1. (20 MARKS) In this problem you consider the two-dimensional data set $\mathcal{D}$, target function f, and linear hypothesis h defined as follows:

   (i) The **unknown** target function f (which we need to learn) labels all points $\mathbf{x} = (x_1, x_2)$ such that $x_2 \geq 0$ belong to class +1 and all those such that $x_2 < 0$ belong to class −1.

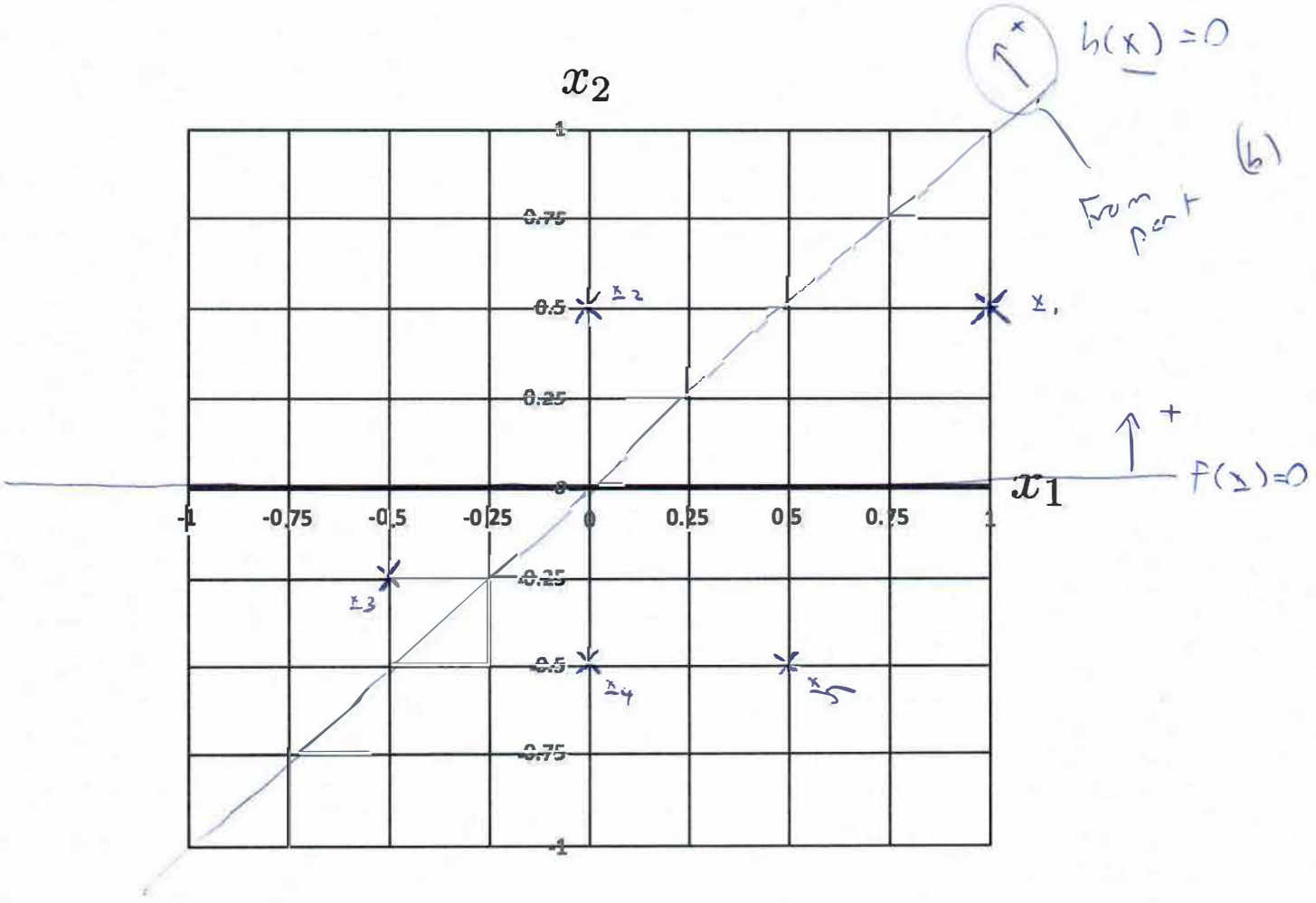
   (ii) The boundary of the linear hypothesis h is the 45-degree line, connecting $(-1,-1)$ to the origin to $(+1,+1)$.

