FIRST NAME: ______________ LAST NAME: ______________

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

ECE 521S — Inference Algorithms and Machine Learning Final Examination

April 17th, 2018
6:30 p.m. – 9:00 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 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!
- All logarithms are to the base e unless otherwise noted.
- 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) 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.

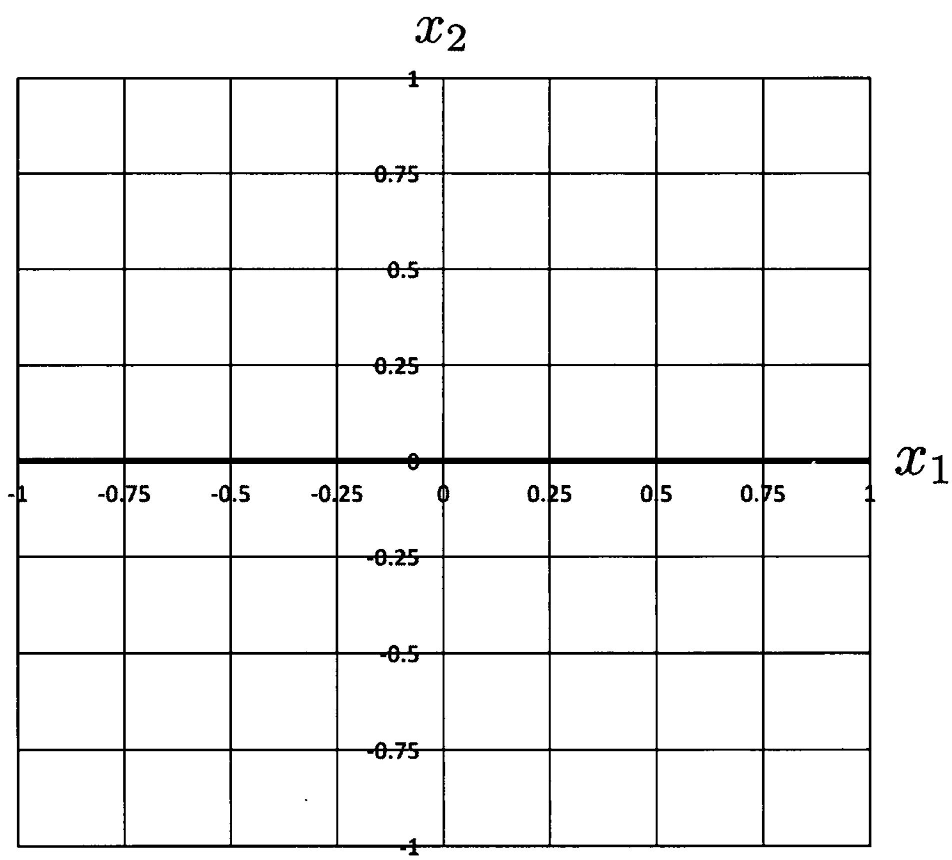
   (ii) The boundary of the linear hypothesis h is the 45-degree line, connecting $(-1,-1)$ to the origin to $(+1,+1)$.

   (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 training set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5\}$ 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.

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?

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 $|x_1| \le 1$ and $|x_2| \le 1$, what is $E_{\text{OUT}}(\mathbf{w}_0)$?

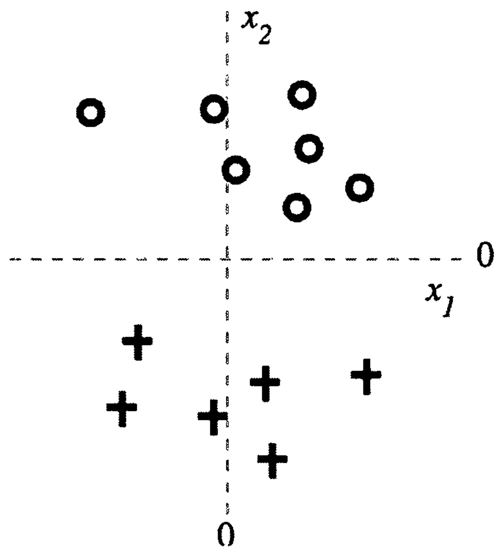
5 marks (e) If, instead, $P(\mathbf{x})$ is defined as

