Automated Learning and Data Analysis

Team G40

TEAM MEMBER 1 Sujith Tumma,stumma2

TEAM MEMBER 2 Vishnu Challa,vchalla2

TEAM MEMBER 3 Indu Chenchala,ichench

April 26, 2022

1 SVM

1. 1a) Support vector machines (SVM) learn a decision boundary leading to the largest margin between classes. In this question, you will train a SVM on a tiny dataset with 4 data points, shown in Figure 1. This dataset consists of two points with Class 1 (y = -1) and two points with Class 2 (y = 1). Each data point has two non-class attributes: x1 and x2.

   i) (5 points) Find the weight vector w and bias w0 for the decision boundary of a hard-margin SVM. What is the equation corresponding to this decision boundary? Show your work, including the equations you used to derive your answer

1) SVM

1a) Given

Margin Plane Constraint

$$y_i(w^T x_i + w_0) - 1 \geq 0$$

Class - 1 $(y = (-1))$

Class - 2 $(y = 1)$

Two closest points in the plane for $y = (-1)$, $P = (5, 2)$

$y = 1$, $P = (2, 4)$

$$\Rightarrow \quad -1(5w_1 + 2w_2 + w_0) - 1 = 0 \rightarrow (7)$$

$$1(2w_1 + 4w_2 + w_0) - 1 = 0 \rightarrow (8)$$

On solving both the above equations

$$5w_1 + 2w_2 + w_0 = (-1)$$

$$2w_1 + 4w_2 + w_0 = 1$$

$$(-) \quad (-) \quad (-)$$

$$3w_1 - 2w_2 = (-2) \longrightarrow (1)$$

The margin is always perpendicular to the slope of line joining our support vectors.

slope of line joining SVs $\Rightarrow \dfrac{(y_2 - y_1)}{(x_2 - x_1)} \Rightarrow \dfrac{(4-2)}{(2-5)}$

$m_1 = \left(\frac{-2}{3}\right)$

Let slope of margin be m_2

$\therefore m_1 \times m_2 = -1$

$-\frac{2}{3} \times m_2 = -1$

$$\boxed{m_2 = \frac{3}{2}} \rightarrow (2)$$

Since the equation of Margin is

$w^T x + w_0 = 0$

$w_1 x_1 + w_2 x_2 + w_0 = 0$

$$x_2 = \left(-\frac{w_1}{w_2}\right) x_1 + \left(-\frac{w_0}{w_2}\right) \rightarrow (3)$$

i.e it is in $y = mx + c$ format. Combining (2) & (3)

$$-\frac{w_1}{w_2} = \frac{3}{2}$$

$$\boxed{w_1 = \left(-\frac{3}{2}\right) w_2} \rightarrow (4)$$

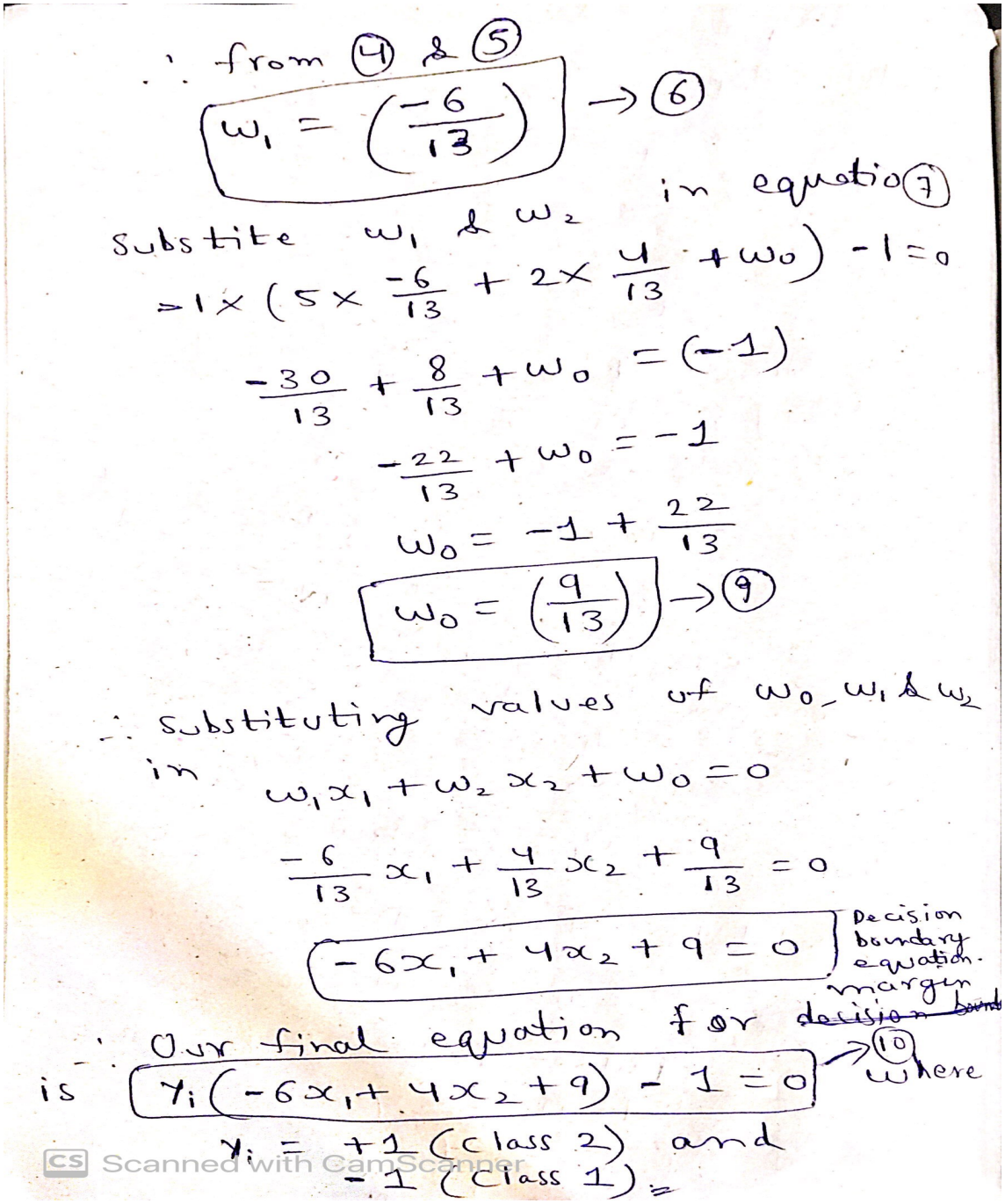
Substitute (4) in (1)

$$3 \times \frac{-3}{2} w_2 - 2 w_2 = -2$$

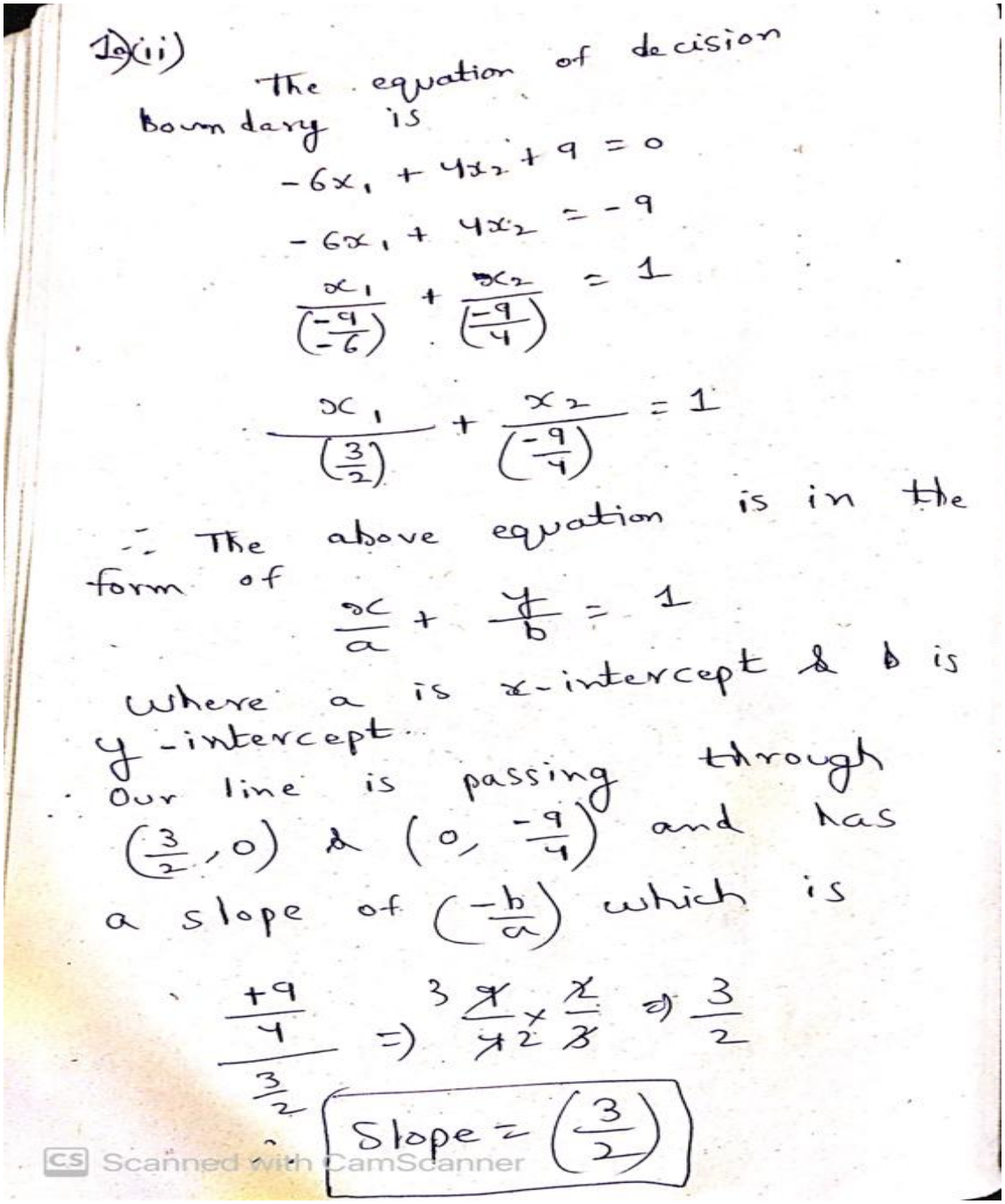
$$-9 w_2 - 4 w_2 = -4$$

$$-13 w_2 = -4$$

$$\boxed{w_2 = \frac{4}{13}} \rightarrow (5)$$



ii) (4 points) Circle the support vectors and draw the decision boundary.