   (iii) All data points $\mathbf{x} \in \mathcal{D}$ have coordinate magnitudes at most one, i.e., $|x_1| \leq 1$ and $|x_2| \leq 1$. The data set $\mathcal{D}$ consists of five data points (so $|\mathcal{D}| = 5$) as is tabulated below

n	$\mathbf{x}_n$	$y_n = f(\mathbf{x}_n)$
1	(1, 0.5)	+1
2	(0, 0.5)	+1
3	(-0.5, -0.25)	-1
4	(0, -0.5)	-1
5	(0.5, -0.5)	-1

2 marks

(a) Sketch (and label) the boundary of f, the boundary of h, and all data points from $\mathcal{D}$ on the figure provided below.



total/2

3 marks

(b) Now, consider a linear classification problem. If $h(\mathbf{x}_2) = +1$ then which of the following is the correct form of $h(\mathbf{x})$?

(i) $h(\mathbf{x}) = \text{sign}(\mathbf{w_0}^T\mathbf{x})$ where $\mathbf{w_0} = (1, 1)$ and $\mathbf{x} = (x_1, x_2)$.
(ii) $h(\mathbf{x}) = \text{sign}(\mathbf{w_0}^T\mathbf{x})$ where $\mathbf{w_0} = (1, 1, 1)$ and $\mathbf{x} = (1, x_1, x_2)$.
(iii) $h(\mathbf{x}) = \text{sign}(\mathbf{w_0}^T\mathbf{x})$ where $\mathbf{w_0} = (0, -1, 1)$ and $\mathbf{x} = (1, x_1, x_2)$.
(iv) $h(\mathbf{x}) = \text{sign}(\mathbf{w_0}^T\mathbf{x})$ where $\mathbf{w_0} = (0, 1, -1)$ and $\mathbf{x} = (1, x_1, x_2)$.

In the space below, indicate your answer, (i)–(iv), and justify your choice.

Normal vector to boundary line $h(\underline{x}) = 0$ is $(-1, +1)$

Since $F(\underline{x}_2) = +1$

Boundary line crosses through origin so bias $= 0$

Therefore (iii) $h(\underline{x}) = \text{sign}\left(\begin{bmatrix} 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}\right)$

chk: $h(\underline{x}_2) = \text{sign}\left(\begin{bmatrix} 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}\right) = \text{sign}(0.5) = +1$

5 marks

(c) Using your form for $h(\mathbf{x})$ from above, what is $E_{\text{IN}}(\mathbf{w_0})$, the **Classification Error** for the data set $\mathcal{D}$ for the linear classification problem?

Points correctly classified: $\underline{x}_2, \underline{x}_4, \underline{x}_5$

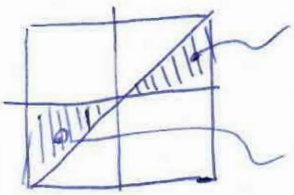
Points incorrectly classified: $\underline{x}_1, \underline{x}_3$

$$E_{IN}(\underline{w}_0) = \frac{\#\text{ incorrect}}{\text{total}} = \frac{2}{5}$$

total/8

5 marks

(d) Assuming that $P(\mathbf{x})$ is uniform, i.e., $P(\mathbf{x}) = 0.25$ for all $\mathbf{x}$ such that $|x_1| \leq 1$ and $|x_2| \leq 1$, what is $E_{\text{OUT}}(\mathbf{w_0})$?



mis-classified region A, area = $\frac{1}{2}$

misclassified region B, area = $\frac{1}{2}$

E_{out} for mis-classification error is area of region weighted by PDF in that region

$$E_{out}(w_0) = \int_{x \in A} \underbrace{P_x(x)}_{1/4} dx + \int_{x \in B} \underbrace{P_x(x)}_{1/4} dx$$

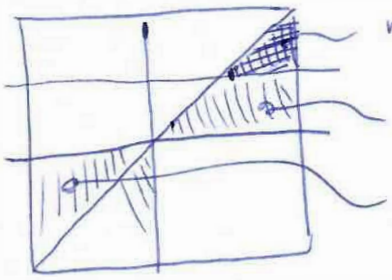
$$= \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \boxed{\frac{1}{4} = E_{out}(w_0)}$$

