FIRST NAME: MohammadReza LAST NAME: Ebrahimi

STUDENT NUMBER: ____________________

ECE 421S/ECE1513S — Introduction to Machine Learning
Final Examination

April 17th, 2019
6:30 p.m. – 9:00 p.m.

Instructors: Ashish Khisti and Ben Liang and Amir Ashouri

Instructions

- Please read the following instructions carefully.
- You have 2 hour 30 minutes to complete the exam.
- Please make sure that you have a complete exam booklet.
- Please answer *all* questions. Read each question carefully.
- The value of each question is indicated. Allocate your time wisely!
- $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian density function with mean μ and variance σ^2.
- No additional pages will be collected beyond this answer book.
- This examination is closed-book; One 8.5 × 11 aid-sheet is permitted. A non-programmable calculator is also allowed.
- Good luck!

1. (20 MARKS) Consider linear regression with training data obtained from a noisy target function $y = f(\mathbf{x}) = \mathbf{c}^T\mathbf{x} + \epsilon$, where $y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^{d+1}$ (with bias term $x_0 = 1$ as usual), $\mathbf{c} \in \mathbb{R}^{d+1}$, and **the noise term ϵ is independently generated with zero mean and variance σ^2**. Suppose we have a set of N such training examples, $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$, *i.e.*, $y_n = \mathbf{c}^T\mathbf{x}_n + \epsilon_n$ for $n = 1, 2, \dots, N$.

Let $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$, and $X = \begin{bmatrix} \mathbf{x}_1{}^T \\ \mathbf{x}_2{}^T \\ \vdots \\ \mathbf{x}_N{}^T \end{bmatrix}$. We assume the columns of X are linearly independent, so that X^TX is invertible. Let $\mathbf{w}_{\text{LS}}$ be the linear least-squares estimator, *i.e.*, the in-sample error $E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\|X\mathbf{w} - \mathbf{y}\|^2$ is minimized when $\mathbf{w} = \mathbf{w}_{\text{LS}}$.

5 marks

(a) Show that $\mathbf{w}_{\text{LS}} = X^\dagger\mathbf{y}$, where $X^\dagger$ is the pseudo-inverse of X, *i.e.*, $X^\dagger = (X^TX)^{-1}X^T$. You may use without proof the fact that the gradient of $E_{\text{in}}(\mathbf{w})$ with respect to $\mathbf{w}$ is $\frac{2}{N}X^T(X\mathbf{w} - \mathbf{y})$.

$$\nabla_W E_{in}(W)\Big|_{W=W_{LS}} = 0$$

$$\Rightarrow \frac{2}{N} X^T(X W_{LS} - y) = 0 \quad \textcircled{1}$$

$$\Rightarrow X^TX W_{LS} - X^T y = 0$$

$$\Rightarrow W_{LS} = (X^TX)^{-1}X^T y \quad \textcircled{4}$$

$$= X^+ y$$

total/5

5 marks

(b) Find a simplified expression for $E_{\text{in}}(\mathbf{w}_{\text{LS}})$ as a function of N, X, $X^\dagger$, and ϵ. You may use without proof the fact that $(I - XX^\dagger)^k = I - XX^\dagger$ for any positive integer k.

$$E_{in}(W_{LS}) = \frac{1}{N} \| X W_{LS} - y \|^2$$

$$= \frac{1}{N} \| XX^\dagger y - y \|^2$$

$$= \frac{1}{N} \| (XX^\dagger - I) y \|^2$$

$$= \frac{1}{N} \| (XX^\dagger - I)(Xc + \epsilon) \|^2$$

2

$$= \frac{1}{N} \| \boxed{\underbrace{(XX^\dagger - I) Xc}_{0}} + (XX^\dagger - I)\epsilon \|^2$$

$$= \boxed{\frac{1}{N} \| (XX^\dagger - I)\epsilon \|^2}$$ 3

Note: $= \frac{1}{N} \| (I - XX^\dagger)\epsilon \|^2$

$$= \frac{1}{N} \epsilon^T \underbrace{(XX^\dagger - I)^T}_{\text{symmetric}} (XX^\dagger - I)\epsilon$$

$$= \frac{1}{N} \epsilon^T (XX^\dagger - I)^2 \epsilon$$

$$= \frac{1}{N} \epsilon^T (I - XX^\dagger) \epsilon$$ ← This is acceptable too.

