CS 6375 Machine Learning Midterm Examination

University of Texas at Dallas

10/12/2020

Name: Midterm Solutions

NetID: _____

Question	Topic	**Points**
1	Short Answers	20
2	True/False Questions	8
3	Naive Bayes	10
4	Perceptrons	10
5	SVMs	24
6	Logistic Regression	18
7	Decision Trees	10
Total		100

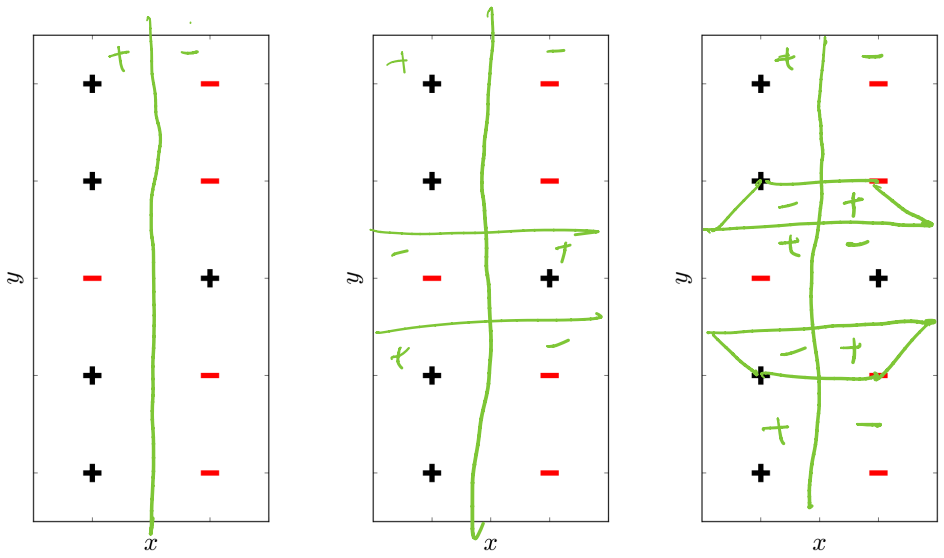
Instructions:

1. This examination contains 14 pages, including this page.
2. You have **two (2) hours** to complete the examination.
3. Either you can use this paper or separate set of sheets to fill in your answers. Write clearly so we can understand your handwriting. Please scan the answers once you are done (you can also take a photograph once you are done) and upload the scanned copy to eLearning. You can also use a tablet device (like an Ipad) to write if you prefer.
4. Please do not search online for answers to the questions. If the answers are similar to something available online, you will get zero points on this examination.
5. The mid-term examination has to be done individually by everyone. If someone copies, the entire group of students involved will get a zero.
6. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult one.
7. All the Best!!

Question 1: Short Answers

[20 pts] Please provide short and clear answers for the questions below.

(a) Given a 2 dimensional dataset below, draw the decision boundaries obtained by the following classifiers (4 points: a and b are 1 point each and c is for 2 points)



(a) Logistic regression ($\lambda = 0$) (b) 1-NN (c) 3-NN

(b) (6 points: each sub-part is 2 points each) A random variable follows an exponential distribution with parameter $\lambda : \lambda > 0$, and has the following density:

$$p(t) = \lambda e^{-\lambda t}, t \in [0, \infty] \tag{1}$$

This distribution models waiting times between events. Given a iid data: $T = (t_1, \cdots, t_n)$, where each t_i is modeled as drawn from the exponential distribution with parameter λ. Then:

- Compute the log-likelihood $p(T|\lambda)$
- Solve for λ_{MLE}
- Suppose we have a prior distribution: $p(\lambda) \propto e^{-\mu\lambda}$. Obtain λ_{MAP}. Compare λ_{MLE} and λ_{MAP} as $n \to \infty$.