5 marks

(e) If, instead, $P(\mathbf{x})$ is defined as

$$P(\mathbf{x}) = \begin{cases} \frac{1}{3} & \text{if} \quad |x_1| \leq 1 \text{ and } 0.5 \leq x_2 \leq 1 \\ \frac{2}{9} & \text{if} \quad |x_1| \leq 1 \text{ and } -1 \leq x_2 < 0.5 \end{cases},$$

what is $E_{\text{OUT}}(\mathbf{w_0})$?



region A, area = $\frac{1}{8}$

region B, area = $\frac{3}{8}$

region C, area = $\frac{4}{8}$

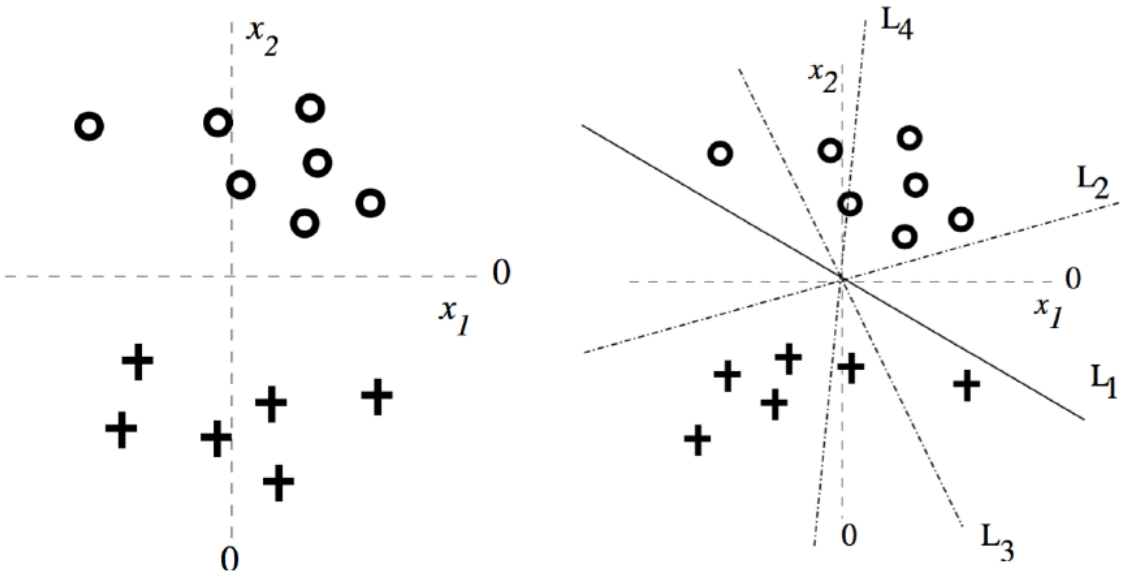
$$E_{out}(w_0) = \int_{x \in A} \underbrace{P_x(x)}_{1/3} dx + \int_{x \in B} \underbrace{P_x(x)}_{2/9} dx + \int_{x \in C} \underbrace{P_x(x)}_{2/9} dx$$

$$= \frac{1}{3} \cdot \frac{1}{8} + \frac{2}{9} \cdot \frac{3}{8} + \frac{2}{9} \cdot \frac{4}{8}$$

$$= \frac{3}{72} + \frac{6}{72} + \frac{8}{72} = \boxed{\frac{17}{72} = E_{out}(w_0)}$$

total/10

2. (10 MARKS) Consider a binary classification problem on a two-dimensional dataset in the (x_1, x_2) plane with $N = 13$ training points shown below. The symbol o represents the label $y = -1$ while the symbol '+' represents the label $y = +1$.



In the figure on the right, $L_1, \ldots, L_4$ indicate four different linear decision boundaries in the (x_1, x_2) plane, that can be used for classification. Throughout this problem we consider only those decision boundaries that pass through the origin represented by: $w_1 x_1 + w_2 x_2 = 0$, where $\mathbf{w} = (w_1, w_2) \in \mathbb{R}^2$.

Furthermore we consider a simple logistic regression model where the output given $\mathbf{x} = (x_1, x_2)$ is given by:

$$p_{\mathbf{w}}(y = 1|\mathbf{x}) = \phi(w_1 x_1 + w_2 x_2) = \frac{1}{1 + e^{-(w_1 x_1 + w_2 x_2)}}.$$

Recall that $\phi(s) = \frac{1}{1+e^{-s}}$ is the sigmoid function. Also note that $p_{\mathbf{w}}(y = -1|\mathbf{x}) = 1 - p_{\mathbf{w}}(y = 1|\mathbf{x})$.

