INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Stamp / Signature of the Invigilator

EXAMINATION (Mid Semester)	**SEMESTER (Spring 2023-2024)**

Roll Number										**Section**		**Name**	

Subject Number	C	S	6	0	0	5	0	**Subject Name**	*MACHINE LEARNING*

Department / Center of the Student		**Additional sheets**	

Important Instructions and Guidelines for Students

1. You must occupy your seat as per the Examination Schedule/Sitting Plan.
2. Do not keep mobile phones or any similar electronic gadgets with you even in the switched off mode.
3. Loose papers, class notes, books or any such materials must not be in your possession, even if they are irrelevant to the subject you are taking examination.
4. Data book, codes, graph papers, relevant standard tables/charts or any other materials are allowed only when instructed by the paper-setter.
5. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, exchange of these items or any other papers (including question papers) is not permitted.
6. Write on both sides of the answer script and do not tear off any page. **Use last page(s) of the answer script for rough work.** Report to the invigilator if the answer script has torn or distorted page(s).
7. It is your responsibility to ensure that you have signed the Attendance Sheet. Keep your Admit Card/Identity Card on the desk for checking by the invigilator.
8. You may leave the examination hall for wash room or for drinking water for a very short period. Record your absence from the Examination Hall in the register provided. Smoking and the consumption of any kind of beverages are strictly prohibited inside the Examination Hall.
9. Do not leave the Examination Hall without submitting your answer script to the invigilator. **In any case, you are not allowed to take away the answer script with you.** After the completion of the examination, do not leave the seat until the invigilators collect all the answer scripts.
10. During the examination, either inside or outside the Examination Hall, gathering information from any kind of sources or exchanging information with others or any such attempt will be treated as **'unfair means'**. Do not adopt unfair means and do not indulge in unseemly behavior.

Violation of any of the above instructions may lead to severe punishment.

Signature of the Student

To be filled in by the examiner

Question Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Marks Obtained											

Marks obtained (in words)	**Signature of the Examiner**	**Signature of the Scrutineer**

Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

Mid-Semester Examination **Machine Learning (CS60050)** **Spring Semester 2023-2024**

February 2024 **Answer *all* questions.** Maximum Marks: 60

— Write your answers at indicated places inside the question paper. —

— This page is kept blank intentionally. —

— The question paper starts from the next page. —

Q1. [Concept Learning] **6 marks**

Consider the following set of 4 training examples to train a robot janitor to predict whether or not an office contains a recycling bin.

Designation	Floor	Department	Office Size	Recycle Bin
Faculty	Second	CS	Medium	Yes
Faculty	Second	EE	Medium	Yes
PostDoc	Second	CS	Small	No
Faculty	First	CS	Medium	Yes

Here, each of the four attributes can take two domain values as mentioned in the table: **Designation =** *Faculty / PostDoc*; **Floor =** *First / Second*; **Department =** *CS / EE*; and **Office Size =** *Small / Medium*.

Consider the space H of conjunctive hypotheses, which, for each attribute, indicates by either '?' (any value acceptable), or a specific value (e.g., *Second* for **Floor**), or 'ϕ' (no value 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). Answer the following.

(a) Calculate 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 = $(2+1)\times(2+1)\times(2+1)\times(2+1)+1 = 82$.

(b) Give a sequence of S and G boundary sets computed by CANDIDATE-ELIMINATION algorithm when the examples are taken in the same order as presented in the above table. **(4)**

Solution:

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

$$S_1 = \langle \mathit{Faculty}, \mathit{Second}, \mathit{CS}, \mathit{Medium} \rangle$$
$$G_1 = \langle ?, ?, ?, ? \rangle \quad = G_0$$

$$S_2 = \langle \mathit{Faculty}, \mathit{Second}, ?, \mathit{Medium} \rangle$$
$$G_2 = \langle ?, ?, ?, ? \rangle \quad = G_1$$

$$S_3 = \langle \mathit{Faculty}, \mathit{Second}, ?, \mathit{Medium} \rangle \quad = S_2$$
$$G_3 = \langle \mathit{Faculty}, ?, ?, ? \rangle \qquad \langle ?, ?, ?, \mathit{Medium} \rangle$$

$$S_4 = \langle \mathit{Faculty}, ?, ?, \mathit{Medium} \rangle$$
$$G_4 = \underset{(G_4^1)}{\langle \mathit{Faculty}, ?, ?, ? \rangle} \qquad \underset{(G_4^2)}{\langle ?, ?, ?, \mathit{Medium} \rangle} \quad = G_3$$

Q2. [Decision Tree Learning] **6 marks**

Right is a dataset of the 2201 passengers and crew aboard the RMS Titanic, which disastrously sunk on April 15th, 1912. For every combination of three attribute variables (**Class, Gender, Age**), we have the counts of how many people survived and did not. Here, we are primarily interested in predicting the outcome variable, **Survival (S)** (*No / Yes*), from the input attributes, **Class (C)**, **Gender (G)** and **Age (A)** by building a decision tree.