total/10

5 marks (c) Observe the result in Part (b) for the extreme case of $d = 0$, *i.e.*, $X = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$. Find the expected in-sample error $\mathbb{E}_\epsilon(E_{\text{in}}(\mathbf{w}_{\text{LS}}))$ as a function of N and σ. (A side comment that can be safely ignored: a similar conclusion can be drawn for general d values, but you do not need to prove that.)

$$E_{in}(W_{LS}) = \frac{1}{N} \epsilon^T (I - XX^\dagger) \epsilon$$

$$= \frac{1}{N} tr\left(\epsilon^T (I - XX^\dagger) \epsilon\right)$$

$$= \frac{1}{N} tr\left((I - XX^\dagger) \epsilon\epsilon^T\right)$$

$$\Rightarrow \mathbb{E}_\epsilon\left[E_{in}(W_{LS})\right] = \frac{1}{N} tr\left((I - XX^\dagger) \underbrace{\mathbb{E}[\epsilon\epsilon^T]}_{\sigma^2 I}\right)$$

$$= \frac{\sigma^2}{N} tr(I - XX^\dagger)$$

$$tr(XX^\dagger) = tr(X^\dagger X) = tr(I_{d+1}) = d+1$$

$$\Rightarrow \mathbb{E}_\epsilon\left[E_{in}(W_{LS})\right] = \frac{\sigma^2}{N}(N - d - 1) = \sigma^2\left(1 - \frac{d+1}{N}\right)$$

solution for general d values.

total/5

5 marks

(c) Observe the result in Part (b) for the extreme case of $d = 0$, *i.e.*, $X = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$. Find the expected in-sample error $\mathbb{E}_\epsilon[E_{\text{in}}(\mathbf{w}_{\text{LS}})]$ as a function of N and σ. (Side comment: you should see that $\mathbb{E}_\epsilon[E_{\text{in}}(\mathbf{w}_{\text{LS}})]$ strictly increases in N and converges to σ^2. A similar conclusion holds for general d values, but you do not need to prove that.)

$$E_{in}(W_{LS}) = \frac{1}{N}\|(XX^{\dagger}-I)\epsilon\|^2$$

$$X^{\dagger} = (X^TX)^{-1}X^T = (N)^{-1}X^T = \frac{1}{N}[1,\dots,1]$$

$$\Rightarrow E_{in}(W_{LS}) = \frac{1}{N}\|(\frac{1}{N}XX^T-I)\epsilon\|^2 = \frac{1}{N^3}\left\|\begin{bmatrix}\epsilon_1+\dots+\epsilon_N-N\epsilon_1\\ \epsilon_1+\dots+\epsilon_N-N\epsilon_2\\ \vdots\\ \epsilon_1+\dots+\epsilon_N-N\epsilon_N\end{bmatrix}\right\|^2$$

$$= \frac{1}{N^3}\left[(\epsilon_1+\dots+\epsilon_N-N\epsilon_1)^2+\dots+(\epsilon_1+\dots+\epsilon_N-N\epsilon_N)^2\right]$$

$$\Rightarrow \mathbb{E}_\epsilon(E_{in}(W_{LS})) = \frac{1}{N^3}\left[var((1-N)\epsilon_1+\epsilon_2+\dots+\epsilon_N)+\dots+var(\epsilon_1+\dots+(1-N)\epsilon_N)\right]$$

$$= \frac{1}{N^3}\left[N((1-N)^2+(N-1))\sigma^2\right] = \underline{\sigma^2(1-\frac{1}{N})}$$

5

other solutions w/ same answer are ok

5 marks

(d) For any **out-of-sample** data point $\mathbf{x}$, the expected out-of-sample error is $\mathbb{E}_\epsilon[(f(\mathbf{x}) - \mathbf{w}_{\text{LS}}{}^T\mathbf{x})^2]$. Show that it is at least σ^2.