We consider the standard log-loss penalty function so that:

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} -\log p_{\mathbf{w}}(y_n|\mathbf{x}_n)$$

where $(\mathbf{x}_n, y_n)$ denotes a training point in the above figure and $N = 13$.

2 marks (**a**) Is the training set linearly separable? Briefly explain your answer.

Yes. L_1 separates the training points.

total/2

5 marks

(**b**) Suppose we wish to minimize the following regularized expression:

$$\min_{\mathbf{w}=(w_1,w_2)\in\mathbb{R}^2} \{E_{\text{in}}(\mathbf{w}) + \lambda \cdot w_2^2\}$$

where λ is a **large positive** constant. Note that only the component w_2 is regularized above. For each of the decision boundaries: L_2, L_3 and L_4 in the figure on the previous page circle **yes** if it can result from minimizing the above expression and **no** otherwise. Briefly explain each case. No calculations are needed.

1 pt
- L_2: **yes** **no**

1 pt
- L_3: **yes** **no**

1 pt
- L_4: **yes** **no**

1 pt { Due to regularization $|w_2| << |w_1|$
So the slope should be large.
$\Rightarrow$ L_2 cannot happen.

1 pt { Reflecting L_4 wrt x_2-axis gives better error $\therefore$ L_4 cannot happen

3 marks

(**c**) Suppose we wish to minimize the following regularized expression:

$$\min_{\mathbf{w}=(w_1,w_2)\in\mathbb{R}^2} \{E_{\text{in}}(\mathbf{w}) + \lambda \cdot (w_1^2 + w_2^2)\}$$

where λ is a **large positive** constant. Select which of the following three cases is the most likely case satisfied by the optimal solution (select only one):

1 pt

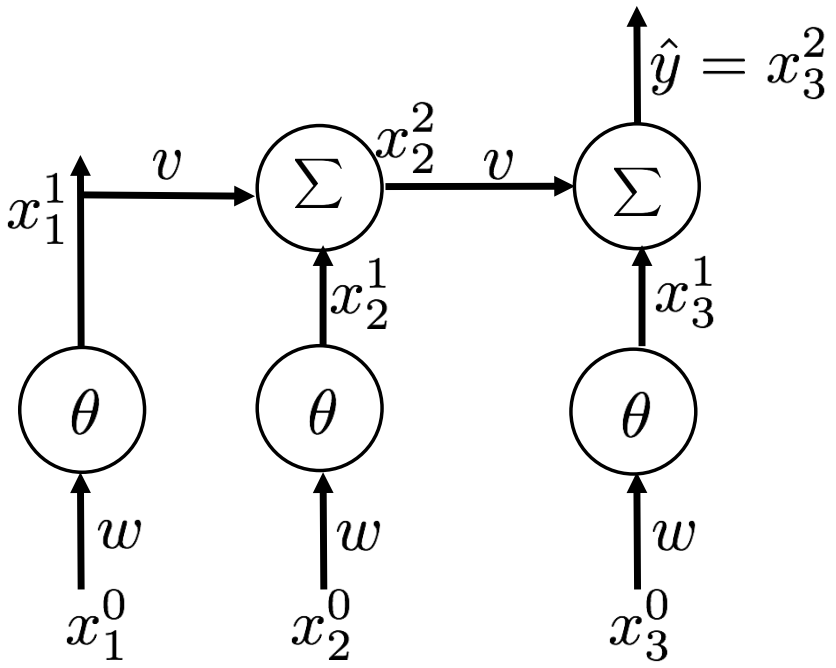
a. Both w_1 and w_2 are small and w_1/w_2 is less than 1
b. Both w_1 and w_2 are small and w_2/w_1 is less than 1
c. Both w_1 and w_2 are small and $w_1/w_2 = 1$.
d. Neither of the above.

Justify your answer. No calculations are needed.

2 pt { since $x_2 = 0$ (horizontal axis) achieves perfect classification, when λ is large $w_1 \approx 0$. w_2 will also be small, but +ve
thus $\frac{w_1}{w_2} < 1$

total/8

3. (30 MARKS) In this problem we consider a neural network shown in the figure below.



