University of Toronto
ECE421: Introduction to Machine Learning, Sample Midterm,
Duration: 110 minutes
Aids: No aid-sheet is permitted. No electronic or mechanical computing devices are permitted.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO

- The University of Toronto and you, as a student, share a commitment to academic integrity. You are reminded that you may be charged with an academic offence for possessing any unauthorized aids during the writing of an exam. Clear, sealable, plastic bags have been provided for all electronic devices with storage, including but not limited to: cell phones, smart watches, SMART devices, tablets, laptops, and calculators. Please turn off all devices, seal them in the bag provided, and place the bag under your desk for the duration of the examination. You will not be able to touch the bag or its contents until the exam is over. If, during an exam, any of these items are found on your person or in the area of your desk other than in the clear, sealable, plastic bag, you may be charged with an academic offence. A typical penalty for an academic offence may cause you to fail the course.
- There are **5 questions** and **Y pages** in this exam, including this one. When you receive the signal to start, please make sure that your copy of the examination is complete.
- Answer each question directly on the examination paper, in the space provided.

Q1 [10 pts] Short Answer Questions

1.a [3 pts] Consider a one-dimensional dataset with 4 examples: $x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 8$. By hand, apply K-Means clustering until convergence, assuming that the initial centroids are $\mu_1[0] = 0$ and $\mu_2[0] = 4$. For each iteration l, report the one-hot-encoding of the cluster assignments $\underline{r}_1[l], \underline{r}_2[l], \underline{r}_3[l]$, and $\underline{r}_4[l]$, where $\underline{r}_n[l] = (r_{n1}[l], r_{n2}[l])$. Moreover, report the new values of the centroids in each iteration, *i.e.*, $\mu_1[l]$ and $\mu_2[l]$.

[**Write your answer to 1.a here.**]

$\underline{r}_1[1] = (1,0)$ and $\underline{r}_2[1] = \underline{r}_3[1] = \underline{r}_4[1] = (0,1)$

$\mu_1[1] = 1$, $\mu_2[1] = \frac{17}{2} = 8.5$

$\underline{r}_1[2] = \underline{r}_2[2] = (1,0)$, $\underline{r}_3[2] = \underline{r}_4[2] = (0,1)$

$\mu_1[2] = \frac{1+3}{2} = 2$, $\mu_2[2] = \frac{6+8}{2} = 7$

$\underline{r}_1[3] = \underline{r}_2[3] = (1,0)$

$\underline{r}_3[3] = \underline{r}_4[3] = (0,1)$

Converged!

1.b [4 pts] Circle True or False for each of the following statements about k-nearest neighbours classification with N training points.

1.b.i As k grows from 1 to N, the test classification accuracy consistently increases.

True / (False)

1.b.ii For small values of k, the model is underfitting.

True / (False)

1.b.iii The decision boundary is smoother with small values of k.

