

# CS M146 Midterm

Atibhav Mittal

TOTAL POINTS

**97 / 100**

QUESTION 1

Short Questions 40 pts

1.1 True/False 21 / 21

✓ - 0 pts all correct

- 3 pts a incorrect
- 3 pts b incorrect
- 3 pts c incorrect
- 3 pts d incorrect
- 3 pts e incorrect
- 3 pts g incorrect
- 3 pts h incorrect

1.2 Model Evaluation 9 / 9

✓ - 0 pts Correct

- 3 pts One incorrect answer
- 6 pts Two incorrect answers
- 1 pts One incorrect explanation
- 2 pts Two incorrect explanations
- 0 pts [Click here to replace this description.](#)

1.3 Decision Boundaries 10 / 10

✓ - 0 pts Correct

- 5 pts one incorrect answer
- 10 pts Two incorrect answers
- 1 pts No mark which region is positive/negative
- 2 pts 1 minor mistake

QUESTION 2

Perceptron 20 pts

2.1 algorithm 12 / 12

- 12 pts 4 wrong answers
- 9 pts 3 wrong answers
- 6 pts 2 wrong answers
- 3 pts 1 wrong answer

✓ - 0 pts Correct

2.2 seperability 2 / 4

- 0 pts Correct (answer no)

✓ - 2 pts Answer true and show the data is linearly seperable

- 4 pts incorrect answer

2.3 data augmentation 4 / 4

✓ - 0 pts Correct: either describe how to extend  $w$  and  $x$ ; or provide a correct 3-d weight vector.

- 2 pts minor error: either forget to describe how to extend either  $w$  or  $x$ ; or provide an incorrect 3-d weight vector with some explanation

- 4 pts Incorrect answer: provide 2-d weight vector; or provide an incorrect 3-d weight vector with no explanation

QUESTION 3

Decision Tree 18 pts

3.1 H(Passed) 4 / 4

✓ - 0 pts Correct

- 2 pts minor mistake
- 4 pts incorrect
- 1 pts forget negative sign in the entropy
- 0 pts Correct formulations, but Incorrect calculation

3.2 G(passed, GPA) 4 / 4

✓ - 0 pts Correct

- 2 pts minor mistake
- 4 pts incorrect answer
- 1 pts tiny mistake
- 0 pts Correct formulations, but Incorrect calculation

3.3 G(passed, study) 3 / 4

- 0 pts Correct
- 2 pts minor mistake
- 4 pts incorrect

✓ - 1 pts tiny mistake

- 0 pts Correct formulation, but Incorrect calculation

☞ Mistake in H(entropy) calculation

### 3.4 Tree 6 / 6

✓ - 0 pts Correct

- 6 pts incorrect

- 2 pts split when labels are pure

- 2 pts Split on attribute with lower information gain

- 4 pts Only one split

### QUESTION 4

## Linear Regression 20 pts

### 4.1 application 6 / 6

✓ - 0 pts Correct

- 1 pts minor mistake

- 2 pts description is unclear

- 6 pts incorrect

### 4.2 optimization algorithm 6 / 6

✓ - 0 pts Correct (GD,SGD, or analytic solution)

- 2 pts missing or incorrect gradient

- 6 pts incorrect

- 4 pts no attempt at gradient

### 4.3 global optimality 8 / 8

✓ - 0 pts Correct

- 4 pts Incomplete answer, with arguments and derivation attempt

- 1 pts Almost correct (with a missing step)

- 6 pts Argues PSD or squared function

- 8 pts Incorrect

### QUESTION 5

## 5 Name & ID 2 / 2

✓ - 0 pts Correct

- 2 pts No name and ID

## Midterm

Feb. 13<sup>th</sup>, 2018

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **four** problems.
- You have 90 minutes to earn a total of 100 points.
- Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.

Good Luck!

Name and ID: (2 Point)

ATIBHAV MITTAL  
804598 987

Name		/2
Short Questions		/40
Perceptron		/20
Decision Tree		/18
Regression		/20
<b>Total</b>		<b>/100</b>

Short Questions [40 points]

1. [21 points] True/False Questions (Add 1 sentence to justify your answer if the answer is "False".)