Given an input $(\mathbf{x}, y)$ where $\mathbf{x} = (x_1^0, x_2^0, x_3^0) \in \mathbb{R}^3$ the neural network computes $\hat{y}$, an approximation to y, as shown in the figure. More specifically the intermediate computations are given as follows:

$$x_1^1 = \theta(w \cdot x_1^0) \qquad x_2^2 = x_2^1 + v \cdot x_1^1$$
$$x_2^1 = \theta(w \cdot x_2^0) \qquad x_3^2 = x_3^1 + v \cdot x_2^2$$
$$x_3^1 = \theta(w \cdot x_3^0) \qquad \hat{y} = x_3^2$$

Note that v and w are the shared weights on the horizontal edges and vertical edges as shown in the figure. Assume that $\theta(\cdot)$ is some arbitrary activation function with derivative denoted by $\theta'(\cdot)$. For the input $(\mathbf{x}, y)$ and model parameters $\Omega = (w, v)$ of the neural network, we assume that the loss is given by:

$$e(\Omega) = (\hat{y} - y)^2.$$

The above neural network preserves the order of the elements x_1^0, x_2^0 and x_3^0 in the input $\mathbf{x}$. It is a simplified version of a recurrent neural network.

total/–

5 marks

(**a**) Find an expression for $\frac{de}{dv}$ where $e(\Omega)$ in the squared loss function on the previous page. Express your answer in terms of the following variables: x_1^1, v, x_2^2 and $\Delta = \hat{y} - y$.

$$e(\Omega) = (\hat{y} - y)^2$$

$$\frac{\partial e}{\partial v} = \frac{\partial e}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial v} = 2\Delta \frac{\partial \hat{y}}{\partial v}$$ 1pt

$$\hat{y} = x_3^1 + v \cdot x_2^{(2)}$$

$$\frac{\partial \hat{y}}{\partial v} = \frac{\partial x_3^1}{\partial v} + \frac{\partial}{\partial v}(v\, x_2^{(2)}) = x_2^{(2)} + v \cdot \frac{\partial x_2^{(2)}}{\partial v}$$ 1pt

$\frac{\partial x_3^1}{\partial v} = 0$

$$x_2^{(2)} = x_2^{(1)} + v \cdot x_1^{(1)} \Rightarrow \frac{\partial x_2^{(2)}}{\partial v} = x_1^{(1)}$$ 2pt, 1pt

$$\Rightarrow \partial \hat{y}/\partial v = x_2^{(2)} + v \cdot x_1^{(1)} \Rightarrow \frac{\partial e}{\partial v} = 2\Delta (x_2^{(2)} + v x_1^{(1)})$$

5 marks

(**b**) Find expressions for $\frac{de}{dx_2^2}$, $\frac{de}{dx_1^1}$, $\frac{de}{dx_2^1}$ and $\frac{de}{dx_3^1}$. Express your answer in the simplest possible form (with as few variables as possible).

$$\frac{\partial e}{\partial x_2^2} = \frac{\partial e}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial x_2} = 2\Delta \frac{\partial \hat{y}}{\partial x_2}$$ 1

$$\hat{y} = x_3^1 + v\, x_2^2 \Rightarrow \frac{\partial \hat{y}}{\partial x_2^2} = v$$ 1pt

$$\frac{\partial e}{\partial x_2^2} = 2\Delta v$$ 1pt

$$\frac{\partial e}{\partial x_1^{(1)}} = \frac{\partial e}{\partial x_2^{(2)}} \cdot \frac{\partial x_2^{(2)}}{\partial x_1^{(1)}} = 2\Delta v^2$$ 1pt

$$\frac{\partial e}{\partial x_2^{(1)}} = \frac{\partial e}{\partial x_2^{(2)}} \cdot \frac{\partial x_2^{(2)}}{\partial x_2^{(1)}} = 2\Delta v$$ 1pt

$$\frac{\partial e}{\partial x_3^{(1)}} = \frac{\partial e}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial x_3^{(1)}} = 2\Delta$$ 1pt

total/10

5 marks (**c**) Using parts (a) and (b) find an expression for $\frac{de}{dw}$.