Attributes			Survival		Count of
Class	**Gender**	**Age**	*No*	*Yes*	**Passengers**
1st	Male	Child	0	5	5
1st	Male	Adult	118	57	175
1st	Female	Child	0	1	1
1st	Female	Adult	4	140	144
2nd	Male	Child	35	24	59
2nd	Male	Adult	1211	281	1492
2nd	Female	Child	17	27	44
2nd	Female	Adult	105	176	281
		Total:	1490	711	2201

Which attribute should you choose at the root of your decision tree? Show the detailed calculations leveraging the entropy-based information gain measures obtained for each attribute you choose. **(6)**

Solution:

Class	(Survival) *No*	*Yes*	Total
1st	122	203	325
2nd	1368	508	1876

Gender	(Survival) *No*	*Yes*	Total
Male	1364	367	1721
Female	126	344	470

Age	(Survival) *No*	*Yes*	Total
Child	52	57	109
Adult	1438	654	2092

For **Class**:

$$\begin{aligned} Gain[C] &= Entropy(S) - \left[\left(\frac{325}{2201}\right)\cdot Entropy(S \mid C = 1st) + \left(\frac{1876}{2201}\right)\cdot Entropy(S \mid C = 2nd)\right] \\ &= Entropy(S) - \left[\frac{325}{2201}\cdot\left(-\left(\frac{203}{325}\right)\cdot\log_2\left(\frac{203}{325}\right) - \left(\frac{122}{325}\right)\cdot\log_2\left(\frac{122}{325}\right)\right)\right. \\ &\qquad \left. + \frac{1876}{2201}\cdot\left(-\left(\frac{508}{1876}\right)\cdot\log_2\left(\frac{508}{1876}\right) - \left(\frac{1368}{1876}\right)\cdot\log_2\left(\frac{1368}{1876}\right)\right)\right] \\ &= Entropy(S) - 0.595518 \end{aligned}$$

For **Gender**:

$$\begin{aligned} Gain[G] &= Entropy(S) - \left[\left(\frac{1721}{2201}\right)\cdot Entropy(S \mid G = Male) + \left(\frac{470}{2201}\right)\cdot Entropy(S \mid G = Female)\right] \\ &= Entropy(S) - \left[\frac{1721}{2201}\cdot\left(-\left(\frac{367}{1721}\right)\cdot\log_2\left(\frac{367}{1721}\right) - \left(\frac{1364}{1721}\right)\cdot\log_2\left(\frac{1364}{1721}\right)\right)\right. \\ &\qquad \left. + \frac{470}{2201}\cdot\left(-\left(\frac{344}{470}\right)\cdot\log_2\left(\frac{344}{470}\right) - \left(\frac{126}{470}\right)\cdot\log_2\left(\frac{126}{470}\right)\right)\right] \\ &= Entropy(S) - 0.530438 \end{aligned}$$

For **Age**:

$$\begin{aligned} Gain[A] &= Entropy(S) - \left[\left(\frac{109}{2201}\right)\cdot Entropy(S \mid A = Child) + \left(\frac{2092}{2201}\right)\cdot Entropy(S \mid A = Adult)\right] \\ &= Entropy(S) - \left[\frac{109}{2201}\cdot\left(-\left(\frac{57}{109}\right)\cdot\log_2\left(\frac{57}{109}\right) - \left(\frac{52}{109}\right)\cdot\log_2\left(\frac{52}{109}\right)\right)\right. \\ &\qquad \left. + \frac{2092}{2201}\cdot\left(-\left(\frac{654}{2092}\right)\cdot\log_2\left(\frac{654}{2092}\right) - \left(\frac{1438}{2092}\right)\cdot\log_2\left(\frac{1438}{2092}\right)\right)\right] \\ &= Entropy(S) - 0.624692 \end{aligned}$$

Since $Entropy(S) = -\frac{711}{2201}\log_2\left(\frac{711}{2201}\right) - \frac{1490}{2201}\log_2\left(\frac{1490}{2201}\right) > 0$,

Gender attribute will have highest Information Gain and will be choosen as root of the decision tree.

