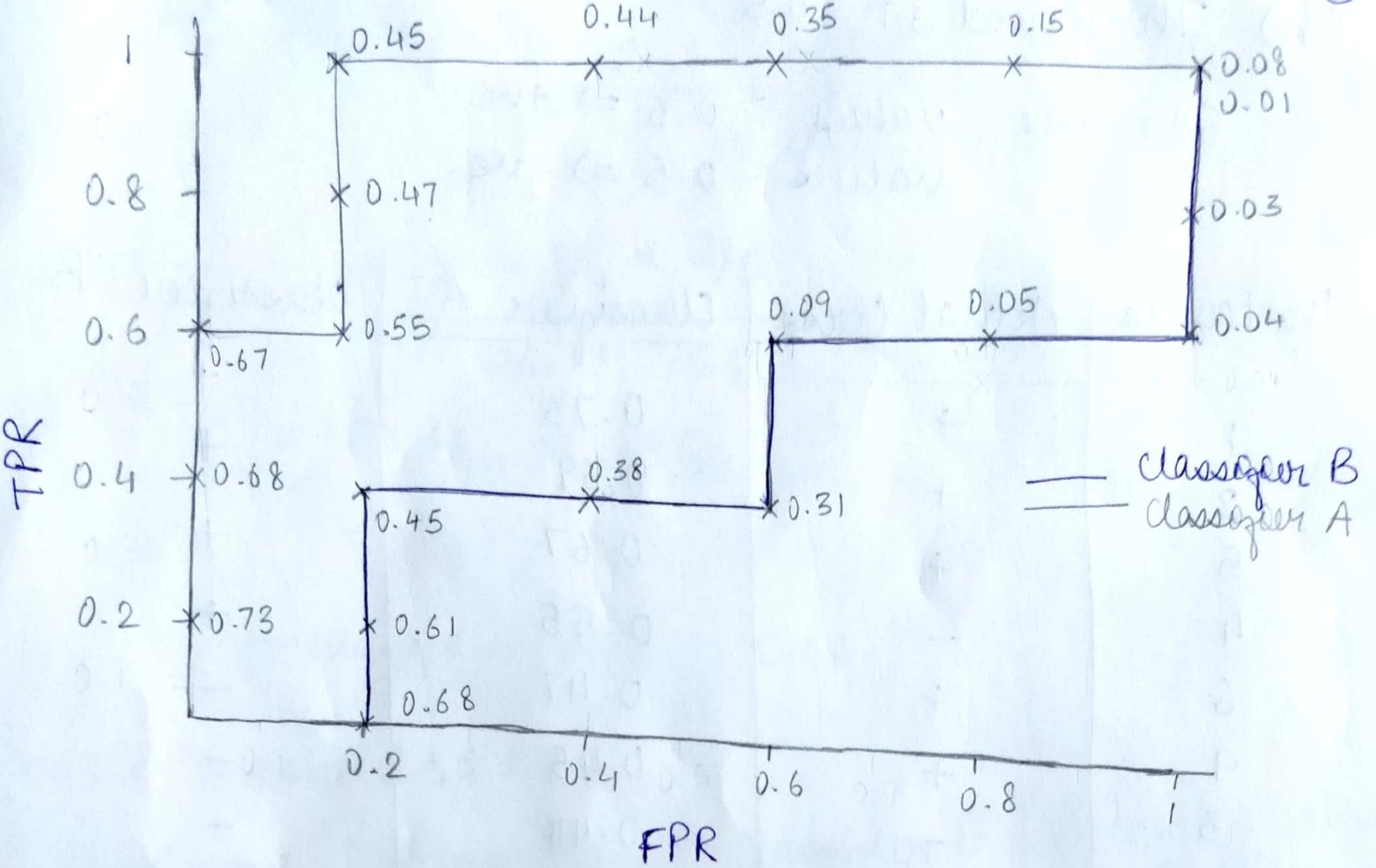
Instances	Actual Class	Classifier A	Classifier B
1	+	0.73	+
2	+	0.69	+
5	+	0.67	+
4	−	0.55	+
6	+	0.47	−
9	+	0.45	−
3	−	0.44	−
10	−	0.35	−
8	−	0.15	−
7	−	0.08	−

• Confusion Matrix:

		Predicted −ve	Predicted +ve
Actual	−ve	4	1
	+ve	2	3

• Precision $= \frac{TP}{TP+FP} = \frac{3}{3+1} = 3/4$ $\boxed{= 75\%}$

• Recall $= \frac{TP}{P} = \frac{3}{3+2} = 3/5$ $\boxed{= 60\%}$



$$TPR = \frac{TP}{\#P} \qquad FPR = \frac{FP}{\#N}$$

We have total 10 instances < 5 +ve, 5 -ve

∴ Classifier A is better than B as it covers more area.

