FIRST NAME: ______________________ LAST NAME: ______________________

STUDENT NUMBER: ______________________

ECE 421S/ECE1513S — Introduction to Machine Learning
Makeup 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 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.

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 $g_2(\cdot)$.

4 marks

(**c**) Suppose we train the hypothesis class $\mathcal{M}$ defined in part (b) over a training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ (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 $1 - \delta$, $E_{\text{out}}(m) \leq E_{\text{in}}(m) + \Omega$, where you should provide a bound on Ω using the VC dimension theory.

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., $E_{\text{in}}(f)$ is minimized.

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.

total/6

2. (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 $\mathcal{D}$, to minimize the squared error..

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

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

total/10

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

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 $\mathcal{D}$?

total/10

3. (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 $l_1 \le l_2 \le l_3 \le \ldots \le l_N$.

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:

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

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 $\mathcal{B}_2 = \{l_1, l_2, \ldots, l_K\}$ denotes one cluster of lakes and $\mathcal{B}_1 = \{l_{K+1}, \ldots, l_N\}$ denotes it's complement.

total/5

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.

total/5

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

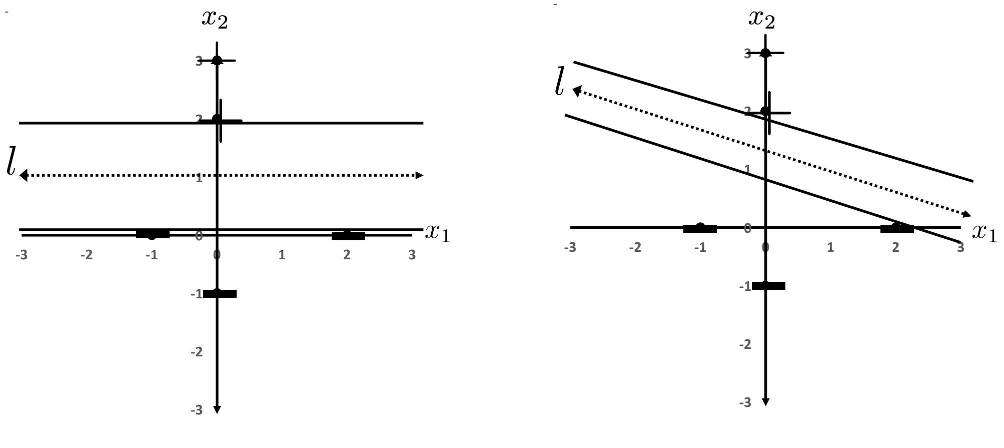
total/5

4. (20 MARKS) Consider a binary linear classification problem where the data vectors $\mathbf{x} \in \mathbb{R}^2$ and the labels $y \in \{-1, +1\}$. Our dataset consists of the following labeled data-points:

$X_1 = (-1, 0) \quad y_1 = -1$
$X_2 = (2, 0) \quad y_2 = -1$
$X_3 = (0, -1) \quad y_2 = -1$
$X_4 = (0, 2) \quad y_3 = +1$
$X_5 = (0, 3) \quad y_4 = +1$

We consider linear classifiers of the form $\hat{y} = \text{sign}(w_1 x_1 + w_2 x_2 + b)$, where $\mathbf{w} = (w_1, w_2)$ is the weight vector and b is the bias.

Consider the two classifiers shown in the figure below. The classification boundary is shown by the dotted line. In addition, we also show two lines parallel to the classification boundary passing through the nearest training points in each class.



(a) (b)

3 marks **(a)** Which of the two classifiers has a larger margin?

5 marks **(b)** Using the given data points, write down a quadratic programming problem to find the maximum margin classifier. You can express your optimization as minimizing an objective function $f(\mathbf{w}, b)$ subject to constraints of the form $g_n(\mathbf{w}, b) \geq 1$ for $n = 1, 2, \ldots, 5$ and specify the functions $f(\cdot)$ and $g_n(\cdot)$.

total/8

7 marks **(c)** Show/argue that the classifier in figure (a) in the previous page achieves optimal margin for the given points. Also specify the support vectors.

5 marks **(d)** Provide an upper bound on the leave-one-out cross validation for the classification rule in part (c).

total/12

5. (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, $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, *i.e.*, $y_n = \mathbf{c}^T\mathbf{x}_n + \epsilon_n$ for $n = 1, 2, \ldots, N$.

Let $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$, 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 $\frac{2}{N}X^T(X\mathbf{w} - \mathbf{y})$.

5 marks

(**b**) Find an expression for $E_{\text{in}}(\mathbf{w}_{\text{LS}})$ as a function of only N, X, $X^\dagger$, and $\boldsymbol{\epsilon}$.

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 $\mathbb{E}_{\boldsymbol{\epsilon}}[E_{\text{in}}(\mathbf{w}_{\text{LS}})]$ as a function of N and σ. (Side comment: you should see that $\mathbb{E}_{\boldsymbol{\epsilon}}[E_{\text{in}}(\mathbf{w}_{\text{LS}})]$ 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.)

5 marks

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

total/10