Name : Shiska Raut
ID : 1001526329

CSE 5334 : Data Mining
Homework 1 (100 points)

SUBMISSION DEADLINE : 03/31/2023, 11:59 PM

REQUIREMENTS : Homework must be handwritten and converted to pdf format for submission. We will NOT accept typed submissions.

1. Vector Space Model (25 Points)

Suppose you want to know which of Shakespeare's novels are the most similar based on the occurrence of certain characters (Anthony, Brutus, Caesar etc). Table 1 shows the count matrix (the number of occurrence of a certain term/character in a document/novel).

Table 1

	Anthony & Cleopetra	Julias Caesar	The Tempest	Hamlet
Anthony	157	73	50	100
Brutus	4	157	10	2
Caesar	232	2	227	100
Calpurnia	2	10	0	5
Cleopetra	5	0	10	6

Table 2 represents the document frequency of each term/character in your collection. Assume the total number of novels in your collection is 1000.

Table 2

Term	Document Frequency
Anthony	500
Brutus	50
Caesar	600
Calpurnia	50
Cleopetra	20

Find out which novel is most similar to Anthony & Cleopetra. Use log weighted TF(term frequency), log weighted IDF(Inverse Document Frequency) and cosine similarity as the similarity measure.

1. VSM

To find out which Novel is most similar to Anthony & Cleopatra, we need to calculate the tf-idf vectors for each document.

Weighing scheme : ltn

let,

Anthony & cleopatra = d_1

Julias Caesar = d_2

The Tempest = d_3

Hamlet = d_4

Also,

Anthony = t_1

Brutus = t_2

Caesar = t_3

Calpurnia = t_4

cleopatra = t_5

1) **Weighted tf**

	d_1	d_2	d_3	d_4
t_1	3.1958	2.8633	2.6989	3
t_2	1.6020	3.1958	2	1.3010
t_3	3.3654	1.3010	3.3560	3
t_4	1.3010	2	0	1.6989
t_5	1.6989	0	2	1.7781

2) Weighted idf (N = 4)

$t_1 = \log_{10}(1000/500) \Rightarrow 0.3010$

$t_2 = \log_{10}(1000/50) \Rightarrow 1.3010$

$t_3 = \log_{10}(1000/600) \Rightarrow 0.2218$

$t_4 = \log_{10}(1000/50) \Rightarrow 1.3010$

$t_5 = \log_{10}(1000/20) \Rightarrow 1.6989$

3) tf*idf for each doc.

	d_1	d_2	d_3	d_4
t_1	0.9619	0.8618	0.8123	0.903
t_2	2.0842	4.1577	2.602	1.6926
t_3	0.7464	0.2885	0.7463	0.6654
t_4	1.6926	2.602	0	2.2102
t_5	2.8862	0	3.3978	3.0208

4) Cosine similarity (non-normalized)

Since we are interested in finding out which doc is most similar to anthony & cleopatra (d_1), we calculate its similarity w.r.t each doc.

$$\cos(\vec{d_1}, \vec{d_2}) = (0.9619 * 0.8618) + (2.0842 * 4.1577) + (0.7464 * 0.2885) + (1.6926 * 2.602) + (2.8862 * 0)$$

$$= 14.1441$$

$$\cos(\vec{d_1}, \vec{d_3}) = (0.9619 * 0.812) + (2.0842 * 2.602) + (0.7464 * 0.7463) + (1.6926 * 0) + (2.8862 * 3.3978)$$

$$= 16.5685$$

$$\cos(\vec{d_1}, \vec{d_4}) = (0.9619 * 0.903) + (2.0842 * 1.6926) + (0.7464 * 0.6659) + (1.6926 * 2.2102) + (2.8862 * 3.0208)$$

$$= 17.353$$

5) Normalized cosine similarity

Calculate magnitude

$$\|d_1\| = \sqrt{(0.9619)^2 + (2.084)^2 + (0.746)^2 + (1.692)^2 + (2.8862)^2}$$

$$= 4.1257$$

$$\|d_2\| = \sqrt{(0.861)^2 + (4.1577)^2 + (0.2885)^2 + (2.602)^2 + 0}$$

$$= 4.988$$

$$\|d_3\| = \sqrt{(0.812)^2 + (2.602)^2 + (0.746)^2 + 0 + (0.397)^2}$$

$$= 4.4195$$

$$\|d_4\| = \sqrt{(0.903)^2 + (1.692)^2 + (0.6654)^2 + (2.21)^2 + (3.02)^2}$$

$$= 4.2583$$

$$\cos(d_1, d_2) = \frac{\vec{d_1} \cdot \vec{d_2}}{\|d_1\| \|d_2\|} = \frac{14.1141}{4.1257 * 4.988}$$