1a(ii)

The equation of decision boundary is

$$-6x_1 + 4x_2 + 9 = 0$$

$$-6x_1 + 4x_2 = -9$$

$$\frac{x_1}{\left(\frac{-9}{-6}\right)} + \frac{x_2}{\left(\frac{-9}{4}\right)} = 1$$

$$\frac{x_1}{\left(\frac{3}{2}\right)} + \frac{x_2}{\left(\frac{-9}{4}\right)} = 1$$

∴ The above equation is in the form of

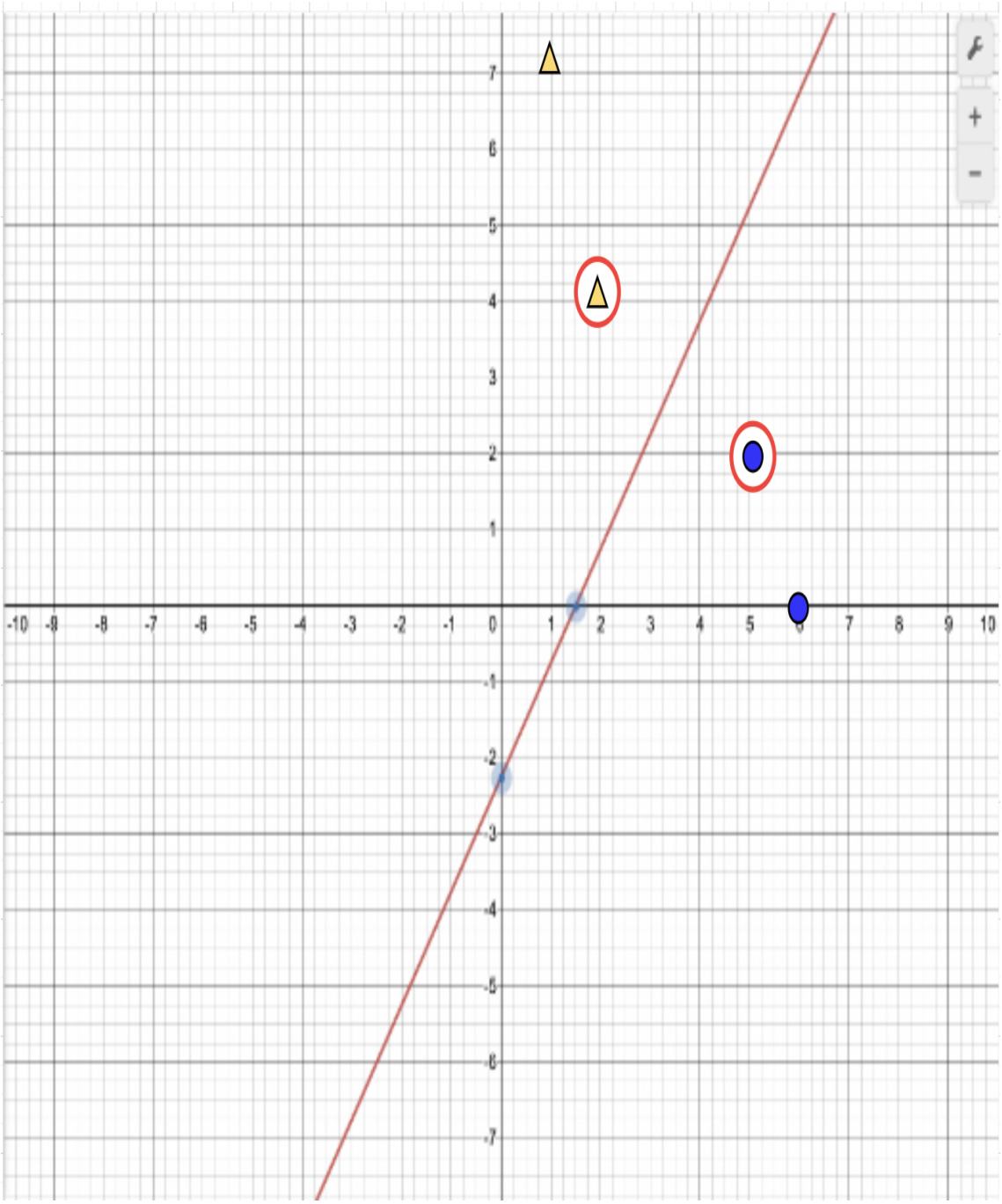
$$\frac{x}{a} + \frac{y}{b} = 1$$

where a is x-intercept & b is y-intercept.

Our line is passing through $\left(\frac{3}{2}, 0\right)$ & $\left(0, \frac{-9}{4}\right)$ and has a slope of $\left(\frac{-b}{a}\right)$ which is

$$\frac{+9}{\frac{4}{3}} \Rightarrow 3\cancel{4}^{2} \times \frac{2}{\cancel{3}} \Rightarrow \frac{3}{2}$$

∴ Slope = $\left(\frac{3}{2}\right)$



2. 1b)This whole question is required for CSC522; Extra Credit for CSC 422: You are given 1- dimensional data points Xi , i [1, 2, 3, 4, 5, 6, 7] as shown in Table 1 ,also shown in Figure 2 in this question.

i)(1 point) Calculate the equation for the decision boundary of a hard-margin SVM, or if this is not possible, explain why in 1-2 sentences.

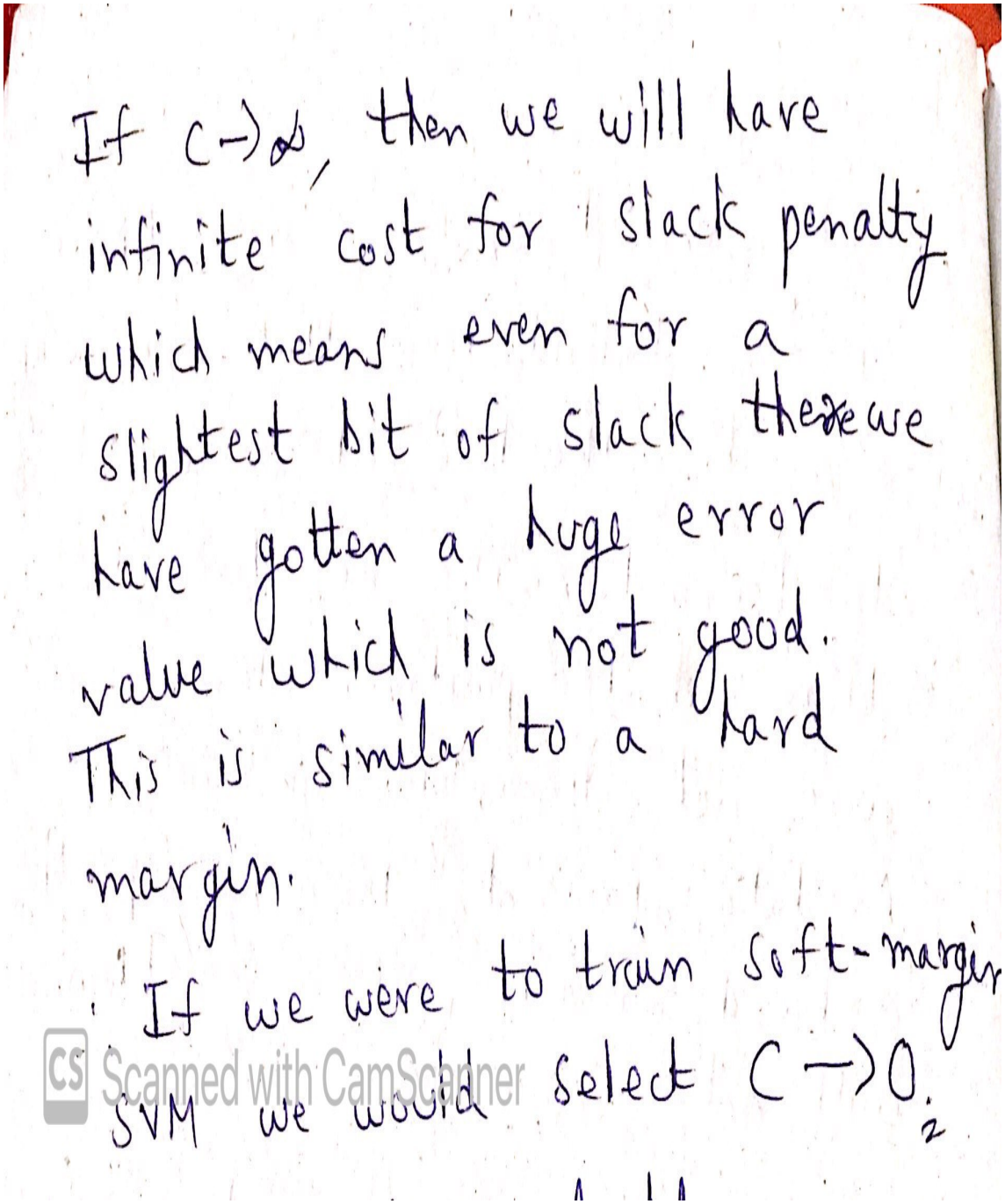
In one-dimensional space, there is no straight line that rigorously separates the two classes, according to observation. The negative class blue and the positive class orange are not linearly separable. As data is not linearly separable, a hard margin SVM algorithm will not converge.

ii) (1 point) If you were to train a soft-margin SVM on this data, would you select a C value where C $\rightarrow$ 0 or C $\rightarrow \infty$. Explain why in 1 sentence.