{Larger area of Classifier → Better classifier}

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$= \frac{2 * 3/4 * 3/5}{3/4 + 3/5}$$

$$= 9/10 * 20/27$$

$$= 2/3$$

$$\text{F-measure} = 2/3 = 0.66$$

(c) Threshold, $t = 0.5$

Instance	Actual Class	Classifier B	Predicted class
3	−	0.68	+
1	+	0.61	+
5	+	0.49	−
7	−	0.38	−
4	−	0.31	−
6	+	0.09	−
8	−	0.05	−
10	−	0.04	−
2	+	0.03	−
9	+	0.01	−

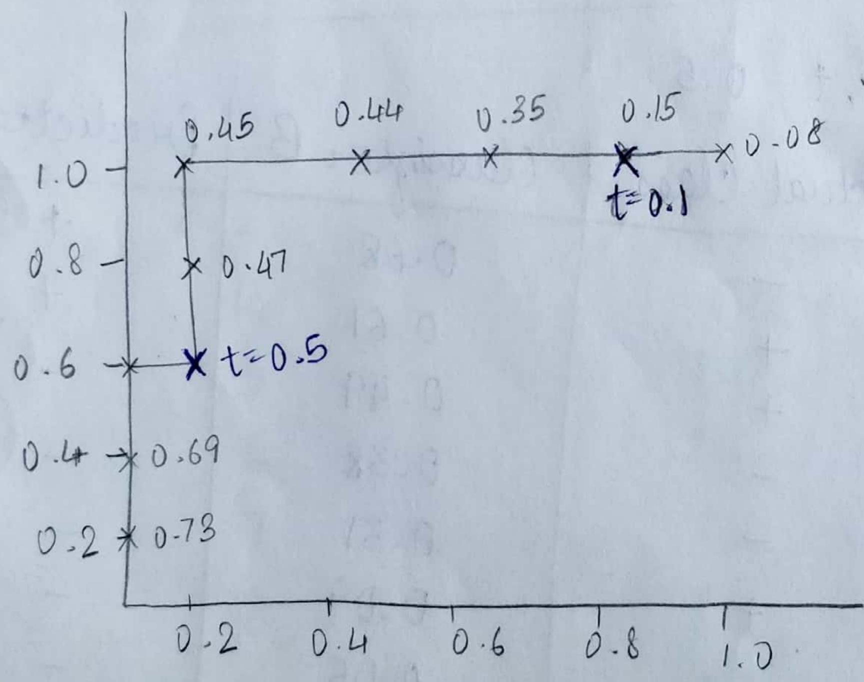
- Confusion Matrix :

		Predicted	
		−ve	+ve
Actual	−ve	4	1
	+ve	4	1

- Precision = $\frac{1}{1+1} = 1/2 = 50\%$
- Recall = $1/5 = 20\%$
- F-measure = $\frac{2 \times 1/2 \times 1/5}{1/2 + 1/5} = 2/7 = 0.28$

⇒ Classifier A is more better as it's F-measure has a higher value compared to B

(d)



Real life application when t=0.5

Used to evaluate clinical test. It is useful when considering value of test to a clinical and for prevalence of disease in medical application like breast cancer, HIV etc.

Real life application when t=0.1

When you have to make not a very important decision like selecting a cloth, movie or ice-cream.

Question-4

Apriori Algorithm

(a)

Transaction ID	Items bought
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

C1: Itemset => {a} {d}
{b} {e}
{c}

Scanned by CamScanner

F1:

Item set	Support Count	Frequent (F)
{a}	5	F
{b}	7	F
{c}	5	F
{d}	9	F
{e}	6	F

C2: {a, b} {a, c} {a, d} {a, e} {b, c} {b, d} {b, e} {c, d} {c, e} {d, e}

F2:

Item set	Support Count	Frequent (F)
{a, b}	3	F
{a, c}	2	I
{a, d}	4	F
{a, e}	4	F
{b, c}	3	F
{b, d}	5	F
{b, e}	4	F
{c, d}	4	F
{c, e}	2	I
{d, e}	6	F

Here I = Infrequent items (ie < 3) → not considered

C3: {a, b, d}
{a, b, e}
{a, d, e}
{b, c, d}
{b, d, e}

F3:

Item set	Support Count	Frequent
{a, b, d}	2	I
{a, b, e}	2	I
{a, d, e}	4	F
{b, c, d}	2	I
{b, d, e}	4	F
{a, b, c}	1	N
{a, c, d}	1	N
{a, c, e}	1	N
{b, c, e}	1	N
{c, d, e}	2	N

→ Now we generate F4; but we see that itemset is infrequent and so we won't consider it.

C4: {a, b, d, e}

F4:

Item set	Support count	Frequent
{a, b, d, e}	~~{a, b, d, e}~~ 2	N

(b) • Percentage of frequent item sets including Null

$= 16/32 \times 100$

$= 50\%$

• Percentage of frequent item sets without Null

$= 15/31 \times 100$

$= 48.38\%$

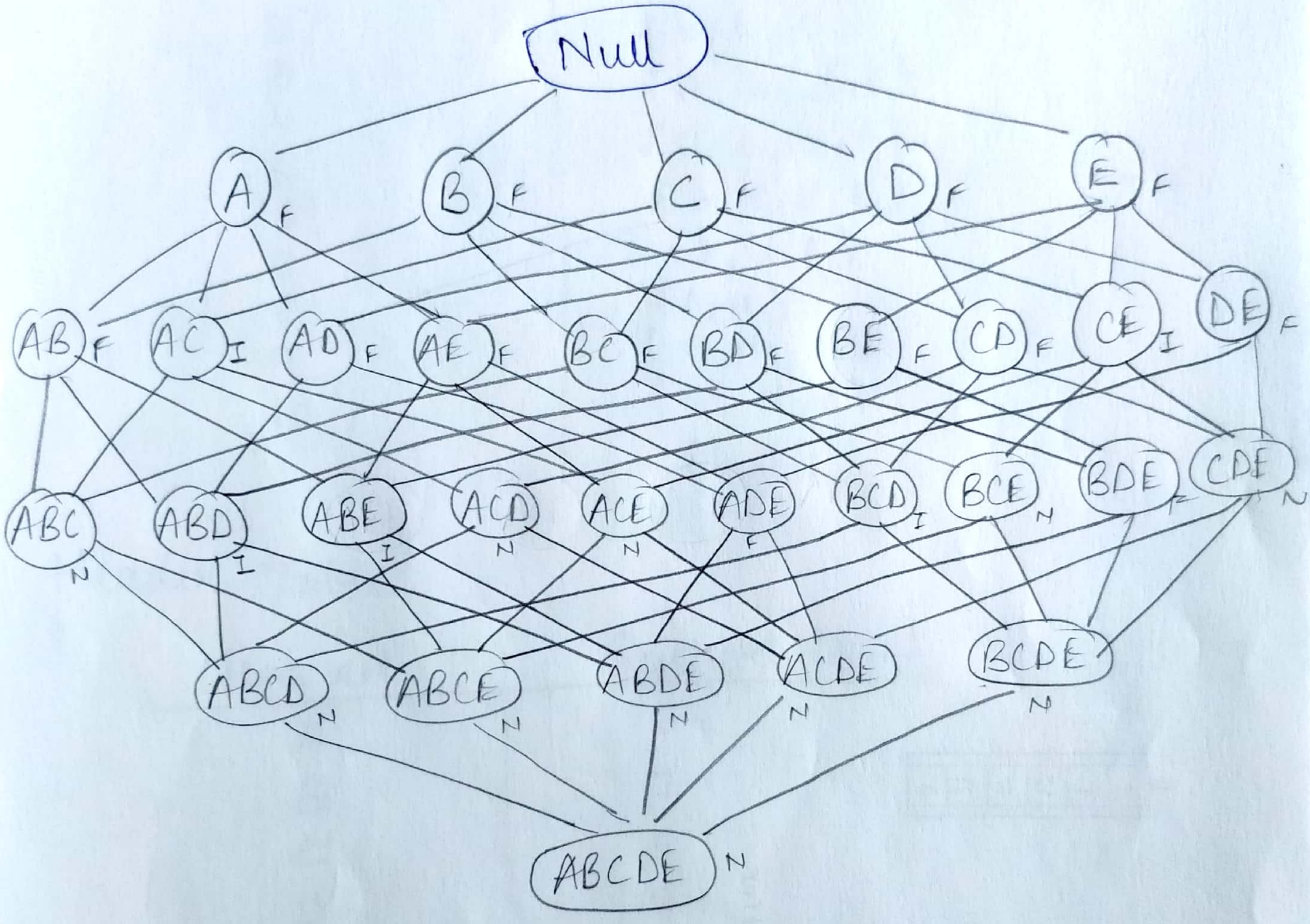
(c) Pruning Ratio

• Counting only N $= \frac{11}{32} \times 100 = 34.4\%$

• Counting I & N $= \frac{16}{32} \times 100 = 50\%$

(d) False Alarm Rate

No of infrequent itemsets = 5/32 = 15.6%



Question-5

FP Trees

min - sup - count = 3

Transaction ID	Items Bought
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