Log-Likelihood

$$p(T|\lambda) = \prod_{i=1}^{n} p(t_i)$$

$$\Rightarrow \log p(T|\lambda) = \sum_{i=1}^{n} \log(\lambda e^{-\lambda t_i})$$

$$= \sum_i (\log\lambda - \lambda t_i)$$

$$= n\log\lambda - \lambda \sum_{i=1}^{n} t_i$$

$$\frac{d \log p(T|\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

$$\Rightarrow \lambda_{MLE} = \frac{n}{\sum_{i=1}^{n} t_i}$$

MAP Estimn

$$p(\lambda) \propto e^{-\mu\lambda}$$

$$\Rightarrow \log p(\lambda|T) = -\mu\lambda + \sum_{i=1}^{n} \log\lambda e^{-\lambda t_i}$$

$$= n\log\lambda - \lambda \sum_{i=1}^{n} t_i - \mu\lambda$$

$$= 0$$

$$\Rightarrow \lambda_{MAP} = \frac{n}{\sum_{i=1}^{n} t_i + \mu}$$

as $n \to \infty$

$\lambda_{MAP} \to \lambda_{MLE}$

(c) (6 points) Decision Trees: Using the dataset below, we want to build a decision tree which classifies Y as T or F given the binary variables A, B, C.

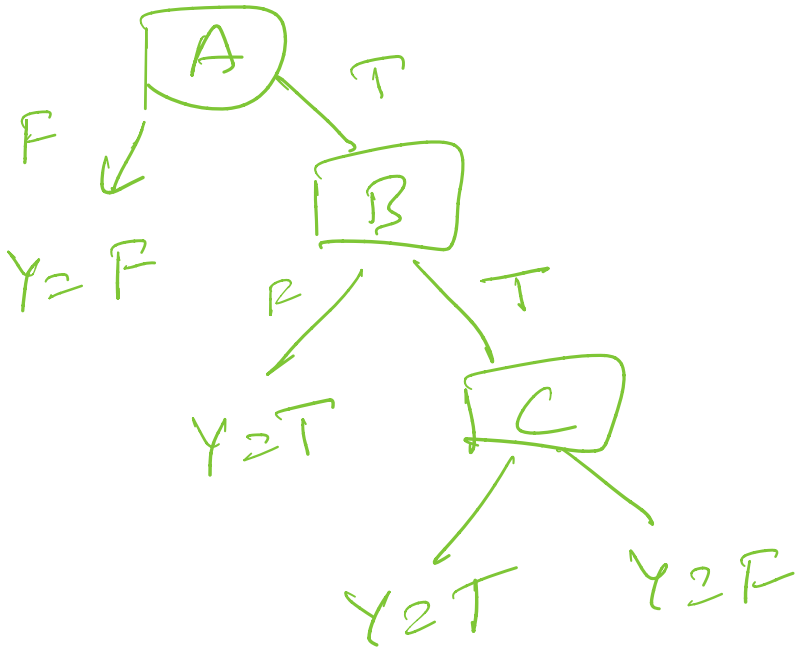
(Part 1) Draw the tree that would be learned by the greedy algorithm with zero training error. You do not need to show any computation.

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

(Part 2) Is this tree optimal (i.e. does it get zero training error with minimal depth)? Explain in less

than two sentences. If it is not optimal, draw the optimal tree as well.

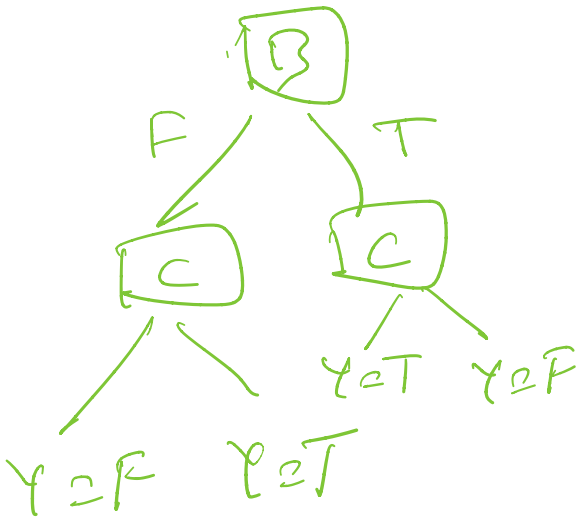
Part 1



Part II

Note: $Y = B \text{ xor } C$

∴ Depth 2 tree



(d) (4 points) Compute the dual of the following problem. Minimize $x^2 + 1$ such that $(x-4)(x-12) \leq 10$.