2pt $\frac{\partial e}{\partial w} = \frac{\partial e}{\partial x_3^{(1)}} \cdot \frac{\partial x_3^{(1)}}{\partial w} + \frac{\partial e}{\partial x_2'} \frac{\partial x_2'}{\partial w} + \frac{\partial e}{\partial x_1^{(1)}} \cdot \frac{\partial x_1^{(1)}}{\partial w}$

$2\Delta\, \theta'(w x_3^0) \cdot x_3^0$ } 1 pt

$+ \; 2\Delta V\, \theta'(w x_2^0) \cdot x_2^0$ } 1 pt

$+ \; 2\Delta V^2\, \theta'(w x_1^0) \cdot x_1^0$ } 1 pt

(don't penalize if $\frac{\partial e}{\partial x_i^{(1)}}$ are wrong in part b)

5 marks (**d**) Compute $\frac{de}{dx_i^0}$ for $i = 1, 2, 3$.

$\frac{\partial e}{\partial x_1^0} = \frac{\partial e}{\partial x_1'} \cdot \frac{\partial x_1^{(1)}}{\partial x_1^0} = 2\Delta V^2\, \theta'(w x_1^0) \cdot w$

1 pt (1 pt)

$\frac{\partial e}{\partial x_2^0} = \frac{\partial e}{\partial x_2'} \cdot \frac{\partial x_2'}{\partial x_2^0} = 2\Delta V\, \theta'(w x_2^0) \cdot w$

1 pt (1 pt)

$\frac{\partial e}{\partial x_3^0} = \frac{\partial e}{\partial x_3'} \cdot \frac{\partial x_3'}{\partial x_3^0} = 2\Delta \cdot \theta'(w \cdot x_3^0)\, w$

1 pt

total/10

5 marks

(e) Suppose that $\mathbf{x} = (1, -1, 1)$ and $y = 1$. Assuming that $w = v = 1$ and $\theta(s) = \max(0, s)$, find numerical values for $e(\Omega)$, $\frac{de}{dv}$ and $\frac{de}{dw}$.

2.5 marks if they have generalized

$z_1^{(1)} = 1$, $z_2^{(1)} = 0$ $z_3^{(1)} = 1$

$z_2^{(2)} = 1$ $z_3^{(2)} = 2$

$e = 1$ (1pt)

$\frac{\partial e}{\partial v} = 2\Delta(z_2^{(2)} + v \cdot z_1^{(1)})$ $\Delta = 1$

$= 4.$ } 2pt

$\frac{\partial e}{\partial w} = 2\{1 + 0 + 1) = 4$ } 2pt

5 marks

(f) Suppose that the training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ consists of N training examples where each $\mathbf{x}_n \in \mathbb{R}^3$ and $y_n \in \mathbb{R}$. Write the pseudocode for training the neural network to minimize

$$\frac{1}{N}\sum_{n=1}^{n}(y_n - \hat{y}_n)^2 + \lambda(v^2 + w^2),$$

using stochastic gradient descent. Here $\lambda > 0$ is a fixed constant. Assume that you already have functions to compute $\frac{de}{dv}$ and $\frac{de}{dw}$.

for $t = 1, 2, \ldots$ — 1 pt

- Select (x_n, y_n) at random.
- do forward pass. (1 pt)
- Compute $\frac{\partial e}{\partial v}$, $\frac{\partial e}{\partial w}$ (1 pt)
- $v_{t+1} = v_t - \epsilon \frac{\partial e}{\partial v} - 2v\lambda$ (1 pt)

$w_{t+1} = w_t - \epsilon \frac{\partial e}{\partial w} - 2w\lambda$ (1 pt)

end.

total/10

4. (15 MARKS) Consider a regression problem where the training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_{100}, y_{100})\}$ consists of 100 points. Each $x_i \in \mathbb{R}$. However each $y_i \in \{0, 1\}$ is **binary valued**. Assume that the dataset $\mathcal{D}$ is generated as follows:

$$x_i = i/100, \quad 1 \le i \le 100$$
$$y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$$

Note that $\Pr(y_i = 1) = p$ and $\Pr(y_i = 0) = 1 - p$ and each y_i is sampled independently of all other labels.

We will consider two learning algorithms:

- **Algorithm NN**: Use 1-Nearest Neighbor Algorithm. In case of a tie use the datapoint to the left of the input.
- **Algorithm Zero**: Always predict zero.

For parts (a) and (b) we will use the Mean Squared Training Error:

$$E_{\text{in}} = \frac{1}{100} \sum_{i=1}^{100} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the output of the algorithm on training point x_i.

3 marks

(**a**) What is the expected Mean Squared Training Error: $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}]$, for Algorithm Zero?