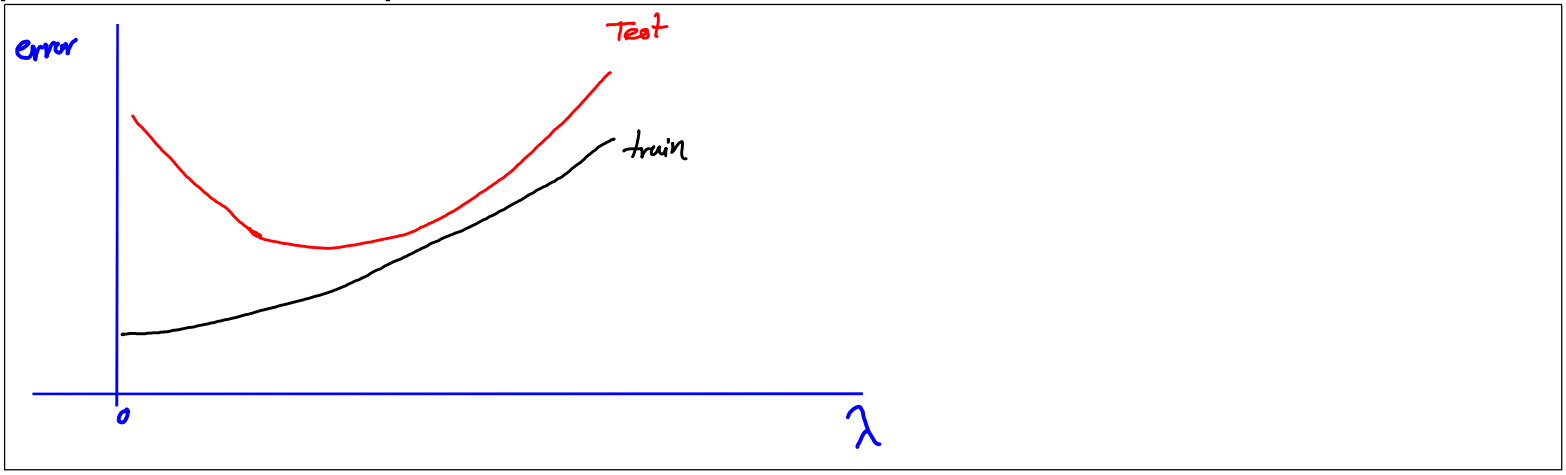
True / (False)

1.b.iv If the data is not linearly separable, k-NN cannot achieve 100% training accuracy.

True / (False)

1.c [3 pts] Suppose we are regularizing a linear regression model by adding $\lambda \|\underline{w}\|_2^2$ to the cost function where the $\underline{w}$ is the weight vector. Plot a typical graph of both the training and test error (y-axis) versus λ (x-axis).

[**Write your answer to 1.c here.**]



Q2 **[15 pts]** Consider a binary linear classification problem where the datapoints are two-dimensional, *i.e.*, $\underline{x} = (x_1, x_2) \in \mathbb{R}^2$, and the labels $y \in \{-1, +1\}$. We are given a dataset with the following four points:

$$\mathcal{D} = \{(\underline{x}_1, y_1), (\underline{x}_2, y_2), (\underline{x}_3, y_3), (\underline{x}_4, y_4)\}$$

where the input data vectors are given by

$$\underline{x}_1 = (-1, -1), \quad \underline{x}_2 = (-1, +1), \quad \underline{x}_3 = (+1, -1), \quad \underline{x}_4 = (+1, +1),$$

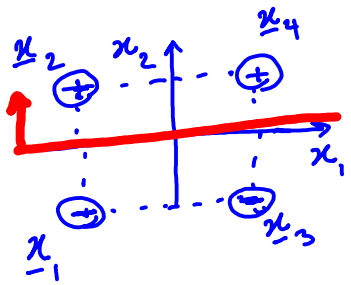
and the associated labels are given by

$$y_1 = -1, \quad y_2 = +1, \quad y_3 = -1, \quad y_4 = +1.$$

We hope to find a perceptron with weight vector $\underline{w} = (w_0, w_1, w_2)$ that classifies this dataset.

2.a **[5 pts]** Let $\hat{y}_n$ be the output of the perceptron for input $\underline{x}_n$, and let $E_{\text{in}}(\underline{w}) = \frac{1}{4} \sum_{n=1}^{4} \mathbb{I}(y_n \neq \hat{y}_n)$ be the in-sample error with respect to $\mathcal{D}$, where $\mathbb{I}(.)$ is the indicator function. Is there a weight vector $\underline{w}$ that correctly classifies all datapoints, *i.e.*, a $\underline{w}$ that gives $E_{\text{in}}(\underline{w}) = 0$? If your answer is yes, find such $\underline{w}$. If your answer is no, explain why not.