Cons: $x^2 - 16x + 38 \leq 0$

$L(x, \lambda) = x^2 + 1 + \lambda(x^2 - 16x + 38)$

$= x^2(\lambda + 1) - 16\lambda x + 38\lambda + 1$

Dual: $g(\lambda) = \min_x L(x, \lambda)$

$$L(x,\lambda) = x^2(\lambda+1) - 16\lambda x + 3\beta\lambda + 1$$

$$\Rightarrow \frac{\partial L}{\partial x} = 2x(\lambda+1) - 16\lambda = 0$$

$$\Rightarrow x = \frac{8\lambda}{\lambda+1}$$

$$L(\lambda) = \frac{64\lambda^2}{\lambda+1} - \frac{128\lambda^2}{\lambda+1} + \frac{304\lambda^2}{\lambda+1} + 1$$

Dual: $\max_{\lambda \geq 0} L(\lambda)$

Question 2: True or False with Explanations

[8 pts] Each question below is for 2 points each. Please do not just write true or false. You also need to provide explanations. Just true or false will yield no points.

(a) Given a function f, the set of sub-gradients is always non-empty.

False. Set of subgrads is non-empty only for Cvx

(b) k-nearest neighbor and decision tree models can be used for regression.

True. Reg: Take avg of k preds

(c) No classifier can do better than a naive Bayes classifier if the distribution of the data is known.

False: NB is restricted b/c of cond Ind.

(d) Maximizing the likelihood of linear regression yields multiple local optimums

logistic

Min Linear-Reg loss

True: Single local optima only if LR is Strictly convex

$\Rightarrow$ Data matrix is psd.

Convex

↓

Global optima = Local optima

(Multiple Local / Global optima)

↓

Strictly Convex

↓

Unique local / global optima.

Question 3: Naive Bayes

[10 pts] Given the following training data points, provide the output of naive bayes classifier for test data points $z1$ and $z2$:

$x1 = (0, 0, 0, 1, 0, 0, 1),\ y1 = 1$
$x2 = (0, 0, 1, 1, 0, 0, 0),\ y2 = 1$
$x3 = (1, 1, 0, 0, 0, 1, 0),\ y3 = -1$
$x4 = (1, 0, 0, 0, 1, 1, 0),\ y4 = -1$
$x5 = (1, 1, 1, 1, 1, 1, 1),\ y5 = 1$
$x6 = (0, 0, 0, 0, 0, 0, 0),\ y6 = 1$
$x7 = (1, 1, 1, 1, 1, 1, 1),\ y7 = -1$
$z1 = (1, 0, 0, 0, 0, 1, 0)$
$z2 = (0, 1, 1, 0, 0, 1, 1)$

$p(y=1) = \frac{4}{7}$

$p(y=-1) = \frac{3}{7}$

$p(x=1 \mid y)$

	x_a	x_b	x_c	x_d	x_e	x_f	x_g
$y=1$	$\frac{4}{4}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{4}{4}$	$\frac{2}{4}$
$y=-1$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$p(y=1 \mid z1) = \left[\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{2}{4}\right] \cdot \frac{4}{7}$