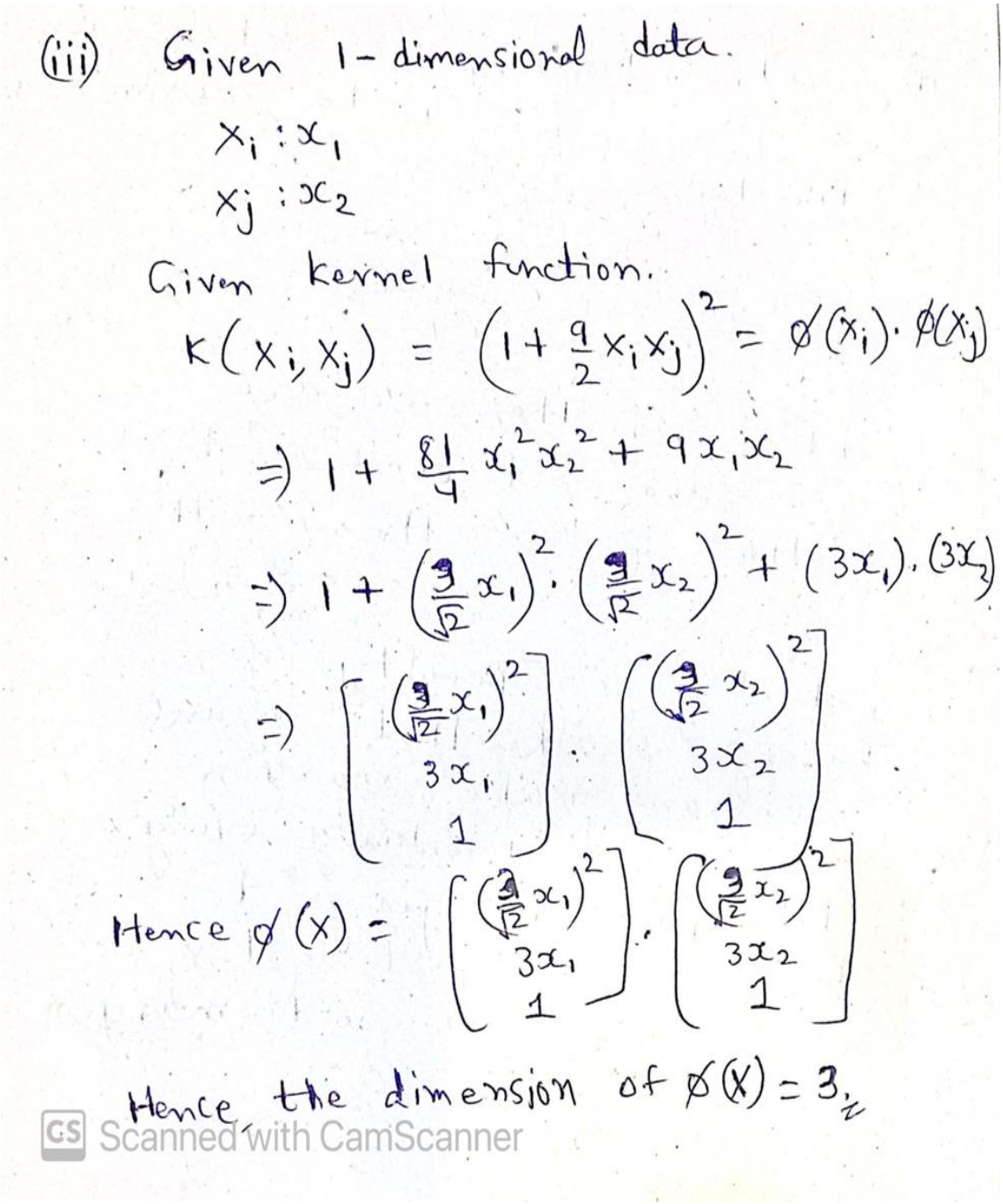
(ii) While training a soft-margin SVM our goal is to minimize the below equation.

$$\frac{\|w\|^2}{2} + C\sum_{i=1}^{n} \xi_i$$

If $C \to 0$ then we are cancelling out the 2nd term of the equation (the one with $C\sum_{i=1}^{n} \xi_i$) indicating no penalty for having slack. That means we can make margin maximized. Eventually every data point is going to be a SVM support vector. Similar to a Soft-margin.



iii)(1 point) Imagine you want to transform the 7 given data points to a higher dimensional space. You decide to use the kernel function $K(X_i, X_j) = ((1 + (9/2)\ X_i X_j)\wedge 2)$

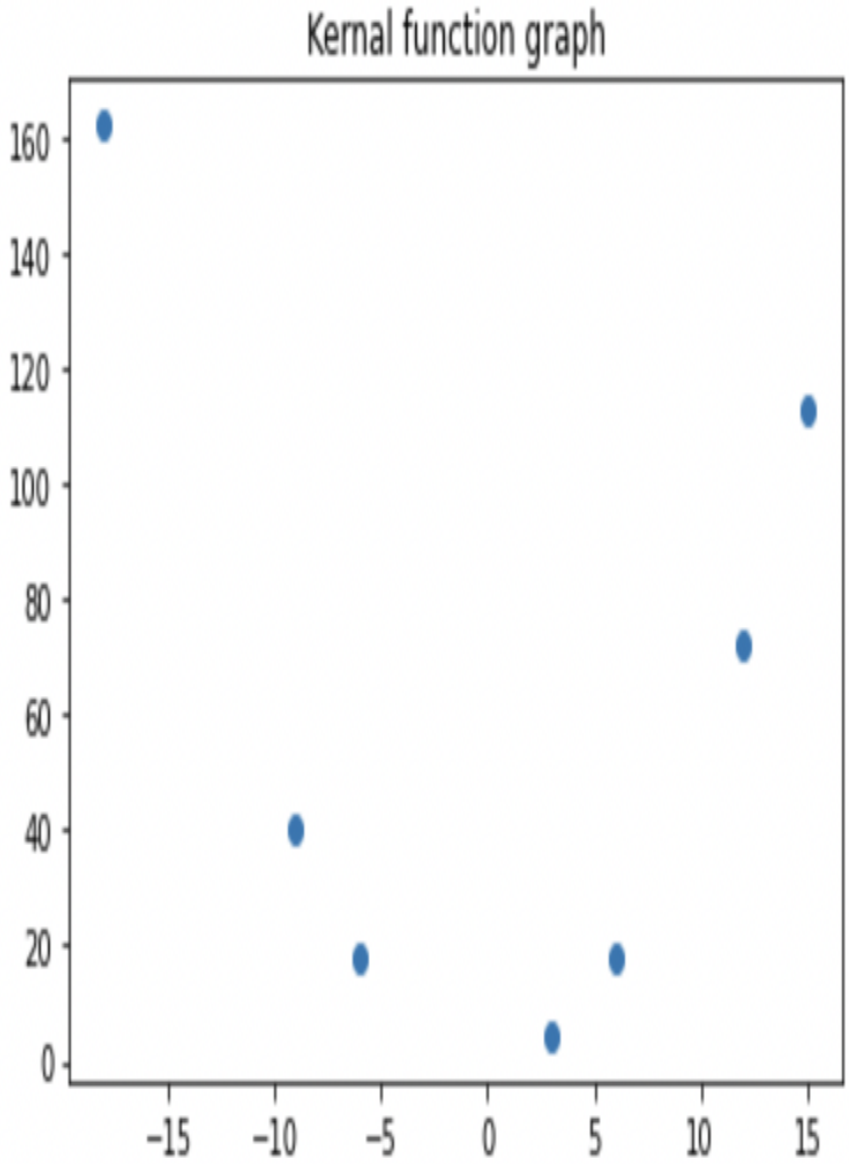


iv) (1 point) Use the function Φ(X) to calculate Φ(Xi) for i ϵ [1, 2, 3, 4, 5, 6, 7]. Graph these data points in the higher-dimensional space. (Hint: If the data is more than 2-dimensional, can you simplify your visualization to show it in 2D?).