[**Write your answer to 2.a here.**]

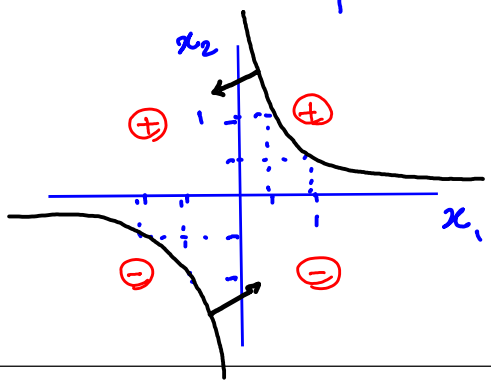


yes. $\underline{w} = (0, 0, 1)$ gives $E_{in}(\underline{w}) = 0$

2.b **[5 pts]** We perform non-linear transformation of each datapoint $\underline{x} \in \mathcal{D}$ to the (augmented) vector $\underline{z} = (1, x_1, x_2, x_1 x_2)$. Consider a perceptron in the three-dimensional $\underline{z}$-space with new weight vector $\underline{w} = (\frac{1}{2}, 0, 0, -1)$. Plot the decision boundary of this perceptron in the two-dimensional $\underline{x}$-space and find $E_{\text{in}}(\underline{w})$.

[**Write your answer to 2.b here.**]

The decision boundary is $\frac{1}{2} \times 1 + 0 \times x_1 + 0 \times x_2 + (-1) x_1 x_2 = 0$, i.e., $x_1 x_2 = 1/2$



$E_{in}(\underline{w}) = \frac{1}{4}(2) = 1/2$.

2.c [5 pts] Suppose we now consider binary logistic regression to classify the points on $\mathcal{D}$ in the $\underline{z}$-space, with the following sigmoid functyion for likelihood:

$$\mathbb{P}[y = +1 \mid \underline{z}] = \theta(\underline{w}^T \underline{z}) = \frac{\exp(\underline{w}^T \underline{z})}{1 + \exp(\underline{w}^T \underline{z})}.$$

Assume that we use the log-loss to measure training error. Find $E_{\text{in}}(\underline{w})$ for $\underline{w} = (\frac{1}{2}, 0, 0, -1)$.

[**Write your answer to 2.c here.**]

$$E_{in}(\underline{w}) = \frac{1}{4}\sum_{n=1}^{4} -\log\left(\hat{P}(y_n \mid \underline{z}_n)\right) = \frac{1}{4}\sum_{n=1}^{4} \log\left(1 + \exp\left(y_n \underline{w}^T \underline{z}_n\right)\right)$$

	y_n	$\underline{x}_n$	$\underline{z}_n$	$y_n \underline{w}^T \underline{z}_n$
$n=1$	-1	$(-1,-1)$	$(1,-\frac{1}{2},-1,1)$	$(-1)(\frac{1}{2}-1) = 1/2$
$n=2$	$+1$	$(-1,+1)$	$(1,-\frac{1}{2},+1,-1)$	$(+1)(\frac{1}{2}+1) = 3/2$
$n=3$	-1	$(1,-1)$	$(1,\frac{1}{2},-1,-1)$	$(-1)(\frac{1}{2}+1) = -3/2$
$n=4$	$+1$	$(1,1)$	$(1,\frac{1}{2},1,1)$	$(+1)(\frac{1}{2}-1) = -1/2$

$$= \frac{1}{4}\left[\log\left(1+e^{1/2}\right) + \log\left(1+e^{3/2}\right) + \log\left(1+e^{-3/2}\right) + \log\left(1+e^{-1/2}\right)\right]$$

Q3 **[15 pts]** Consider a regression problem with data points $x \in \mathbb{R}$ and labels $y \in \mathbb{R}$. We are given a dataset with the following three points:

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\} = \{(0, 0), (1, 2), (3, 0)\}.$$

The in-sample error is $E_{\text{in}} = \frac{1}{3} \sum_{n=1}^{3} (y_n - \hat{y}_n)^2$, where $\hat{y}_n$ is the predicted label for x_n, for all $n \in \{1, 2, 3\}$.

3.a **[5 pts]** Find the best constant regression model, in the form $\hat{y} = w_0$, that minimizes E_{in}. What is E_{in} in this case?