Q3. [Bayesian Learning] **7 marks**

Consider the dataset shown to the right. It has 4 features $\mathbf{X} = (x_1, x_2, x_3, x_4)$ and the outcome can take any of 3 labels, $y \in \{+1, 0, -1\}$. Assume that, the probabilities, $Pr(\mathbf{X} \mid y)$ and $Pr(y)$, are both Bernoulli distributions. Answer the questions that follow under the Naive Bayes assumption.

x_1	x_2	x_3	x_4	y
1	1	0	1	+1
0	1	1	0	+1
1	0	1	1	0
0	1	1	1	0
0	1	0	0	−1
1	0	0	1	−1
0	0	1	1	−1

(a) Compute the Maximum Likelihood Estimates for $Pr(x_i = 1 \mid y)$, for all $i \in \{1,2,3,4\}$ and for all $y \in \{+1, 0, -1\}$. **(3)**

Solution:

$Pr(x_i = 1 \mid y)$	$y = +1$	$y = 0$	$y = -1$
$x_1 = 1$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$
$x_2 = 1$	1	$\frac{1}{2}$	$\frac{1}{3}$

$Pr(x_i = 1 \mid y)$	$y = +1$	$y = 0$	$y = -1$
$x_3 = 1$	$\frac{1}{2}$	1	$\frac{1}{3}$
$x_4 = 1$	$\frac{1}{2}$	1	$\frac{2}{3}$

(b) Compute the Maximum Likelihood Estimates for the prior probabilities, $Pr(y = +1)$, $Pr(y = 0)$, and $Pr(y = -1)$. **(1)**

Solution:

$Pr(y = +1) = \boxed{\frac{2}{7}}$, $Pr(y = 0) = \boxed{\frac{2}{7}}$, $Pr(y = -1) = \boxed{\frac{3}{7}}$.

(c) Use the values computed in the above two parts to classify the data point $(x_1 = 1,\ x_2 = 1,\ x_3 = 1,\ x_4 = 1)$ as belonging to class $+1$, 0 or -1. **(3)**

Solution:

According to Naive Bayes assumption, attributes are independent given y, thus we can write the conditional joint probability as,

$$\begin{aligned} Pr(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1) &= Pr(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1 \mid y) \cdot Pr(y) \\ &= Pr(y) \cdot \prod_{i=1}^{4} Pr(x_i = 1 \mid y). \end{aligned}$$

We calculate the probability given different values of y and pick the one with highest probability:

$$Pr(y = +1) \cdot \prod_{i=1}^{4} Pr(x_i = 1 \mid y = +1) = \frac{2}{7} \cdot \left(\frac{1}{2} \times 1 \times \frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{28}$$

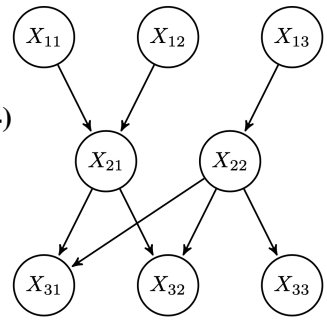
$$Pr(y = 0) \cdot \prod_{i=1}^{4} Pr(x_i = 1 \mid y = 0) = \frac{2}{7} \cdot \left(\frac{1}{2} \times \frac{1}{2} \times 1 \times 1\right) = \frac{1}{14}$$

$$Pr(y = -1) \cdot \prod_{i=1}^{4} Pr(x_i = 1 \mid y = -1) = \frac{3}{7} \cdot \left(\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3}\right) = \frac{2}{189}$$

Since $y = 0$ yields the largest value, we classify the given data point as $\hat{y} = 0$.

Q4. [Bayesian Network] **5 marks**

Consider the Bayesian network with Boolean variables shown to the right.



(a) Answer whether the following statements are True / False providing proper justification using D-separation method. **(4)**

(i) $X_{31} \perp\!\!\!\perp X_{33} \mid X_{32}$,
i.e., "X_{31} is conditionally independent of X_{33} given X_{32}".

(ii) $X_{21} \perp\!\!\!\perp X_{33} \mid \{X_{11}, X_{12}\}$,
i.e., "X_{21} is conditionally independent of X_{33} given X_{11} and X_{12}".

