CS M146 Midterm

MARK GUEVARA

TOTAL POINTS

80 / 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 integral p(x) \leq 1 (but we are asking p(x), not integral p(x))

QUESTION 2

Short Question 23 pts

2.1 (a)-(d) 13 / 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) 5 / 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 3 / 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) 2 / 4

- **- 4 pts** Totally wrong
- ✓ **- 2 pts Partially Correct**
- **- 0 pts** Correct

4.2 (b) **4 / 8**

✓ **- 4 pts did mention yx 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) **6 / 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 w0w1w2 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**

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) **6 / 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^Tx= 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

- 5 pts Not correct.

QUESTION 6

6 name **2 / 2**

✓ - 0 pts Correct

CM146: Introduction to Machine Learning	Fall 2018

Midterm

Nov. 5^{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 **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) Mark Guevara 704962920

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 $0 \le p(x) \le 1$ for all x.

True; a probability always lies in this range

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

False; k-NN is non-linear

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

True; however, most models will use the minimum negative log likelihood in place of maximum likelihood as it is more efficient

(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; with 98% accuracy, around 20 patients will be misclassified as having cancer when they do not.

(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; a classifier with 100% accuracy mave have over fit the training set and could perform poorly on test data

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

False; it is learned by maximizing information gain

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 require more iterations to converge, why?

Gradient descent calculates the gradient of the entire set when determining a step, while stochastic estimates it by using a single data point. Stochastic is slower to converge as a result because its steps are not always in the correct direction.

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

Many models have hyperparameters that need to be chosen. Having a dev set allows for those parameters to be chosen independent of the data set to prevent overfitting.

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

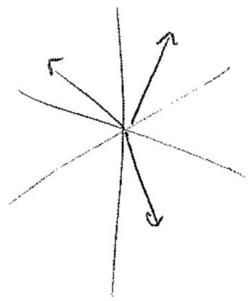
Logistic regression involves finding the probability of y_i having a certain value given x_i, while linear regression is a method for finding an approximate linear model for a data set using a measurement like least-mean-squares.

(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 would be best because it is the best at modeling probabilistic models. The other models are better for categorization (although a decision tree could be used to an extent)

(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 $sign(w \cdot x)$. Hint: a set of vectors are linearly independent if no vector in the set can be defined as a linear combination of the others.

Select two points with opposite classification. The + can be used to define a hyperplane that separates it from −. Add another + (or by symmetry −) and create a hyperplane using points of the same as a basis. Because the points are linearly independent, the − point does not lie on this hyperplane. Repeat, adding each + to one hyperplane and each − to another. All of the points on the + hyperplane cannot define the − points, so this hyperplane behaves as a linear classifier.



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 $G(S, A) = H(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} H(S_v)$, 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

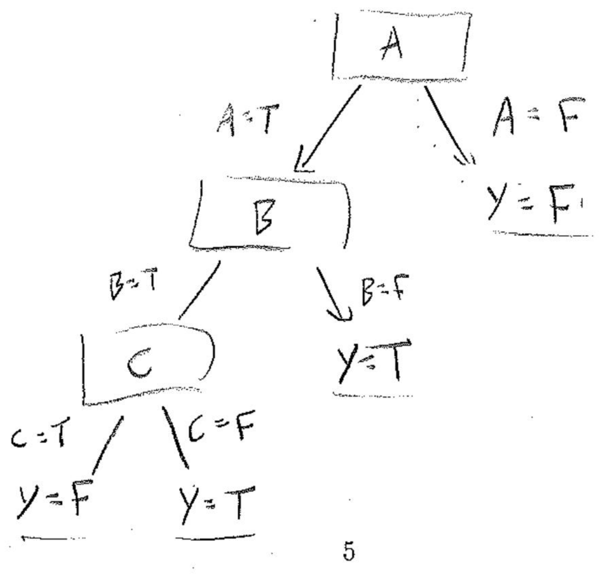
i. **(2 pts)** Calculate the entropy of the label y.

$$H(Y) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}$$

$$= -\log\frac{1}{2} \quad = 1$$

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

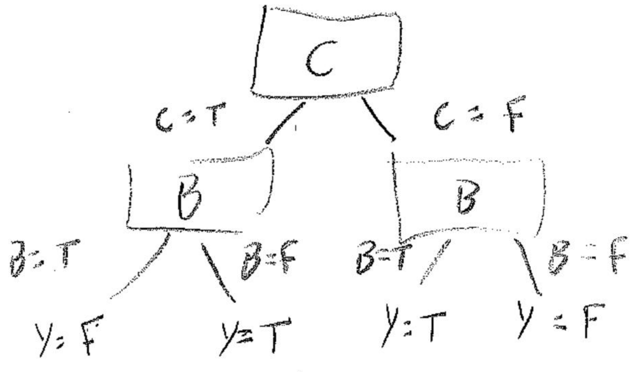
A reduces entropy the most



★ note: B and C have the same information gain, so could be swapped on the tree

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.

No; there is a solution that doesn't use the ID3 algorithm's method but only has a depth of 2 instead of 3. Both have 0 training error.



(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?