$$\mathbb{E}_\epsilon((f(x)-W_{LS}^Tx)^2) = \mathbb{E}_\epsilon((c^Tx+\epsilon-W_{LS}^Tx)^2)$$

$$= \mathbb{E}_\epsilon[((c-W_{LS})^Tx+\epsilon)^2]$$

$$= [(c-W_{LS})^Tx]^2 + \mathbb{E}(\epsilon^2)$$

$$\boxed{= \underbrace{[(c-W_{LS})^Tx]^2}_{\geqslant 0} + \sigma^2 \geqslant \sigma^2}$$

5

total/10

2. (20 MARKS) Consider a binary classification problem in two dimensions. Let $\mathcal{H}_0$ denote the hypothesis class of all linear classifiers in two dimensions passing **through the origin**. Thus any $h \in \mathcal{H}_0$ can be expressed as:

$$h(\mathbf{x}) = \text{sign}(w_1 x_1 + w_2 x_2)$$

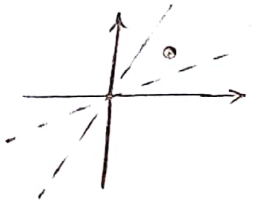
where $\mathbf{x} = (x_1, x_2)$ is the data point and $\mathbf{w} = (w_1, w_2) \in \mathbb{R}^2$.

Furthermore let $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ be arbitrary classification functions such that $g_i : \mathbb{R}^2 \to \{-1, +1\}$ and assume that neither $g_1(\cdot)$ nor $g_2(\cdot)$ are in $\mathcal{H}_0$.

6 marks (a) Let $m_{\mathcal{H}_0}(N)$ denote the growth function of the hypothesis class $\mathcal{H}_0$. Compute the growth function for $N = 1, 2, 3, 4$. Provide a brief justification for each computation.

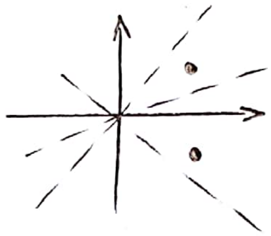
1

− $m_{\mathcal{H}_0}(1) = 2$



1

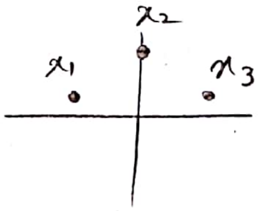
− $m_{\mathcal{H}_0}(2) = 4$



2

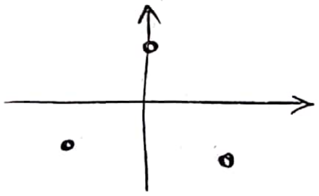
− $m_{\mathcal{H}_0}(3) = 6$

① if the points are on one side of a line



cannot generate −+−/+−+

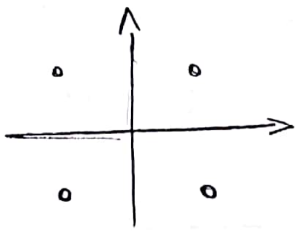
② otherwise



cannot generate −−−/+++

2

− $m_{\mathcal{H}_0}(4) = 8$



cannot select only one point. → 4 x 2 possibilities cannot be generated.

if no justification provided −2

total/6

4 marks

(b) Let us define a new hypothesis class $\mathcal{M} = \mathcal{H}_0 \cup \{g_1, g_2\}$. Let $m_{\mathcal{M}}(N)$ denote the growth function of $\mathcal{M}$. What is the maximum possible value $m_{\mathcal{M}}(N)$ for $N = 1,2,3,4$ over all choices of $g_1(\cdot)$ and $g_2(\cdot)$.

g_1 and g_2 can at most add one element to $M(\underline{x}_1,\dots,\underline{x}_N)$ each, as long as $|M(\underline{x}_1,\dots,\underline{x}_N)| < 2^N$

$$\Rightarrow \max_{g_1,g_2} m_M(N) = \begin{cases} 2 & N=1 \quad 1 \\ 4 & N=2 \quad 1 \\ 8 & N=3 \quad 1 \\ 10 & N=4 \quad 1 \end{cases}$$