(a) When the hypothesis space is richer, over-fitting is more likely.

True

(b) Nearest neighbors is more efficient at training time than logistic regression.

True

(c) Perception algorithms can always stop after seeing  $\gamma^2/R^2$  number of examples if the data is linearly separable, where  $\gamma$  is the size of the margin and  $R$  is the size of the largest instance.

False, it can stop after seeing  $\frac{R^2}{\gamma^2}$  number of examples

(d) Instead of maximizing a likelihood function, we can minimize the corresponding negative log-likelihood function.

True

(e) If data is not linearly separable, decision tree can not reach training error zero.

False, XOR is not linearly separable but a decision tree can reach zero error on it.

(g) If data is not linearly separable, logistic regression can not reach training error zero.

True

(h) To predict the probability of an event, one would prefer a linear regression model trained with squared error to a classifier trained with logistic regression.

False, Logistic Regression is preferred because the range of the logistic regression is  $[0,1]$ , which is what the range of probability laws is. Linear Regression on the other hand has range all real numbers, which is not ideal for probability events.

2. [9 points] You are a reviewer for the International Conference on Machine Learning, and you read papers with the following claims. Would you accept or reject each paper? Provide a one sentence justification if your answer is "reject".

- accept/reject] "My model is better than yours. Look at the training error rates!"

Reject because training error could be extremely low due to overfitting.

- accept/reject "My model is better than yours. After tuning the parameters on the test set, my model achieves lower test error rates!"

Reject because test set is used to see how well a model generalizes and should not be looked at for training/tuning hyper parameters

- accept/reject "My model is better than yours. After tuning the parameters using 5-fold cross validation, my model achieves lower test error rates!"

Accept

3. [10 points] On the 2D dataset of Fig. 1, draw the decision boundaries learned by logistic regression and 1-NN (using two features  $x$  and  $y$ ). Be sure to mark which regions are labeled positive or negative, and assume that ties are broken arbitrarily.

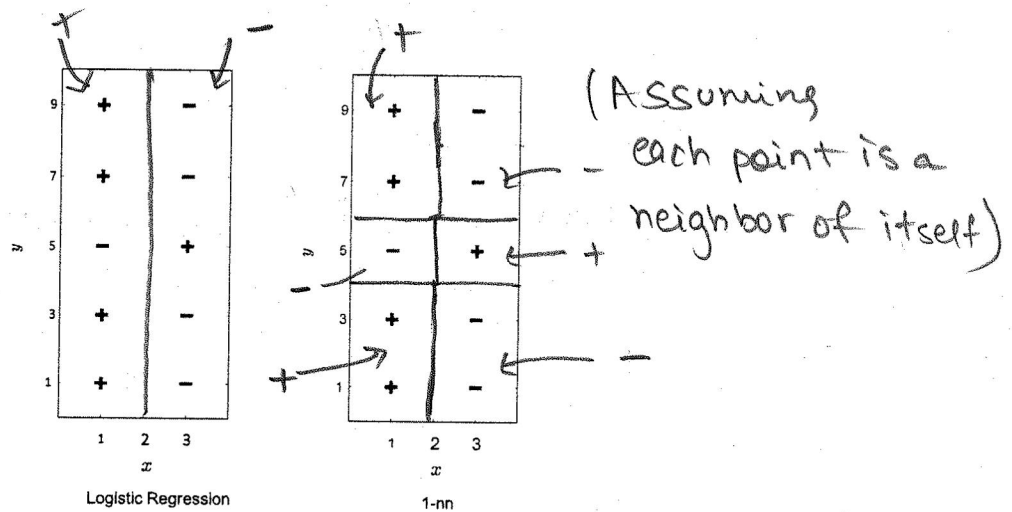


Figure 1: Example 2D dataset for question

Perceptron [20 points]

Recall that the Perceptron algorithm makes an updates when the model makes a mistake. Assume now our model makes prediction using the following formulation:

$$y = \begin{cases} 1 & \text{if } w^T x \geq 1, \\ -1 & \text{if } w^T x < 1. \end{cases} \quad (1)$$

1. [12 points] Finish the following Perceptron algorithm by choosing from the following options.