$\frac{36}{2^{12} \cdot 7}$

$p(y=-1 \mid z1) = \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{7}$

$= \frac{2^4}{3^5 \times 7}$

$\frac{36}{2^{12}\pi} < \frac{24}{3^{5}\pi}$ 양변에 π를 곱하여 정리

$z_1 : y = -1$

For z_2, observe $p(z_2 \mid y=-1) = 0$

$\therefore$ z_2 classified as $y=1$

$(p(z_2 \mid y=1) > 0)$

$w_0 =$ zero vector
$\alpha = 1$

Question 4: Perceptrons

[10 pts] Demonstrate how the perceptron without bias (i.e. we set the parameter $b = 0$ and keep it fixed) updates its parameters given the following training sequence:

$$x1 = (0, 0, 0, 1, 0, 0, 1),\ y1 = 1$$
$$x2 = (1, 1, 0, 0, 0, 1, 0),\ y2 = -1$$
$$x3 = (0, 0, 1, 1, 0, 0, 0),\ y3 = 1$$
$$x4 = (1, 0, 0, 0, 1, 1, 0),\ y4 = -1$$
$$x5 = (1, 0, 0, 0, 0, 1, 0),\ y5 = -1$$

Start $w = (0, 0, 0, 0, 0, 0, 0)$

$y_1 w^T x_1 = 0 \leq 0 \rightarrow$ yes

$w = w + y_1 x_1 = (0, 0, 0, 1, 0, 0, 1)$

$y_2 (w^T x_2) = 0 \leq 0 \rightarrow$ yes

$w = w + y_2 x_2$

$= (-1, -1, 0, 1, 0, -1, 1)$

$y_3 w^T x_3 = 1 \leq 0 \rightarrow$ no

no update.

$y_4 w^T x_4 = 2 \leq 0$ No

No update.

$y_5 w^T x_5 = 2 \leq 0$ No

No update.

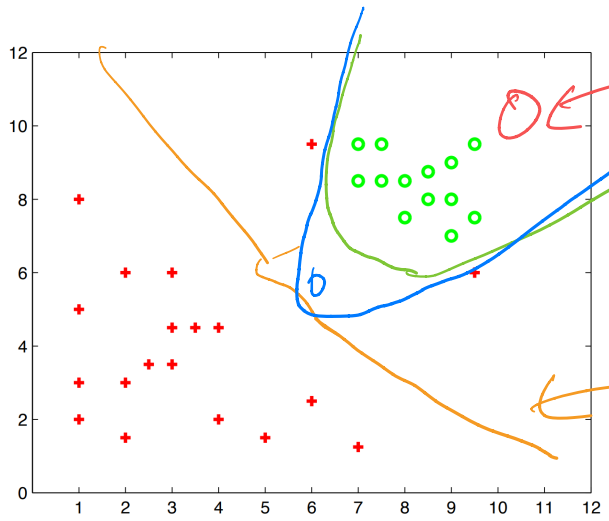
$w = (-1, -1, 0, 1, 0, -1, 1)$

$y_i (w^T x_i) \geq 0 \quad \forall i \Rightarrow$ Train error $= 0$

Question 5: SVMs

[24 pts] This question will cover support vector machines.

(a) (15 points) Given the following dataset (shown in the figure below), assume we are training the SVM with a quadratic kernel. The slack penalty C will determine the location of the separating hyper-plane. Each question below is for 3 points each.



part C

C

case 1

case 2

- Where would the decision boundary be if $C \to \infty$?

Perfectly separate Data

- For $C \approx 0$, where would the decision boundary be?

$C \approx 0$: no penalty on misclassf

$\Rightarrow$ Maximize margin

- If we know that we cannot fully trust the obtained data points, which of the scenarios would we

prefer to train the model? (Large C or Small C)

Small C since data can be wrong (high reg)

- Given C is large, draw an additional data point which *will not* change the decision boundary. Justify.

Draw a (not support vec)

- Given C is large, draw an additional data point which *will* change the decision boundary. Justify.

