FIRST NAME: ____________________ LAST NAME: ____________________

STUDENT NUMBER: ____________________ Section (Circle One): **Draper** **Khisti**

ECE 521S — Inference Algorithms and Machine Learning MidTerm Examination

Friday March 2nd, 2018
4:10 p.m. – 6:00 p.m.

Instructor: Ashish Khisti and Stark Draper

Circle your tutorial section:

1. TUT0101 Wed 10:00-12:00(LM155)
2. TUT0102 Thu 9:00-11:00(BA2175)
3. TUT0103 Wed 10:00-12:00(HA410)
4. TUT0104 Wed 12:00-14:00(HS106)
5. TUT0105 Tue 15:00-17:00(BA2175)

Instructions

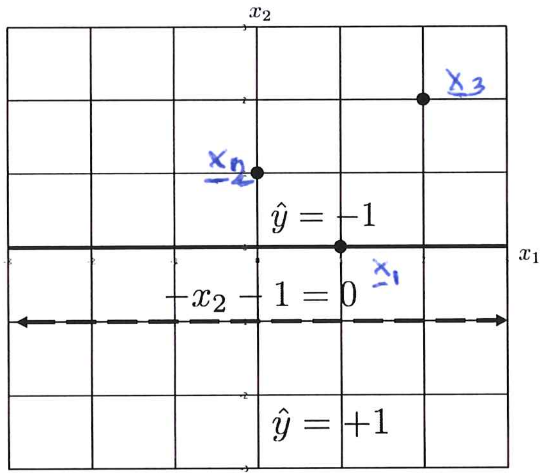
- Please read the following instructions carefully.
- You have 1 hour fifty minutes (1:50) to complete the exam.
- Please make sure that you have a complete exam booklet.
- Please answer *all* questions. Read each question carefully.
- The value of each question is indicated. Allocate your time wisely!
- No additional pages will be collected beyond this answer book. You may use the reverse side of each page if needed to show additional work.
- This examination is closed-book; One 8.5 × 11 aid-sheet is permitted. A non-programmable calculator is also allowed.
- Good luck!

1. (10 MARKS) Consider a binary classification problem where the data points are two dimensional, i.e., $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ and the labels $y \in \{-1, 1\}$. Throughout this problem consider the following three points:

$$\mathbf{x}_1 = (1,0)^T, \quad \mathbf{x}_2 = (0,1)^T, \quad \mathbf{x}_3 = (2,2)^T.$$

4 marks

(a) Suppose that the hypothesis set $\mathcal{H}$ consists of all linear classifiers whose decision boundary is a **horizontal line** in the (x_1, x_2) plane. As one example, the classifier $h(\mathbf{x}) = \text{sign}(-x_2 - 1)$ belongs to the set $\mathcal{H}$. The decision boundary of $h(\mathbf{x})$ is the dashed horizontal line shown in the figure below. Note that in this classifier all points below this line are classified as $\hat{y} = +1$ while all points above the horizontal line are classified as $\hat{y} = -1$ by the hypothesis $h(\mathbf{x})$



List all dichotomies in $\mathcal{H}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ that can be achieved. (Recall that a dichotomy in this problem will be a vector of length three whose elements are either $+1$ or -1, and is achieved by applying some hypothesis in $\mathcal{H}$ to the points $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ in that order.)

What is the VC dimension of $\mathcal{H}$? (No justification is needed for this)

Achievable Dichotomies:

x_1	x_2	x_3
+1	+1	+1
-1	+1	+1
-1	-1	+1
-1	-1	-1
1	-1	-1
1	1	-1

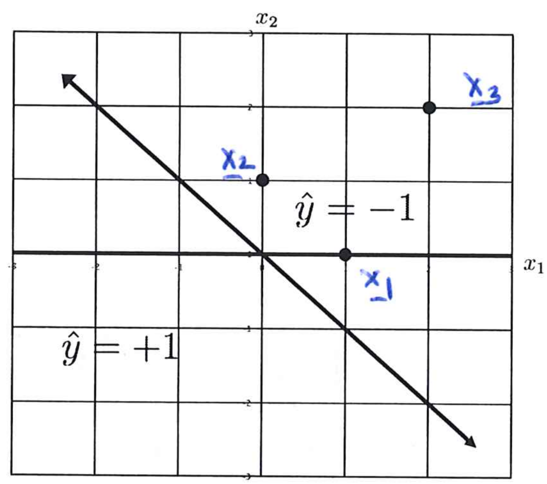
0.5 pt for each correct dichotomy
- 0.5 for each incorrect

$\underbrace{d_{vc}(H) = 1}_{1 \text{ pt}}$

total/4

4 marks

(b) Suppose that the hypothesis set $\mathcal{G}$ consists of all linear classifiers **passing through the origin**. As one example the classifier $g(\mathbf{x}) = \text{sign}(-x_1 - x_2)$ belongs to the set $\mathcal{G}$. Its decision boundary is shown by the solid line passing through the origin in the figure below. Note that all points below the decision boundary are classified as $\hat{y} = +1$ and all points above this line are classified as $\hat{y} = -1$.