From the above kernel function 1 is vector constant. So we will be discarding 1 and will only represent ((9/2)*(X))∧2 and 3X in the graph. Below is the coordinates table and the graph.

X	3X	$(9/2)*X^2$
-6	-18	162
-3	-9	40.5
-2	-6	18
1	3	4.5
2	6	18
4	12	72
5	15	112.5

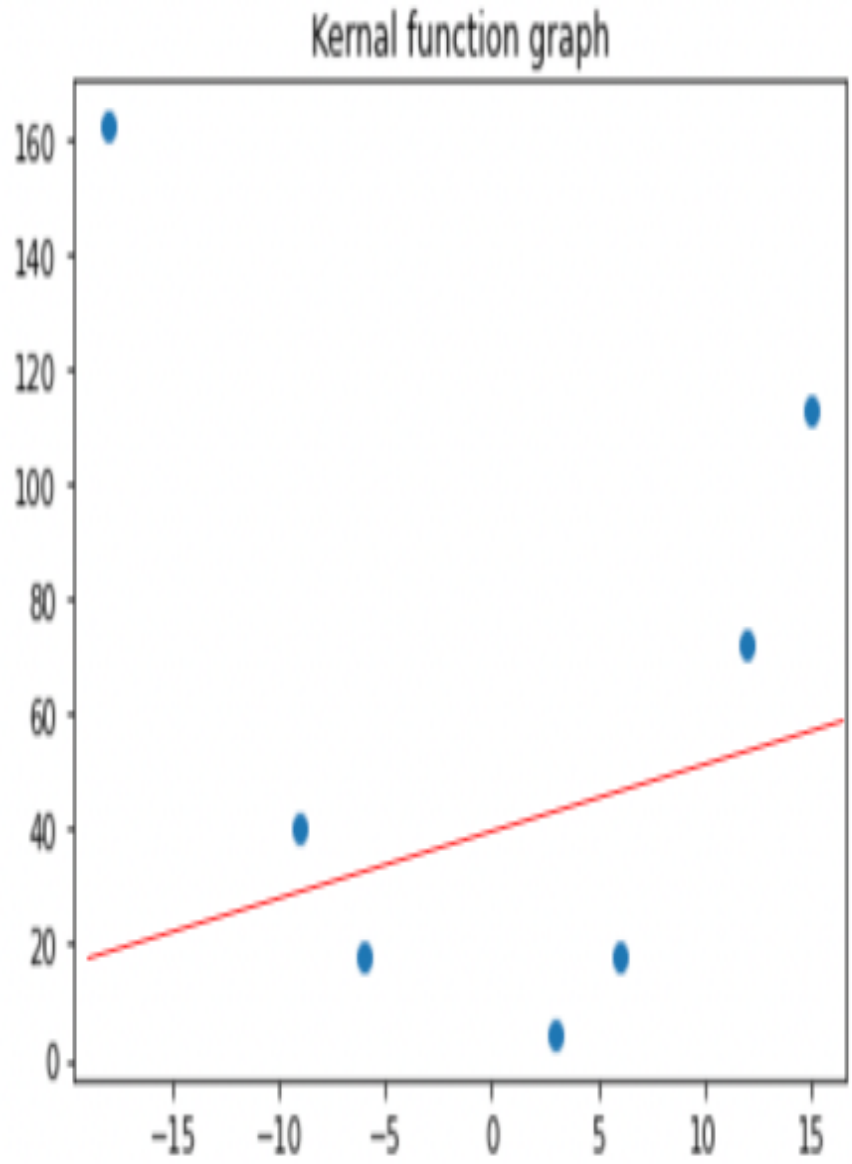
```
<matplotlib.collections.PathCollection at 0x7f0ce29a8cc0>
```



v) (1 point) Is it possible to linearly separate the data in the higher-dimensional space? If so, draw the decision boundary in your graph. If not, explain why. Note: You do not have to calculate the weights, just draw the decision boundary.

Yes, they are linearly separable. The decision boundary is shown below.

<matplotlib.collections.PathCollection at 0x7f0ce29a8cc0>

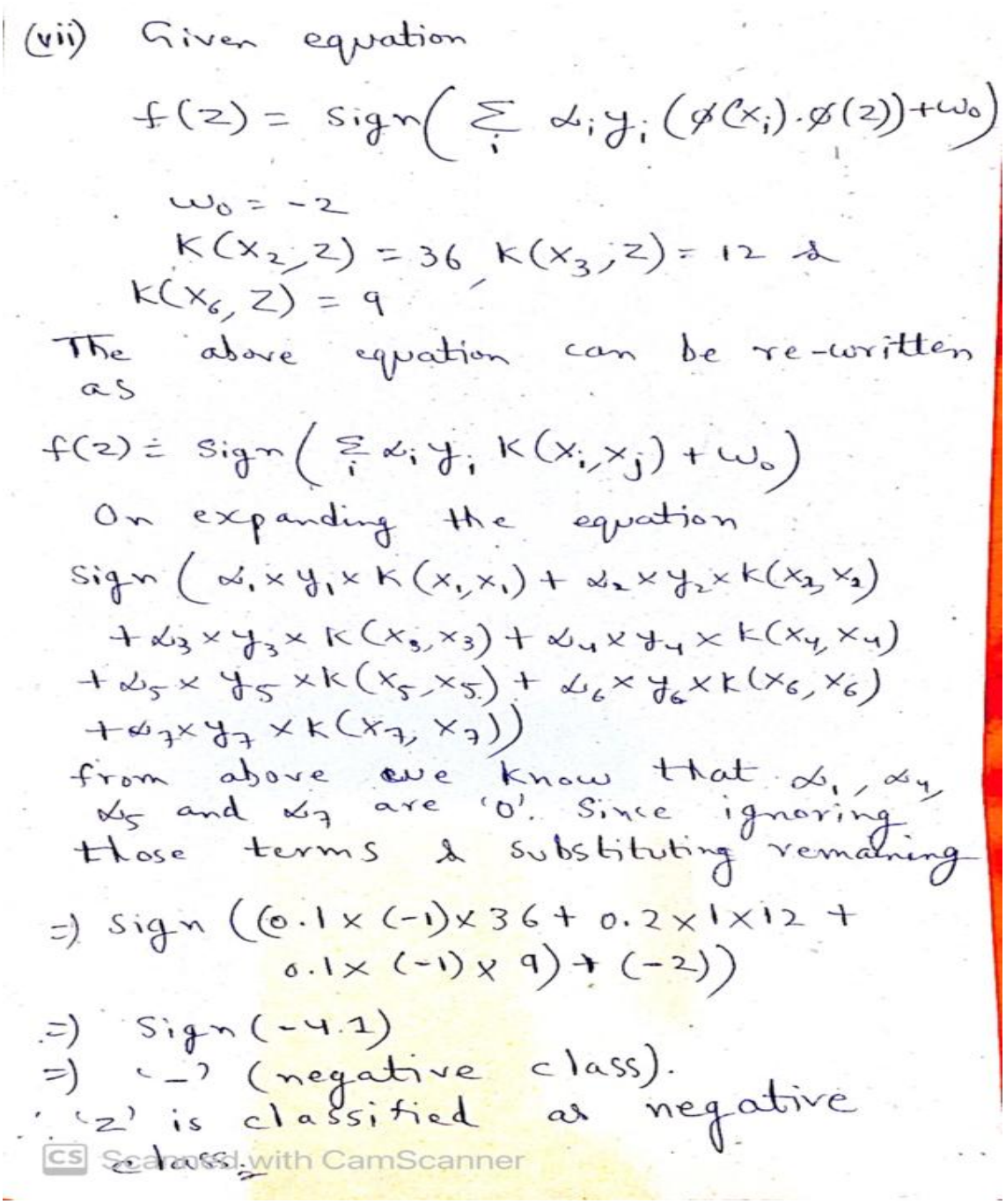


vi)1 point) You train a hard-margin SVM on the higher-dimensional data using a library and it gives you the following Lagrange multiplier for your data1 : $\alpha 2 = 0.1$, $\alpha 3 = 0.2$, $\alpha 6 = 0.1$. What are the remaining Lagrange multipliers, $\alpha 1$, $\alpha 4$, $\alpha 5$ and $\alpha 7$? Justify your answer in 1-2 sentences. (Hint: This should not require any math to calculate.)

According to the graph above X2, X3 and X6 are support vectors as they are on the hard margin of the decision boundary. Since they are the support vectors they have a lagrange multiplier value α¿0 and for the rest of the points the lagrange multiplier value is α=0 as they are not the support vectors. So $\alpha 1$=0, $\alpha 4$=0, $\alpha 5$=0 and $\alpha 7$=0.

vii)(1 point) Recall that the SVM's prediction (using the Kernel transformation) for a data point Z can be defined with the following equation: f(Z) = sign(Σi αiyi(ϕ(Xi) $\cdot$ ϕ(Z)) + w0). You are given w0 = 2. You are now asked to classify a new test data point, Z, using the SVM defined earlier by the Lagrange

multipliers. You do not know what Z's attributes are, but you do know: K(X2, Z) = 36, K(X3, Z) = 12, K(X6, Z) = 9. Classify Z using the SVM. (Hint: If you find yourself trying to solve for Z's x value, you are doing it wrong.)