Solution:

(i) *False*

Both the possible paths (shown below) from X_{31} to X_{33} are active:

- $X_{31} \leftarrow X_{21} \rightarrow X_{32}$ (observed) $\leftarrow X_{22} \rightarrow X_{33}$ (all three overlapping triples are active triples)
- $X_{31} \leftarrow X_{22} \rightarrow X_{33}$ (this is an active triple)

(ii) *True*

Both the possible paths (shown below) from X_{21} to X_{33} are inactive:

- $X_{21} \rightarrow X_{32} \leftarrow X_{22} \rightarrow X_{33}$ (blocking collider at X_{32}).
- $X_{21} \rightarrow X_{31} \leftarrow X_{22} \rightarrow X_{33}$ (blocking collider at X_{31}).

(b) Write the joint probability $Pr(X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{31}, X_{32}, X_{33})$ factored according to the Bayes network and conditional probabilities. **(1)**

Solution:

$$\begin{aligned} Pr(X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{31}, X_{32}, X_{33}) \;=\; & Pr(X_{11}) \cdot Pr(X_{12}) \cdot Pr(X_{13}) \cdot \\ & Pr(X_{21} \mid X_{11}, X_{12}) \cdot Pr(X_{22} \mid X_{13}) \cdot \\ & Pr(X_{31} \mid X_{21}, X_{22}) \cdot Pr(X_{32} \mid X_{21}, X_{22}) \cdot Pr(X_{33} \mid X_{22}) \end{aligned}$$

Q5. [Instance-based Learning] **6 marks**

Consider the 2-dimensional data set shown in the following table.

(x,y)	$(1,1)$	$(2,1)$	$(2,2)$	$(3,0)$	$(3,2)$	$(3,3)$	$(3.4,1)$	$(4,3)$	$(5,2)$
Class	$-$	$+$	$-$	$+$	$+$	$+$	$-$	$-$	$-$

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

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

Solution:

- 1-*nearest neighbour:* Class of $(3,1)$ will be $-$
  [since the nearest data point, $(3.4,1)$, has '$-$' label]
- 3-*nearest neighbour:* Class of $(3,1)$ will be $+$
  [since any 3 nearest data points, i.e. $\langle\,(2,1),(3,0),(3.4,1)\,\rangle$ or $\langle\,(3,0),(3,2),(3.4,1)\,\rangle$ or $\langle\,(2,1),(3,2),(3.4,1)\,\rangle$, have two '$+$' and one '$-$' labels]
- 5-*nearest neighbour:* Class of $(3,1)$ will be $+$
  [since 5 nearest data points, $\langle\,(2,1),(2,2),(3,0),(3,2),(3.4,1)\,\rangle$, have three '$+$' and two '$-$' labels]
- 9-*nearest neighbour:* Class of $(3,1)$ will be $-$
  [since 9 nearest data points include all points, we have four '$+$' and five '$-$' labels]

(b) Again classify the same data point $(x,y) = (3,1)$ according to its 1-, 3-, 5-, and 9- nearest neighbours (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 neighbour:* Class of $(3,1)$ will be $-$
  [since the nearest data point, $(3.4,1)$, has '$-$' label]
- 3-*nearest neighbour:* Class of $(3,1)$ will be $-$
  [since any 3 nearest data points, i.e. $\langle\,(2,1),(3,0),(3.4,1)\,\rangle$ or $\langle\,(3,0),(3,2),(3.4,1)\,\rangle$ or $\langle\,(2,1),(3,2),(3.4,1)\,\rangle$, have two '$+$' and one '$-$' labels, the combined distance-weight with two '$+$' labeled points (which is, $\frac{1}{1}+\frac{1}{1}=2$) is less than the distance-weight with the '$-$' labeled point (which is, $\frac{1}{0.4}=2.5$)]
- 5-*nearest neighbour:* Class of $(3,1)$ will be $-$
  [since 5 nearest data points, $\langle\,(2,1),(2,2),(3,0),(3,2),(3.4,1)\,\rangle$, have three '$+$' and two '$-$' labels, the combined distance-weight with three '$+$' labeled points (which is, $\frac{1}{1}+\frac{1}{1}+\frac{1}{1}=3$) is less than the combined distance-weight with two '$-$' labeled points (which is, $\frac{1}{\sqrt{2}}+\frac{1}{0.4}\approx 3.21$)]
- 9-*nearest neighbour:* Class of $(3,1)$ will be $-$
  [since 9 nearest data points include all points, we have four '$+$' and five '$-$' labels; the combined distance-weight with four '$+$' labeled points (which is, $\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{2}=3.5$) is less than the combined distance-weight with five '$-$' labeled points (which is, $\frac{1}{2}+\frac{1}{\sqrt{2}}+\frac{1}{0.4}+\frac{1}{\sqrt{5}}+\frac{1}{\sqrt{5}}\approx 4.6$)]

