Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

Machine Learning (CS60050) **Mid-Semester Examination** **Spring Semester, 2022-2023**

Date: 15-Feb-2023 (Wed, AN) **Answer *all* questions.** **Maximum Marks:** 60

Q1. [Concept Learning] **10 marks**

Consider the following set of attributes with their listed domain values:

- **Fever:** { *High, Moderate, None* }
- **Cough:** { *Intense, Mild* }
- **RunningNose:** { *Yes, No* }
- **Weakness:** { *Extreme, Slight* }
- **Headache:** { *Yes, No* }
- **Saturation:** { *Good, Bad* }

Suppose, based on the values of the above mentioned attributes, you are trying to learn the concept whether someone has **Covid** or not. You are given with the following training data set (4 examples):

Example	Fever	Cough	RunningNose	Weakness	Headache	Saturation	Covid
1	High	Mild	No	Extreme	No	Bad	*Yes*
2	High	Mild	No	Slight	No	Bad	*Yes*
3	None	Intense	No	Slight	No	Good	*No*
4	High	Mild	No	Extreme	Yes	Good	*Yes*

Consider the space H of conjunctive hypotheses, which, for each attribute, either:

- indicates by a '?' that any value is acceptable; or
- specifies a single required value (e.g., *Mild* for **Cough**); or
- indicates by a 'ϕ' that no value is acceptable.

Let a version space (a subset of consistent hypotheses in H) be represented by an S set (specific boundary, at the top) and a G set (general boundary, at the bottom). Suppose the 4 training examples above are presented in order. Answer the following.

(a) What is the total size (cardinality) of the possible hypothesis space? **(2)**

Solution:

Each attribute can take the mention values as well as '?'. Additionally, there is one more hypothesis which takes nothing into consideration (all 'ϕ'). So, the total size (cardinality) of the possible hypothesis space = $(3+1) \times (2+1) \times (2+1) \times (2+1) \times (2+1) \times (2+1) + 1 = 973$.

(b) Applying *Candidate-Elimination* algorithm, draw a diagram showing the evolution of the version space for concept **Covid** given the training examples, by clearly expressing S_1, G_1, S_2, G_2, S_3, G_3, S_4, G_4. If the G set does not change given a new example, just write $G_{i+1} = G_i$ $(1 \leq i < 4)$ next to the drawing of G_i (similarly for S set as well). **(4)**

Solution:

$$S_0 = \langle \phi, \phi, \phi, \phi, \phi, \phi \rangle$$
$$G_0 = \langle ?, ?, ?, ?, ?, ? \rangle$$

$$S_1 = \langle \textit{High}, \textit{Mild}, \textit{No}, \textit{Extreme}, \textit{No}, \textit{Bad} \rangle$$
$$G_1 = \langle ?, ?, ?, ?, ?, ? \rangle \quad = G_0$$
$$S_2 = \langle \textit{High}, \textit{Mild}, \textit{No}, ?, \textit{No}, \textit{Bad} \rangle$$
$$G_2 = \langle ?, ?, ?, ?, ?, ? \rangle \quad = G_1$$

$$
\begin{array}{lcllll}
S_3 & = & \langle \, High,\ Mild,\ No,\ ?,\ No,\ Bad \, \rangle \quad = S_2 & & \\
G_3 & = & \langle \, High,\ ?,\ ?,\ ?,\ ?,\ ? \, \rangle & \langle \, ?,\ Mild,\ ?,\ ?,\ ?,\ ? \, \rangle & \langle \, ?,\ ?,\ ?,\ ?,\ ?,\ Bad \, \rangle \\
S_4 & = & \langle \, High,\ Mild,\ No,\ ?,\ ?,\ ? \, \rangle & & \\
G_4 & = & \langle \, High,\ ?,\ ?,\ ?,\ ?,\ ? \, \rangle & \langle \, ?,\ Mild,\ ?,\ ?,\ ?,\ ? \, \rangle & \\
 & & \qquad (G_4^1) & \qquad (G_4^2) &
\end{array}
$$

(c) Write down all the hypotheses in the final version space (the ones that lie between S_4 and G_4 according to the partial ordering relation *Less-Specific-Than)*. **(2)**