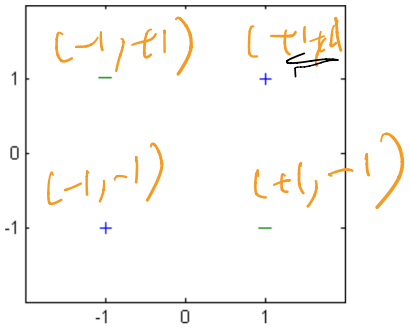
Draw a support vec: will change Dec bound

(b) (9 points) Consider the data set below. Under which of the following feature vectors is the data linearly separable? For full credit, you must justify your answer by either providing a linear separator or explaining why such a separator does not exist. Each part is 1.8 points each

- $\phi(x_1, x+2) = [x_1 + x_2, x_1 - x_2]$
- $\phi(x_1, x_2) = [x_1^2, x_2^2, x_1 x_2]$
- $\phi(x_1, x_2) = [\exp(x_1), \exp(x_2)]$
- $\phi(x_1, x_2) = [x_1 \sin x_2, x_1]$
- $\phi(x_1, x_2) = [x_1 x_2, x_1]$

$\phi(x_1, x_2)$

$x_1 x_2 = 0$



① $(x_1 + x_2, x_1 - x_2)$ ✗

will not separate

coz orig. data is not LinSep

② $(x_1^2, x_2^2, x_1 x_2)$ ✓✓

$(0, 0, 0)(x_1^2 x_2^2 x_1 x_2) = 0$

✓ Dec. boundary

③ $(\exp(x_1), \exp(x_2))$ → NO

since this is mon. funct.

$(1/e, 1/e), (\frac{1}{e}, e)$ $(e, \frac{1}{e})$ (e, e) ✗

(4) $[x_1 \sin x_2, x_1) \to$ yes

since $[x_1 \sin x_2, x_1](1, 0) = 0$

is sep hyperplane.

(5) $[x_1 x_2, x_1) \to$ yes

since $(x_1 x_2, x_1)[1, 0) = 0$

is sep hyperplane.

Question 6: Logistic Regression

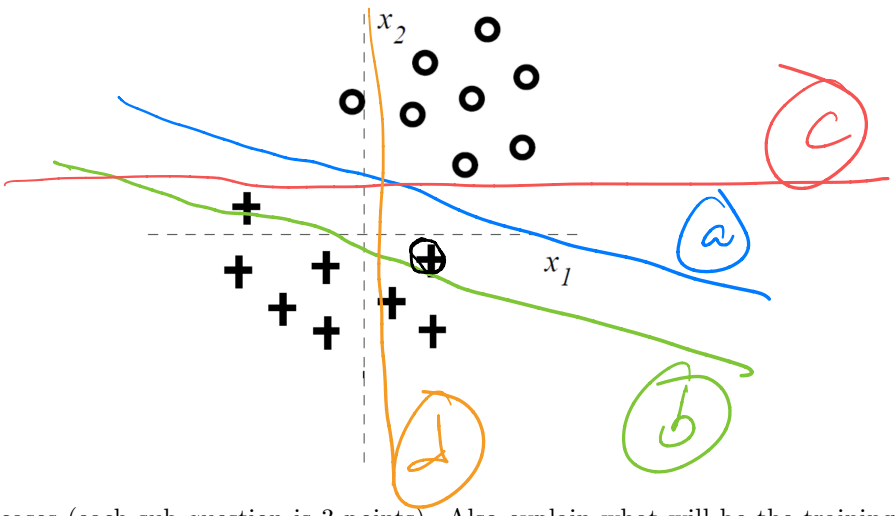
[18 pts] Given a two dimensional dataset shown in the figure below, we attempt to solve a simple binary classification using a linear logistic regression. The model is:

$$p(y = 1|x, w_0, w_1, w_2) = \frac{1}{1 + \exp(-w_0 - w_1 x_1 - w_2 x_2)} \tag{2}$$

Consider training a regularized linear logistic regression model where we try to maximize:

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C_0 w_0^2 - C_1 w_1^2 - C_2 w_2^2 \tag{3}$$

Note we have different regularization parameters for each coordinate. Draw the approximate decision bound-



aries in the following cases (each sub-question is 3 points). Also explain what will be the training error in each case.