Q6. [Linear Classification – Perceptron Learning] **6 marks**

The table shown to the right is a list of sample points in the 2-dimensional space ($\mathbb{R}^2$). Suppose that, we run the perceptron learning algorithm on these sample points (as per the mentioned order). We record the total number of times each point participates in a stochastic gradient descent step because it is misclassified (refer to the rightmost column), throughout the run of the algorithm. Answer the following.

x_1	x_2	y	# Misclassified
-3	2	$+1$	0
-1	1	$+1$	0
-1	-1	-1	2
2	2	-1	1
1	-1	-1	0

(a) Suppose that the learning rate is $\eta = 1$ and the initial weight vector is $\mathbf{w}^{(0)} = [-3, 2, 1]$, where the last component is the weight (w_0) for the threshold / bias term ($x_0 = 1$). What is the equation of the separating line found by the algorithm, in terms of the features x_1 and x_2? **(4)**

Solution:

At each iteration, the weights are updated by picking a misclassified point and applying the update rule. The learned weights are $\mathbf{w} = \mathbf{w}^{(0)} + \eta \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \mathbf{X}_i$, where the variable α_i is the number of times the i^{th} point is misclassified.

Recall that, we augment each sample point $\mathbf{X}_i$ with default $x_0 = 1$ for the bias, i.e. $\mathbf{X}_i = [x_1, x_2, x_0]^{(i)}$. Since the weight update happens only for the misclassified points, thus we have:

$$\mathbf{w} \quad = \quad [-3,2,1] \; + \; 2 \cdot -1 \cdot [-1,-1,1] \; + \; 1 \cdot -1 \cdot [2,2,1] \quad = \quad [-3,2,-2].$$

Therefore, the equation of the separating line is: $-3x_1 + 2x_2 - 2 = 0$.

(Alternatively, $-3x_2 + 2x_1 - 2 = 0$)

(b) Will our result (that is, execution and outcome of the perceptron learning algorithm) differ if we add an additional training point $(2, -2)$ having the label '+1'? Explain. **(2)**

Solution:

The data would no longer be linearly separable, so the perceptron algorithm would not terminate.

Q7. [Logistic Regression and Gradients] **12 marks**

For a logistic regression model f having two output class lebels, $y \in \{0,1\}$, let the logistic loss be defined as: $L(x,y;\mathbf{w}) = -y\log\big(f(x;\mathbf{w})\big) - (1-y)\log\big(1-f(x;\mathbf{w})\big)$, where f has a range of $[0,1]$.

In this problem, we want to pick a suitable logistic function to approximate f for the loss. Suppose, you are unsure between chosing a sigmoid function, $g(x;\mathbf{w})$, or a shifted tanh function, $h(x;\mathbf{w})$, where:

$$g(x;\mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T.x}} \qquad \text{and} \qquad h(x;\mathbf{w}) = \frac{1}{2}\tanh(\mathbf{w}^T.x) + \frac{1}{2} \quad \text{with} \quad \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

In particular, answer the following.

(a) Before picking f, is it possible to compute the gradient of the logistic loss function (i.e., differentiate L with respect to $\mathbf{w}$)? If yes, derive what is $\frac{\partial L(x,y;\mathbf{w})}{\partial \mathbf{w}}$. If no, justify. **(2)**

Solution:

Yes! We use the chain rule:

$$\begin{aligned}\frac{\partial L(x,y;\mathbf{w})}{\partial \mathbf{w}} &= -y\cdot\frac{1}{f(x;\mathbf{w})}\cdot\frac{\partial f(x;\mathbf{w})}{\partial \mathbf{w}} + (1-y)\cdot\frac{1}{1-f(x;\mathbf{w})}\cdot\frac{\partial f(x;\mathbf{w})}{\partial \mathbf{w}} \\ &= \left(\frac{f(x;\mathbf{w})-y}{f(x;\mathbf{w})\cdot\big(1-f(x;\mathbf{w})\big)}\right)\cdot\frac{\partial f(x;\mathbf{w})}{\partial \mathbf{w}}\end{aligned}$$

(b) Next, if you wish to substitute g or h for f, then first derive what are $\frac{\partial g(x;\mathbf{w})}{\partial \mathbf{w}}$ and $\frac{\partial h(x;\mathbf{w})}{\partial \mathbf{w}}$. **(4)**