Solution:

$$
\begin{array}{lcl}
H_1 & = & \langle \, High,\ ?,\ No,\ ?,\ ?,\ ? \, \rangle \\
H_2 & = & \langle \, High,\ Mild,\ ?,\ ?,\ ?,\ ? \, \rangle \\
H_3 & = & \langle \, ?,\ Mild,\ No,\ ?,\ ?,\ ? \, \rangle
\end{array}
$$

(d) In the final version space, draw lines between hypotheses that are related by this relation. For example, there should be a line between $\langle ?, Mild, ?, ?, ? \rangle$ and $\langle ?, Mild, ?, Extreme, ?, ? \rangle$. **(2)**

Solution:

$$\begin{tikzcd}
 & & S_4 & & \\
H_1 \arrow[urr] & & H_2 \arrow[u] & & H_3 \arrow[ull] \\
 & G_4^1 \arrow[ul] \arrow[ur] & & G_4^2 \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Q2. **[Decision-Tree Learning]** $\boxed{\textbf{10 marks}}$

For a binary classification problem, consider the training examples shown in the following table.

Instance	A_1	A_2	A_3	Target Class
1	*True*	*True*	1.0	+
2	*True*	*True*	6.0	+
3	*True*	*False*	5.0	−
4	*False*	*False*	4.0	+
5	*False*	*True*	7.0	−
6	*False*	*True*	3.0	−
7	*False*	*False*	8.0	−
8	*True*	*False*	7.0	+
9	*False*	*True*	5.0	−

The attributes, A_1 and A_2, can take either *True* or *False* values, whereas A_3 is a *continuous* attribute. The **Target Class** can be either + (positive) or − (negative). Answer the following.

(a) What is the entropy of this collection of training examples with respect to positive (+) class? **(2)**

Solution:

There are 4 positive (+) examples and 5 negative (−) examples. Thus, $\mathbb{P}_{(+)} = \frac{4}{9}$ and $\mathbb{P}_{(-)} = \frac{5}{9}$.

The entropy w.r.t. the positive (+) class of the training examples is $= -\frac{4}{9}\log_2\left(\frac{4}{9}\right) = 0.52$.

The entropy w.r.t. the negative (−) class of the training examples is $= -\frac{5}{9}\log_2\left(\frac{5}{9}\right) = 0.47$.

The entropy of the training examples is $= -\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right) = 0.9911$.

(b) What are the information gains of A_1 and A_2 relative to these training examples? **(3)**

Solution:

For attribute A_1, the corresponding counts and probabilities are:

A_1	$+$	$-$
True	3	1
False	1	4

The entropy for A_1 is $= \frac{4}{9}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] + \frac{5}{9}\left[-\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right] = 0.7616$.

Therefore, the information gain for A_1 is $(0.9911 - 0.7616) = 0.2294$.

For attribute A_2, the corresponding counts and probabilities are:

A_2	$+$	$-$
True	2	3
False	2	2

The entropy for A_2 is $= \frac{5}{9}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right] + \frac{4}{9}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] = 0.9839$.

Therefore, the information gain for A_2 is $(0.9911 - 0.9839) = 0.0072$.

(c) For A_3, which is a continuous attribute, compute the information gain for every possible split. **(4)**

Solution:

A_3	**Class Label**	**Split Point**	**Entropy**	**Information Gain**
1.0	$+$	2.0	0.8484	0.1427
3.0	$-$	3.5	0.9885	0.0026
4.0	$+$	4.5	0.9183	0.0728
5.0	$-$			
5.0	$-$	5.5	0.9839	0.0072
6.0	$+$	6.5	0.9728	0.0183
7.0	$+$			
7.0	$-$	7.5	0.8889	0.1022

(d) According to the information gain, which is the best split point considering only A_3 attribute? **(1)**

Solution:

The best split for A_3 occurs at split point 2.0.

(e) What is the best attribute (among A_1, A_2, A_3) to split according to the information gain? **(1)**

Solution:

According to information gain, A_1 produces the best split.

Note: There is an error in the marks breakup, so it is possible to score 11 in this question!!

Q3. **[Bayesian Leaning]** **10 marks**

Consider the data set shown in the following table.

Instance	A	B	C	**Class**
1	0	0	1	$-$
2	1	0	1	$+$
3	0	1	0	$-$
4	1	0	0	$-$
5	1	0	1	$+$
6	0	0	1	$+$
7	1	1	0	$-$
8	0	0	0	$-$
9	0	1	0	$+$
10	1	1	1	$+$

The attributes, A, B and C, can take two values (either 1 or 0) and the **Class** can be either $+$ or $-$. Answer the following.