$$= 0.6858$$

$$\cos(d_1, d_3) = \frac{\vec{d_1} \cdot \vec{d_3}}{\|d_1\| \|d_3\|} = \frac{16.5685}{4.1257 * 4.4195}$$

$$= 0.9086$$

$$\cos(d_1, d_4) = \frac{\vec{d_1} \cdot \vec{d_4}}{\|d_1\| \|d_4\|} = \frac{17.353}{4.1257 * 4.2583}$$

$$= 0.9877$$

∴ Using normalized cosine similarity, anthony & cleopatra (d1) is most similar to Hamlet (d4)

2. Decision Tree (25 Points)

Suppose, you are a robot in a lumber yard, and must learn to discriminate Oak wood from Pine wood. You choose to learn a Decision Tree classifier. You are given the following examples:

ID	Density	Grain	Hardness	Class
1	Heavy	Small	Hard	Oak
2	Heavy	Large	Hard	Oak
3	Heavy	Small	Hard	Oak
4	Light	Large	Soft	Oak
5	Light	Large	Hard	Pine
6	Heavy	Small	Soft	Pine
7	Heavy	Large	Soft	Pine
8	Heavy	Small	Soft	Pine

a) Which attribute would information gain choose as the root of the tree? Show calculations. (10 points)

b) Draw the decision tree that would be constructed by recursively applying information gain to select roots of sub-trees. Show calculations. (10 points)

c) Classify the following two new examples as Oak or Pine using your decision tree above. (5 points)
[Density=Light, Grain=Small, Hardness=Hard] ?
[Density=Light, Grain=Small, Hardness=Soft] ?

2.

$N = 8$

Attributes = Density, Grain, Hardless

To find which split would result in max information gain we must calculate entropy w.r.t. each attribute split.

let, C0 : Oak

C1 : Pine

Entropy of parent E(P)

C0 : 4

C1 : 4

$E(P) = \log_2(n_c) = 1$

↑ maximum entropy

a) Splitting on Density

Heavy

C0 : 3

C1 : 3

$E = \log_2(n_c) = 1$

↑ maximum entropy

Light

C0 : 1 $\quad E = \log_2(n_c) = 1$

C1 : 1 ← maximum entropy

$E(\text{Density}) = \left(\frac{6}{8} \times 1\right) + \left(\frac{2}{8} \times 1\right) = 1$

Information gain = $E(P) - E(\text{Density})$

$= 1 - 1$

$= 0$

b) Splitting on Grain

Small

C0 : 2 $\quad E = \log_2[n_c) = 1$

C1 : 2 ← max entropy

large

C0 : 2 $\quad E = \log_2(n_c) = 1$

C1 : 2 ← max entropy

$E(\text{Grain}) = \left(\frac{4}{8} \times 1\right) + \left(\frac{4}{8} \times 1\right) = 1$

Information gain = $E(P) - E(\text{Grain})$

$= 1 - 1$

$= 0$

c) Splitting on Hardness

Hard

C0 : 3

C1 : 1

$E = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right)$

$= 0.81$

Soft

C0 : 1

C1 : 3

$E = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right)$

$= 0.81$

$E(\text{Hardness}) = \left(\frac{4}{8} \times 0.81\right) + \left(\frac{4}{8} \times 0.81\right)$