Header Table

Item set	Support Count
A	5
B	7
C	5
D	9
E	6

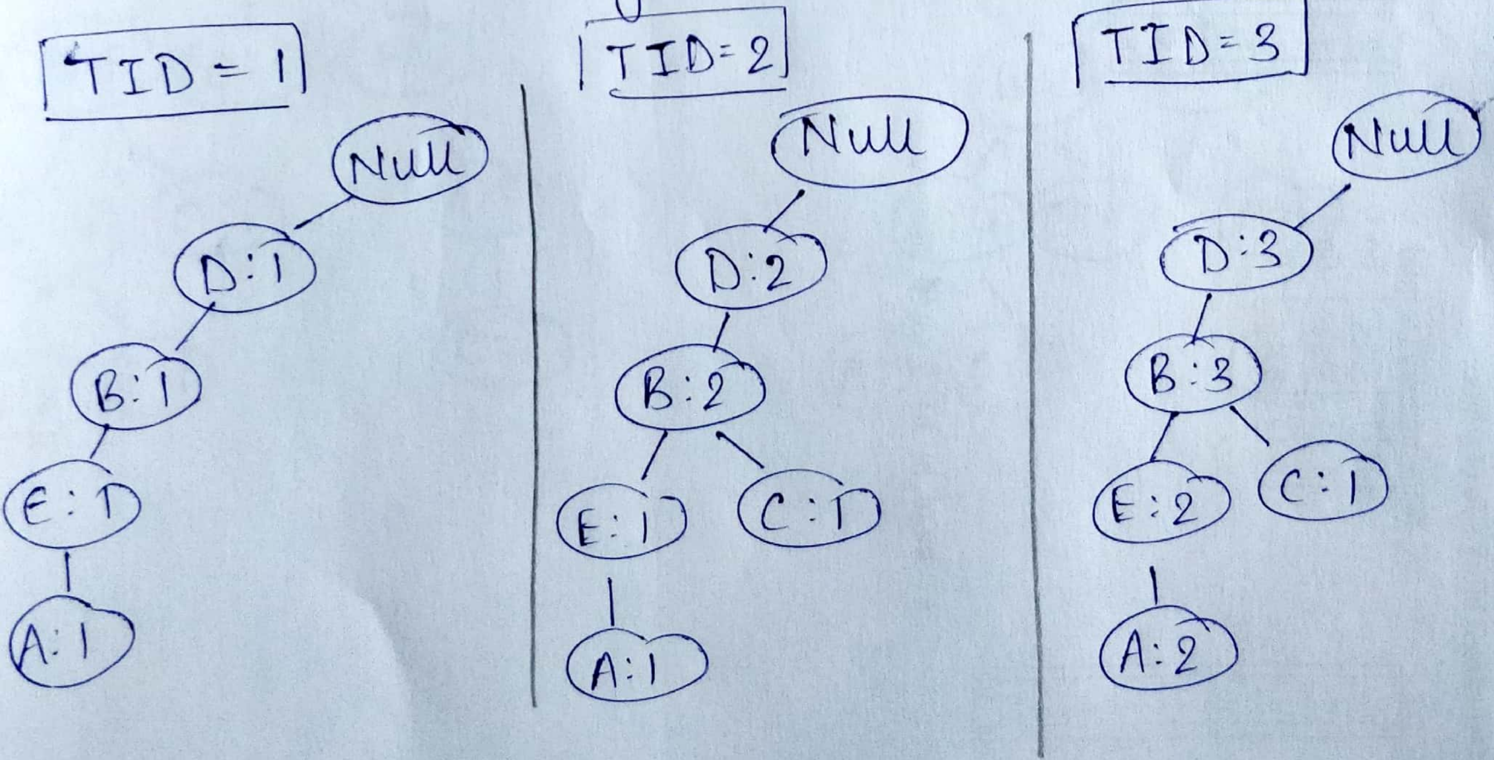
Sorted Header Table

Item Set	Support Count
D	9
B	7
E	6
A	5
C	5

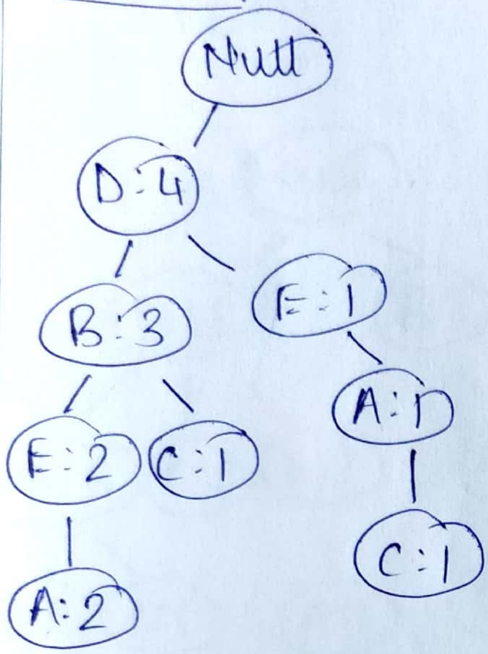
Arranging the items in descending order

Transaction ID	Items Bought
1	{d, b, e, a}
2	{d, b, c}
3	{d, b, e, a}
4	{d, e, a, c}
5	{d, b, e, c}
6	{d, b, e}
7	{d, c}
8	{b, a, c}
9	{d, e, a}
10	{d, b}