[**Write your answer to 3.a here.**]

$E_{in} = \frac{1}{3}\left((0-w_0)^2 + (2-w_0)^2 + (0-w_0)^2\right)$

$\frac{dE_{in}}{dw_0} = 0 \Rightarrow 2w_0 + 2(2-w_0)(-1) + 2w_0 = 0 \Rightarrow 2w_0 = 2 - w_0 \Rightarrow w_0 = 2/3$

With $w_0 = 2/3$, $E_{in} = \frac{1}{3}\left(2\times(\frac{2}{3})^2 + (\frac{4}{3})^2\right)$

3.b **[5 pts]** Find the best quadratic regression model, in the form $\hat{y} = w_0 + w_1 x + w_2 x^2$, that minimizes E_{in}. What is E_{in} in that case?

[**Write your answer to 3.b here.**]

We have a system of linear equations with 3 unknowns and 3 equations.

$$\left.\begin{aligned} w_0 + w_1(0) + w_2(0) &= 0 \\ w_0 + w_1(1) + w_2(1) &= 2 \\ w_0 + w_1(3) + w_2(9) &= 0 \end{aligned}\right\} \Rightarrow w_0 = 0,\ w_1 = 3,\ \text{and}\ w_2 = -1$$

With such $\underline{w}$, $\hat{y}_n = y_n$ for any $n \in \{1, 2, 3\}$. Thus, $E_{in}(\underline{w}) = \frac{1}{3}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 = 0$

3.c **[5 pts]** Now, suppose we are further given a single validation data point $(1, 0)$. Which of the above regression models is the best, according to this validation data point?

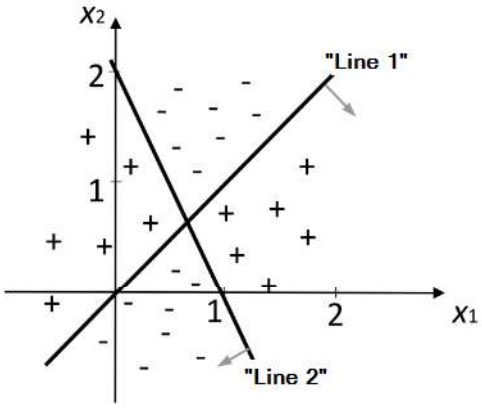
[**Write your answer to 3.c here.**]

Validation error from model in part a: $E_{val} = (0 - 2/3)^2 = 4/9$

Validation error from model in part b: $E_{val} = (0 - (3-1))^2 = 4$

a is better.

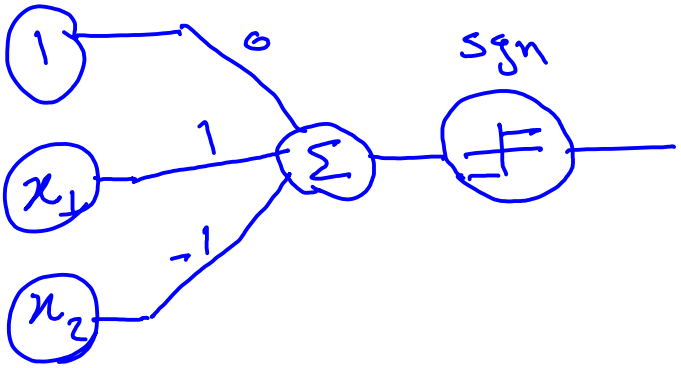
Q4 **[20 pts]** Consider a binary linear classification problem where the data points are two-dimensional, *i.e.*, $\underline{x} = (x_1, x_2) \in \mathbb{R}^2$, and the labels $y \in \{-1, +1\}$. We wish to build a multi-layer perceptron to classify dataset $\mathcal{D}$ as shown below, where the "+" and "-" signs indicate examples with labels $+1$ and -1, respectively. The lines are $x_2 = x_1$ and $x_2 = -2x_1 + 2$.



