CS M146 Midterm

Jingyue Shen

TOTAL POINTS

95 / 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 **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 **0 / 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**

You provide a 4-d weight vector!

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 **7 / 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

Need to prove for all z

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) Jingyue Shen 704797256

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.

   ?

   (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

   Perceptron will make at most $t \le \frac{R^2}{\gamma^2}$ errors so it can stop after making $\frac{R^2}{\gamma^2}$ mistakes

   (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.

   For example, the XOR function is not linear separable, but decision tree can learn it with zero 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 the conditional probability of an event, while linear regression does not.

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. lower training error rates does not necessarily indicate low test error rate. The model might overfit the training data.

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

  Reject. Test set should not be used to tune parameters. We should use a dev set to tune them instead, or it would be cheating.

- **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.

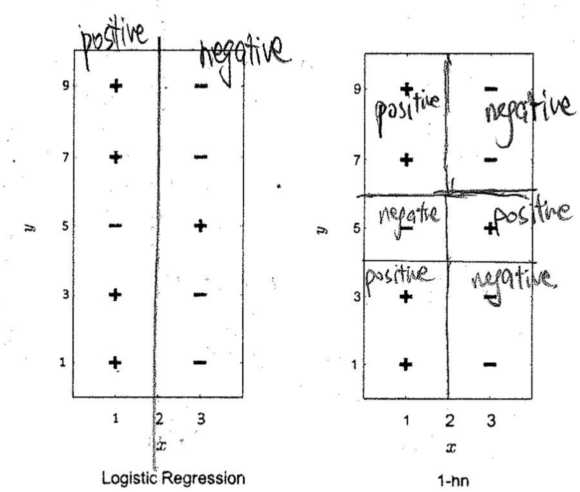


Figure 1: Example 2D dataset for question

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:

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

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 ___(g)___

$w \leftarrow$ ___(m/k)___

if ___(d)___

$w \leftarrow$ ___(l)___

Return w

y
$w = w + x_i$

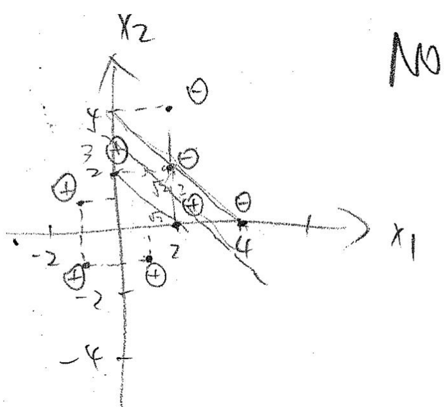
$w + y_i x_i$

$w + y_i x_i$ $-y_i$ $-y_i x_i$

$w \leftarrow w - x_i$

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)



No. Since the margin of data set is $\frac{\sqrt{2}}{2} < 1$ ∴ cannot separate the dataset

$x_2 = -x_1 + 3$

$x_1 + x_2 - 3 = 0$ $= 1$

$11 - 3$

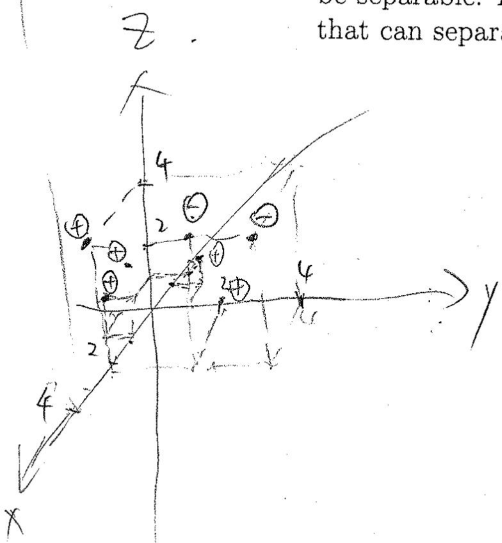
$2 + 2 - 3 = 1$

(2,0,4) (2,4,4) (-1,1,1) (1,-1,1) (-1,-1,1) (4,0,16) (2,2,4) (0,2,0)

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)

4 20 2 2 2 16 8 4

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.



make the thrid feature x_3 to be $x_1^2+x_2^2$.

and extend $w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ b \end{pmatrix}$, where b correspondly to the bias term

So now each data point looks like $(x_1, x_2, x_3, 1)$

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

1

1

0

$-\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}$

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)?