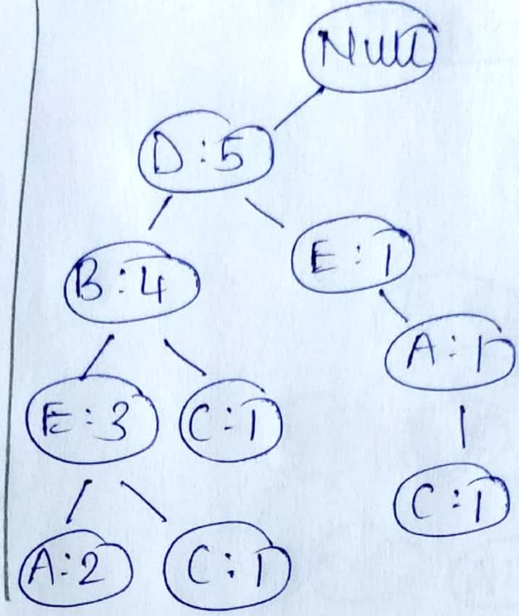
Now, constructing FP trees



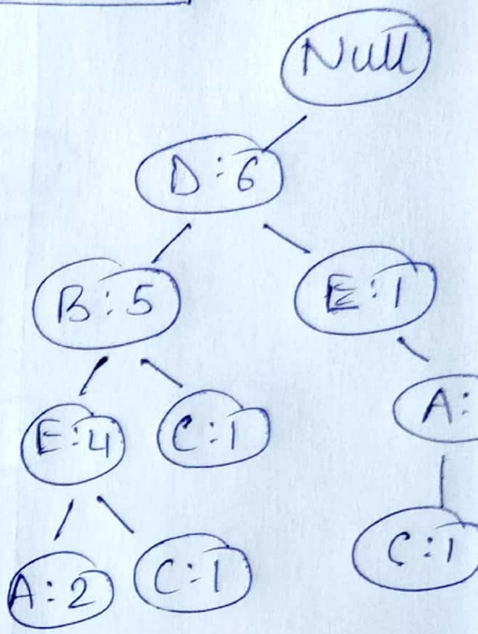
TID=4



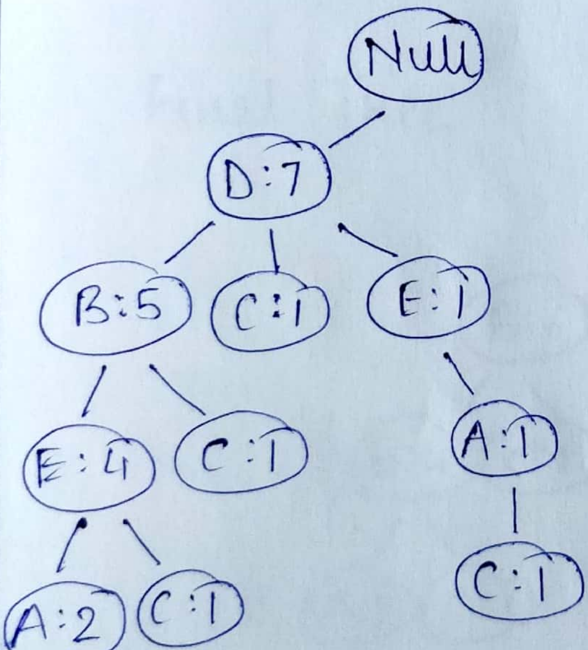
TID=5



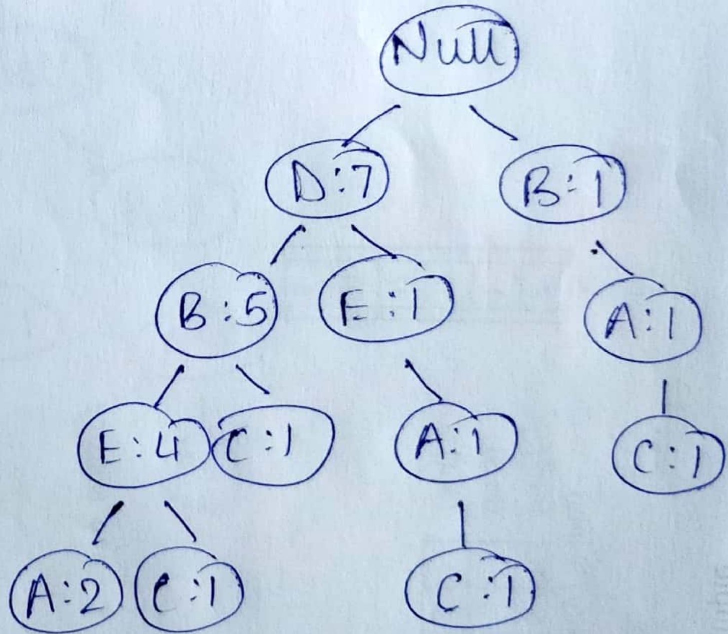
TID=6



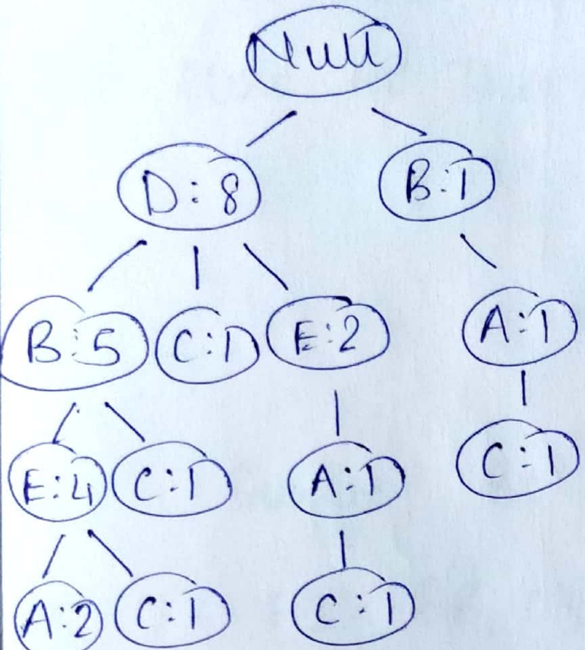
TID=7



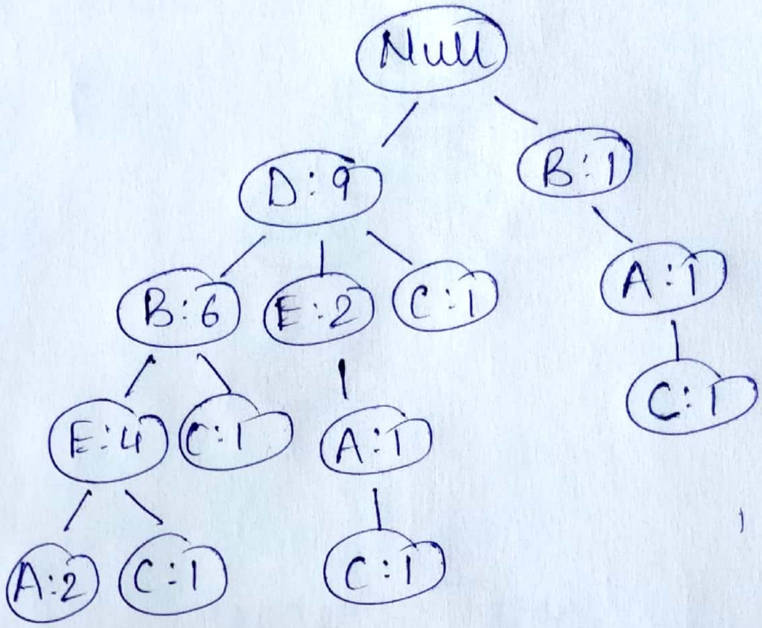
TID=8



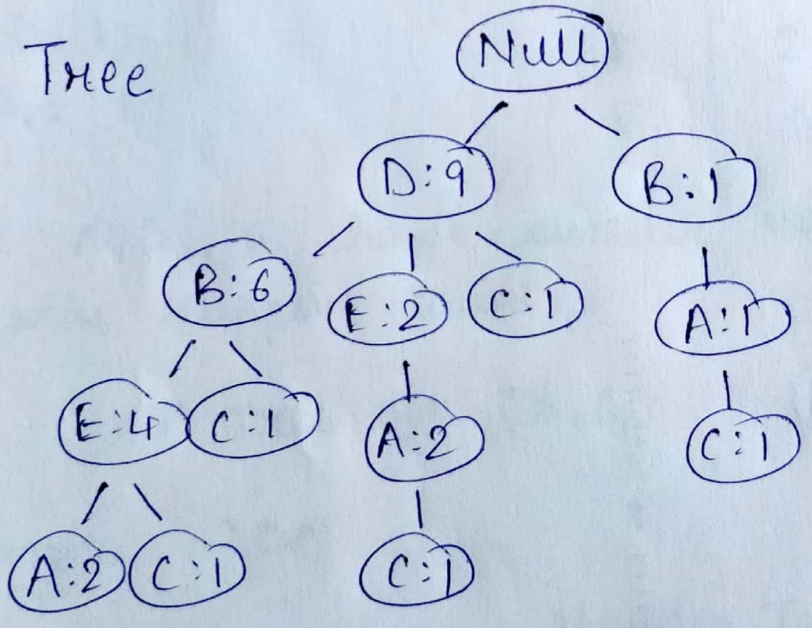
TID=9



TID=10



Final Tree



FI= {C}

B, A : 1

D, B : 1

D, E, A : 1

D, B, E : 1

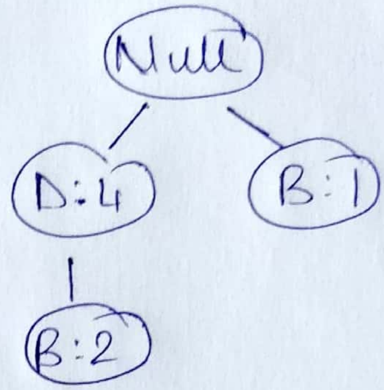
D : 1

Header Table

D	4
B	3
E	2
A	2

→ Here E & A have smaller value than minimum support count

Here FP Tree for C :



Suffix 'BC'

FI = {B, C}

A : 1
D : 1
D, E : 1

Header Table

A	1
D	2
E	1

→ Here A, D & E have smaller value than min support count

∴ FP Tree for {B, C} : Null

Suffix 'DC'

B : 1
A, E : 1
B, E : 1

Header Table

E	2
B	2
A	1

∴ <u>Pruned</u>

Frequent item sets

{c}:5

{b,c}:3

{d,c}:4

→ Now, FI = {A}

D, B, E : 1