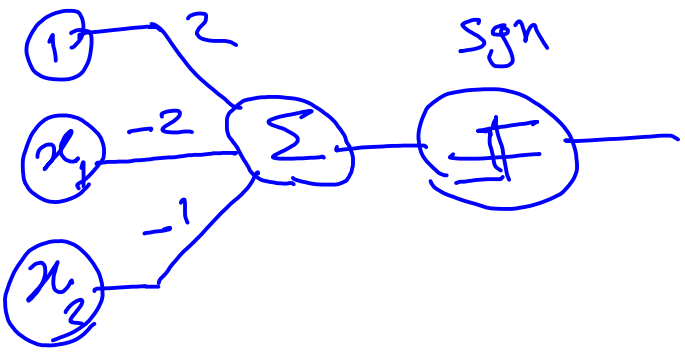
4.a **[5 pts]** Design and **draw** two perceptrons to implement the two lines shown in the figure above, with the requirement that each of your perceptrons must classify the side indicated by the arrow as $+1$.

[**Write your answer to 4.a here.**]

Line 1: $\underline{w}_1 = (0, 1, -1)$



Line 2: $\underline{w}_2 = (2, -2, -1)$

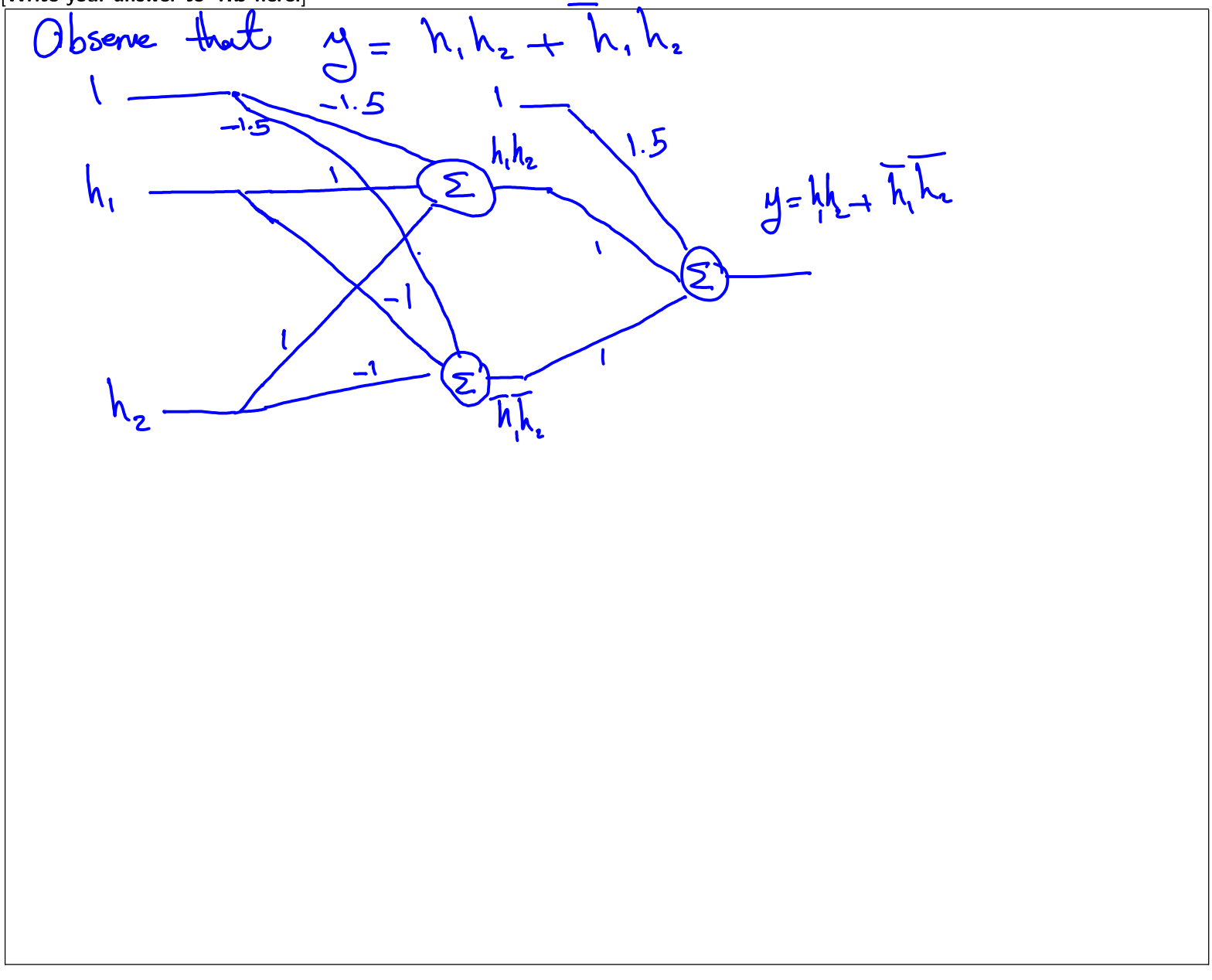


4.b **[10 pts]** Design and draw a multi-layer perceptron to implement the Boolean function $y = f(h_1, h_2)$ defined by the table below:

h_1	h_2	y
-1	-1	$+1$
-1	$+1$	-1
$+1$	-1	-1
$+1$	$+1$	$+1$

[**NOTE:** y must be the direct output of the last layer of multi-layer perceptron, *i.e.*, you are not allowed to multiply any number, such as -1, to the output to obtain y.]

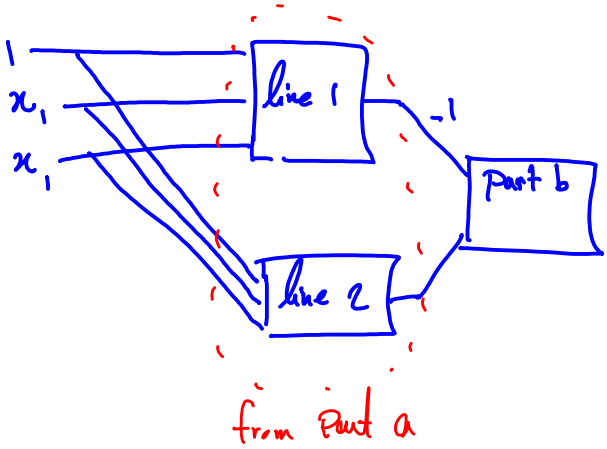
[**Write your answer to 4.b here.**]



4.c [5 pts] Use only the perceptrons in parts **4.a** and **4.b** to build a multi-layer perceptron to classify the dataset $\mathcal{D}$. Draw your design and clearly label all edges and weights.

[**Write your answer to 4.c here.**]

One way to do this is by setting $h_1 = -y_1$ and $h_2 = y_2$



Q5 **[15 pts]** In ancient times, there was a village surrounded by hundreds of lakes. Each lake was either poisonous or healthy. Anyone who ate fish from a poisonous lake would die immediately, while anyone who ate fish from a healthy lake would survive. All fish looked and tasted identical, and villagers knew no other way of knowing whether a lake was poisonous or healthy, which of course was a huge problem for the villagers.

