CS M146 Midterm

Michael Donald Wu

TOTAL POINTS

90 / 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 6 / 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 4 / 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) 4 / 4

✓ **- 0 pts Correct**

- 2 pts minor mistake

- 4 pts incorrect

- 1 pts tiny mistake

- **0 pts** Correct formulation, but Incorrect 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 4 / 8

- **0 pts** Correct

✓ **- 4 pts Incomplete answer, with arguements 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

CM146: Introduction to Machine Learning	Winter 2018
Midterm	
Feb. 13^{th}, 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) Michael Wu 404751542

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

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, no need for fitting incrementally

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

   false, $t \leq \gamma^2/R^2$ where t is number of mistakes, not examples

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

   True

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

   False, a xor b is not lin. sep. but decision tree can have 0 training error

   (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 models probability whereas linear regression tries to achieve lowest error from line

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!" training error minimization is not our goal, we want test error minimization
   - **accept/reject** "My model is better than yours. After tuning the parameters on the test set, my model achieves lower test error rates!" This train/test set may be a fluke that is not repeatable, we need to train/test multiple times
   - **accept/reject** "My model is better than yours. After tuning the parameters using 5-fold cross validation, my model achieves lower test error rates!"

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.

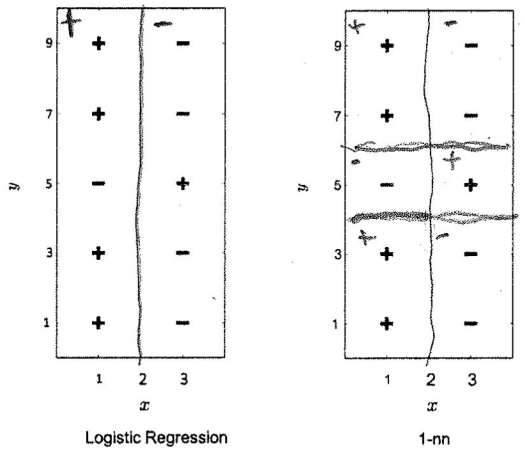


Figure 1: Example 2D dataset for question

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:

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

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)

   $w \leftarrow$ (k)

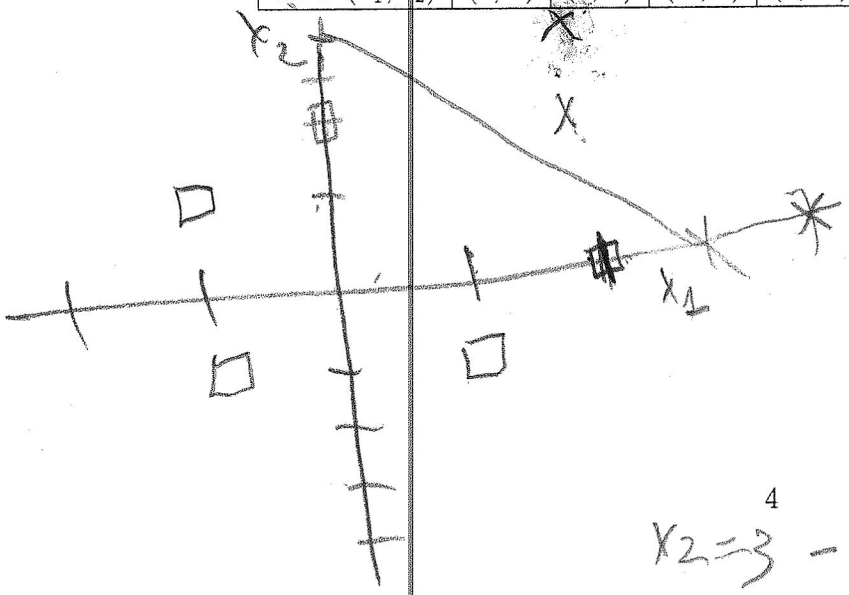
   if (g)

   $w \leftarrow$ (l)

   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)



$y \equiv$ { 1 if $-x_1 - x_2 <$
-1 if $-x_1 - x_2 <$

$x_2 = 3 - x_1$

NO perceptron must have weight threshold

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.

extend w to have bias term

$w = (w_1, w_2, b)$

extend data to $(x_1, x_2, 1)$

Then if $w = (-1, -1, 4)$

$$y = \begin{cases} 1 & \text{if } w^T x_i \geq 1 \\ -1 & \text{if } w^T x_i < -1 \end{cases}$$

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 $G(S, A) = H(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} H(S_v)$, where S_v is the subset of S for which A has value v.