$= 0.81$

Information gain = $E(P) - E(\text{Hardness})$

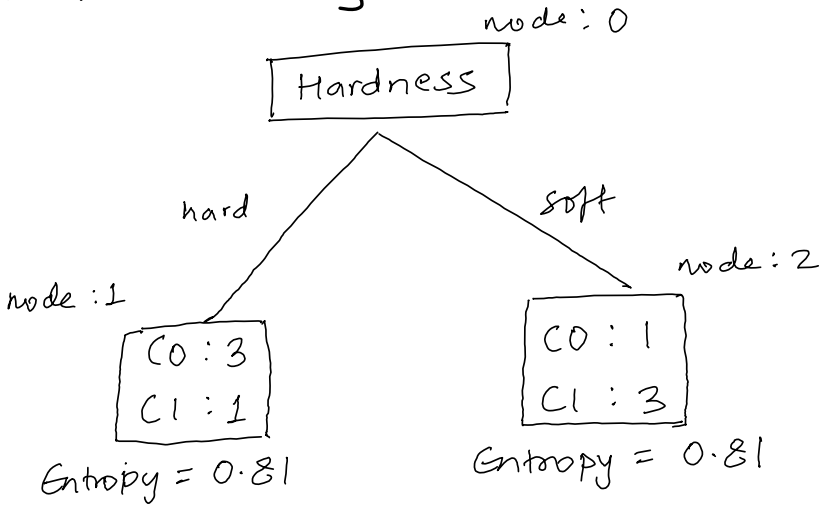
$= 1 - 0.81$

$= 0.18$

Ans a)

Info. gain would choose Hardness as the root of the tree.

b) After choosing 'Hardness' as the root



Remaining attributes : Density, Grain

For node : 1,

i) Splitting on density

heavy

C0 : 3
C1 : 0
} pure (oak)

E = 0 (min entropy)

light

C0 : 0
C1 : 1
} pure (pine)

E = 0 (min entropy)

Info gain = E(P) − E(Density) = 0.81 − 0

= 0.81

ii) Splitting on Grain

small		large	
C0 : 2	pure (oak)	C0 : 1	max entropy
C1 : 0		C1 : 1	

Entropy = 0

Entropy = $\log_2(2) = 1$

$$E(\text{Grain}) = 0 + \frac{2}{4} \times 1 = 0.5$$

Info gain = $E(P) - E(\text{Grain}) = 0.81 - 0.5$

$= 0.31$

∴ node 2 should be split on Density as it results in max info gain.