Fortunately, a famous chemist visited the village and was told of this dilemma. The chemist suggested using the pH level of water to determine whether a given lake was poisonous or healthy, and hypothesized that lakes with poisonous fish would have higher pH values than healthy lakes. Accordingly, the chemist visited each lake and collected the pH value of the water in each lake. The data set is denoted by $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$, where x_i denotes the pH level of the lake $i \in \{1, 2, \ldots, N\}$. Assume that $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_N$.

You are hired as a machine learning scientist to help determine the probability that a randomly selected lake is poisonous given its pH value. In order to do this, you propose to use a Gaussian Mixture Model (GMM) as follows:

- $\mathbb{P}[\text{lake is poisonous}] = \pi_1$
- $\mathbb{P}[\text{lake is healthy}] = \pi_2$
- $f(\text{pH} = x \,|\, \text{lake is poisonous})$ is $\mathcal{N}(x \,|\, \mu_1, \sigma_1^2)$
- $f(\text{pH} = x \,|\, \text{lake is healthy})$ is $\mathcal{N}(x \,|\, \mu_2, \sigma_2^2)$
- $\mu_1 \geq \mu_2$ as the pH value for poisonous lakes will be higher on average

Here, $f(\cdot)$ denotes the conditional density function for the pH value. You decide to use the EM algorithm to train the above GMM on the dataset $\mathcal{D}$.