Solution:

$$\begin{aligned}\frac{\partial g(x;\mathbf{w})}{\partial \mathbf{w}} &= -\big(1+e^{-\mathbf{w}^T.x}\big)^{-2}\cdot\frac{\partial}{\partial \mathbf{w}}\big(1+e^{-\mathbf{w}^T.x}\big) \\ &= \frac{x\cdot e^{-\mathbf{w}^T.x}}{\big(1+e^{-\mathbf{w}^T.x}\big)^2} \qquad \text{(this is a valid answer)} \\ &= x\cdot\frac{1}{\big(1+e^{-\mathbf{w}^T.x}\big)}\cdot\frac{e^{-\mathbf{w}^T.x}}{\big(1+e^{-\mathbf{w}^T.x}\big)} \\ &= x\cdot g(x;\mathbf{w})\cdot\big(1-g(x;\mathbf{w})\big)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial h(x;\mathbf{w})}{\partial \mathbf{w}} &= \frac{1}{2}\cdot\frac{\partial\tanh(\mathbf{w}^T.x)}{\partial \mathbf{w}} \\ &= \frac{1}{2}\cdot\frac{\partial}{\partial \mathbf{w}}\left(\frac{e^{\mathbf{w}^T.x}-e^{-\mathbf{w}^T.x}}{e^{\mathbf{w}^T.x}+e^{-\mathbf{w}^T.x}}\right) \\ &= \frac{1}{2}\left[x\cdot\frac{\big(e^{\mathbf{w}^T.x}+e^{-\mathbf{w}^T.x}\big)\cdot\big(e^{\mathbf{w}^T.x}+e^{-\mathbf{w}^T.x}\big)-\big(e^{\mathbf{w}^T.x}-e^{-\mathbf{w}^T.x}\big)\cdot\big(e^{\mathbf{w}^T.x}-e^{-\mathbf{w}^T.x}\big)}{\big(e^{\mathbf{w}^T.x}+e^{-\mathbf{w}^T.x}\big)^2}\right] \\ &= \frac{1}{2}\cdot x\cdot\left[1-\left(\frac{e^{\mathbf{w}^T.x}-e^{-\mathbf{w}^T.x}}{e^{\mathbf{w}^T.x}+e^{-\mathbf{w}^T.x}}\right)^2\right] \qquad \text{(this is a valid answer)} \\ &= \frac{1}{2}\cdot x\cdot\big(1-\tanh^2(\mathbf{w}^T.x)\big)\end{aligned}$$

(c) Revisit Part (a) again to *prove that* the gradient of the logistic loss function by substituting g for f and then h for f, respectively, can be derived as follows:

$$\frac{\partial L(x,y;\mathbf{w})}{\mathbf{w}}\Big|_g = x\cdot\big(g(x;\mathbf{w})-y\big) \qquad \text{and} \qquad \frac{\partial L(x,y;\mathbf{w})}{\mathbf{w}}\Big|_h = 2\cdot x\cdot\big(h(x;\mathbf{w})-y\big)$$

You can use the results that you have derived in Part (b). **(4)**

Solution:

$$\begin{aligned}\frac{\partial L(x,y;\mathbf{w})}{\partial \mathbf{w}}\Big|_g &= \left(\frac{g(x;\mathbf{w})-y}{g(x;\mathbf{w})\cdot\big(1-g(x;\mathbf{w})\big)}\right)\cdot\frac{\partial g(x;\mathbf{w})}{\partial \mathbf{w}}\\ &= \left(\frac{g(x;\mathbf{w})-y}{g(x;\mathbf{w})\cdot\big(1-g(x;\mathbf{w})\big)}\right)\cdot x\cdot g(x;\mathbf{w})\cdot\big(1-g(x;\mathbf{w})\big)\\ &= x\cdot\big(g(x;\mathbf{w})-y\big) \qquad [Proved]\end{aligned}$$

and

$$\begin{aligned}\frac{\partial L(x,y;\mathbf{w})}{\partial \mathbf{w}}\Big|_h &= \left(\frac{h(x;\mathbf{w})-y}{h(x;\mathbf{w})\cdot\big(1-h(x;\mathbf{w})\big)}\right)\cdot\frac{\partial h(x;\mathbf{w})}{\partial \mathbf{w}}\\ &= \left(\frac{h(x;\mathbf{w})-y}{\frac{1}{2}\big(\tanh(\mathbf{w}^T.x)+1\big)\cdot\big(1-\frac{1}{2}\big(\tanh(\mathbf{w}^T.x)+1\big)\big)}\right)\cdot\frac{1}{2}\cdot x\cdot\big(1-\tanh^2(\mathbf{w}^T.x)\big)\\ &= \left(\frac{h(x;\mathbf{w})-y}{\big(\tanh(\mathbf{w}^T.x)+1\big)\cdot\frac{1}{2}\cdot\big(1-\tanh(\mathbf{w}^T.x)\big)}\right)\cdot\frac{1}{2}\cdot x\cdot\big(1-\tanh^2(\mathbf{w}^T.x)\big)\\ &= 2\cdot x\cdot\big(h(x;\mathbf{w})-y\big) \qquad [Proved]\end{aligned}$$