$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}] = \underbrace{E[y^2]}_{2\text{ pt}} = \underbrace{p}_{1\text{ pt}}$

2 marks

(**b**) What is the expected Mean Squared Training Error: $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}]$, for Algorithm NN?

$E_{in} = 0.$ as nearest nbr $\Rightarrow y_i = \hat{y}_i$

$\mathbb{E}_{\mathcal{D}}[E_{in}] = 0$

Binary Marking

2 / 0.

total/5

For parts (c) and (d) we will use the leave one out cross validation as discussed in class.

2 marks (**c**) What is the expected leave one out cross validation error for Algorithm Zero?

1 pt Since $\hat{y} = 0$, the Alg. does not use D

Same as part a: $E_D[E_{in}] = p$ } 1 pt

6 marks (**d**) What is the expected leave one out cross validation error for Algorithm NN?

4 pt for simplification $= E_D\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]$

$= E_{y,y'}\left[(y-y')^2\right]$ $\quad y, y' \overset{iid}{\sim} \text{Ber}(p)$

4 pt for calculation $= E[y^2] + E[y'^2] - 2E[y]E[y']$

$= 2p - 2p^2 = 2p(1-p)$

2 marks (**e**) Based on the expected leave one out cross validation error, explain for what values of p will you choose Algorithm Zero over Algorithm NN?

total/10

5. (10 MARKS) In this problem you consider an already-trained Gaussian mixture model (GMM). The GMM was trained to fit data on student performance in an introduction-to-machine learning class. The GMM was trained using two components ($K = 2$) as the class consisted of two categories of students, undergraduate students (category 1) and graduate students (category 2). The learned parameters of the GMM are as follows:

   - The weights of the two categories are $w_1 = 2/3$ and $w_2 = 1/3$.
   - The distribution of scores in category 1 is $\mathcal{N}(x; 70, 10^2)$.
   - The distribution of scores in category 2 is $\mathcal{N}(x; 80, 5^2)$.

5 marks

(a) According to your model, what is the probability that an arbitrarily selected student scores greater than 80%? That is, compute $\Pr[x \geq 80]$. (In your computation, use the approximation that for a zero-mean σ^2-variance random variable x, i.e., $x \sim \mathcal{N}(x; 0, \sigma^2)$, $\Pr[|x| \leq \sigma] = 2/3$.)

$\Pr[x \geq 80] = \Pr[x \geq 80 \mid \text{class} = 1] \Pr[\text{class} = 1]$

$+ \Pr[x \geq 80 \mid \text{class} = 2] \Pr[\text{class} = 2]$

$= \frac{1}{6} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{1}{3}$

↑ b/c 80 is one std. deviation above mean for class 2 & by approximation 1/6 of prob mass above one std

↳ b/c 80 is the mean for class 2 and 1/2 the probability mass is above the mean for Gaussians

$= \frac{2}{18} + \frac{3}{18} = \boxed{\frac{5}{18} = \Pr[x \geq 80]}$

total/5

5 marks

(b) If a particular student has a score greater than 80, what is the probability that they are from category 1 (undergraduates)? That is, compute $\Pr[\text{class} = 1 | x \geq 80]$. (Use the same approximation as in the previous part.)

Use Bayes' Rule:

$$\Pr[\text{class}=1 \mid x \geq 80] \quad \text{(Bayes)}$$

$$= \frac{\Pr[\text{class}=1,\ x \geq 80]}{\Pr[x \geq 80]} \quad \text{(Bayes)}$$

$$= \frac{\Pr[x \geq 80 \mid \text{class}=1]\ \Pr[\text{class}=1]}{\Pr[x \geq 80]}$$

given & part(a)

$$= \frac{\frac{1}{6} \cdot \frac{2}{3}}{\frac{5}{18}}$$

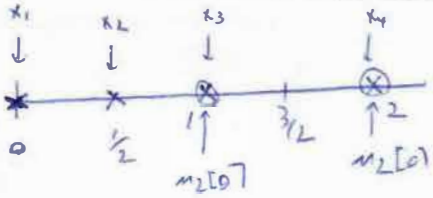
$$= \frac{2}{18} \cdot \frac{18}{5} = \boxed{\frac{2}{5} = \Pr[\text{class}=1 \mid x \geq 80]}$$

total/5

6. (10 MARKS) In this problem you consider the K – *means* algorithm. In this problem $K = 2$ and you have four data points x_n in your data set $\mathcal{D}$ all of which lie on the real line $x_i \in \mathbb{R}$. Your data set is $\mathcal{D} = \{0, 0.5, 0.5 + \Delta, 1.5 + \Delta\}$ where $\Delta \geq 0$ is a problem parameter.