(a) Estimate the conditional probabilities for the following:
$\mathbb{P}(A=0 \mid +)$, $\mathbb{P}(B=1 \mid +)$, $\mathbb{P}(C=1 \mid +)$, $\mathbb{P}(A=0 \mid -)$, $\mathbb{P}(B=1 \mid -)$, $\mathbb{P}(C=1 \mid -)$. **(3)**

Solution:

$$\begin{aligned} \mathbb{P}(A=0 \mid +) &= \frac{2}{5} & \qquad \mathbb{P}(A=0 \mid -) &= \frac{3}{5} \\ \mathbb{P}(B=1 \mid +) &= \frac{2}{5} & \qquad \mathbb{P}(B=1 \mid -) &= \frac{2}{5} \\ \mathbb{P}(C=1 \mid +) &= \frac{4}{5} & \qquad \mathbb{P}(C=1 \mid -) &= \frac{1}{5} \end{aligned}$$

(b) Use the conditional probabilities in part (a) to predict the class label for a given test sample, $(\mathsf{A}=0,\ \mathsf{B}=1,\ \mathsf{C}=1)$, using the Naive Bayes approach. **(4)**

Solution:

$$\mathbb{P}(+ \mid \mathsf{A}=0,\mathsf{B}=1,\mathsf{C}=1) = \mathbb{P}(+).\mathbb{P}(\mathsf{A}=0,\mathsf{B}=1,\mathsf{C}=1 \mid +)$$

$$= \frac{\mathbb{P}(+).\mathbb{P}(\mathsf{A}=0 \mid +).\mathbb{P}(\mathsf{B}=1 \mid +).\mathbb{P}(\mathsf{C}=1 \mid +)}{\mathbb{P}(+).\mathbb{P}(\mathsf{A}=0 \mid +).\mathbb{P}(\mathsf{B}=1 \mid +).\mathbb{P}(\mathsf{C}=1 \mid +)+\mathbb{P}(-).\mathbb{P}(\mathsf{A}=0 \mid -).\mathbb{P}(\mathsf{B}=1 \mid -).\mathbb{P}(\mathsf{C}=1 \mid -)}$$

$$= \frac{\left(\frac{1}{2}\right).\left(\frac{2}{5}\right).\left(\frac{2}{5}\right).\left(\frac{4}{5}\right)}{\left(\frac{1}{2}\right).\left(\frac{2}{5}\right).\left(\frac{2}{5}\right).\left(\frac{4}{5}\right)+\left(\frac{1}{2}\right).\left(\frac{3}{5}\right).\left(\frac{2}{5}\right).\left(\frac{1}{5}\right)} = \frac{8}{11}.$$

$$\therefore\ \mathbb{P}(- \mid \mathsf{A}=0,\mathsf{B}=1,\mathsf{C}=1) = 1-\mathbb{P}(+ \mid \mathsf{A}=0,\mathsf{B}=1,\mathsf{C}=1) = \frac{3}{11}.$$

Since $\mathbb{P}(+ \mid \mathsf{A}=0,\mathsf{B}=1,\mathsf{C}=1) > \mathbb{P}(- \mid \mathsf{A}=0,\mathsf{B}=1,\mathsf{C}=1)$, therefore the predicted class label will be '$+$'.

(c) Are the variables, A and B, independent with values, $\mathsf{A}=1$ and $\mathsf{B}=1$? **(1.5)**

Solution:

$\mathbb{P}(\mathsf{A}=1)=\frac{1}{2}$ and $\mathbb{P}(\mathsf{B}=1)=\frac{2}{5}$.
Since $\mathbb{P}(\mathsf{A}=1,\mathsf{B}=1)=\frac{1}{5}=\mathbb{P}(\mathsf{A}=1).\mathbb{P}(\mathsf{B}=1)$, therefore A and B are independent.

(d) Are these variables, A and B, conditionally independent with values, $\mathsf{A}=1$ and $\mathsf{B}=1$, given the class '$+$'? **(1.5)**

Solution:

$\mathbb{P}(\mathsf{A}=1 \mid +)=\frac{3}{5}$ and $\mathbb{P}(\mathsf{B}=1 \mid +)=\frac{2}{5}$.
Since $\mathbb{P}(\mathsf{A}=1,\mathsf{B}=1 \mid +)=\frac{1}{5}\neq\frac{6}{25}=\mathbb{P}(\mathsf{A}=1 \mid +).\mathbb{P}(\mathsf{B}=1 \mid +)$, therefore A and B are not conditionally independent given the class '$+$'.