D, E : 1

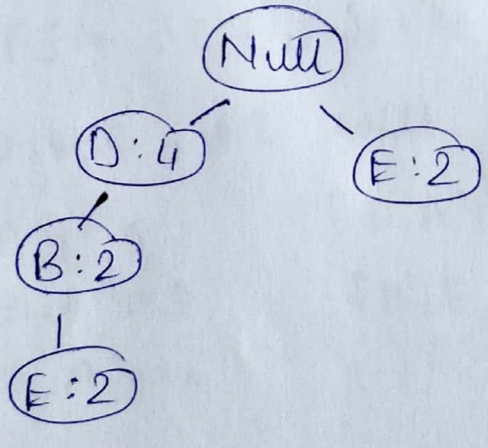
B : 1

D, E : 1

Header Table

D	4
E	4
B	3

Conditional FP tree for {A}



FI = {A} = 5

→ FI = {B, A}

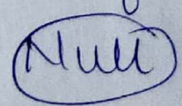
Header Table

D	2
B	2

D, E : 1

D, E : 1

Conditional Tree for {B, A}



→ FI = {E, A}

Header Table

D	4

D:1
D:1
D:1
D:1

Conditional tree for {E, A}:

Null
/
D:4

→ FI = {D, A} = 4

Null

{D, E, A} will be removed 4 times

→ So, FI = {D, E, A} = 4

Therefore, FI will be:

{A} = 5
{B, A} = 3
{E, A} = 4
{D, A} = 4
{D, E, A} = 4

FI = {E} :

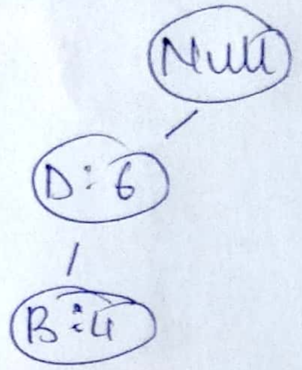
TID	
1	D, B : 1
3	D, B : 1
4	D : 1
5	D, B : 1
6	D, B : 1
9	D : 1

Header Table

D	6
B	4

Scanned by CamScanner

Conditional FP tree for 'E' would be:



FI = {E} = 6

→ FI = {B, E}

TID : 1 D : 1
3 D : 1
5 D : 1
6 D : 1

Header Table

D | 4

FP Tree : Null

→ FI = {D, E}

FP Tree : Null

Header Table

Null

→ FI = {D, B, E}

Header Table

Null

→ Suffix 'B'