(vii) Given equation

$$f(z) = \text{sign}\left(\sum_i \alpha_i y_i (\phi(x_i) \cdot \phi(z)) + w_0\right)$$

$w_0 = -2$

$K(x_2, z) = 36$, $K(x_3, z) = 12$ &

$K(x_6, z) = 9$

The above equation can be re-written as

$$f(z) = \text{sign}\left(\sum_i \alpha_i y_i K(x_i, x_j) + w_0\right)$$

On expanding the equation

$$\text{sign}\left(\alpha_1 \times y_1 \times K(x_1, x_1) + \alpha_2 \times y_2 \times K(x_2, x_2) + \alpha_3 \times y_3 \times K(x_3, x_3) + \alpha_4 \times y_4 \times K(x_4, x_4) + \alpha_5 \times y_5 \times K(x_5, x_5) + \alpha_6 \times y_6 \times K(x_6, x_6) + \alpha_7 \times y_7 \times K(x_7, x_7)\right)$$

from above we know that α_1, α_4, α_5 and α_7 are '0'. Since ignoring those terms & substituting remaining

$$\Rightarrow \text{sign}\left((0.1 \times (-1) \times 36 + 0.2 \times 1 \times 12 + 0.1 \times (-1) \times 9) + (-2)\right)$$

$$\Rightarrow \text{Sign}(-4.1)$$

$\Rightarrow$ '−' (negative class).

'z' is classified as negative class.

2 K-Means Clustering

1. (8 points) After each iteration of k-means, report the coordinates of the new centroids and which cluster each data point belongs to. Stop when the algorithm converges and clearly label on the graph where the algorithm converges. To report your work, give your answer in tabular format with the following attributes: Round (e.g. Round 1, 2, etc), Points (e.g. A, B, C), and Cluster ID (order does not matter). Also report the centroids for each cluster after each round. Please followed the example table format in Table 3.

ANS

(2) K-Means Clustering

(2a) Given Initial seed Centroids are at points C, A, E.

Assigning Each Data points B, D, F to nearest centroid.

Round 1:-

Point B ⇒ forms cluster with Centroid "E" since this is nearest comparatively with A, C.

So, {BE} forms 1 cluster.

Calculating Centroid for B,E = B(1,4), E(2,1)

$\left(\frac{1+2}{2}, \frac{4+1}{2}\right) = (1.5, 2.5)$

Points D,F ⇒ forms cluster with centroid "C"

Centroid: C(7,4) D(1,9) F(5,6)

$\left(\frac{7+1+5}{3}, \frac{4+9+6}{3}\right) \Rightarrow \left(\frac{13}{3}, \frac{19}{3}\right) = (4.3, 6.3)$

Round	Points	Cluster_ID	Centroids
1	{A}	1	[9,5]
1	{E,B}	2	[1.5, 2.5]
1	{C,D,F}	3	[4.33, 6.33]

Round 2 :-

Points{C, A} form cluster because centroid (9,5) is nearest to point C.

Now centroids changes for clusters 1 & 3.

(A, C) ⇒ (9,5) and (7,4) ⇒ $\left(\frac{16}{2}, \frac{9}{2}\right)$ ⇒ (8,4.5)

(D, F) ⇒ (1,9) and (5,6) ⇒ $\left(\frac{6}{2}, \frac{15}{2}\right)$ ⇒ (3, 7.5)

Round	Points	Cluster-ID	Centroids
2	{A,C}	1	[8, 4.5]
2	{B,E}	2	[1.5, 2.5]
2	{D, F}	3	[3, 7.5]

Round 3 :- ~~All clusters~~ remain same

Round	Points	Cluster-ID	Centroids
3	{A,C}	1	[8, 4.5]
3	{B,E}	2	[1.5, 2.5]
3	{D,F}	3	[3, 7.5]

On Round 3, All Data Points are in the same cluster Since the centroids are not changing. This states Algorithm converged.

Overall Table.

Round	Points	cluster-ID	centroids
1	A	1	[9,5]
	E,B	2	[1.5, 2.5]
	C, D, F	3	[4.33, 6.33]
2	A,C	1	[8, 4.5]
	B,E	2	[1.5, 2.5]
	D, F	3	[3, 7.5]
3	A,C	1	[8, 4.5]
	B,E	2	[1.5, 2.5
	D, F	3	[3, 7.5]

able format in Table 3

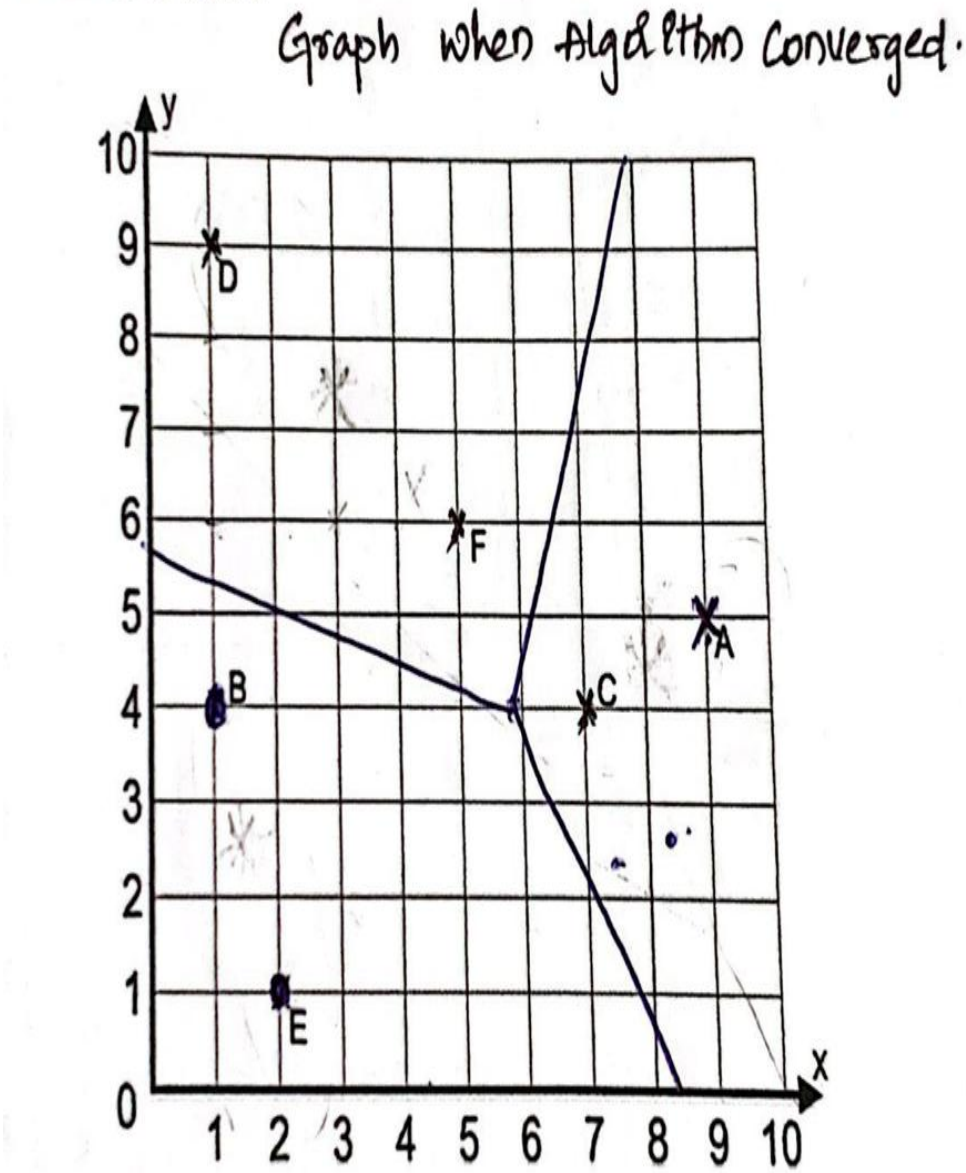


Figure 3: K-means Clustering (a)

2. 2b) (2 points) How many rounds are needed for the K-means clustering algorithm to converge?
   ANS

(2b) 3 rounds are needed for the K-means Algorithm to converge. We can conclude this with the last iteration. On the last iteration if none of the points changes its cluster. As we saw in round 3 no point changed its cluster so we can terminate the Algorithm

2c)(4 points) Calculate the Sum of Squared Errors (SSE) for the k-means clustering.
ANS

(2c)

From Round 3.

Cluster 1 : A, C and centroid (C1)

Distance (A, C1) and Distance (C, C1)

$(A, C1) = \sqrt{(9-8)^2 + (5-4.5)^2} = \sqrt{1+(0.5)^2} = \sqrt{1.25}$

$(C, C1) = \sqrt{(7-8)^2 + (4-4.5)^2} = \sqrt{1+(0.5)^2} = \sqrt{1.25}$

Sum of Squared Error [SSE1] = $(\sqrt{1.25})^2 + (\sqrt{1.25})^2 \Rightarrow 2.50$.