Q4. [Instance-based Learning]

6 marks

Consider the one-dimensional data set shown in the following table.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	$-$	$-$	$+$	$+$	$+$	$-$	$-$	$+$	$-$	$-$

Here, x can take continuous values and y has two labels ($+$ and $-$). Answer the following.

(a) Classify the data point $x = 5.0$ according to its 1-, 3-, 5-, and 9- nearest neighbors (using majority voting). Briefly explain your results. **(2)**

Solution:

- 1-nearest neighbor: $+$
  $\big[$ since the nearest data point, $x = 4.9$, has '$+$' label $\big]$
- 3-nearest neighbor: $-$
  $\big[$ since 3 nearest data points, $x = 4.9, 5.2, 5.3$, have one '$+$' and two '$-$' labels $\big]$
- 5-nearest neighbor: $+$
  $\big[$ since 5 nearest data points, either $x = 4.6, 4.9, 5.2, 5.3, 5.5$ or $x = 4.5, 4.6, 4.9, 5.2, 5.3$ (both cases) have three '$+$' and two '$-$' labels $\big]$
- 9-nearest neighbor: $-$
  $\big[$ since 9 nearest data points include all points except either $x = 0.5$ or $x = 9.5$ and in both cases, we have four '$+$' and five '$-$' labels $\big]$

(b) Again classify the same data point $x = 5.0$ according to its 1-, 3-, 5-, and 9- nearest neighbors (using distance-weighted voting). Briefly explain your results.

Note: In distance-weighted scheme, the weights are inversely proportional to the Euclidean distances between two data points. **(4)**

Solution:

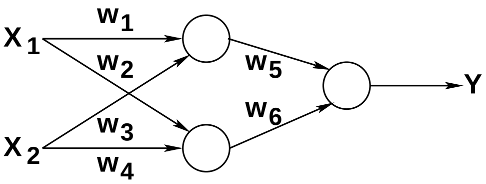
- 1-nearest neighbor: $+$
  $\big[$ the nearest data point, $x = 4.9$, has '$+$' label $\big]$
- 3-nearest neighbor: $+$
  $\big[$ since 3 nearest data points, $x = 4.9, 5.2, 5.3$, have one '$+$' and two '$-$' labels, the combined distance-weight with the '$+$' labeled point (which is, $\frac{1}{0.1} = 10$) is more than the combined distance-weight with the '$-$' labeled points (which is, $\frac{1}{0.2} + \frac{1}{0.3} = 8.67$) $\big]$
- 5-nearest neighbor: $+$
  $\big[$ since 5 nearest data points, either $x = 4.6, 4.9, 5.2, 5.3, 5.5$ or $x = 4.5, 4.6, 4.9, 5.2, 5.3$ (both cases) have three '$+$' and two '$-$' labels, the combined distance-weight with the '$+$' labeled points (which is, $\frac{1}{0.1} + \frac{1}{0.4} + \frac{1}{0.5} = 14.5$) is more than the combined distance-weight with the '$-$' labeled points (which is, $\frac{1}{0.2} + \frac{1}{0.3} = 8.67$) $\big]$
- 9-nearest neighbor: $+$
  $\big[$ since 9 nearest data points include all points except either $x = 0.5$ or $x = 9.5$ and in both cases, we have four '$+$' and five '$-$' labels; the combined distance-weight with the '$+$' labeled points (which is, $\frac{1}{0.1} + \frac{1}{0.4} + \frac{1}{0.5} + \frac{1}{0.5} = 16.5$) is more than the combined distance-weight with the '$-$' labeled points (which is, $\frac{1}{0.2} + \frac{1}{0.3} + \frac{1}{2.0} + \frac{1}{2.0} + \frac{1}{4.5} = 12.56$) $\big]$

Q5. [Perceptrons]

4 marks

Suppose we have a multi-layer perceptron network (shown below) with linear activation units. In other words, the output of each unit is a constant C multiplied by the weighted sum of inputs.