- (a)  $w^T x_i \geq 0$     (b)  $y_i = 1$     (c)  $w^T x \geq 0$  and  $y_i = 1$     (d)  $w^T x \geq 0$  and  $y_i = -1$   
 (e)  $w^T x_i < 0$     (f)  $y_i = -1$     (g)  $w^T x < 0$  and  $y_i = 1$     (h)  $w^T x < 0$  and  $y_i = -1$   
 (i)  $x_i$     (j)  $-x_i$     (k)  $w + x_i$     (l)  $w - x_i$   
 (m)  $y_i(w + x_i)$     (n)  $-y_i(w + x_i)$     (o)  $w^T x_i$     (p)  $-w^T x_i$

Given a training set  $D = \{x_i, y_i\}_{i=1}^m$

Initialize  $w \leftarrow 0$ .

For  $(x_i, y_i) \in D$ :

if (d)  $w^T x \geq 0$  and  $y_i = -1$

$w \leftarrow$  (l)  $w - x_i$

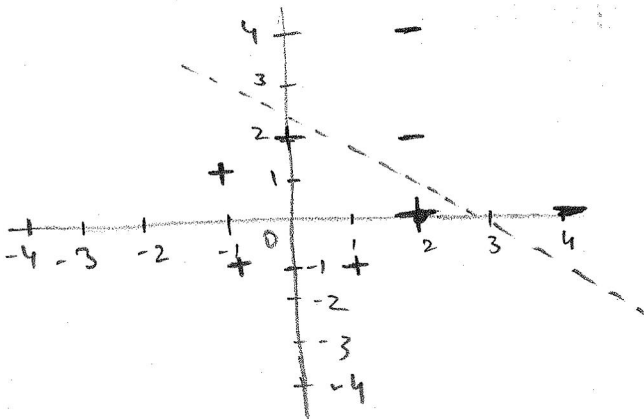
if (g)  $w^T x < 0$  and  $y_i = 1$

$w \leftarrow$  (k)  $w + x_i$

Return  $w$

2. [4 points] Let  $w$  to be a two dimensional vector. Given the following dataset, can the function described in (1) separate the dataset?

Instance	1	2	3	4	5	6	7	8
Label $y$	+1	-1	+1	+1	+1	-1	-1	+1
Data $(x_1, x_2)$	(2, 0)	(2, 4)	(-1, 1)	(1, -1)	(-1, -1)	(4, 0)	(2, 2)	(0, 2)



Yes, the data is linearly separable, so the function in (1) can separate the dataset.

(Assuming that one of the features in  $x_{(i)}$  is constant and is 1)

Instance	1	2	3	4	5	6	7	8
Label $y$	+1	-1	+1	+1	+1	-1	-1	+1
Data $(x_1, x_2)$	(2, 0)	(2, 4)	(-1, 1)	(1, -1)	(-1, -1)	(4, 0)	(2, 2)	(0, 2)

3. [4 points] If your answer to the previous question is "no", please describe how to extend  $w$  and data points  $x$  into 3-dimensional vectors, such that the data can be separable. If your answer to the previous question is "yes", write down the  $w$  that can separate the data.

$x$ : intercept  $\sim 3$

$y$ : intercept  $\sim 2.5$

$$\frac{x_1}{3} + \frac{x_2}{2.5} = 1$$

if  $\frac{x_1}{3} + \frac{x_2}{2.5} < 1$ , predict  $y = +1$

$\frac{x_1}{3} + \frac{x_2}{2.5} \geq 1$ , predict  $y = -1$

i.e.,  $-\frac{x_1}{3} - \frac{x_2}{2.5} + 2 \geq 1$ , predict  $y = +1$

$-\frac{x_1}{3} - \frac{x_2}{2.5} + 2 < 1$ , predict  $y = -1$

Hence,  $w^T = \left[ -\frac{1}{3} \quad -\frac{1}{2.5} \quad 2 \right]$

(Assuming that there is a bias term in the features)

### Decision Tree [18 points]

We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

For this problem, you can write your answers using  $\log_2$ , but it may be helpful to note that  $\log_2 3 \approx 1.6$  and entropy  $H(S) = -\sum_{v=1}^K P(S=v) \log_2 P(S=v)$ . The information gain of an attribute  $A$  is  $G(S, A) = H(S) - \sum_{v \in \text{Value}(A)} \frac{|S_v|}{|S|} H(S_v)$ , where  $S_v$  is the subset of  $S$  for which  $A$  has value  $v$ .