For node : 2,

i) Splitting on Density

Heavy		light	
C0 : 0	pure (pine)	C0 : 1	pure (oak)
C1 : 3		C1 : 0	

E (Density) = 0

Info gain = $E(P) - E(\text{Density}) = 0.81 - 0$

$= 0.81$

ii) Splitting on Grain,

small
C0 : 0
C1 : 2
} pure

large
C0 : 1
C1 : 1
} max entropy

Entropy = 0

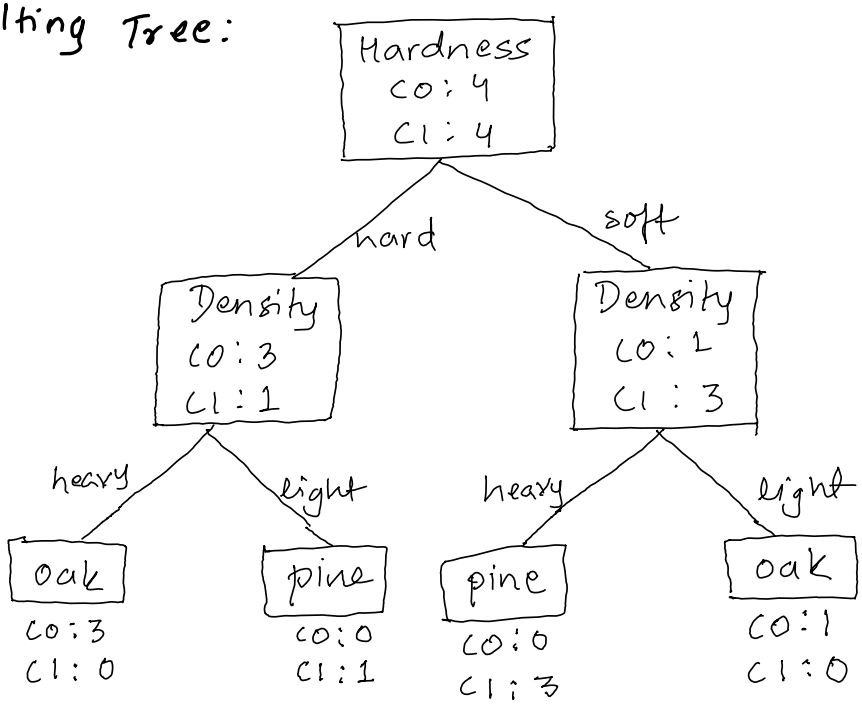
Entropy = $\log_2(2) = 1$

$E(\text{Grain}) = 0 + \frac{2}{4} \times 1 = 0.5$

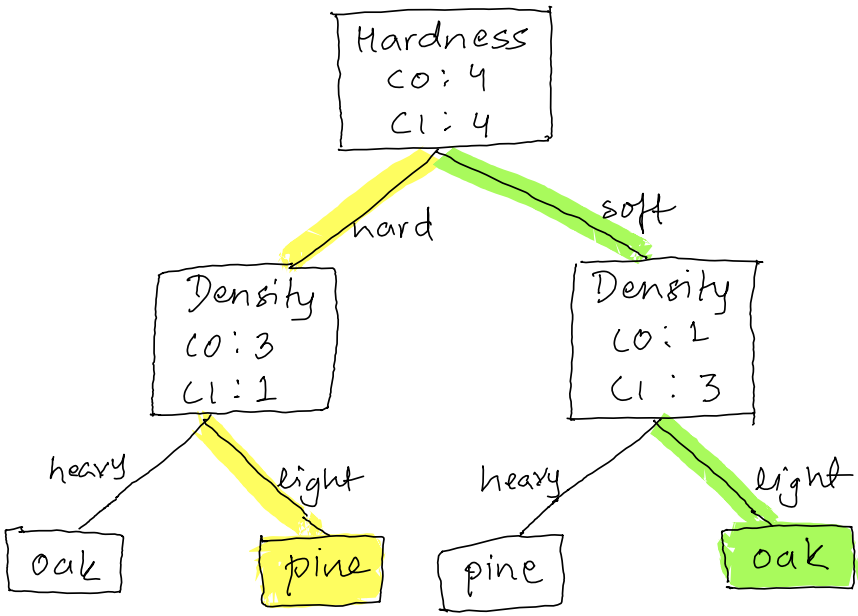
Info gain : $E(P) - E(\text{Grain}) = 0.81 - 0.5$
$= 0.31$

∴ Node : 2 should be split on Density as it results in max info gain.

Resulting Tree:



c) Using the decision tree from b):



1) Density = light
Hardness = Hard
Grain = Small
would be classified as "Pine".

2) Density = light
Grain = Small
Hardness = Soft
Would be classified as "Oak".

3. Naive Bayes (25 points)

Given the weather dataset where the attributes (Outlook, Temperature, Humidity, Windy) shows the weather condition on a particular day and whether or not Golf was played on that day.

Instance	Outlook	Temperature	Humidity	Windy	Play Golf
1	Rainy	Hot	High	FALSE	No
2	Rainy	Hot	High	TRUE	No
3	Overcast	Hot	High	FALSE	Yes
4	Sunny	Mild	High	FALSE	Yes
5	Sunny	Cool	Normal	FALSE	Yes
6	Sunny	Cool	Normal	TRUE	No
7	Overcast	Cool	Normal	TRUE	Yes
8	Rainy	Mild	High	FALSE	No
9	Rainy	Cool	Normal	FALSE	Yes
10	Sunny	Mild	Normal	FALSE	Yes
11	Rainy	Mild	Normal	TRUE	Yes
12	Overcast	Mild	High	TRUE	Yes
13	Overcast	Hot	Normal	FALSE	Yes
14	Sunny	Mild	High	TRUE	No

Using the weather dataset table, classify the following test instance using Naive Bayes Classifier:

Instance	Outlook	Temperature	Humidity	Windy	Play Golf
Test#1	Rainy	Cool	High	TRUE	?

4. Support Vector Machine Classifier (25 Points)

Consider 8 training samples. The positive class has 4 points: (2,5),(2,2),(5,2) and (4,4). The negative class has 4 points: (-2,-2),(-4,0),(-4,-4) and (-8,-4).

a. What are the coefficients of the support vectors? **(7 points)**
b. Based on the coefficients of the support vectors, compute the maximum margin and then give the weights of the corresponding linear model. **(12 points)**
c. Determine the class that (-1,-1) belongs to. Show your calculation to explain. **(6 points)**

3. Naive Bayes

$n = 14$

Test #1

To classify the instance we need to calculate $P(\text{class} \mid \text{Data})$ i.e.