5.a **[5 pts]** Using the above GMM, provide an expression of the probability that a randomly selected lake is poisonous given that its pH level is measured to be x.

[**Write your answer to 5.a here.**]

$$\mathbb{P}[\text{lake is poisonous} \,|\, \text{pH} = x] = \frac{\mathbb{P}[\text{pH} = x \,|\, \text{lake is poisonous}]\, \mathbb{P}[\text{lake is poisonous}]}{\mathbb{P}[\text{pH} = x \,|\, \text{lake is poisonous}]\, \mathbb{P}[\text{lake is poisonous}] + \mathbb{P}[\text{pH} = x \,|\, \text{lake is healthy}]\, \mathbb{P}[\text{lake is healthy}]}$$

$$= \frac{\pi_1 \, \mathcal{N}(x \,|\, \mu_1, \sigma_1^2)}{\pi_1 \, \mathcal{N}(x \,|\, \mu_1, \sigma_1^2) + \pi_2 \, \mathcal{N}(x \,|\, \mu_2, \sigma_2^2)}$$

5.b [5 pts] Write down the pseudocode for the EM algorithm with hard decisions that finds the parameters of the GMM. Assume that we initialize the algorithm in such a way that

$$\underline{r}_n = \begin{cases} (0,1), \text{ if } 1 \le n \le K, \\ (1,0), \text{ if } K+1 \le n \le N. \end{cases}$$

[**Write your answer to 5.b here.**]

Initialize $\underline{r}_n$ as above

Iterate until Convergence:

For $k = 1, 2$

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} r_{nk}, \quad \mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}, \quad \sigma_k^2 = \frac{1}{\sum_{n=1}^{N} r_{nk}} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)^2$$

For $n = 1, \ldots, N$:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\max_j \pi_j \mathcal{N}(x_n | \mu_j, \sigma_j^2) \\ 0 & \text{otherwise} \end{cases}$$

5.c [5 pts] Suppose that $N = 5$ and we observe $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 6$, and $x_5 = 10$. Assume that we initialize the algorithm in such a way that $\underline{r}_1 = \underline{r}_2 = \underline{r}_3 = (0,1)$ and $\underline{r}_4 = \underline{r}_5 = (1,0)$. Execute the hard decision EM algorithm and compute the parameters of the GMM.

[**Write your answer to 5.c here.**]

Initial r_n's: $\underline{r}_1 = \underline{r}_2 = \underline{r}_3 = (0,1)$ and $\underline{r}_4 = \underline{r}_5 = (1,0)$

Updating π_k's, μ_k's, and σ_k^2's: $\pi_1 = \frac{2}{5}$, $\pi_2 = \frac{3}{5}$

$$\mu_1 = \frac{1}{2} \times (6+10) = 8, \quad \mu_2 = \frac{1}{3}(1+2+3) = 2$$

$$\sigma_1^2 = \frac{1}{2}\left((6-8)^2 + (10-8)^2\right) = 4, \quad \sigma_2^2 = \frac{1}{3}\left((1-2)^2 + (2-2)^2 + (3-2)^2\right) = 2/3$$

Updating r_{nk}'s: $\pi_1 \mathcal{N}(x_1 | \mu_1, \sigma_1^2) < \pi_2 \mathcal{N}(x_1 | \mu_2, \sigma_2^2) \Rightarrow \underline{r}_1 = (0,1)$

Similarly we can show that $\underline{r}_2 = (0,1)$, $\underline{r}_3 = (0,1)$, $\underline{r}_4 = \underline{r}_5 = (1,0)$

Converged!