1. [4 points] What is the entropy  $H(\text{Passed})$ ?

$$H(\text{Passed}) = -\frac{4}{6} \log\left(\frac{4}{6}\right) - \frac{2}{6} \log\left(\frac{2}{6}\right) = -\frac{2}{3}(1-1.6) + \frac{1}{3}(1.6) = \frac{1.6}{3} + 0.4 = \frac{2.8}{3}$$

2. [4 points] What is the entropy  $G(\text{Passed}, \text{GPA})$ ?

$$H(\text{GPA}=\text{low}) = -\frac{1}{2} \log\left(\frac{1}{2}\right) - \frac{1}{2} \log\left(\frac{1}{2}\right) = 1, \quad H(\text{GPA}=\text{M}) = 1, \quad H(\text{GPA}=\text{H}) = 1 \log(1) + 0 \log(0) = 0$$

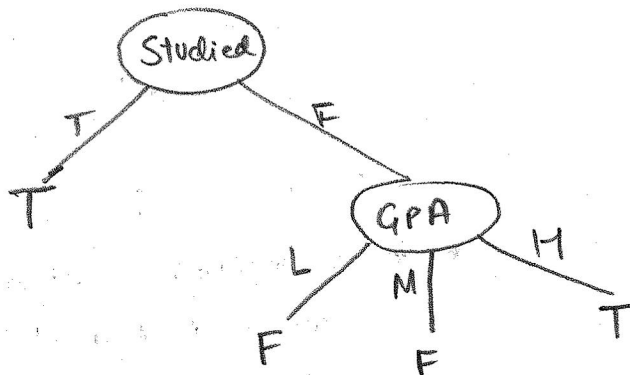
$$\text{Conditional Entropy (GPA)} = \frac{1}{3}(1) + \frac{1}{3}(1) + \frac{1}{3}(0) = \frac{2}{3} \approx 0.66 \quad | \quad G(\text{Passed}, \text{GPA}) = \frac{2.8}{3} - \frac{2}{3} = \frac{0.8}{3} \approx 0.27$$

3. [4 points] What is the entropy  $G(\text{Passed}, \text{Studied})$ ?

$$H(\text{Studied}=\text{T}) = -\frac{3}{3} \log\left(\frac{3}{3}\right) - 0 \log(0) = 0 \quad H(\text{Studied}=\text{F}) = -\frac{1}{3} \log\left(\frac{1}{3}\right) - \frac{2}{3} \log\left(\frac{2}{3}\right) = \frac{1.6}{3} + \frac{0.4}{3} = \frac{2}{3}$$

$$\text{Conditional Ent. (Studied)} = \frac{1}{2}(0) + \frac{1}{2}\left(\frac{2}{3}\right) = \frac{1}{3} \quad | \quad G(\text{Passed}, \text{Studied}) = \frac{2.8}{3} - \frac{1}{3} = \frac{1.8}{3} = 0.6$$

4. [6 points] Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.





## Linear Regression [20 points]

- [6 points] Describe one application of linear regression. Please define clearly what are your input, output, and features.

One application of linear regression could be predicting a person's height based on their parent's heights and their gender.

Features = { Father's height, Mother's Height, Gender }

Input = Dataset with above features and labelled children's heights.

Output = Person's height based on above features.

- [6 points] Given a dataset  $\{(x^{(i)}, y^{(i)})\}_{i=1}^M$  in a two dimensional space. The objective function of linear regression with square loss is

$$J(w_1, w_2) = \frac{1}{2} \sum_{i=1}^M (y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)}))^2, \quad (2)$$

where  $w_1$  and  $w_2$  are feature weight to be learned. Write down one optimization procedure that can learn  $w_1$  and  $w_2$  from data. Please be as explicit as possible.