   $H(passed) = -\frac{4}{6}\times\log_2\frac{4}{6} - \frac{2}{6}\times\log_2\frac{2}{6} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

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

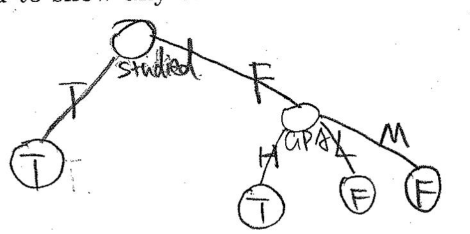
   $G(passed, GPA) = H(passed) - (\frac{1}{3}\times 1 + \frac{1}{3}\times 1 + \frac{1}{3}\times 0) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}$

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

   $G(passed, Studied) = H(passed) - [\frac{1}{2}\times 0 + \frac{1}{2}\times(-\frac{1}{3}\times\log_2\frac{1}{3} - \frac{2}{3}\times\log_2\frac{2}{3})]$

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

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.

Predict house price with respect to room size, the number of rooms

input: each entry in training set has room size in m^2, number of rooms, the house price as input. the input for prediction are room size and the number of rooms

features: room size, number of rooms

output: the predicted house price.

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:

1. set $\tilde{w}_0 = (0, 0)$

Repeat:

2. calculate the gradient $\nabla J(w_{1t}, w_{2t})$ for round t. $= \left(-\sum_{i=1}^{M}[y_i - (w_{1t}x_1^{(i)} + w_{2t}x_2^{(i)})]x_1^{(i)}, -\sum_{i=1}^{M}[y_i - (w_{1t}x_1^{(i)} + w_{2t}x_2^{(i)})]x_2^{(i)}\right)$

3. update w_t:

$$w_{t+1} = \hat{w}_t - \alpha \nabla J(w_{1t}, w_{2t})$$, where α is learning rate

until convergence or less-than some threshold.

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.)

We need to prove that $J(w_1, w_2)$ is convex.

$$\because \frac{\partial J(w_1,w_2)}{\partial w_1^2} = \frac{\partial -\sum_{i=1}^{M}[y_i - (w_1 x_1^{(i)} + w_2 x_2^{(i)})]\, x_1^{(i)}}{\partial w_1} = -\sum x_1^{(i)}(-x_1^{(i)}) = \sum [x_1^{(i)}]^2.$$

$$\frac{\partial J(w_1,w_2)}{\partial w_2^2} = -\sum_{i=1}^{M} x_2^{(i)}(-x_2^{(i)}) = \sum_{i=1}^{M}[x_2^{(i)}]^2.$$

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

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

$\therefore$ Hessian
$$H = \begin{bmatrix} \frac{\partial J(w_1,w_2)}{\partial w_1^2} & \frac{\partial J(w_1,w_2)}{\partial w_2 w_1} \\ \frac{\partial J(w_1,w_2)}{\partial w_1 w_2} & \frac{\partial J(w_1,w_2)}{\partial w_2^2} \end{bmatrix} = \begin{bmatrix} \sum [x_1^{(i)}]^2 & \sum x_1^{(i)} x_2^{(i)} \\ \sum x_1^{(i)} x_2^{(i)} & \sum [x_2^{(i)}]^2 \end{bmatrix}$$

$\because$ ~~if~~ if H is semi-definite, then ~~it is~~ the function J is Convex

take $z = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$$z^t H z = \left(\sum x_1^{(i)} x_2^{(i)} \quad \sum [x_2^{(i)}]^2\right)\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sum [x_2^{(i)}]^2 \geq 0$$

$\therefore$ it is semi-definite

$\therefore$ J is convex and has a global optimal solution