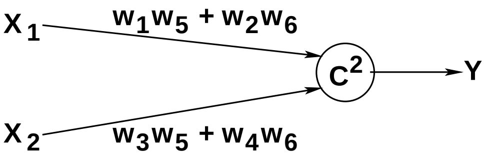
Answer the following.



(a) Can any function that is represented by the above network also be represented by a single unit perceptron? If yes, draw the equivalent perceptron detailing the weights and the activation function. Otherwise, briefly explain why not possible. **(2)**

Solution:

Yes. We can use C^2 as the activation function to be multiplied with the weighted sum of inputs. Here, the weights of the inputs, X_1 and X_2 will be, $(w_1.w_5 + w_2.w_6)$ and $(w_3.w_5 + w_4.w_6)$, respectively. Below is the schematic description of the same.



(b) Can the space of functions that is represented by the above network also be represented by linear regression? If yes, present the linear regression function detailing the coefficients. Otherwise, briefly explain why not possible. **(2)**

Solution:

Yes. Any function in the given network has the following form:

$$Y \quad = \quad C^2.(w_1.w_5 + w_2.w_6).X_1 + C^2.(w_3.w_5 + w_4.w_6).X_2 \quad = \quad \beta_1.X_1 + \beta_2.X_2.$$

This is linear regression on inputs, X_1 and X_2, with constant coefficients,

$$\beta_1 = C^2.(w_1.w_5 + w_2.w_6) \qquad \textit{and} \qquad \beta_2 = C^2.(w_3.w_5 + w_4.w_6).$$

Q6. **[Logistic Regression and Neural Network]** **10 marks**

For a binary logistic regression model with input attribute set x and an output y, having an internal sigmoid activation function (of the form $\sigma(z) = \frac{1}{1+e^{-z}}$, where $z = w^T.x$ with weight vector w), we predict the output $y = 1$ when $\mathbb{P}(y = 1 \mid x\, ;\, w) \geq \frac{1}{2}$.

(a) Prove that, this logistic regression model is also a linear classifier. **(4)**

Solution:

Using the parametric form for $\mathbb{P}(y = 1 \mid x\, ;\, w)$:

$$\begin{aligned} \mathbb{P}(y = 1 \mid x\, ;\, w) \geq \frac{1}{2} \quad &\Longrightarrow \quad \frac{1}{1+e^{-w^T.x}} \geq \frac{1}{2} \\ &\Longrightarrow \quad 1+e^{-w^T.x} \leq 2 \\ &\Longrightarrow \quad e^{-w^T.x} \leq 1 \\ &\Longrightarrow \quad -w^T.x \leq 0 \\ &\Longrightarrow \quad w^T.x \geq 0 \end{aligned}$$

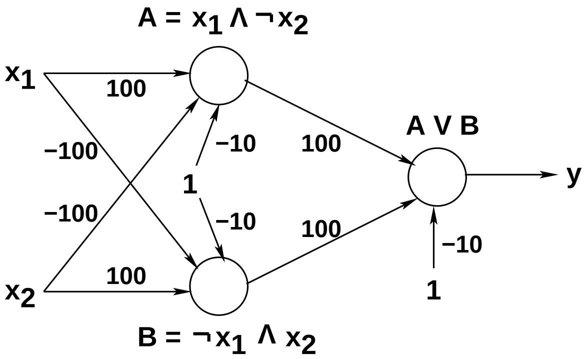
Therefore, we predict $y = 1$ if $w^T.x \geq 0$. [Proved]

(b) Consider the XOR function $y = (x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)$. We can alternatively express this as, $y = \begin{cases} \geq \frac{1}{2}, & \text{if } x_1 \neq x_2 \\ < \frac{1}{2}, & \text{otherwise} \end{cases}$. Using the above-mentioned binary logistic regrassion model as a unit having binary inputs $x_0 (= 1)$, x_1 and x_2 (i.e. $x = [1, x_1, x_2]^T$), and output y, draw a fully connected three-unit Neural Network that realizes the function $y = (x_1 \texttt{ XOR } x_2)$. Show the suitable weight vector, $w = [w_0, w_1, w_2]^T$, for each unit clearly. **(6)**

Solution:

Let $y = (A \vee B)$, where $A = (x_1 \wedge \neg x_2)$ and $B = (\neg x_1 \wedge x_2)$.

In the Neural Netwrk, we shall have the first layer implementing A and B (two AND formulas) through two units and the last layer/unit implementing y (the OR formula).