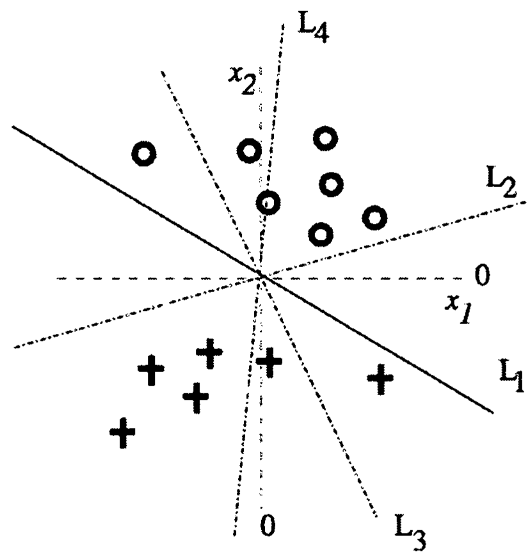
$$P(\mathbf{x}) = \begin{cases} \frac{1}{3} & \text{if} \quad |x_1| \le 1 \text{ and } 0.5 \le x_2 \le 1 \\ \frac{2}{9} & \text{if} \quad |x_1| \le 1 \text{ and } -1 \le x_2 < 0.5 \end{cases},$$

what is $E_{\text{OUT}}(\mathbf{w}_0)$?

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 figures above, the horizontal axis corresponds to x_1 and the vertical axis corresponds to x_2. 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, and represented by: $w_1x_1 + w_2x_2 = 0$, where $\mathbf{w} = (w_1, w_2) \in \mathbb{R}^2$.

Furthermore we consider a simple logistic regression model. The outputs given $\mathbf{x} = (x_1, x_2)$ are:

$$p_{\mathbf{w}}(y = 1|\mathbf{x}) = \phi(w_1x_1 + w_2x_2) = \frac{1}{1 + e^{-(w_1x_1+w_2x_2)}}$$
$$p_{\mathbf{w}}(y = -1|\mathbf{x}) = \phi(-w_1x_1 - w_2x_2) = \frac{1}{1 + e^{(w_1x_1+w_2x_2)}}$$

Recall that $\phi(s) = \frac{1}{1+e^{-s}}$ is the sigmoid function.

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.

total/2

5 marks

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

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

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.

a. L_2: **yes** **no**

b. L_3: **yes** **no**

c. L_4: **yes** **no**

3 marks

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

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

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):

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

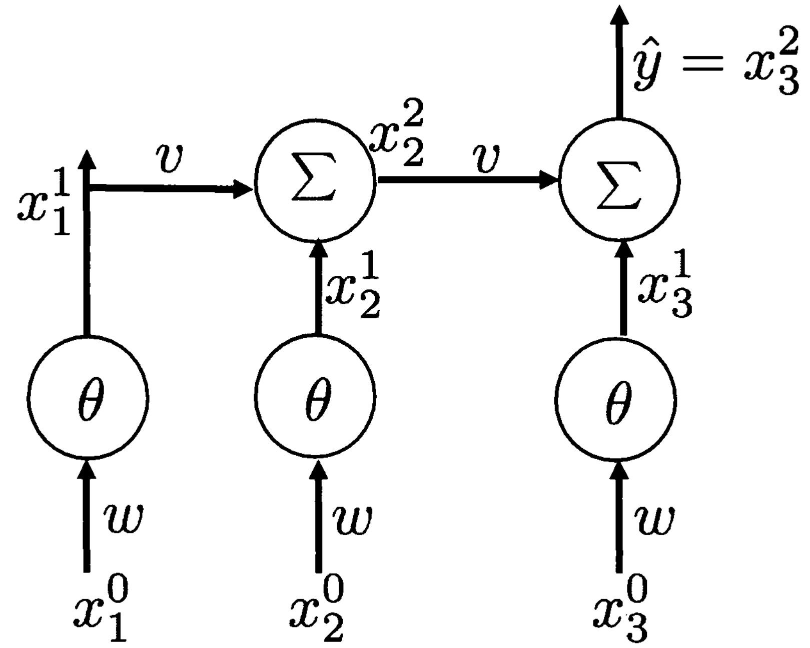
c. Both w_1 and w_2 are small and $w_1/w_2 = 1$.

d. Neither of the above.

Justify your answer. No calculations are needed.

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:

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

Note that v and w are the shared weights on the respective 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 $\Delta = \hat{y} - y$.

total/5

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).

total/5

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

total/5

5 marks

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

total/5

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 $\frac{de}{dw}$.

total/5

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 $\frac{de}{dw}$.

total/5

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 Classification. 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?

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

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?

8 marks **(d)** What is the expected leave one out cross validation error for 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., $x \sim \mathcal{N}(x; 0, \sigma^2)$, then we have that: $\Pr[|x| \leq \sigma] = 2/3$.)

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.)

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.

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 positive 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.

total/6