TID : 1 D:1
2 D:1
3 D:1
5 D:1
6 D:1
9 D:1
10 D:1

Header Table

D | 7

FI = {B}
= 7

→ Suffix 'DB'

TID: 1 : 1
2 : 1
3 : 1
5 : 1
6 : 1
10 : 1

Header Table

Null

FP Tree: Null

FI = {D, B}
= 6

→ Suffix 'D':

TID: 1 : 1
2 : 1
3 : 1
4 : 1
5 : 1
6 : 1
7 : 1
9 : 1
10 : 1

Header Table

Null

FP Tree: Null

→ The final list of frequent item sets would be as follows:

{c}: 5	{d, e, a}: 4	{b}: 7
{b, c}: 3	{d, a}: 4	{d, b}: 6
{d, c}: 4	{e}: 6	{d}: 9
{a}: 5	{b, e}: 4	
{b, a}: 3	{d, e}: 6	
{e, a}: 4	{d, b, e}: 4	

Question-6

→ Single Link Clustering

Minimum distance formula:

min(distance (P1, P2), P3)

≅ max (similarity (P1, P2), P3)

Now, we will merge P2 & P5 at point

(P2, P5) = 0.98

w.r.t. P1 : max (similarity (P2, P5), P1)

max (0.10, 0.35)

0.35

w.r.t. P3 : max (similarity (P2, P5), P3)

max (0.64, 0.85)

0.85

w.r.t. P4 : max (similarity (P2, P5), P4)

max (0.47, 0.76)

0.76

After merging, max similarity is:

((P2, P5), P3) = 0.85

Now updating the matrix we will get:

w.r.t. P1

max(similarity (P2, P5), P3), P1)

max (0.35, 0.41)

0.41

w.r.t. P4

max (0.76, 0.44)

0.76 ⇒ max similarity

ie [(P2, P5, P3), P4] = 0.76

Updating Matrix :

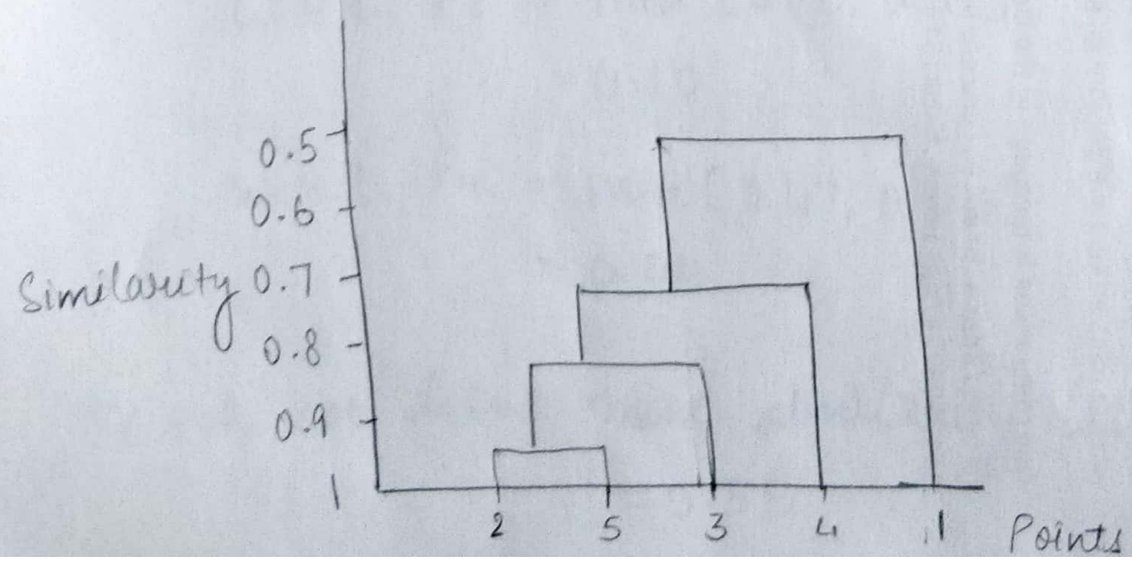
w.r.t. P1

max(simularity ((P2, P5, P3), P4), P1)

max(0.41, 0.55)

0.55

Now, we will construct Dendrogram



→ Complete Link Clustering

Applying Maximum distn formula:

= min (similarity (P1, P2), P3)

max similarity is at pt (P2, P5) = 0.98

w.r.t. P1 = min [sum (P2, P5), P1]

= min [0 (0, 0.35)]

= 0.10

w.r.t. P3 = min [0.64, 0.85]

= 0.64 ⇐ max similarity

w.r.t. P4 = min [0.47, 0.76]

= 0.47

We have max similarity at [(P2, P5), P3]

= 0.64

Now, we update the matrix:

w.r.t. P1 = min [0.10, 0.41]

= 0.10

w.r.t. P4 = min [0.47, 0.44]

= 0.44

Now, we have max similarity at point [P4, P1]