Cluster 2 : B, E and centroid (C2)

Distance B/w (B, C2) and (E, C2)

$(B, C2) = \sqrt{(1-1.5)^2 + (4-2.5)^2} = \sqrt{2.5}$

$(E, C2) = \sqrt{(2-1.5)^2 + (1-2.5)^2} = \sqrt{2.5}$

Sum of Squared Error SSE2 $\Rightarrow (\sqrt{2.5})^2 + (2.5)^2 \Rightarrow 5$

Cluster 3 :- (D, C3) & (F, C3)

$(D, C3) = \sqrt{(3-1)^2 + (9-7.5)^2} = \sqrt{6.25}$

$(F, C3) = \sqrt{(5-3)^2 + (6-7.5)^2} \Rightarrow \sqrt{4+2.25}$

SSE3 = $(\sqrt{6.25})^2 + (\sqrt{6.25})^2 \Rightarrow 12.50$

SSE $\Rightarrow$ SSE1 + SSE2 + SSE3 $\Rightarrow$ 2.50 + 5 + 12.50 $\Rightarrow$ 20

3 Hierarchical Clustering

1. (4 points) Perform single link hierarchical clustering. Show your work at each iteration by giving the inter-cluster distances. Report your results by drawing a corresponding dendrogram. The dendrogram should clearly show the order and the height in which the clusters are merged. If possible, use a program to construct your dendrogram (e.g. PowerPoint, LucidChart2 , or VisualParadigm3). Scanned hand drawings will also be accepted if they are very clear. NOTE: There may be ties (i.e. two clusters have the same distance). In this case, you can choose any order to merge in, and ensure that this is reflected in your dendrogram.

   ANS

(3) Hierarchical clustering

Given Distance Matrix

	B	C	D	E	F
A	8.06	2.24	8.94	8.06	4.12
B	0	6	5	3.16	4.47
C	6	0	7.81	5.83	2.83
D	5	7.81	0	8.06	5
E	3.16	5.83	8.06	0	5.83

(3a) Single link hierarchical clustering

Iteration - 1 :-

Choosing minimum value from matrix (i.e) 2.24

So points AC can be clustered. Now Euclidean matrix becomes

→ need to update matrix using minimum value.

	B	D	E	F
AC	6	7.81	5.83	2.83
B	0	5	3.16	4.47
D	5	0	8.06	5
E	3.16	8.06	0	5.83

Eq'n :- A + C = 2.24

Again choosing minimum from matrix 2.83 and updating matrix with minimum value

	B	D	E
ACF	4.47	5	5.83
B	0	5	3.16
D	5	0	8.06

Eq'ns: A + C = 2.24

DC + F = 2.83.

Iteration 3 :- choosing min value: 3.16 & updating matrix with min value.

	BE	D
ACF	4.47	5
BE	0	5

Eq'n: A + C = 2.24

AC + F = 2.83

B + E = 3.16.

Iteration: 4 :-

ACFBE	D
	5

Eq'n

A + C = 2.24

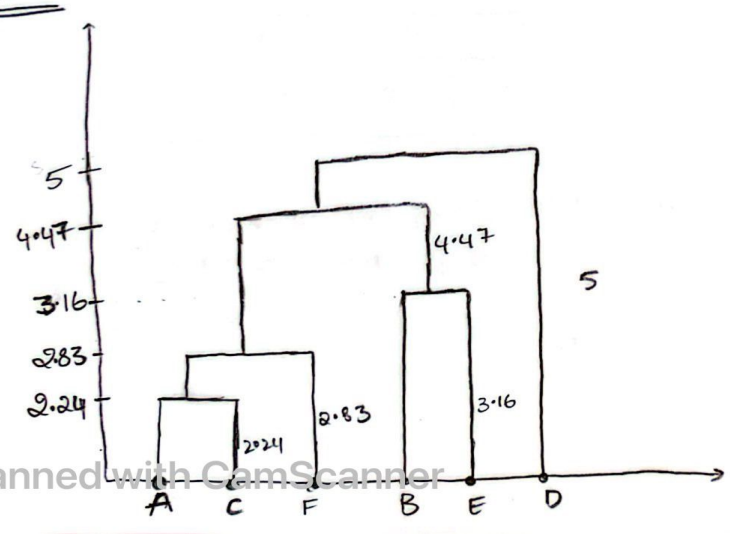
AC + F = 2.83

B + E = 3.16

ACF + BE = 4.47

ACFBE + D = 5

DenDrogram



3b) (4 points) Perform complete link hierarchical clustering on the dataset. As above, show your calculations and report the corresponding dendrogram.

(3b) Complete Link hierarchical clustering

Given Euclidean matrix

	B	C	D	E	F
A	8.06	2.24	8.94	8.06	4.12
B	0	6	5	3.16	4.47
C	6	0	7.81	5.83	2.83
D	5	7.81	0	8.06	5
E	3.16	5.83	8.06	0	5.83

Iteration-1

In Complete link clustering, choose minimum value cluster two Points. For updating matrix consider max value

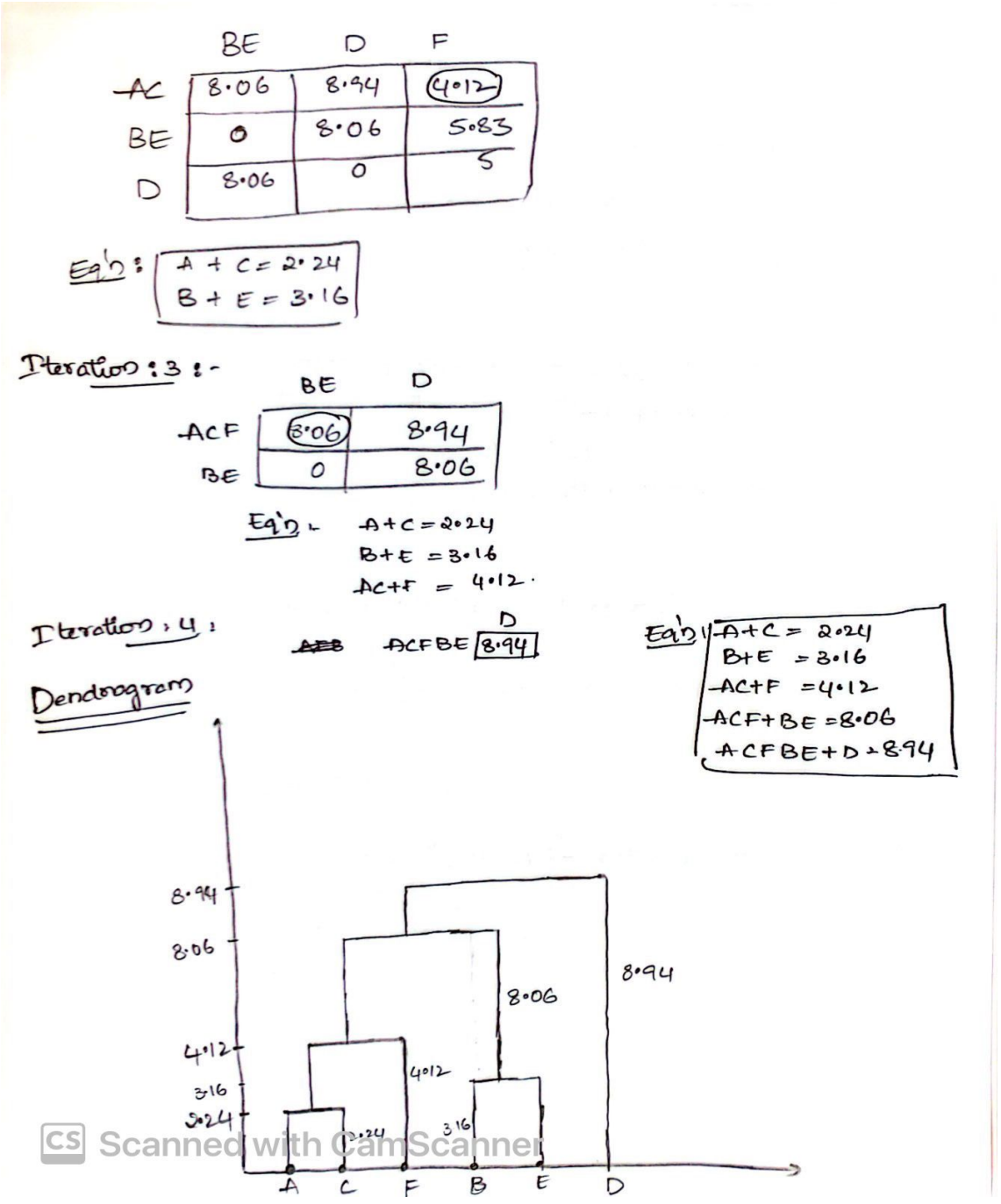
minvalue is 2.24 Points A, C can be clustered