4 marks

(c) Suppose we train the hypothesis class $\mathcal{M}$ defined in part (b) over a training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ (where $\mathbf{x}_i \overset{iid}{\sim} p_{\mathbf{x}}(\cdot)$ and $y_i = f(\mathbf{x}_i)$) and produce an output hypothesis $m(\cdot) \in \mathcal{M}$ that minimizes the training error. Let $E_{\text{in}}(m)$ denote the average classification error of $m(\cdot)$ on the training set $\mathcal{D}$ and $E_{\text{out}}(m)$ denote the test error for this hypothesis. Assume that $g_1(\cdot)$ and $g_2(\cdot)$ are **fixed** hypothesis selected before observing $\mathcal{D}$. Find a relation between $E_{\text{in}}(m)$ and $E_{\text{out}}(m)$ in the following sense:

With Probability at-least $1-\delta$, $E_{\text{out}}(m) \le E_{\text{in}}(m) + \Omega$, where you should provide a bound on Ω using the VC dimension theory.

VC bound : $\Omega = \sqrt{\frac{8}{N}\log\left(\frac{4\, m_M(2N)}{\delta}\right)}$

$m_M(N) \le N^{d_{vc}(M)} + 1$

part b $\longrightarrow$ $\boxed{d_{vc}(M) = 3}$ 2

$$\Rightarrow \underbrace{\Omega \le \sqrt{\frac{8}{N}\log\left(\frac{4((2N)^3+1)}{\delta}\right)}}_{2}$$

total/8

6 marks

(d) In this part, suppose that the hypothesis $g_1(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{R}^2$. However $g_2(\cdot)$ is selected **after** observing the dataset $\mathcal{D}$ (stated in part (c)) as follows:

Let $\mathcal{F}$ denote the class of all hypothesis where the decision boundaries are horizontal in the $x_1 - x_2$ plane i.e.,

$$f(\mathbf{x}) = \text{sign}(a \cdot x_2 - c)$$

where $c \in \mathbb{R}$ and $a \in \{-1, +1\}$. We select $g_2(\cdot)$ by finding $f \in \mathcal{F}$ that minimizes the average classification error over $\mathcal{D}$, i.e., $E_{\text{in}}(f)$

The hypothesis $g_2(\cdot)$, thus selected, is included in $\mathcal{M}$ along with $g_1(\cdot)$. We train the hypothesis class $\mathcal{M}$ on the **same** dataset $\mathcal{D}$ (used to select $g_2(\cdot)$), and output $m \in \mathcal{M}$ that minimizes the training error.

Find a relation between $E_{\text{in}}(m)$ and $E_{\text{out}}(m)$ in the following sense: with probability at-least $1 - \delta$, $E_{\text{out}}(m) \leq E_{\text{in}}(m) + \Omega$, where you should provide a bound on Ω using the VC dimension theory.

This is equivalent to use hypothesi class $\mathcal{H}_1 = \mathcal{H}_0 \cup \mathcal{F}$ 1

$m_{\mathcal{H}_1}(3) = 8$

$m_{\mathcal{H}_1}(4) \leqslant 10 \longrightarrow$ from part b

$\Big\} \Rightarrow \boxed{d_{vc}(\mathcal{H}_1) = 3}$ 4

$$\Rightarrow \Omega \leqslant \sqrt{\frac{8}{N} \log\left(\frac{4((2N)^3+1)}{\delta}\right)}$$ 1

if used upper bound on d_{vc}:

$d_{vc}(\mathcal{H}_0 \cup \mathcal{F}) \leqslant 1 + d_{vc}(\mathcal{H}_0) + d_{vc}(\mathcal{F}) = 5$

assign half mark. 3/6

total/6

3. (20 MARKS) Consider a target function $f(x) = x^2$ where the domain of input x is the interval $[-1,1]$. The training data set contains only one example: $\mathcal{D} = \{(U, U^2)\}$, where U is sampled uniformly from $[-1,1]$. Our hypothesis set $\mathcal{H}$ consists of all lines of the form $h(x) = ax$, for $a \in \mathbb{R}$. Let $g^{\mathcal{D}}(x)$ be the output hypothesis given training data set $\mathcal{D}$.