(a) When $C_0, C_1, C_2 \approx 0$

No reg : Train error → 0

(b) When C_0 is very large and $C_1, C_2 \approx 0$

No bias term : Train error > 0

T.E = 1

(c) When C_1 is very large and $C_0, C_2 \approx 0$

T.E = 0

C_1 is large $\Rightarrow w_1 \to 0$

$\therefore$ D.B will depend less on x_1 & horizontal.

(d) When C_2 is very large and $C_0, C_1 \approx 0$

C_2 is large $\Rightarrow w_2 \to 0$

DB will be vertical

T.E ≥ 0

(e) Consider the case when C_1 and C_2 are very large and $C_0 \approx 0$. What is the value of w_0 that we expect to obtain?

$$\min_{w_0} N\left[\log\left(1+\exp(w_0)\right)+\log\left(1+\exp(-w_0)\right)\right]$$

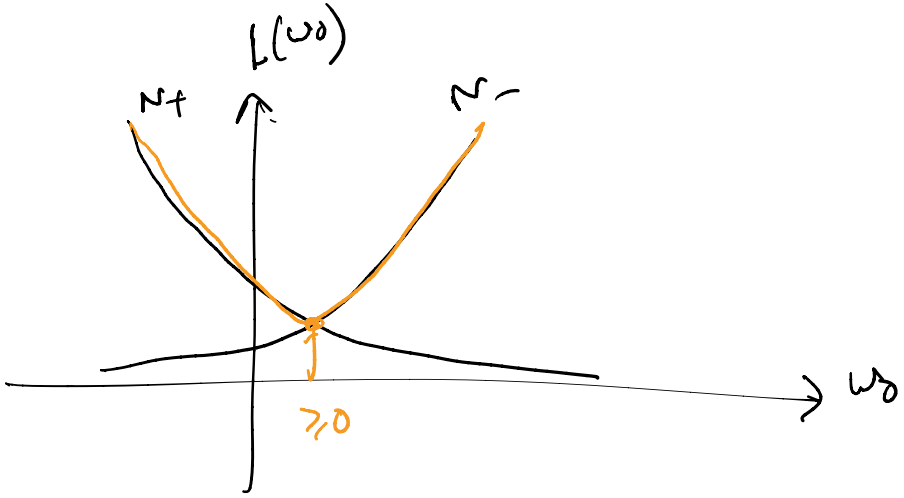
$$\Rightarrow w_0 = 0$$

(f) In the above case, assume we add a few more positives to this dataset (i.e. make it imbalanced). What is the value of w_0 we expect? (you can give a range of values of w_0 is you prefer).

$$N_+ \log\left(1+\exp(-w_0)\right) \longleftarrow \min$$

$$+ N_- \log\left(1+\exp(w_0)\right)$$

Dominating



N_+ Dominates over N_-

so $s_0|r$ will be ≥ 0

Question 7: Decision Trees

[10 pts] Consider the dataset shown below. We will use this dataset 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. For this problem, assume that $\log_2 3 \approx 1.6$ (it is ok if you leave the answers in terms of the logs as well).

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

(a) (6 points) Compute $H(\text{Passed}), H(\text{Passed} \mid \text{GPA})$ and $H(\text{Passed} \mid \text{Studied})$

$$H(\text{Passed}) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}$$

$$= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

$$= \log_2 3 - \frac{2}{3} \simeq 0.92$$

Sim. $H(\text{Passed} \mid \text{GPA}) = \frac{2}{3} \simeq 0.66$

$$H(\text{Passed} \mid \text{Studied}) = \frac{1}{2}\log_2 3 - \frac{1}{3}$$

$$\simeq 0.46$$

(b) (4 points) Draw the full decision tree that would be learned for this dataset.