	B	D	E	F
AC	8.06	8.94	8.06	4.12
B	0	5	3.16	4.47
D	5	0	8.06	5
E	3.16	8.06	0	5.83

Eqn: A+C: 2.24

Iteration 2: choose min value (i.e) point which is nearer = 3.16

So B, E can be clustered & update the matrix with max values

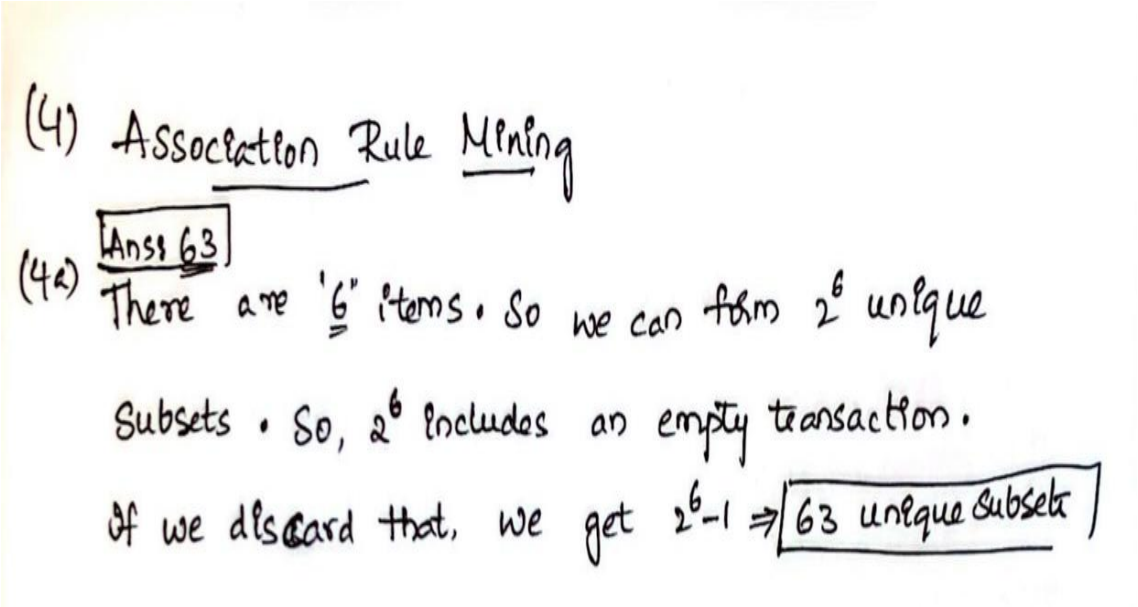


3c) (2 points) If we assume there are two clusters, will the single or complete link approach give a better clustering? Justify your answer. (Hint: you should not need to calculate SSE to answer the question.)
ANS
In case of only two clusters we think the complete link approach gives a better clustering. Because complete link clustering forms equal number of clusters in each round of iteration. Whereas in single link clustering if one item is cluster in one cluster and all the remaining ones go to the others. There is no uniform split in single link clustering.

4 Association Rule Mining

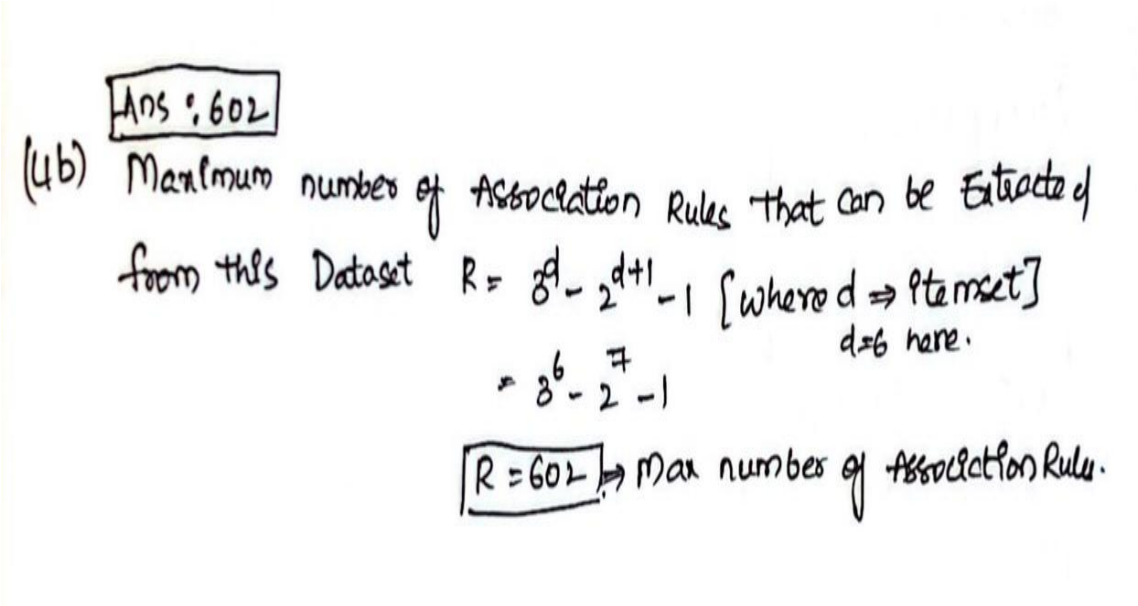
1. 4a) (1 point) What is the maximum number of unique itemsets that can be extracted from this data set (including itemsets that have zero support)? Briefly explain your answer in 1-2 sentences

ANS



4b) (1 point) What is the maximum number of association rules that can be extracted from this data set (including rules that have zero support)? Briefly explain your answer in 2-3 sentences

ANS



4c) (1 point) Compute the support of the itemset: Chicken, Cookie?

ANS

(4c) Suppat of Itemset {chicken, cookie} = 4/10

4d) (2 points) Compute the support and confidence of association rule: Cheese → Butter?

ANS

(4d) Suppat of Itemset({~~chicken~~ cheese} → {Butter}) = σ({cheese, butter}) / |T|

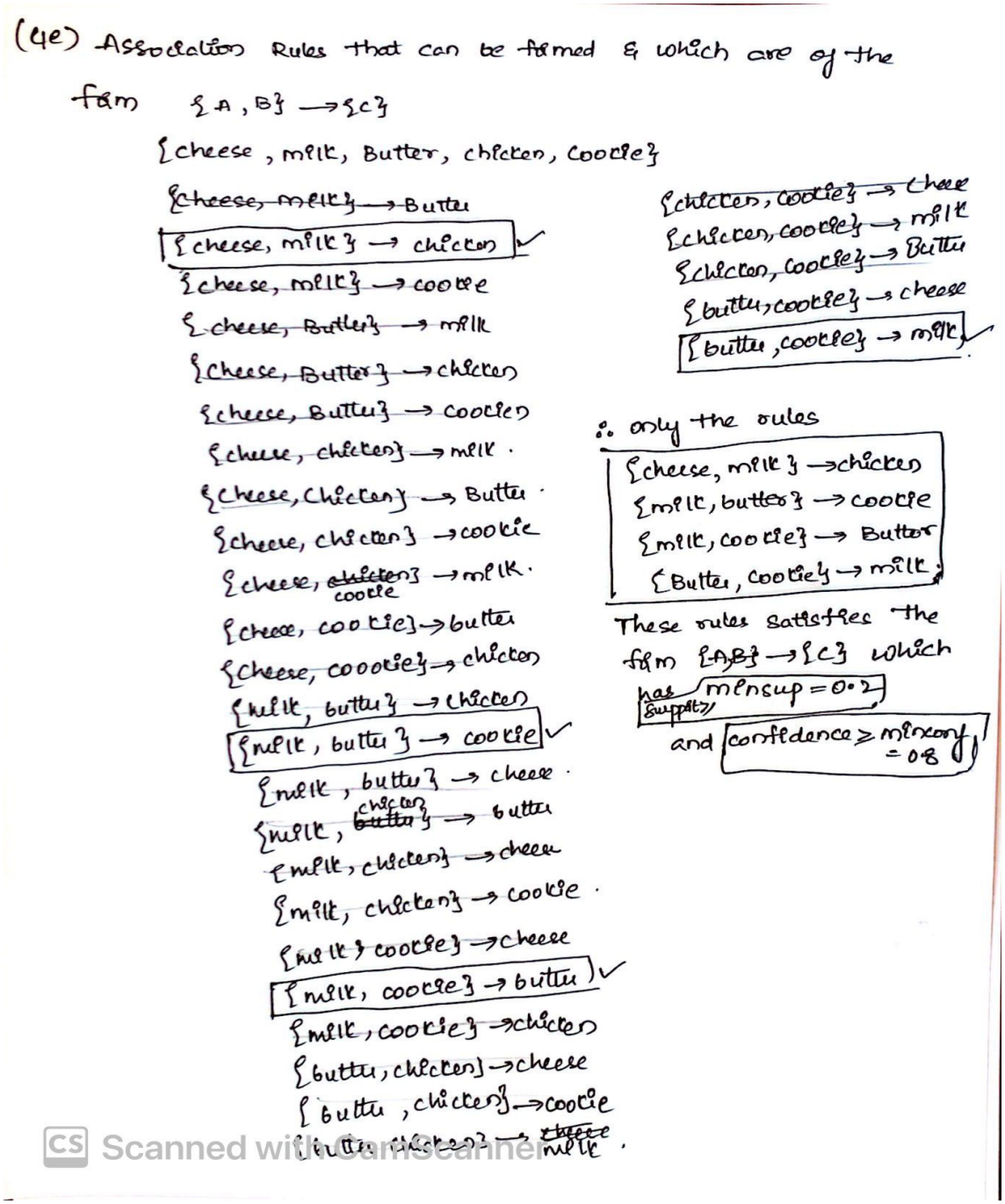
Suppat {cheese → Butter} = 2/10

Confidence of Itemset({cheese} → {butter}) = σ{cheese, butter} / σ({cheese})

Confidence {cheese} → {butter} = 2/4

4e) (3 points) Given min support = 0.2 and min confidence = 0.8, identify all valid association rules of the form A, B → C.

ANS



4f) (2 points) In a different dataset, the support of the rule a → b is 0.62, and the support of the rule a, c → b, d is 0.31. What can we say for sure about the support of the rule a → b, d. Explain in 1-2 sentences.

ANS

(4f) Given,

Support of rule $\{a\} \rightarrow \{b\} = 0.62$

Support of rule $\{a,c\} \rightarrow \{b,d\} = 0.31$.

find support of rule $\{a\} \rightarrow \{b,d\} \Rightarrow ?$

From the ~~Anti~~monotone Property:-

Property states that support of an itemset never exceeds the support of its subsets.