5 marks (a) Show that the average hypothesis, *i.e.*, the expectation of $g^{\mathcal{D}}(x)$ over $\mathcal{D}$, is $\bar{g}(x) = 0$.

$$g^{\mathcal{D}}(u) = u^2 \Rightarrow a^{\mathcal{D}} u = u^2 \Rightarrow a^{\mathcal{D}} = u$$

$$\Rightarrow \boxed{g^{\mathcal{D}}(x) = ux}$$ 2

$$\bar{g}(x) = \mathbb{E}_u(ux) = \underbrace{\mathbb{E}_u(u)}_{0}\, x = 0 \Rightarrow \underline{\bar{g}(x) = 0}$$ 3

5 marks (b) Find bias(x) and var(x).

$$\text{bias}(x) = (\bar{g}(x) - f(x))^2 = (0 - x^2)^2 = \underline{x^4}$$ 2

$$\text{var}(x) = \mathbb{E}_{\mathcal{D}}\left[(g^{\mathcal{D}}(x) - \bar{g}(x))^2\right]$$

$$= \mathbb{E}_u\left[(u - \bar{u})^2 x^2\right] = \text{var}(u)\, x^2$$

$$= \underline{\frac{1}{3} x^2}$$ 3

total/10

5 marks (c) Find the expected out-of-sample error $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{\mathcal{D}})]$.

Solution 1

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})] = \mathbb{E}_x[bias(x) + var(x)] \quad (2)$$

$$= \mathbb{E}_x\left(\frac{1}{3}x^2 + x^4\right)$$

$$= \int_{-1}^{1} \frac{1}{2}\left(\frac{1}{3}x^2 + x^4\right)dx = 1/9 + 1/5 = 14/45 \quad (3)$$

Solution 2

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})] = \mathbb{E}_u\left[\mathbb{E}_x\left((g^{\mathcal{D}}(x) - f(x))^2\right)\right]$$

$$= \mathbb{E}_{u,x}\left[(ux - x^2)^2\right] = \cdots$$

acceptable if answer is 14/45

5 marks (d) Suppose we actually have two data examples: $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$, where x_1 and x_2 are independently and uniformly sampled from $[-1, 1]$. However, we use the leave-one-out cross validation method, so that the training data set always contains only one example. What is the expectation of the cross validation estimation E_{CV} over $\mathcal{D}$?

Solution 1

$$\mathbb{E}_{\mathcal{D}}[E_{cv}] = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{2}(e_1 + e_2)\right] = \mathbb{E}_{\mathcal{D}}[e_1]$$

$$e_1 = \left(f(x_1) - g_1^-(x_1)\right)^2$$

$$= (x_1^2 - x_2x_1)^2 = x_1^4 + x_2^2x_1^2 - 2x_2x_1^3 \quad (2)$$

$$\mathbb{E}_{\mathcal{D}}[e_1] = \mathbb{E}_{x_1,x_2}\left[x_1^4 + x_2^2x_1^2 - 2x_2x_1^3\right]$$

$$= \mathbb{E}(x_1^4) + \mathbb{E}(x_1^2)^2 - 2\cancelto{0}{\mathbb{E}(x_2)}\,\mathbb{E}(x_1^3)$$

$$= \frac{1}{5} + \left(\frac{1}{3}\right)^2 = \frac{14}{45} \quad (3)$$

Sol2 E_{cv} is an unbiased estimator of average out of sample error.

$\Rightarrow$ part (c) $\rightarrow$ $\mathbb{E}^{\mathcal{D}}(E_{cv}) = \mathbb{E}^{\mathcal{D}}[E_{out}] = 14/45$ ✓

total/10

4. (20 MARKS) In ancient times, there was a village surrounded by hundreds of lakes. Each lake was either poisonous (P) or healthy (H). 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 identical and tasted identical and there was no way of knowing whether a lake was poisonous or healthy.

Fortunately a famous chemist visited the village and was told of this dilemma. She suggested using the pH level of water to determine whether a given lake was poisonous or healthy. She hypothesized that lakes with poisonous fish would have higher pH value than healthy lakes. Accordingly she visited each lake and collected the pH value of the water in each lake. The data set is denoted by:

$$\mathcal{D} = \{l_1, l_2, \ldots, l_N\}$$

where l_i denotes the pH level of lake $i \in \{1, 2, \ldots, N\}$. We assume that $l_1 \leq l_2 \leq l_3 \leq \ldots \leq l_N$.