Start entropy $H(S) = 1$

Gain of split 1:

$$1 - \frac{400}{800}\left(-\frac{300}{400}\log\frac{300}{400} - \frac{100}{400}\log\frac{100}{400}\right) - \frac{400}{800}\left(\frac{100}{400}\log\frac{100}{400} - \frac{300}{400}\log\frac{300}{400}\right.$$

$$= 1 - \left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right) \approx 1 - \left(-\frac{3}{4}(-0.4) - \frac{1}{4}(-2)\right) \approx 1 - 0.3 - 0.5$$

$$\approx 0.2$$

Gain of split 2:

$$1 - \frac{600}{800}\left(-\frac{200}{600}\log\frac{200}{600} - \frac{400}{600}\log\frac{400}{600}\right) - \frac{200}{800}(0)$$

$$= 1 - \frac{3}{4}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right) \approx 1 - \frac{3}{4}\left(-\frac{1}{3}(-1.6) - \frac{2}{3}(-0.6)\right)$$

$$\approx 1 - \frac{3}{4}(0.53 + 0.4) \qquad 1 - \frac{3}{4}(0.93)$$

$$\approx 0.3 \qquad \approx 0.7$$

The second split results in a larger information gain so it is preferred

4 Perceptron Algorithm [23 pts]

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

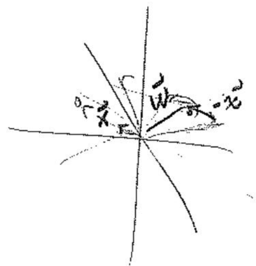
We run the Perceptron algorithm on all the samples once, starting with an initial set of weights $w = (1, 1)$ and bias $b = 0$. On which examples, the model makes an update?

1. $w = (1,1) \Rightarrow y w^T x = 1(10+10) \geq 0$ ✓ no update
2. " $= -1(0+0) \geq 0$ ✓ no update
3. " $1(8+4) \geq 0$ ✓ " "
4. " $-1(3+3) \ngeq 0$ ✗ $\Rightarrow$ update $w = w + yx$ $(1,1) - (3,3$
5. $w = (-2,-2) \Rightarrow 1(-8,-16) \ngeq 0$ ✗ $\Rightarrow$ update   " " $(-2,-2) + (4,8$
6. $w = (2,6) \Rightarrow -1(1, 3) \ngeq 0$ ✗ $\Rightarrow$ update $(2,6) - (0.5, 0.5$
7. $w = (1.5, 5.5)$ $1(+, +) \geq 0$ ✓ no update
8. $w =$ " " $\Rightarrow 1(+, +) \geq 0$ ✓ no update

$\boxed{\text{Updates on } 5, 6, 7}$

(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 :



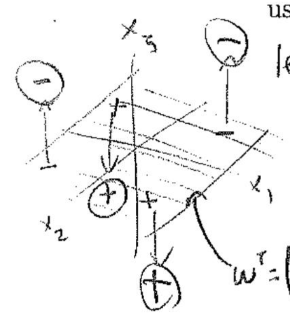
$w \leftarrow w +$ $\|w - x\| y x$?

Let x, y be such that $y(w^T x) < 0$.
Therefore (x, y) create an incorrect prediction and the model is updated s.t. $w \leftarrow w + \|w - x\| y x$.
$\|w - x\|$ is always greater than $\|w\|$, as x is opposite the direction of w, so adding $\|w - x\| y x$ will always (?) place x in the right classification side.

(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.



let $x_3 = x_1 x_2$

The data set of (x_1, x_2, x_3) is then linearly separable using $y = -x_3$ or $w^T = (0,0,-1)$

x_1	x_2	x_3
-1	-1	1
-1	1	-1
1	-1	-1
1	1	1

(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?

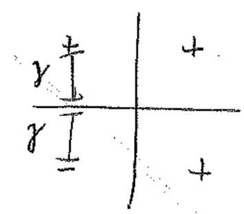
x_1	x_2	y
-1	-1	-1
1	-1	1
1	1	1
-1	1	1

$w^T = (1, 1, 1)$

It is not the only solution, ex any $0 < w_3 < 2$ will work with $w_0 = 1$ $w_2 = 1$

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

The maximum margin is 1, assuming a euclidian distance metric



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 $x = (1, 2)$?

$$P = \sigma(-\ln 4 + (1)\ln 2 - (2)\ln 3)$$
$$= \sigma\left(\ln\left(\tfrac{1}{4}\right) + \ln(2) + \ln\left(\tfrac{1}{9}\right)\right)$$
$$= \sigma\left(\ln\left(\tfrac{1}{18}\right)\right)$$
$$= \frac{1}{1+e^{-\frac{1}{18}}} = \frac{1}{1+\frac{1}{e^{\frac{1}{18}}}} = \frac{1}{1+18} = \boxed{\frac{1}{19}}$$

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

Logistic regression is a linear classifier because its algorithm seeks to find a w vector that can divide the data set. The value of P scales with $w^T x$, but fundamentally the data could be categorized into + and − if, say, $P \geq 0.5$ and $P < 0.5$ respectively, making it a linear model.

(c) **(10 pts)** In the hoemwork, 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.

$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial w_1} \\ \vdots \\ \frac{\partial J}{\partial w_n} \end{bmatrix} \qquad w_1 x_1 + \dots + w_n x_n$$

Where $$\frac{\partial J}{\partial w_j} = \sum_{i=1}^{m} y_i \frac{1}{\cancel{\sigma(w^T x_i)}} \left(\cancel{\sigma(w^T x_i)}\right)\left(1 - \sigma(w^T x_i)\right) x_j$$

$$+ (1 - y_i) \frac{1}{\cancel{1 - \sigma(w^T x_i)}} (-1) \left(\sigma(w^T x_i)\right)\left(\cancel{1 - \sigma(w^T x_i)}\right) x_j$$

$$= \sum_{i=1}^{m} y_i x_j \left(1 - \cancel{\sigma(w^T x_i)}\right) - (1 - y_i) x_j \left(\cancel{\sigma(w^T x_i)}\right)$$

$$= \sum_{i=1}^{m} y_i x_j - x_j \left(\sigma(w^T x_i)\right)$$

$$= \sum_{i=1}^{m} (y_i - 1) x_j \left(\sigma(w^T x_i)\right)$$