4 marks

(a) For this part let $\Delta = 0.5$ and initialize K-means by initializing the two cluster centers at $\mu_1[0] = 1$ and $\mu_2[0] = 2$. Run K-means till convergence. For each iteration ℓ until convergence, describe your set memberships $\{\mathcal{B}_1[\ell], \mathcal{B}_2[\ell]\}$ and cluster centers $\{\mu_1[\ell], \mu_2[\ell]\}$. Make sure you identify the final values of the cluster centers and set memberships at convergence.



$\ell = 0$: $\times \rightsquigarrow$ data

$o \rightsquigarrow$ initialization of μ_i

$\ell = 1$: step (a), membership

$\mathcal{B}_1[1] = \{x_1, x_2, x_3\}$

$\mathcal{B}_2[1] = \{x_4\}$

step (b), cluster center

$\mu_1[1] = \frac{1}{3}\left(0 + \frac{1}{2} + 1\right) = \frac{1}{3}\ \frac{3}{2} = \frac{1}{2}$

$\mu_2[1] = \frac{1}{1}(2) = 2$

$\ell = 2$: At convergence since

$\|x_3 - \mu_2[1]\| = \frac{1}{2} < \|x_3 - \mu_2[1]\| = 3/2$

and x_3 is edge pt

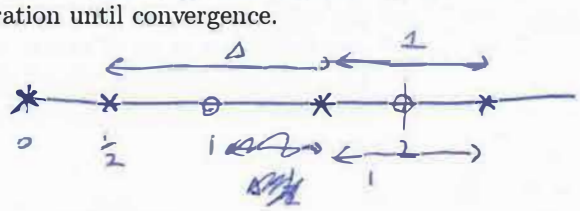
$\mathcal{B}_1^* = \{x_1, x_2, x_3\}$ $\quad \mu_1^* = \frac{1}{2}$

$\mathcal{B}_2^* = \{x_4\}$ $\quad \mu_2^* = 2$

total/4

6 marks

(b) For this part find the smallest positive value of Δ such that K-means, initialized in the same manner as in part (a), i.e., $\mu_1[0] = 1, \mu_2[0] = 2$, converges to a *different* solution from that obtained in part (a). In your solution describe (i) what is this minimum value of Δ and explain your reasoning / derivation, and (ii) as in part (a) run the cluster algorithm, describing the values of cluster centers and set memberships for each iteration until convergence.



IF $\Delta < 1$ Then will cluster w/ B_1 in 1st iteration & get

$\mu_1[1] = \frac{1}{3}\left(0 + \frac{1}{2} + \frac{1}{2} + \Delta\right) = \frac{1}{3} + \frac{\Delta}{3}$

$\mu_2[1] = \frac{3}{2} + \Delta$

note $\left\| \frac{1}{2} + \Delta - \underbrace{\left(\frac{1}{3} + \frac{\Delta}{3}\right)}_{\mu_1[1]} \right\| = \underbrace{\frac{1}{6} + \frac{2\Delta}{3}}_{\frac{1}{6} + \frac{2}{3} = \frac{5}{6}}$ w.l.o.g $\left\| \frac{3}{2} + \Delta - \left(\frac{1}{2} + \Delta\right) \right\| = 1$

so if $\Delta \geq 1$ stay in cluster 1

---

on other hand, if $\Delta = 1 + \varepsilon$, $\varepsilon \geq 0$ arbitrarily small

Then $B_1[1] = \{x_1, x_2\}$ $\mu_1[1] = \frac{1}{4}$

$B_2[1] = \{x_3, x_4\}$ $\mu_2[1] = \frac{1}{2}\left(\frac{3}{2} + \Delta + \frac{5}{2} + \Delta\right) = 2 + \Delta$

$\left\| \frac{1}{2} + \Delta - \frac{1}{4} \right\| = \frac{1}{4} + \Delta$ w.l.o.g $\left| 2 + \Delta - \left(\frac{3}{2} + \Delta\right) \right| = \frac{1}{2}$

since $\Delta > 1$ $> \frac{5}{4}$ $\downarrow \frac{1}{2}$

so, stay init clusters stay fixed

---

smallest positive value: any $\Delta \geq 1$

total/6