THE UNIVERSITY OF TEXAS AT AUSTIN

MIS382N - BUSINESS DATA SCIENCE

FALL 2019

MIDTERM EXAM

TUESDAY, NOVEMBER 12, 2019

Name: ______________________________

Email: ______________________________

- You have 75 minutes for this exam.
- The exam is closed book and closed notes, except for two handwritten pages of notes.
- No electronic device may be used.
- Write your answers in the spaces provided.
- **Please show all of your work. Answers without appropriate justification will receive very little credit.** If you need extra space, use the back of the previous page.

Problem 1 (20 pnts): __________

Problem 2 (20 pnts): __________

Problem 3 (20 pnts): __________

Problem 4 (20 pnts): __________

Problem 5 (20 pnts): __________

Total (100 pnts) : __________

Problem 1 (20 pnts): ________________

You solve a logistic regression with two features, and you use no offset. You find

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

1. Draw the set of points that corresponds to the decision region, i.e., the set of points for which this logistic regression classifier assigns a 50% chance of being 1 and a 50% chance of being 0. Justify/explain your answer.

2. Draw the set of points for which this logistic regression classifier assigns a 4.98% chance of being 1 and a 95.02% chance of being 0. Hint: $\exp(-3) = 0.0498$. Justify/explain your answer.

Problem 2 (20 pnts): ________________

In a (obviously, much less popular) machine learning class on campus, the professor gives 2 midterms. For the second midterm, the professor collects the following information:

Table 1: Midterm 2 Data

Student	Hours studied	Hours slept	Attends OH	Score on MT#1	Score on MT#2
1	21	49	1	27	26
2	18	34	0	17	18
3	28	25	0	13	13
4	15	53	1	27	28
5	13	55	1	29	29

Suppose we model this using a Poisson model, and *we do not use an intercept*. We fit a model to predict Score on MT#2, using Hours studied, Hours slept, and Attends OH, and we find that the corresponding values of β are: $\boldsymbol{\beta} = (.01, .06, .001)$.

(a) What is the probability that student 4 gets an 29 on MT#2, under this model? Write the expression – you do not need to evaluate it.

(b) What change is associated with the expected score of a student on MT#2, who, everything else being equal, sleeps 3 additional hours? Write the expression – you do not need to evaluate it.

(c) Suppose now that you use MT#1 as an exposure variable. Would you expect your values of $\boldsymbol{\beta}$ to be smaller, larger, the same, or generally not comparable? *Justify your answer*.

(d) BONUS: Did you sleep enough this last week?

Problem 3 (15 pnts): ______________

TF and Multiple Choice: circle your answer, *and provide a brief justification*.

1. With an appropriate increase in the regularization coefficient in linear regression, it is possible to decrease the training loss, i.e., to obtain a better fit on the training data. (Never. Always. Only with Ridge Regression. Only with Lasso.)

2. If X_1 and Y are uncorrelated, then we can discard X_1 and we will never hurt training or testing error. True. False.

3. Logarithmic transformations do not change the training loss for decision trees, but they can improve the testing error. True. False.

4. If we use gradient boosting with *too small* a learning rate, we might make the training error worse. True. False.

5. If we use gradient boosting with the best possible learning rate for reducing training error, we will also always improve the testing error. True. False.

Problem 4 (20 pnts): ______________

Consider the following binary classification problem. For this problem, we want to use the exponential loss: $exp(-\hat{y}y)$, where $\hat{y}$ is given by $h(x)$ for the function $h(x)$ of our choice.

Table 2: Data

x(1)	x(2)	y
0.2	0.6	1
0.3	0.6	1
0.7	0.4	1
0.3	0.4	-1
0.6	0.6	-1
0.8	0.6	-1

(a) Suppose we fit a stump, and you split on $x(1) \geq 0.4$. Find the value of the leaves that minimizes the loss function.

(b) Call the stump above h_1. Suppose we wish to use the AdaBoost framework to boost the stump above with a function of the form: $h_2(x) = \beta_1 x(1) + \beta_2 x(2)$. Write a minimization problem for β_1 and β_2. You do not have to evaluate complicated expressions, but be as explicit as possible. Thus your answer should have the form: "minimize: $\sum_{i=1}^{6}$(expression involving β_1 and β_2)"

Problem 5 (20 pnts): ________________

Part A: For a dataset, a model predicts probabilities $\{0.3, 0.4, 0.5, 0.8, 0.9\}$ and the true corresponsing labels are $y = \{1, 0, 0, 0, 1\}$. Draw the ROC curve and compute the AUC for these predictions.

Part B: Consider this dataset

$$\begin{bmatrix} x_1 & x_2 & y \\ 1 & 1 & -1 \\ -1 & -1 & -1 \\ 1 & -1 & +1 \\ -1 & 1 & +1 \end{bmatrix}$$

(a) Plot this dataset on the plane with the labels. Consider a linear classifier: $\hat{y} = \text{sign}(\beta_1 x_1 + \beta_2 x_2 + \beta_0)$. For $\beta_1 = 1, \beta_2 = 1$ and $\beta_0 = 0.5$, draw the region of the plane that is assigned $+1$. (note that $sign(z) = +1$ if $z \geq 0$ and -1 otherwise).

(b) Is it possible that such a linear classifier can correctly classify all the examples in this dataset ?