

CS M146 Midterm

Royson Lin

TOTAL POINTS

61 / 100

QUESTION 1

1 True/false 15 / 18

- 0 pts Correct

✓ - 3 pts (a) incorrect (e.g., saying $p(x)$ is probability)

- 3 pts (b) incorrect

- 3 pts (c) incorrect

- 3 pts (d) incorrect

- 3 pts (e) incorrect

- 3 pts (f) incorrect

- 2 pts (a) partial points for showing how to use integral to get probability from $p(x)$ and arguing $0 \leq \int p(x) dx \leq 1$ (but we are asking $p(x)$, not $\int p(x) dx$)

QUESTION 2

Short Question 23 pts

2.1 (a)-(d) 11 / 13

- 0 pts Correct

- 4 pts (a) incorrect

- 3 pts (b) incorrect

- 3 pts (c) incorrect

- 3 pts (d) incorrect

✓ - 2 pts (a) partially correct

- 1.5 pts (b) partially correct

- 1.5 pts (c) partially correct

- 1.5 pts (d) partially correct

- 0 pts (b) should specify tuning "hyper-parameter"

2.2 (e) 0 / 10

- 0 pts Correct

- 1 pts Answer correct but missed one/two steps while proving

- 2 pts Some minor mistakes/missed a important step

- 5 pts Major mistakes, but mentioned some important points like solving a linear system Xw . E.g.,

trying to solve $Xw = 0$ instead of $Xw = y$ or mention X is invertible

- 8 pts only mentioned definition of linear independence

✓ - 10 pts incorrect

QUESTION 3

Decision tree 15 pts

3.1 (a) i, ii 7 / 7

✓ - 0 pts Correct

- 2 pts a) i. incorrect

- 0.5 pts a) i. partially incorrect

- 5 pts a) ii. incorrect

- 2.5 pts a) ii. partially incorrect

3.2 (a) iii 0 / 3

- 0 pts Correct

- 1.5 pts a) iii. Partially incorrect

✓ - 3 pts a) iii) incorrect

3.3 (b) 5 / 5

✓ - 0 pts Correct

- 2.5 pts partially incorrect

- 5 pts incorrect

QUESTION 4

Perceptron 23 pts

4.1 (a) (answer 2,4,5,6; 4,5,6; 2,4,6; 4,6; are all okay) 4 / 4

- 4 pts Totally wrong

- 2 pts Partially Correct

✓ - 0 pts Correct

4.2 (b) 0 / 8

- 4 pts did mention η or mention learning rate, but got totally wrong with the constraint of the learning rate

- 0 pts correct

- **2 pts** made tiny mistakes on the constraint of the learning rate
- ✓ - **8 pts** did not mention yx or learning rate (yx is the basic and necessary component when updating the weights)

4.3 (c),(d) 5 / 6

- **3 pts** c is wrong
- **3 pts** d is wrong
- **6 pts** both c and d are wrong
- **0 pts** all correct
- ✓ - **1 pts** c is partially correct: mention "adding dimension" without specific solutions or with wrong solutions
- **1 pts** d is partially correct: A. wrong $w_0w_1w_2$
B. neglect the question "only solution"

4.4 (e) 0 / 5

- ✓ - **2 pts** partially correct, e.g. draw a correct diagram
- **0 pts** correct
- **5 pts** wrong
- **3 Point adjustment**

QUESTION 5

19 pts

5.1 (a) 2 / 3

- **0 pts** Correct
 - **1 pts** No Y prediction
 - ✓ - **1 pts** Incorrect Prediction
 - **1.5 pts** Wrong calculation & not finished; no Y prediction
 - **1.5 pts** Incomplete & wrong calculation
 - **0.5 pts** Wrong calculation
 - **0.5 pts** No Y prediction after calculating probabilities
 - **1.5 pts** Wrong calculation & wrong prediction
 - **1 pts** Wrong formula is used
 - **0 pts** Slight mistake in calculation
 - **1.5 pts** Not finished; no Y prediction
 - **1 pts** Your calculation is wrong & how you get Y?
- See solution
- **0.5 pts** You need to show how you get Y

- **1 pts** Wrong calculation & prediction is wrong
- **3 pts** No answer
- **2 pts** Unfinished