Note:

- Many combinations of weight can realize the same XOR function, provided that,
  - ∗ when the weighted sum is $-$ve, the sigmoid output will be $< \frac{1}{2}$;
  - ∗ when the weighted sum is $+$ve, the sigmoid output will be $\geq \frac{1}{2}$.
- If the relative magnitude of weights is skewed, then the output may also be skewed.

Q7. **[Linear Classifier and Support Vector Machine]** **10 marks**

Consider a set of 2-dimensional training data points (x_1, x_2) belonging to two classes '$+1$' and '-1', respectively, as shown below.

- Class '$+1$': $(3,1)$; $(3,-1)$; $(6,1)$; $(6,-1)$
- Class '-1': $(1,0)$; $(0,1)$; $(0,-1)$; $(-1,0)$

We design a linear hard-margin SVM to classify these linearly separable points. Answer the following.

(a) Pictorially (graphically) represent the constellation of data points and the optimal separating hyperplane. Write the equation of the optimal separator and mention the width of the margin (figuring it out manually from the diagram/graph you have shown). **(2)**

Solution:

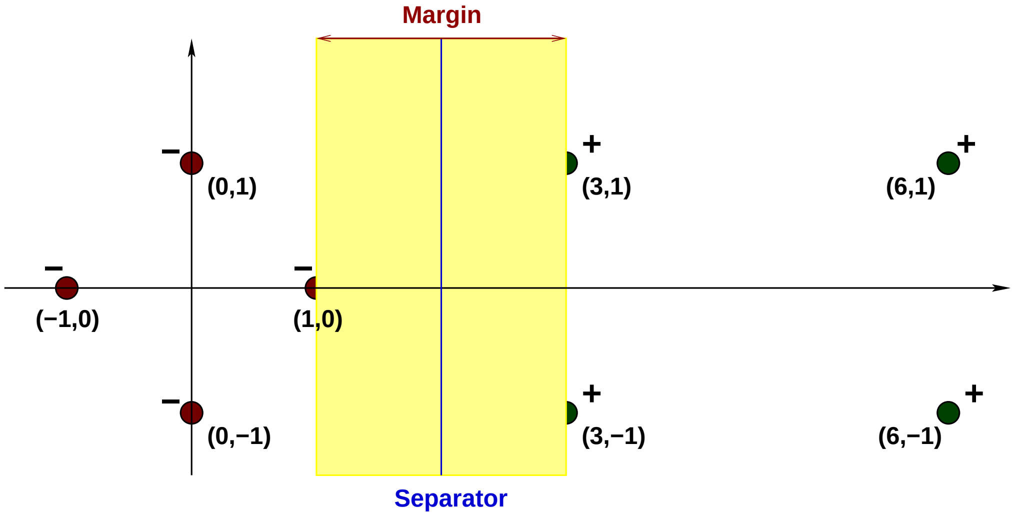
The constellation of data points and the optimal separator (with margin) is presented below.

SVM tries to maximize the margin between two classes of data points. Therefore, the optimal decision boundary crosses the point $(2,0)$ and is parallel to vertical axis. Thus, the equation of optimal separator is given as, $x_1 - 2 = 0$, having the width of the margin $= 2$-units.

(b) Which data points are the support vectors here? **(2)**

Solution:

Support vectors are $(3,1)$, $(3,-1)$ and $(1,0)$. These three points have minimum perpendicular distance from the separator line (Euclidean distance of 1 unit).



(c) What weight vector and threshold (bias) value are being learnt using hard-margin SVM training algorithm with these eight training points? Show the detailed calculations. **(4)**

Solution:

Let the weight vector learnt be of the for $w = [w_1, w_2]^T$ and threshold/bias is b.

From the three support vectors, $(3,1)$, $(3,-1)$ and $(1,0)$ (which are the closest points from the separating line), we get,

$$\begin{aligned} 3w_1 + w_2 + b &= +1 \\ 3w_1 - w_2 + b &= +1 \\ w_1 + b &= -1 \end{aligned}$$

Solving above equations, we get, $w_1 = 1$, $w_2 = 0$, and $b = -2$.

(d) Using the learnt weights and threshold values (in part (c)), what is the margin you get for the optimal classifier? Derive mathematically. **(2)**

Solution:

$$\text{Margin} = \frac{2}{||w||} = \frac{2}{\sqrt{w_1^2 + w_2^2}} = 2.$$

— END —