1. [4 points] What is the entropy H(Passed)?

$$-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$\frac{1.6}{3} + \frac{1.2}{3} \qquad \frac{2.8}{3}$$

2. [4 points] What is the entropy G(Passed, GPA)?

$$H(\text{Passed}) - \frac{1}{3}\left(H(\text{Passed} \mid \text{GPA}=\text{L})\right) + H(\text{Passed} \mid \text{GPA}=\text{M}) + H(\text{Passed} \mid \text{GPA}=\text{H})$$

$$= \left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] - \frac{2}{3}$$

$$= \frac{0.8}{3}$$

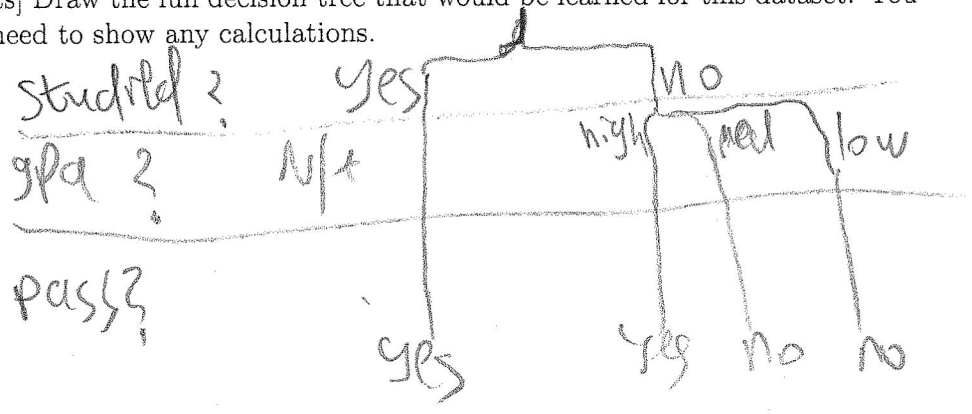
3. [4 points] What is the entropy G(Passed, Studied)?

$$= \left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] - \frac{1}{2}\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)$$

$$= -\frac{1}{6}\log_2\left(\frac{1}{3}\right) - \frac{2}{6}\log_2\left(\frac{2}{3}\right)$$

$$= \frac{1.4}{3}$$

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]

1. [6 points] Describe one application of linear regression. Please define clearly what are your input, output, and features.

to predict height of thrown ball

X input = work done when throwing ball straight up

y output = max height of ball

(1, x) feature = amount of work done, and drag term

2. [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

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

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

gradient $J = \left\langle \frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2} \right\rangle = \left\langle \frac{1}{2}\sum_{i=1}^{m} 2(y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)}))(-x_1^{(i)}), \frac{1}{2}\sum_{i=1}^{m} 2(y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)}))(-x_2^{(i)}) \right\rangle$

$= -\left\langle \sum(x_1^{(i)} y_i - w_1 x_1^{(i)2} - w_2 x_2^{(i)} x_1^{(i)}), \sum(x_2^{(i)} y_i - w_1 x_1^{(i)} x_2^{(i)} - w_2 x_2^{(i)2}) \right\rangle$

η = step-size

repeat

$\langle w_1, w_2 \rangle_{(k+1)} \leftarrow \langle w_1, w_2 \rangle_k + \eta \left\langle \sum_{i=1}^{M}(x_1^{(i)} y_i - w_1^{(k)} x_1^{(i)2} - w_2^{(k)} x_2^{(i)} x_1^{(i)}), \sum_{i=1}^{M}(x_2^{(i)} y_i - w_1^{(k)} x_1^{(i)} x_2^{(i)} - w_2^{(k)} x_2^{(i)2}) \right\rangle$

until grad $J = 0$

return $\langle w_1, w_2 \rangle_{final}$

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.)

basically use the fact that each term is squared to prove that it must be ≥ 0, so a global minimum exists.

let $J_i = (y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)}))^2$ $\quad J_i \geq 0$, continuous

then $J = \frac{1}{2}(J_1 + J_2 + \ldots + J_n) \geq 0$, also continuous

~~Polynomial~~

because J is a continuous, lower bounded function defined for all $(w_1, w_2) \in \mathbb{R}^2$, a global minimum must exist.