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

- Pr(lake is poisonous) = p_1
- Pr(lake is healthy) = p_2
- $f(\text{pH} = l \mid \text{lake is poisonous}) \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- $f(\text{pH} = l \mid \text{lake is healthy}) \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- $\mu_1 \leq \mu_2$ × $\mu_1 \gg \mu_2$

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 marks (a) Using the above GMM, provide an expression of the probability that a lake is poisonous given that its pH level is measured to be l.

$$Pr(P \mid l) = \frac{f(l \mid P)\, P_1}{f(l \mid P) P_1 + f(l \mid H) P_2}$$

5

or in terms of $\mathcal{N}(\mu_i, \sigma_i^2)$

total/5

5 marks

(b) 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 $B_1^0 = \{l_1, l_2, \dots, l_K\}$ denotes one cluster of lakes and $B_2^0 = \{l_{K+1}, \dots, l_N\}$ denotes it's complement.

```
- Initialize $B_1^{(new)}$, $B_2^{(new)}$ by $B_1^0$, $B_2^0$
- $B_1 \leftarrow B_1^{(new)}$, $B_2 \leftarrow B_2^{(new)}$
1 - $\boxed{\mu_1 \leftarrow \frac{1}{|B_1|} \sum_{l_i \in B_1} l_i}$, $\mu_2 \leftarrow \frac{1}{|B_2|} \sum_{l_i \in B_2} l_i$
1 - $\boxed{\sigma_1^2 \leftarrow \frac{1}{|B_1|} \sum_{l_i \in B_1} (l_i - \mu_1)^2}$, $\sigma_2^2 \leftarrow \frac{1}{|B_2|} \sum_{l_i \in B_2} (l_i - \mu_2)^2$
1 - $\boxed{P_1 \leftarrow |B_1|/N}$, $P_2 \leftarrow |B_2|/N$
- $B_1^{(new)}$, $B_2^{(new)} \leftarrow \phi$
- for each $l$ in $B_1 \cup B_2$:
    1 { - if $Pr(P|l) \geq \frac{1}{2}$ : $B_1^{(new)} \leftarrow B_1^{(new)} \cup l$
        - else : $B_2^{(new)} \leftarrow B_2^{(new)} \cup l$
- run untill $[B_1^{(new)}, B_2^{(new)}] = [B_1, B_2]$ 1
```

5 marks

(c) Write down the pseudocode for the EM algorithm with soft decisions that finds the parameters of the GMM. Do state the initialization of your algorithm explicitly.

```
- Initialize $\mu_i \leftarrow mean(B_i^0)$, $\sigma_i^2 \leftarrow var(B_i^0)$, $P_i \leftarrow \frac{|B_i|}{N}$
// E-step:
- for $i$ in 1 to $N$:
    1 $r_i^{(1)} \leftarrow Pr(P|l_i)$, $r_i^{(2)} \leftarrow 1 - Pr(P|l_i)$
  // M-step:
- for $c$ in 1 to 2:
    1 $\mu_c = \frac{\sum_{i=1}^{N} r_i^{(c)} l_i}{\sum_{i=1}^{N} r_i^{(c)}}$, 1 $\sigma_c^2 = \frac{\sum_{i=1}^{N} (l_i - \mu_c)^2 r_i^{(c)}}{\sum_{i=1}^{N} r_i^{(c)}}$, 1 $P_c = \frac{\sum r_i^{(c)}}{N}$
- run untill log-likelihood change < $\epsilon$
```

-1 each if init. or stopping not specified

total/10

5 marks (d) Suppose that N = 5 and we observe $\mathcal{D} = \{1, 2, 3, 6, 10\}$. Let $\mathcal{B}_1 = \{1, 2, 3\}$ and $\mathcal{B}_2 = \{6, 10\}$. Execute the hard decision EM algorithm and compute the parameters of the GMM.