$(\{a\} \rightarrow \{b\}) \subseteq \{a\} \rightarrow \{b,d\} \subseteq \{a,c\} \rightarrow \{b,d\}$.

Therefore:

$$s(\{a\} \rightarrow \{b\}) \geq s(\{a\} \rightarrow \{b,d\}) \geq s(\{a,c\} \rightarrow \{b,d\})$$

$$0.62 \geq s(\{a\} \rightarrow \{b,d\}) \geq s(\{a,c\} \rightarrow \{b,d\}) \quad (0.31)$$

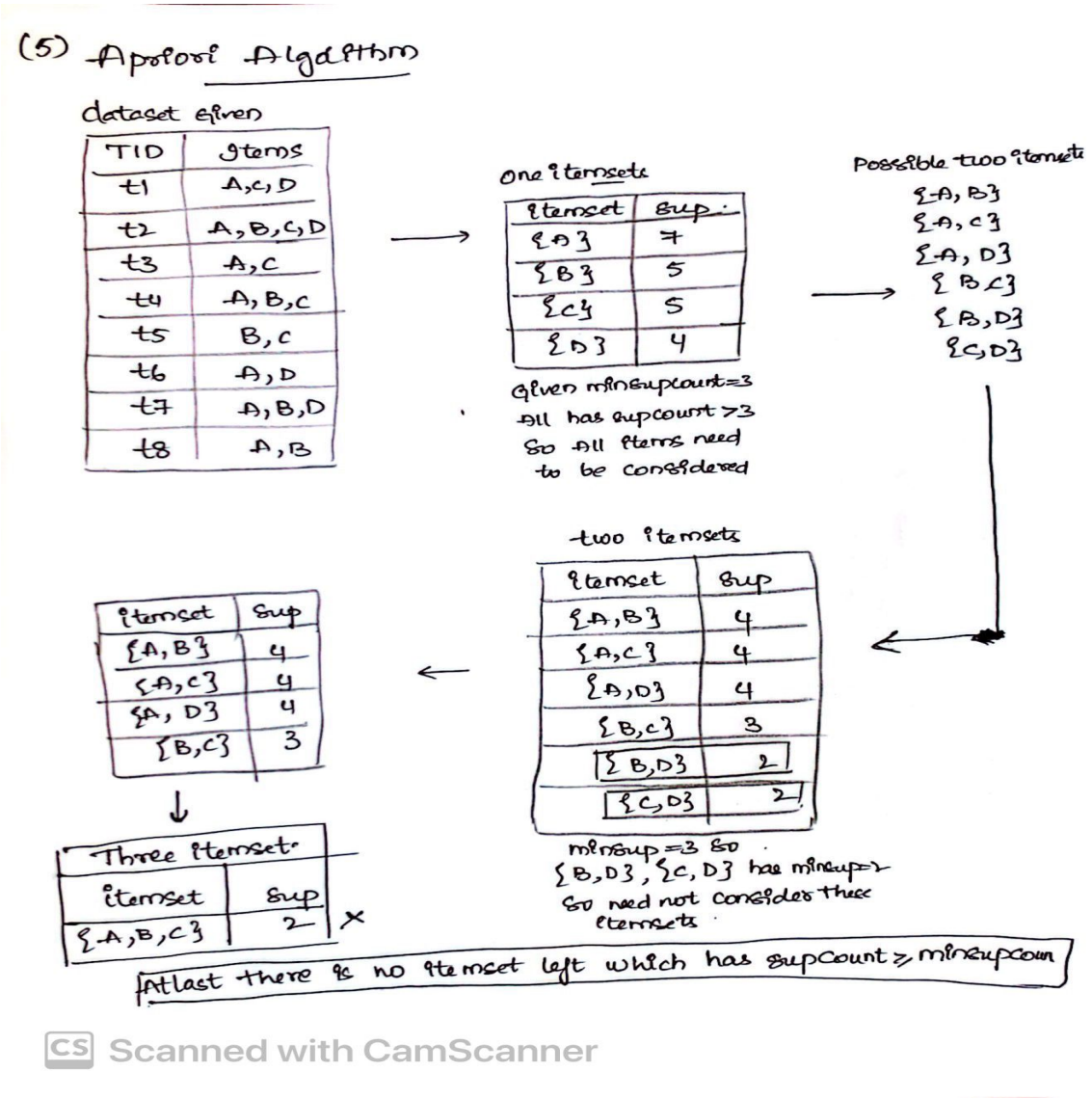
From this we can for sure say that

$s(\{a\} \rightarrow \{b,d\})$ lies b/w 0.62 & 0.31

5 Apriori algorithm

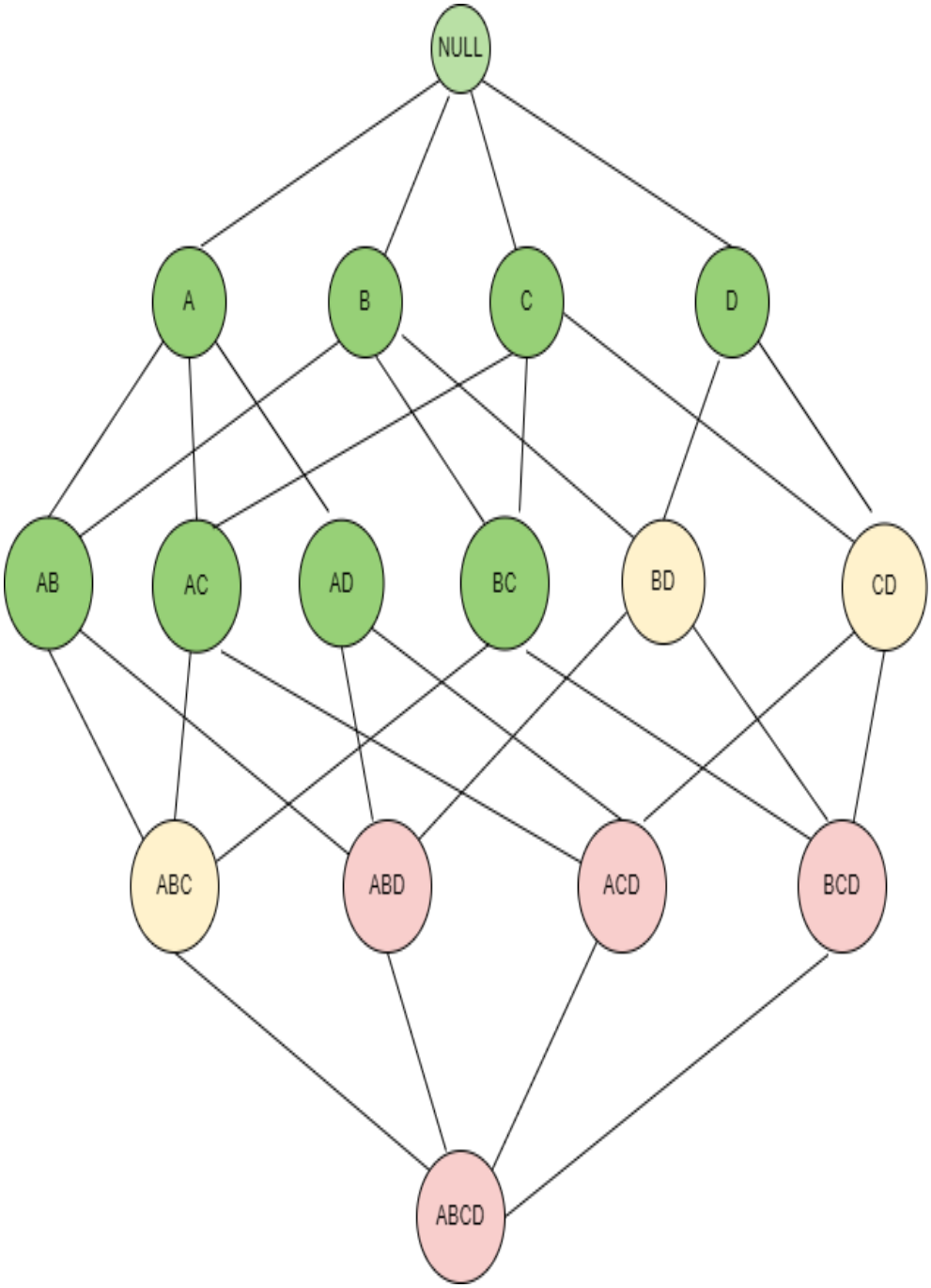
1. 5a) (5 points) Show (compute) each step of frequent itemset generation process using the apriori algorithm, with a minimum support count of 3.

   ANS



   5b) (5 points) Show the lattice structure for the data given in table above, and mark each node in the lattice as either F: Frequent, IC: Infrequent due insufficient support count, or IP: Infrequent due to pruning (we do not need to calculate the support count). (Scanned hand-drawing is acceptable as long as it is clear. Tip (optional): Applications like CamScanner and Adobe Scan improve the quality of photos and reduce their size.)

   ANS
   Color green represents = F: Frequent
   Color yellow represnts =IC: Infrequent due insufficient support count
   color pink represents = IP: Infrequent due to pruning


NULL
A
B
C
D
AB
AC
AD
BC
BD
CD
ABC
ABD
ACD
BCD
ABCD