$P(\text{Yes} \mid \text{Rainy, Cool, High, TRUE})$ and

$P(\text{No} \mid \text{Rainy, Cool, High, TRUE})$

a) $P(\text{Yes}) = 9/14$

$P(\text{No}) = 5/14$

b) Using naive bayes classifier, we know,

$P(\text{Yes} \mid \text{Rainy, Cool, High, TRUE}) = P(\text{Rainy} \mid \text{Yes}) \cdot P(\text{Cool} \mid \text{Yes}) \cdot P(\text{High} \mid \text{Yes}) \cdot P(\text{TRUE} \mid \text{Yes}) \cdot P(\text{Yes})$

$P(\text{Rainy} \mid \text{Yes}) = 2/9$ $\quad P(\text{TRUE} \mid \text{Yes}) = 3/9$

$P(\text{Cool} \mid \text{Yes}) = 3/9$

$P(\text{High} \mid \text{Yes}) = 3/9$

$P(\text{Yes} \mid \text{Rainy}, \text{Cool}, \text{High}, \text{True})$

$= \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14}$

$= \boxed{0.00529}$

$P(\text{No} \mid \text{Rainy}, \text{Cool}, \text{High}, \text{TRUE}) = P(\text{Rainy} \mid \text{No}) \cdot P(\text{Cool} \mid \text{No}) \cdot P(\text{High} \mid \text{No}) \cdot P(\text{TRUE} \mid \text{No}) \cdot P(\text{No})$

$P(\text{Rainy} \mid \text{No}) = 3/5$

$P(\text{Cool} \mid \text{No}) = 1/5$

$P(\text{High} \mid \text{No}) = 4/5$

$P(\text{TRUE} \mid \text{No}) = 3/5$

$P(\text{No} \mid \text{Rainy}, \text{Cool}, \text{High}, \text{True})$

$= \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14}$

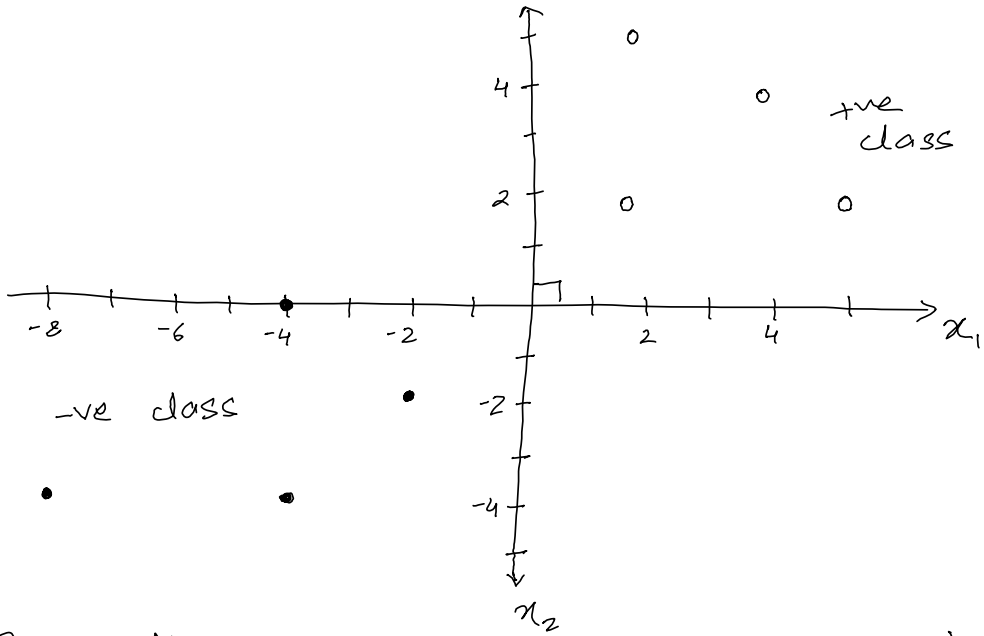
$= \boxed{0.0205}$

∴ Test #1 would be classified as "No", as the probability of belonging to that class is higher.