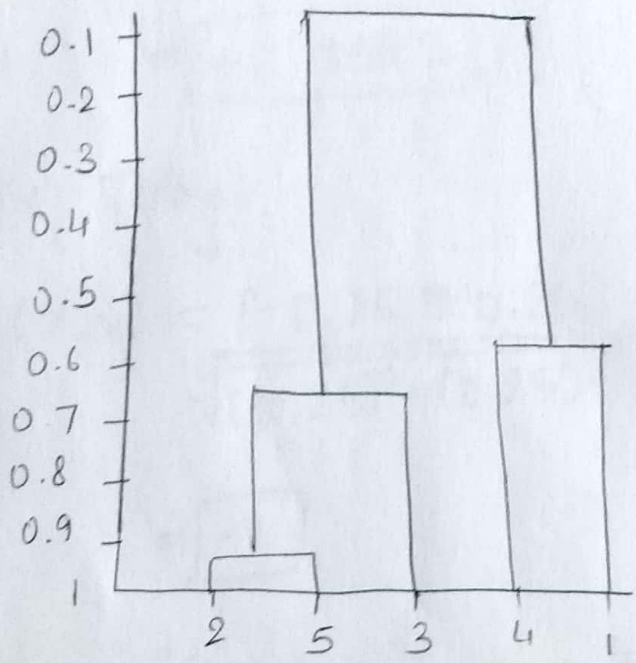
= 0.55

Update matrix w.r.t. P2, P3, P5

min [0.44, 0.10] = 0.10

Hence, max similarity at pt.
[P2, P5, P3, P4, P1] = 0.10

→ Dendrogram :



$\boxed{\text{Question - 7}}$

(a) User-User similarity with Pearson correlation

Formula:

$$sim(x,y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x-\bar{x})^2}\sqrt{\sum_{i=1}^{m}(y-\bar{y})^2}}$$

(i) For Lisa Rose:

$$sim(x,y) = \frac{(-0.75 * -0.625) + (-0.25 * 0.375) * (0.25 * 0.375) + (0.75 * -0.125)}{\sqrt{(0.75)^2 + (-0.25)^2 + (0.25)^2 + (0.75)^2} * \sqrt{(-0.625)^2 + (0.375)^2 + (0.375)^2 + (-0.125)^2}}$$

$$= \boxed{-0.2582}$$

(ii) For Toby:

$$sim(x,y) = \frac{(-0.25 * 0.25) + (0.25 * -0.25)}{\sqrt{(-0.25)^2 + (0.25)^2} \sqrt{(0.25)^2 + (-0.25)^2}}$$

$$= \boxed{-1}$$

(iii) Gene Seymour:

$$sim(x,y) = \frac{(-0.75 * 0.625) + (-0.25 * 0.125) + (0.25 * 1.375) + (0.75 * -0.625)}{\sqrt{(-0.75)^2 + (-0.25)^2 + (0.25)^2 + (0.75)^2} * \sqrt{(-0.625)^2 + (-0.125)^2 + (1.375)^2 + (0.629)}}$$

$$= \boxed{0.2046}$$

(iv) Claudia Puig:

$$Sim(x,y) = \frac{(-0.5 \times -0.5) + 0 + (-0.5 \times -0.5)}{\sqrt{(0.5)^2 + (0.5)^2} * \sqrt{(0.5)^2 + (0.5)^2}}$$

$$= \boxed{1}$$

(v) Jack Matthews:

$$sim(x,y) = \frac{(-0.75 \times -0.75) + (-0.25 \times 0.25) + (0.25 \times 1.25) + (0.75 \times -0.75)}{\sqrt{(-0.75)^2 + (0.25)^2 + (0.25)^2 + (0.75)^2} + \sqrt{(-0.75)^2 + (0.25)^2 + (1.25)^2 + (0.25)^2}}$$

$$= \boxed{0.1348}$$

$$\Rightarrow \hat{r}_{ui} = \frac{\sum_{v \in u_i} r_{v,i} \cdot sim_{v,u}}{\sum_{v \in u_i} sim_{v,u}}$$

$$\hat{r}_{ui} = \frac{(0.4045 * 2.5) + (0.2046 * 3.5) + (1 * 2.5) + (3.5 * 0.1348)}{0.4045 + 0.2046 + 1 + 0.1348}$$

$$= 2.6946$$

$$\boxed{r_{v_i} = 2.7}$$

	You	Me	Dupree
Michael Phelps		2.7	

(b) $\lambda_1 = \lambda_2 = \gamma = 0.1$ (Given)

Initially, we will take $b_{\gamma} = b_i = 0$

Formula: $b_{\gamma} = b_{\gamma} + \gamma(e_{\gamma i} - \lambda b_{\gamma})$

$e_{\gamma i} = r - [\mu + b_r + b_i]$

$b_{\gamma} = b_{\gamma} + \gamma(r - \mu - b_{\gamma} - \lambda b_{\gamma})$

$\mu = \sum_{i=1}^{n} r_i = \frac{113}{35}$

$\boxed{\mu = 3.2285}$

(i) "Lady in the water"

$b_{\gamma} = 0 + 0.1(2.5 - 3.2285 - 0 - 0.180)$

$\boxed{b_{\gamma} = -0.0728}$

(ii) "Snakes on a phone"

$b_{\gamma} = -0.0728 + 0.1(3.5 - 3.2285 + 0.0728 + 0.1 \times 0.7285)$

$\boxed{b_{\gamma} = -0.0377}$

(iii) "Just my luck"

$\boxed{b_r = -0.05641}$

(iv) "Supreme Return"

$\boxed{b_r = -0.0230}$

(v) "You, me and Dupree"

$\boxed{b_r = -0.09338}$

(vi) "The night listener"

$\boxed{b_r = -0.10596}$

Hence, the value after 1st pass would be:

$\boxed{b_r = -0.10596}$