(d) Assume that, you are able to format $\frac{\partial L(x,y;\mathbf{w})}{\partial \mathbf{w}}$ as $c\cdot x\cdot\big(f(x;\mathbf{w})-y\big)$ in the previous part, where c is a constant and f is the corresponding sigmoid function g or tanh function h. Explain why this loss function's gradient is very convenient for backpropagation when you use such logistic regression models within your artificial neural networks. **(2)**

Solution:

This gradient is convenient because we compute all components of it during the forward pass (evaluation of $L(x,y;\mathbf{w})$). The gradient just uses x (our data), y (our label), and $f(x;\mathbf{w})$ which we have to compute anyways when we compute the loss. Thus *there is no extra computation when backpropagating* other than putting the pieces together.

Q8. [Neural Networks and Backpropagation] **12 marks**

(a) Assume that, you are given with two types of Neural Network activation functions, namely:

- Linear function: $y = w_0 + \sum_{i \geq 1} w_i x_i$
- Hard threshold function: $y = \begin{cases} 1, & \text{if } w_0 + \sum_{i \geq 1} w_i x_i \geq 0, \\ 0, & \text{otherwise.} \end{cases}$

Which of the following functions can be exactly represented by a neural network with *one* hidden layer which uses linear and/or hard threshold activation functions?

(i) Polynomials of degree 1

(ii) Polynomials of degree 2

(iii) Hinge loss function, $h(x) = \max(1-x, 0)$

(iv) Piecewise constant functions (in 1-dimension)

For each of the above cases, justify your answer. **(4)**

Solution:

(i) Polynomials of degree 1: *Yes*.

To realize $y = \mathbf{A}.\mathbf{x} + b$, we can use linear threshold activation function directly in its hidden layer node, making all its input weights $w_i = a_i$ for all $i \geq 1$, and $w_0 = b$.

(ii) Polynomials of degree 2: *No*.

To realize $y = \mathbf{A}.\mathbf{x}^2 + \mathbf{B}.\mathbf{x} + c$, we are unable to obtain $\mathbf{A}.\mathbf{x}^2$ as a linear combination of $\mathbf{x}$.

(iii) Hinge loss function, $h(x) = \max(1-x, 0)$: *No*.

In order to realize hinge loss function, we need to combine linear function with threshold (constant) function. However, we may need two hidden layer nodes in series to obtain this, because one hidden layer obtains either linear of threshold function but not both together.

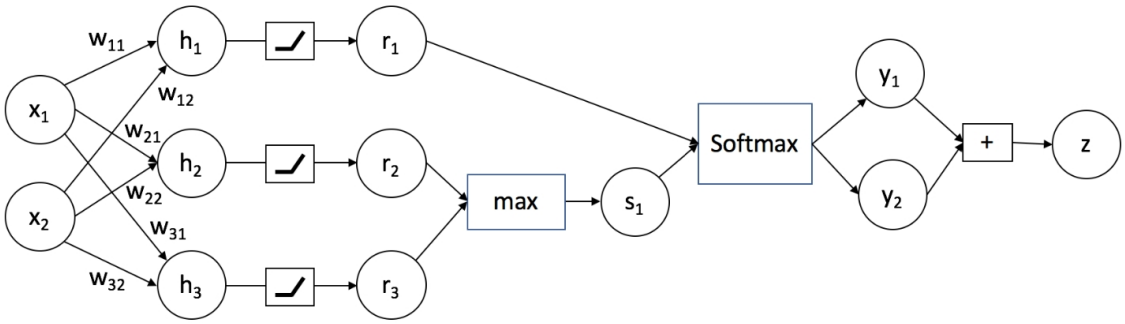
(iv) Piecewise constant functions (in 1-dimension): *Yes*.

Suppose, the given function contains k piecewise constant values, namely, $c_1, c_2, \cdots, c_k$. Every piecewise constant, c_j $(1 \leq j \leq k)$ can get realized (in parallel) from a separate threshold function node, h_j, inside the hidden layer, where for every j we have:

$w_0^{(j)} + w^{(j)}.x \geq 0 \implies x \geq -\frac{w_0^{(j)}}{w^{(j)}}$ (weight framed according to piecewise boundaries of x)

Thus, to match the every c_j, the corresponding hidden layer node outputs $h_j = 1$.