4. SVM

+ve class : (2, 5), (2, 2), (5, 2), (4, 4)

-ve class : (-2, -2), (-4, 0), (-4, -4), (-8, 4)



The equation of our margin is: $w_1x_1 + w_2x_2 + b = 0$

+ve margin → $w_1x_1 + w_2x_2 + b + 1 = 0$

-ve margin → $w_1x_1 + w_2x_2 + b - 1 = 0$

a) let (-2, -2) & (-4, 0) from the -ve class and (2, 2) from the +ve class be the support vectors.

b)

→ plugging (-2, -2) & (-4, 0) in equation of the -ve margin gives us

$$-2w_1 - 2w_2 + b - 1 = 0 \quad \text{—— (i)}$$

$$-4w_1 + 0w_2 + b - 1 = 0 \quad \text{—— (ii)}$$

Multiplying eq (i) by -2 and adding with (iii) gives us,

$$\begin{array}{r} 4w_1 + 4w_2 - 2b + 2 = 0 \\ -4w_1 \quad\quad + \quad b - 1 = 0 \\ \hline 4w_2 - b + 1 = 0 \end{array}$$

$$\text{or, } w_2 = \frac{b-1}{4} \quad \text{—— (iii)}$$

→ Plugging (2, 2) in the margin eqn gives us,

$$2w_1 + 2w_2 + b + 1 = 0 \quad \text{—— (iv)}$$

using (iii) in (iv)

$$2w_1 + 2\left(\frac{b-1}{4_2}\right) + b + 1 = 0$$

$$\text{or, } 4w_1 + b - 1 + 2b + 2 = 0$$

$$\text{or, } 4w_1 + 3b + 1 = 0 \quad \text{—— (v)}$$

Using eq (v) & (ii),

$$\begin{array}{r} 4w_1 + 3b + 1 = 0 \\ -4w_1 + b - 1 = 0 \\ \hline 4b = 0 \end{array}$$

$$\text{or, } b = 0$$

Replacing $b = 0$ in eq (iii)

$$w_2 = \frac{0-1}{4} = -\frac{1}{4}$$

Replacing $b = 0$ & $w_2 = 0$ in eq (i),

$$-2w_1 - 2\left(-\frac{1}{4}\right)_2 + 0 - 1 = 0$$

$$-2w_1 - \frac{1}{2} = 0$$

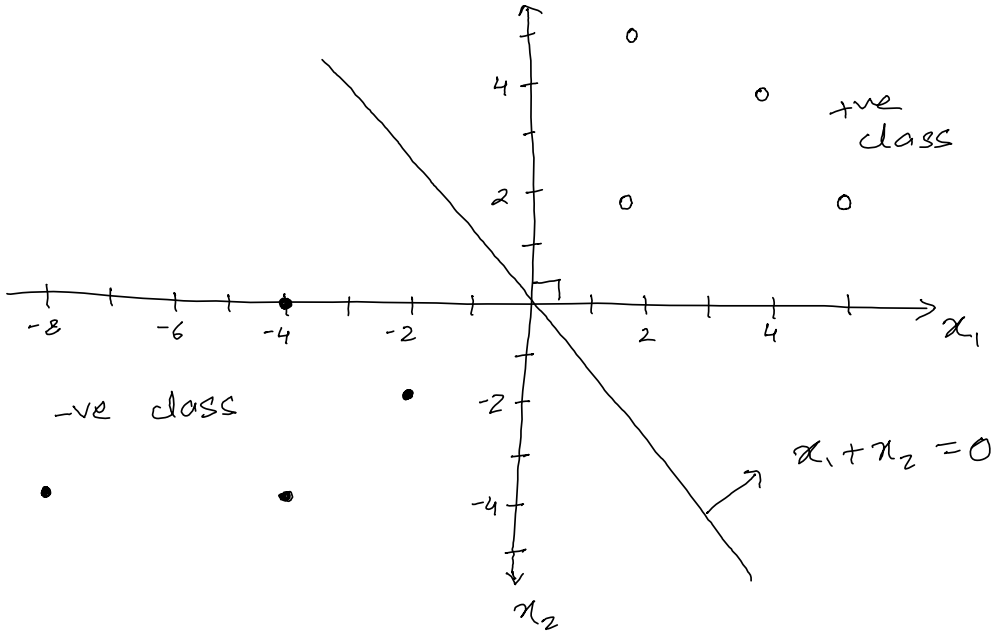
$$w_1 = -\frac{1}{4}$$

∴ Equation of the margin is:

$$-\frac{1}{4}x_1 - \frac{1}{4}x_2 = 0$$

$$\text{or, } x_1 + x_2 = 0$$

The weights of the linear are: $[1, 1, 0]$



c) $(-1, -1)$

Plugging $(-1, -1)$ into equation of the margin given:

$$x_1 + x_2 = (-1) + (-1) = -2$$

$\therefore$ $(-1, -1)$ falls in the -ve class