We can use Gradient Descent to learn  $w_1, w_2$  from data.

For Gradient Descent, we first need the gradient of  $J$  w.r.t.  $w_1, w_2$

$$\frac{\partial J}{\partial w_1} = \sum_{i=1}^M -(y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)})) x_1^{(i)} \quad \frac{\partial J}{\partial w_2} = \sum_{i=1}^M -(y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)})) x_2^{(i)}$$

The Algorithm for learning  $w_1, w_2$  is:

① Initialize  $w_1, w_2 = 0$ ,  $\alpha = \text{StepSize (Hyperparameter)}$

② for each  $((x_1^{(i)}, x_2^{(i)}), y_i) \in D$

$$\bullet w_1 = w_1 - \alpha \frac{\partial J}{\partial w_1}$$

$$\bullet w_2 = w_2 - \alpha \frac{\partial J}{\partial w_2}$$

(Perform updates simultaneously)

③ Return  $w_1, w_2$

Note: the for loop over the set  $D$ , can be repeated multiple times, eg.  $T$  times, in which case  $T$  also becomes a hyperparameter.

3. [8 points] Prove that Eq. (2) has a global optimal solution. (Full points if the proof is mathematically correct. 4 points if you can describe the procedure for proving the claim.)

Use Hessian Matrix to check if  $z^T H z \geq 0 \quad \forall z \in \mathbb{R}^2$

$$H = \begin{bmatrix} \frac{\partial^2 J}{\partial w_1^2} & \frac{\partial^2 J}{\partial w_1 \partial w_2} \\ \frac{\partial^2 J}{\partial w_1 \partial w_2} & \frac{\partial^2 J}{\partial w_2^2} \end{bmatrix} \quad \frac{\partial^2 J}{\partial w_1^2} = \sum_{i=1}^M (-x_1^{(i)})(-x_1^{(i)}) = \sum_{i=1}^M (x_1^{(i)})^2$$

$$\frac{\partial^2 J}{\partial w_2^2} = \sum_{i=1}^M (x_2^{(i)})^2$$

$$\frac{\partial^2 J}{\partial w_1 \partial w_2} = \sum_{i=1}^M x_2^{(i)} x_1^{(i)}$$

$$\frac{\partial^2 J}{\partial w_2 \partial w_1} = \sum_{i=1}^M x_1^{(i)} x_2^{(i)}$$

$$H = \begin{bmatrix} \sum_{i=1}^M (x_1^{(i)})^2 & \sum_{i=1}^M x_1^{(i)} x_2^{(i)} \\ \sum_{i=1}^M x_1^{(i)} x_2^{(i)} & \sum_{i=1}^M (x_2^{(i)})^2 \end{bmatrix} = \begin{bmatrix} k^2 & l \\ l & n^2 \end{bmatrix}$$

$$\left( \begin{array}{l} \text{for } k^2 = \sum_{i=1}^M (x_1^{(i)})^2 \\ l = m = \sum_{i=1}^M x_1^{(i)} x_2^{(i)} \\ n^2 = \sum_{i=1}^M (x_2^{(i)})^2 \end{array} \right)$$

can write  $k^2, n^2$ , because they're both always positive

Consider  $z = \begin{bmatrix} a \\ b \end{bmatrix}$ ,  $z^T = \begin{bmatrix} a & b \end{bmatrix}$

$$z^T H z = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} k^2 & l \\ l & n^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} ak^2 + bl \\ al + bn^2 \end{bmatrix}$$

$$= a^2 k^2 + abl + bal + b^2 n^2 = a^2 k^2 + 2ab l + b^2 n^2$$

$$= \sum_{i=1}^M a^2 (x_1^{(i)})^2 + \sum_{i=1}^M 2ab x_1^{(i)} x_2^{(i)} + \sum_{i=1}^M b^2 (x_2^{(i)})^2$$

$$= \sum_{i=1}^M (ax_1^{(i)} + bx_2^{(i)})^2$$

Since it is a sum of squares, it is always  $\geq 0$ .

$\Rightarrow$  Hessian is positive semi-definite.

$\Rightarrow J(w_1, w_2)$  is convex and thus has a global optimum.