5.2 (b) 0 / 6

- **0 pts** Correct
- **2 pts** But you need to prove it.
- **1 pts** You need to show that the other form of this classifier is $w^T x = 0$
- ✓ - **6 pts** Wrong answer
- **0.5 pts** See the solution in CCLE
- **1 pts** See the solution in CCLE
- **2 pts** Your proof is not correct
- **3 pts** Wrong perception ; see the solution on CCLE
- **2 pts** I did not understand what have you written. Assuming you have written 'linear classifier' I have graded. You need to prove it. Please the the solution on CCLE

5.3 (c) 10 / 10

- ✓ - **0 pts** Correct
- **0 pts** You forgot to mention the sum
- **2 pts** Please see the solution on CCLE
- **10 pts** No answer
- **5 pts** Unfinished
- **8 pts** Wrong answer
- **0 pts** Slight mistake
- **2 pts** How??
- **9 pts** No answer
- **8 pts** Not finished
- **0 pts** Mistake
- **3 pts** Please see the solution on CCLE

QUESTION 6

6 name 2 / 2

- ✓ - **0 pts** Correct

Midterm

Nov. 5th, 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 Five 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) Royson Lin 104790253

Name		/2
True/False Questions		/18
Short Questions		/23
Decision Tree		/15
Perceptron		/23
Regression		/19
Total		/100

1 True/False Questions (Add a 1 sentence justification.) [18 pts]

- (a) (3 pts) For a continuous random variable x and its probability density function $p(x)$, it holds that $0 \leq p(x) \leq 1$ for all x .

True, no x can have any probability outside 0 and 1.

- (b) (3 pts) K-NN is a linear classification model.

False. Not all the boundaries drawn by K-NN are linear, but it is a binary classification model.

- (c) (3 pts) Logistic regression is a probabilistic model and we use the maximum likelihood principle to learn the model parameters.

True. We use probabilistic models when data are not linearly separable.

- (d) (3 pts) Suppose you are given a dataset with 990 cancer-free images and 10 images from cancer patients. If you train a classifier which achieves 98% accuracy on this dataset, it is a reasonably good classifier.

False. Not necessarily because it could be overfitted to the training data, thus doing badly on the test data.

- (e) (3 pts) A classifier that attains 100% accuracy on the training set is always better than a classifier that attains 70% accuracy on the training set.

False; it might even be worse since it is more likely to be overfitted and do worse on test data!

- (f) (3 pts) A decision tree is learned by minimizing information gain.

False. It is learned by maximizing information gain, picking the feature at each node that does so.

2 Short Questions [23 pts]

- (a) (4 pts) What is the main difference between gradient descent and stochastic gradient descent (in one sentence)? Which one requires more iterations to converge, why?

Stochastic gradient is slower. SGD requires more iterations to converge.

- (b) (3 pts) What is the motivation to have a development set?

A development or dev set allows us to fine-tune our hyperparameters, to see which setting of the hyperparameters allows our learning model to perform the best.

- (c) (3 pts) Describe the differences between linear regression and logistic regression (in less than two sentences).

Logistic regression can be used for both binary classification and for regression, whereas linear regression can be used only for regression.

- (d) (3 pts) Consider the models that we have discussed in lecture: decision trees, k -NN, logistic regression, Perceptrons. If you are required to train a model that predicts the probability that the patient has cancer, which of these would you prefer, and why?

Logistic regression because data for cancer is probably not linearly separable and there are too many factors for decision trees and KNN.

- (e) (10 pts) Given n linearly independent feature vectors in n dimensions, show that for any assignment to the binary labels you can always construct a linear classifier with weight vector w which separates the points. Assume that the classifier has the form $\text{sign}(w \cdot x)$. Hint: a set of vectors are linearly independent if no vector in the set can be defined as a linear combination of the others.

Since they are linearly independent, there is no way to "get from" one vector to another. So linearly separable.

3 Decision Trees [15 pts]

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 .

- (a) We will use the dataset below to learn a decision tree which predicts the output Y , given by the binary values of A, B, C .