List all dichotomies in $\mathcal{G}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ that can be achieved.

State without justification the VC dimension of $\mathcal{G}$.

0.5 for each correct
-0.5 for an incorrect

$\underline{x_1}$	$\underline{x_2}$	$\underline{x_3}$
+1	+1	+1
-1	+1	+1
1	-1	+1
-1	-1	-1
1	-1	-1
-1	1	-1

$$\underbrace{d_{VC}(\mathcal{H}) = 2}_{1 \text{ pt}}$$

2 marks

(c) Suppose $\mathcal{M} = \mathcal{H} \cup \mathcal{G}$ is the union of the hypothesis classes in parts (a) and (b). What is the number of dichotomies in $\mathcal{M}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$? Provide a **brief** justification for your answer.

$|\mathcal{M}(\underline{x_1}, \underline{x_2}, \underline{x_3})| = 8$ — 1

the two dichotomy vector missing in $\mathcal{G}$, are found in $\mathcal{H}$. 1

total/6

2. (10 MARKS) Consider a regression problem where the input $x \in [-1, +1]$ and the label y is generated using the following target function:

$$y = f(x) = \begin{cases} 1, & \text{if } 0 \le x \le 1, \\ -1, & \text{if } -1 \le x < 0, \\ 0, & \text{else.} \end{cases} \tag{1}$$

Suppose that the training set consists of $N = 1$ point i.e., $\mathcal{D} = \{(x_0, y_0)\}$, where x_0 is sampled uniformly from $[-1, 1]$ and $y_0 = f(x_0)$.

Consider the hypothesis class $\mathcal{H}_0$ of functions that are constant i.e.,

$$\mathcal{H}_0 = \{h(x) = c : c \in \mathbb{R}\}.$$

Let $g^{\mathcal{D}}(x)$ be the hypothesis in $\mathcal{H}_0$ that will minimize the in-sample training error over $\mathcal{D}$.

2 marks

(a) Find an expression for $g^{\mathcal{D}}(x)$ by separately considering the cases when $x_0 \ge 0$ and $x_0 < 0$.

$$g^{\mathcal{D}}(x) = \begin{cases} +1, & 1 \ge x_0 \ge 0 \quad \text{1 pt} \\ -1, & -1 \le x_0 < 0 \quad \text{1 pt} \end{cases}$$

2 pt

$\left(\text{or } g^{\mathcal{D}}(x) = y_0 \right)$

5 marks

(b) Compute $\bar{g}(x) = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)]$, $\text{bias}(x) = (\bar{g}(x) - f(x))^2$ and $\text{var}(x) = \mathbb{E}_{\mathcal{D}}\left[(g^{\mathcal{D}}(x) - \bar{g}(x))^2\right]$. Here $\mathbb{E}[\cdot]$ denotes the expectation operator.

$\bar{g}(x) = \mathbb{E}[y_0] = 0$ 1 pt

$\text{bias}(x) = (0 - f(x))^2 = 1$ 2 pt

$\text{var}(x) = \mathbb{E}_{\mathcal{D}}[(y_0)^2] = \mathbb{E}_{\mathcal{D}}[y_0^2]$

$= 1$ 2 pt

total/7

3 marks

(c) Let the out of sample test error associated with $\mathcal{D}$ be defined as

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

where x is uniformly distributed over $[-1, 1]$ and is independent of $\mathcal{D}$. Compute the following using your result in part (b):

$$\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{\mathcal{D}})].$$

We have shown that

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})] = \underbrace{E_x[\text{bias}(x)] + E_x[\text{var}(x)]}_{1\text{pt}}$$

$E_x[\text{bias}(x)] = 1$ 1 pt

$E_x[\text{var}(x)] = 1$ 1 pt

$\therefore \mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})] = 2.$

total/3

10 marks **3.** Suppose we are given a sequence of real numbers: $x_1, x_2, x_3, x_4, \ldots$ where $x_i \in \mathbb{R}$. We observe the following values: $x_1 = -1$, $x_2 = 0$, $x_3 = +1$, $x_4 = +1$.