Iteration 1 : $B_1 = \{1,2,3\}$ $B_2 = \{6,10\}$

2 M
$$\begin{cases} \mu_1 = 2 \quad \mu_2 = 8 \\ \sigma_1^2 = 0.667 \quad \sigma_2^2 = 4 \\ P_1 = 3/5 \quad P_2 = 2/5 \end{cases}$$

3 E
$$\begin{cases} Pr(1 \in B_1 \mid \ell = 1) = 0.998 \longrightarrow 1 \in B_1^{(new)} \\ Pr(2 \in B_1 \mid \ell = 2) = 0.997 \longrightarrow 2 \in B_1^{(new)} \\ Pr(3 \in B_1 \mid \ell = 3) = 0.975 \longrightarrow 3 \in B_1^{(new)} \\ Pr(6 \in B_1 \mid \ell = 6) = 0 \longrightarrow 6 \in B_2^{(new)} \\ Pr(10 \in B_1 \mid \ell = 10) = 0 \longrightarrow 10 \in B_2^{(new)} \end{cases}$$

$\Rightarrow B_1^{(new)} = \{1,2,3\}$ $B_2^{(new)} = \{6,10\}$

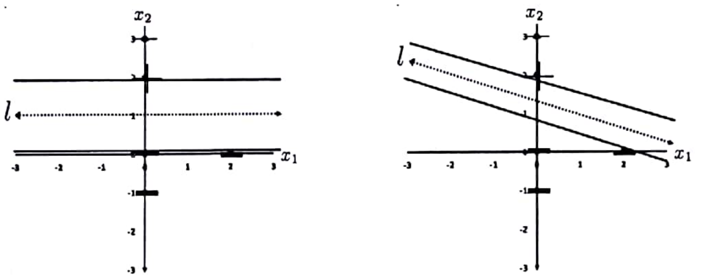
$B_1^{(new)} = B_1 \Rightarrow$ stop.

total/5

1. (20 MARKS) Consider a dataset with 5 data points in $\mathbb{R}^2$

$X_1 = (0,0)$	$y_1 = -1$
$X_1 = (2,0)$	$y_2 = -1$
$X_1 = (0,-1)$	$y_2 = -1$
$X_1 = (0,2)$	$y_3 = +1$
$X_1 = (0,3)$	$y_4 = +1$

Figures below correctly classify the given data using a hyperplane (l)



(a) (b)

3 marks

(a) Given the provided classifiers and the 5 data points above, choose the one that optimally does a maximum margin classification using SVM.

(a) has a larger margin.

3

5 marks

(b) Using the given data points, formulate a Q-P optimization by writing down the objective function and the constraints.

$$\begin{cases} \min\limits_{w,b} \frac{1}{2} w^T w \\ \text{s.t.} \quad y_n (w^T x_n + b) \geqslant 1 \end{cases}$$

$$\equiv \begin{cases} \min\limits_{w_1,w_2,b} \frac{1}{2}(w_1^2 + w_2^2) \longrightarrow \text{(-2) if wrong} \\ \text{s.t.} \quad -b \geqslant 1 \quad (1) \quad 1 \\ \quad -2w_1 - b \geqslant 1 \quad (2) \quad 1 \\ \quad -w_2 - b \geqslant 1 \quad (3) \quad 1 \\ \quad 2w_2 + b \geqslant 1 \quad (4) \quad 1 \\ \quad 3w_2 + b \geqslant 1 \quad (5) \quad 1 \end{cases}$$

total/10

7 marks (c) Show/argue that the solution you chose in (a) is the optimal margin for the given points.

combining (1) and (4) $\longrightarrow w_2 \geq 1$ 2

combining (1) and (2) $\longrightarrow w_1 \leq 0$ 2

$\Rightarrow \underbrace{w_2^* = 1, \; w_1^* = 0, \; b^* = -1}_{2} \Rightarrow$ max-margin hyperplane: 1 $x_2 - 1 = 0$

5 marks (d) Find the support vectors and the E_{CV} bound for the following data.

support vectors: $\overset{1}{x_1}$ $\overset{1}{x_2}$ $\overset{1}{x_4}$

$$E_{CV} \leq \frac{\#\text{ of support vectors}}{N} = 3/5$$ 2

total/10