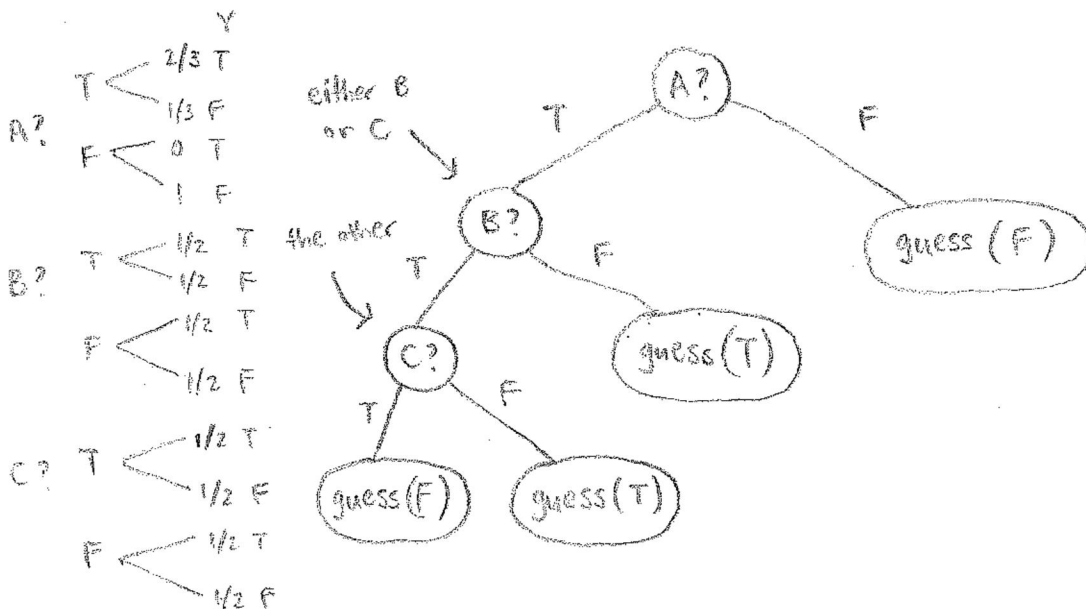
A	B	C	Y
F	F	F	F
T	F	T	T
T	T	F	T
T	T	T	F

- i. (2 pts) Calculate the entropy of the label y .

$$q = \frac{2}{4}$$

$$\begin{aligned}
 H(y) &= - \left(\frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4} \right) \\
 &= - \log_2 \frac{1}{2} = -(-1) = \boxed{1}
 \end{aligned}$$

- ii. (5 pts) Draw the decision tree that will be learned using the ID3 algorithm that achieves zero training error.



- iii. (3 pts) Is this tree optimal (i.e. does it get minimal training error with minimal depth?) explain in two sentences, and if it isn't optimal draw the optimal tree.

$$G(S, A) = H(S) - \sum_{\text{values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

- (b) (5 pts) You have a dataset of 400 positive examples and 400 negative examples. Now suppose you have two possible splits. One split results in (300+, 100-) and (100+, 300-). The other choice results in (200+, 400-), and (200, 0). Which split is most preferable and why?

To figure out which split is preferable, we calculate the information gain of each. Let S = the dataset.

$$H(S) = - \left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2} \right) = - \log \frac{1}{2} = 1$$

Split 1, information gain:

$$\begin{aligned} G(S, A) &= 1 - \left(\frac{400}{800} \left(- \left(\frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4} \right) \right) + \frac{400}{800} \left(- \left(\frac{1}{4} \log \frac{1}{4} + \frac{3}{4} \log \frac{3}{4} \right) \right) \right) \\ &= 1 - \left(- \left(\frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4} \right) \right) = 1 + \frac{3}{4} \log \frac{3}{4} + \frac{1}{4} \log \frac{1}{4} \\ &= 1 + \frac{3}{4} (\log 3 - \log 4) + \frac{1}{4} \log 2^{-2} = 1 + \frac{3}{4} (1.6 - 2) + \frac{1}{4} (-2) \\ &= 1 + \frac{3}{4} (-0.4) - \frac{1}{2} = 1 - \frac{1.2}{4} - \frac{1}{2} = 1 - 0.3 - 0.5 = 0.2 \end{aligned}$$