We wish to select a prediction function of the form $\hat{x}_i = w_0 + w_1 \cdot x_{i-1}$, for $i \geq 2$ that makes a prediction of x_i from the value of x_{i-1}. Our task is to minimize the following in sample training error:

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{3} \sum_{i=2}^{4} (\hat{x}_i - x_i)^2.$$

3 marks (a) Rewrite the above problem specifications to get the problem into the standard form of a least squares problem: $\mathbf{w}^\star = \arg\min_{\mathbf{w} \in \mathbb{R}^2} \frac{1}{3} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$. Specifically specify the data matrix $\mathbf{X}$ and the target vector $\mathbf{y}$, where $\mathbf{w} = (w_0, w_1)^T$.

$$\underline{y} = \begin{bmatrix} x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \longrightarrow 1 \text{ pt}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \longrightarrow 2 \text{ pt}$$

3 marks (b) Find the least squares solution $\mathbf{w}^\star = (w_0^\star, w_1^\star)$ in part (a).

$$\underline{w}^* = (X^T X)^{-1} X^T \underline{y}$$

$$(X^T X)^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \text{ — } 1 \text{ pt}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \text{ — } 1 \text{ pt}$$

$$\underline{w}^* = \begin{bmatrix} 2/3 \\ 1/2 \end{bmatrix} \text{ — } 1 \text{ pt}$$

(alternative solution accepted)

total/6

4 marks

(c) Redo parts (a) and (b) if the prediction function is of the form $\hat{x}_i = w_1 \cdot x_{i-1}$ i.e., we set $w_0 = 0$.

$$\underbrace{\underline{y} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}}_{1\text{ pt}} \qquad \underbrace{X = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}}_{1\text{ pt}}$$

$$(X^T X)^{-1} = \frac{1}{2}$$

$$X^T Y = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = 1$$

1 pt.

$\underline{w}_1^* = \frac{1}{2}$ — 1 pt

(alternative solu^n to $\underline{w}_1^*$ is accepted)

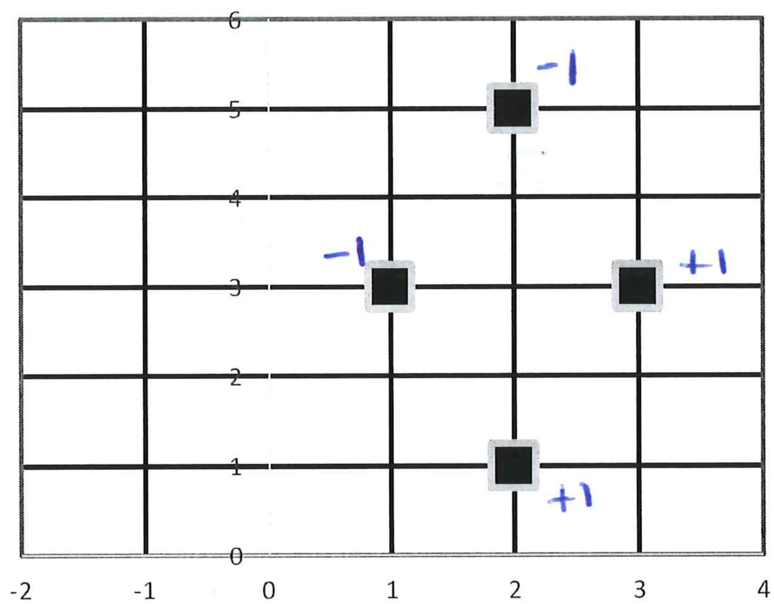
total/4

6 marks

4. Suppose we have a binary linear classification problem with each training point $\mathbf{x} \in \mathbb{R}^2$ and each label $y \in \{-1, 1\}$. Suppose that our training set consists of four points:

$$\mathbf{x}_1 = (1,3)^T, \quad y_1 = -1$$
$$\mathbf{x}_2 = (2,5)^T, \quad y_2 = -1$$
$$\mathbf{x}_3 = (3,3)^T, \quad y_3 = +1$$
$$\mathbf{x}_4 = (2,1)^T, \quad y_4 = +1$$

We implement a **k-Nearest Neighbor classification** rule as in Assignment 1.



3 marks

(a) Suppose that $k = 1$. Specify the output for the following input points: $\tilde{\mathbf{x}}_1 = (1.5, 3)$, $\tilde{\mathbf{x}}_2 = (2.5, 3)$, $\tilde{\mathbf{x}}_3 = (0, 0)$ (no work needed).

$\hat{y}_1 = -1 \quad \hat{y}_2 = +1, \quad \hat{y}_3 = +1$

(1 pt each)

3 marks

(b) Suppose that $k = 3$. Consider the point $\tilde{\mathbf{x}} = (0, 0)$: (1) find the nearest three neighbors and (2) find output label.

2 pt → Nearest Nhbrs: ~~(2, 1)~~ $\underline{x}_4$, $\underline{x}_1$ $\underline{x}_3$

1 pt → output $\hat{y} = +1$

total/6

4 marks

5. Answer the following Tensorflow questions.

2 marks

(a) Suppose we save the following code in a file 'hello.py' and execute it. What will be the output?

```
import tensorflow as tf
a = tf.constant(2)
b = tf.constant (3)
c = tf.add(a,b)
print(c.eval())
```

error. message for print(.)
statement

2pt

2 marks

(b) Modify the code in 'hello.py' so that we print the numerical value of c.

~~do~~' sess = tf. InteractiveSession ()
c = tf. add (a, b)
print (c. eval ())

2pt

'OR' with tf. Session ()
print (c. eval ())

total/4