(b) Below is a neural network with inputs x_1 and x_2. The internal nodes are computed below. All variables take scalar values. Answer the following.



$$h_1 = w_{11}x_1 + w_{12}x_2 \qquad r_1 = \max(h_1, 0)$$
$$h_2 = w_{21}x_1 + w_{22}x_2 \qquad r_2 = \max(h_2, 0)$$
$$h_3 = w_{31}x_1 + w_{32}x_2 \qquad r_3 = \max(h_3, 0)$$
$$s_1 = \max(r_2, r_3) \qquad z = y_1 + y_2$$
$$y_1 = \frac{e^{r_1}}{e^{r_1} + e^{s_1}} \qquad y_2 = \frac{e^{s_1}}{e^{r_1} + e^{s_1}}$$

(i) Given $x_1 = 1, x_2 = -2, w_{11} = 6, w_{12} = 2, w_{21} = 4, w_{22} = 7, w_{31} = 5, w_{32} = 1$, compute the values of all internal nodes, $h_1, h_2, h_3, r_1, r_2, r_3, s_1, y_1, y_2, z$, during *forward propagation*. **(2)**

(ii) During *backpropagation* step, analytically compute the following four gradients: $\frac{\partial y_1}{\partial s_1}$, $\frac{\partial y_1}{\partial r_1}$, $\frac{\partial z}{\partial s_1}$, and $\frac{\partial s_1}{\partial x_2}$. The answer should be an expression of any of the nodes in the network $(x_1, x_2, h_1, h_2, h_3, r_1, r_2, r_3, s_1, y_1, y_2, z)$ or weights $(w_{11}, w_{12}, w_{21}, w_{22}, w_{31}, w_{32})$. Show your calculations in details. **(6)**

Solution:

(i)

h_1	h_2	h_3	r_1	r_2	r_3	s_1	y_1	y_2	z
2	-10	3	2	0	3	3	$\frac{1}{1+e}$	$\frac{e}{1+e}$	1

(ii) Note that, $y_2 = 1 - y_1$ and similarly $y_1 = 1 - y_2$. Therefore, $\frac{\partial y_2}{\partial s_1} = -\frac{\partial y_1}{\partial s_1}$.

$$\frac{\partial y_1}{\partial s_1} = \frac{\partial}{\partial s_1}\left(\frac{1}{1+e^{s_1-r_1}}\right) = -1 \cdot \frac{e^{s_1-r_1}}{\left(1+e^{s_1-r_1}\right)^2}$$
$$= -\frac{e^{r_1}}{e^{r_1}+e^{s_1}} \cdot \frac{e^{s_1}}{e^{r_1}+e^{s_1}} = -y_1 \cdot (1-y_1) = -y_1 \cdot y_2$$

$$\frac{\partial y_1}{\partial r_1} = \frac{\partial}{\partial r_1}\left(\frac{1}{1+e^{s_1-r_1}}\right) = \frac{e^{s_1-r_1}}{\left(1+e^{s_1-r_1}\right)^2}$$
$$= \frac{e^{r_1}}{e^{r_1}+e^{s_1}} \cdot \frac{e^{s_1}}{e^{r_1}+e^{s_1}} = y_1 \cdot (1-y_1) = y_1 \cdot y_2$$

$$\frac{\partial z}{\partial s_1} = \frac{\partial z}{\partial y_1} \cdot \frac{\partial y_1}{\partial s_1} + \frac{\partial z}{\partial y_2} \cdot \frac{\partial y_2}{\partial s_1} = 1 \cdot (-y_1 \cdot y_2) + 1 \cdot (y_1 \cdot y_2) = 0$$

$$\frac{\partial s_1}{\partial x_2} = \begin{cases} \frac{\partial s_1}{\partial r_2} \cdot \frac{\partial r_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial x_2} = 1 \cdot 1 \cdot w_{22} = w_{22}, & \text{if } r_2 \geq r_3 \text{ and } h_2 > 0 \\ \frac{\partial s_1}{\partial r_3} \cdot \frac{\partial r_3}{\partial h_3} \cdot \frac{\partial h_3}{\partial x_2} = 1 \cdot 1 \cdot w_{32} = w_{32}, & \text{if } r_2 < r_3 \text{ and } h_3 > 0 \\ 0, & \text{otherwise } (h_2, h_3 \leq 0) \end{cases}$$

— The question paper ends here. —