Split 2, information gain:

$$\begin{aligned} G(S, A) &= 1 - \left(\frac{600}{800} \left(- \left(\frac{2}{6} \log \frac{2}{6} + \frac{4}{6} \log \frac{4}{6} \right) \right) + \frac{200}{800} \left(- \frac{200}{200} \log \frac{200}{200} \right) \right) \\ &= 1 + \frac{3}{4} \left(\frac{1}{3} \log \frac{1}{3} + \frac{2}{3} \log \frac{2}{3} \right) = 1 + \frac{3}{4} \left(\frac{1}{3} (\log 1 - \log 3) + \frac{2}{3} (\log 2 - \log 3) \right) \\ &= 1 + \frac{3}{4} \left(\frac{1}{3} (-1.6) + \frac{2}{3} (1 - 1.6) \right) = 1 + \frac{3}{4} \left(\frac{-1.6}{3} + \frac{-1.2}{3} \right) = 1 + \frac{3}{4} \left(\frac{-2.8}{3} \right) \\ &= 1 - 0.7 = 0.3 \end{aligned}$$

6

The second split is preferable because it has a higher information gain.

4 Perceptron Algorithm [23 pts]

(a) (4 pts) Assume that you are given training data $(x, y) \in \mathbb{R}^2 \times \{\pm 1\}$ in the following order:

Instance	1	2	3	4	5	6	7	8
Label y	+1	-1	+1	-1	+1	-1	+1	+1
Data (x_1, x_2)	(10, 10)	(0, 0)	(8, 4)	(3, 3)	(4, 8)	(0.5, 0.5)	(4, 3)	(2, 5)

We run the Perceptron algorithm on all the samples once, starting with an initial set of weights $w = (1, 1)$ and bias $b = 0$. On which examples, the model makes an update?

Assume the algorithm runs in order of instance #.

$$w = (1, 1)$$

$$1. \quad +1(10+10) = 20 > 0$$

$$2. \quad -1(0+0) = 0 \text{ (we will not update on 0)}$$

$$3. \quad +1(8+4) = 12 > 0$$

$$4. \quad -1(3+3) = -6 < 0$$

$$\text{Update: } w \leftarrow w + \gamma x = (1, 1) + (-3, -3) = (-2, -2)$$

$$b \leftarrow b + \gamma = -1$$

$$5. \quad +1(-2+2) = 0 < 0$$

$$\text{Update: } w \leftarrow w + \gamma x = (-2, -2) + (4, 8) = (2, 6)$$

$$6. \quad -1(2 \cdot 0.5 + 5 \cdot 0.5) = -4 < 0$$

$$\text{Update: } w \leftarrow w + \gamma x = (2, 6) + (-0.5, -0.5) = (1.5, 5.5)$$

$$b \leftarrow b - 1 = -1$$

$$7. \quad +1(4 \cdot 1.5 + 3 \cdot 1.5) = 9.5 > 0$$

$$8. \quad +1(1.5 \cdot 2 + 5.5 \cdot 5 - 1) = 29.5 > 0$$

The model updates on examples 4, 5, and 6.

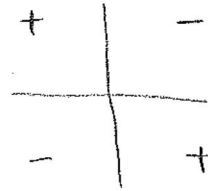
(b) (8 pts) Suggest a variation of the Perceptron update rule which has the following property: If the algorithm sees two consecutive occurrences of the same example, it will never make a mistake on the second occurrence. (Hint: determine an appropriate learning rate that guarantees this property). Prove your answer is correct.

The update rule is :

$$w \leftarrow w + \frac{\gamma}{2} x$$

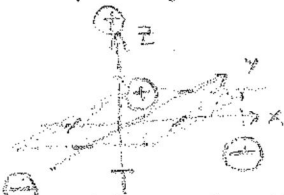
- (c) (3 pts) Linear separability is a pre-requisite for the Perceptron algorithm. In practice, data is almost always inseparable, such as XOR.

x_1	x_2	y
-1	-1	-1
-1	+1	+1
+1	-1	+1
+1	+1	-1



Provide a solution to convert the inseparable data to be linearly separable. The XOR can be used for the illustration.

One solution could be to add another dimension, then shift the positive examples up one unit in that dimension & the negatives down one unit.



Then the data can be separated by the plane $z=0$ (for example).

- (d) (3 pts) Design (specify w_0, w_1, w_2 for) a two-input Perceptron (with an additional bias or offset term) that computes "OR" Boolean functions. Is your answer the only solution?

$$\begin{aligned} w_1 &= 1 \\ w_2 &= 1 \\ (w_0 &= 0) \\ b &= 1 \end{aligned}$$

x_1	x_2	y
-1	-1	-1
1	-1	1
1	1	1
-1	1	1

$x_1 + x_2$
-2
0
2
0

$$\text{sgn}(\vec{w} \cdot \mathbf{x}^{(1)} + b) = \text{sgn}(-1 - 1 + 1) = \text{sgn}(-1) = -1$$

$$\text{sgn}(\vec{w} \cdot \mathbf{x}^{(2)} + b) = \text{sgn}(1 - 1 + 1) = \text{sgn}(1) = 1$$

$$\text{sgn}(\vec{w} \cdot \mathbf{x}^{(3)} + b) = \text{sgn}(1 + 1 + 1) = \text{sgn}(3) = 1$$

$$\text{sgn}(\vec{w} \cdot \mathbf{x}^{(4)} + b) = \text{sgn}(-1 + 1 + 1) = \text{sgn}(1) = 1$$

- (e) (5 pts) What is the maximal margin γ in the above OR dataset.



The maximal margin γ is the greatest distance between a positive and negative example in the dataset.

In the OR dataset it would be the distance between

$$(-1, -1) \text{ and } (1, 1) \text{ which is } \sqrt{(1+1)^2 + (1+1)^2} = \sqrt{8} = 2\sqrt{2}.$$

5 Logistic Regression [19 pts]

Considering the following model of logistic regression for a binary classification, with a sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$:

$$P(Y = 1|X, w_0, w_1, w_2) = \sigma(w_0 + w_1X_1 + w_2X_2)$$

- (a) (3 pts) Suppose we have learned that for the logistic regression model, $(w_0, w_1, w_2) = (-\ln(4), \ln(2), -\ln(3))$. What will be the prediction ($y = 1$ or $y = -1$) for the given $x = (1, 2)$?

$$\begin{aligned} P(Y = 1 | X, w_0, w_1, w_2) &= \sigma(w_0 + w_1X_1 + w_2X_2) \\ &= \sigma(-\ln(4) + (\ln(2))(1) + (-\ln(3))(2)) \\ &= \sigma(-2\ln(2) + \ln(2) - 2\ln(3)) = \sigma(-\ln(2) - 2\ln(3)) < 0.5 \end{aligned}$$

So the prediction will be $y = -1$.

- (b) (6 pts) Is logistic regression a linear or non-linear classifier? Prove your answer.

(c) (10 pts) In the homework, we mention an alternative formulation of learning a logistic regression model when $y \in \{1, 0\}$

$$\arg \min_w \sum_{i=1}^m y_i \log \sigma(w^T x_i) + (1 - y_i) \log(1 - \sigma(w^T x_i))$$

. Derive its gradient.

Each component of the gradient will be of the following form with i replaced with actual numbers from 1 to m .

$$\frac{\partial}{\partial w_i} \arg \min_w \sum_{i=1}^m y_i \log \sigma(w_i x_i) + (1 - y_i) \log(1 - \sigma(w_i x_i))$$

$$= \quad " \quad " \quad y_i \frac{1}{\sigma(w_i x_i) (1 - \sigma(w_i x_i))} (1 - \sigma(w_i x_i)) + (1 - y_i) \frac{1}{1 - \sigma(w_i x_i)} (-\sigma(w_i x_i) (1 - \sigma(w_i x_i)))$$

$$= \quad " \quad " \quad y_i (1 - \sigma(w_i x_i)) + (1 - y_i) \sigma(w_i x_i)$$

$$= y_i - y_i \sigma(w_i x_i) + \sigma(w_i x_i) - y_i \sigma(w_i x_i)$$

$$= y_i + \sigma(w_i x_i) - 2y_i \sigma(w_i x_i)$$

