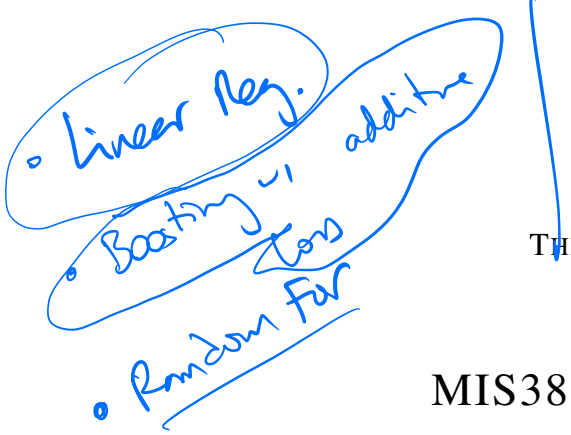


Important but not tested on midterm.

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): _____

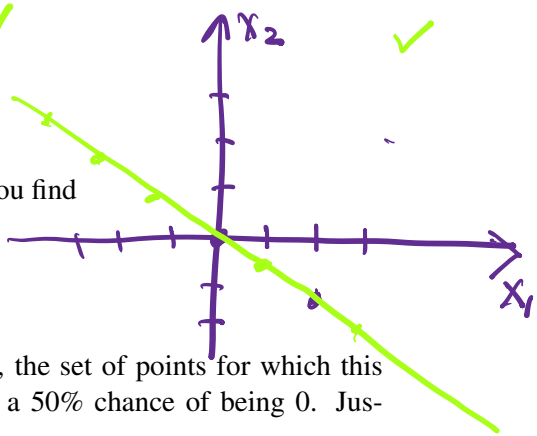
Total (100 pnts) : _____

Sigmoid (0) = 1/2

$\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$

Problem 1 (20 pnts): _____

$x_1 + 2x_2 = 0$ ✓



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

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

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.

Data: X, y. $y \in \{-1, +1\}$ Goal: Model for $P(Y=1 \mid X) \in [0,1]$

linear regression: $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ — pass through sigmoid fn: $S(z) = \frac{1}{1+\exp(-z)}$

LR model: $P(Y=1 \mid X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) = \text{Sigmoid}(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$.

model.fit(X, y) produces coefficients: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ — chosen to maximize likelihood of data we saw.

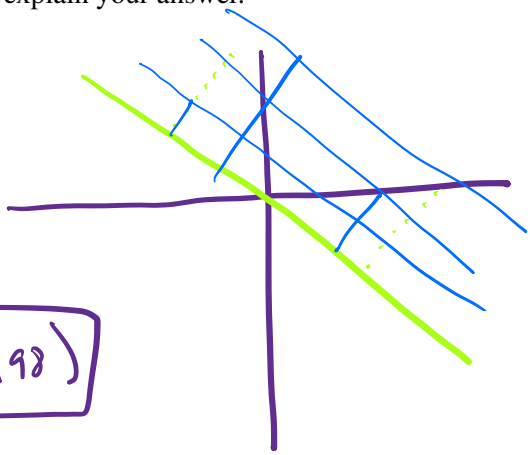
2. Draw the set of points for which this logistic regression classifier assigns a 4.98% chance of being 1 and a 95.02% chance of being 0. Hint: $\exp(-3) = 0.0498$. Justify/explain your answer.

$$.0498 = \frac{1}{1 + \exp\{-x_1 - 2x_2\}}$$

$$\exp\{-x_1 - 2x_2\} = \frac{1 - .0498}{0.0498}$$

$$+ x_1 + 2x_2 = -\left[\ln(.9502) + \ln(0.0498)\right]$$

$$x_1 + 2x_2 = \text{const}$$



Solution: ① $P(Y=1 \mid X=x) = \frac{1}{1+\exp\{-\hat{\beta}^T x\}}$. Setting $= 1/2$, we find we need $\exp\{-\hat{\beta}^T x\} = 1$, or equivalently, $\hat{\beta}^T x = 0$, i.e., $x_1 + 2x_2 = 0$ This is the decision boundary, and the plot is given above.

② Similar logic: $P(Y=1 \mid X=x) = \frac{1}{1+\exp\{-\hat{\beta}^T x\}} = 0.0498$. Solving we find

$\exp\{-x_1 - 2x_2\} = \frac{1 - 0.0498}{0.0498} \Rightarrow$ $x_1 + 2x_2 = \ln\left(\frac{1}{.9502}\right) + \ln\left(\frac{1}{.0498}\right)$

Need to plot this.

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.)

min: Train Error + ||·||

Solution: Regularization can improve testing error, but can only increase training error.

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.

Y $X_1 = Y + \text{Noise}$

$X_2 = -\text{Noise}$

$X_1 - X_2 = Y$

False. Noise columns can help us overfit and hence could reduce training error.

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

$X_i > \alpha$ $\log X_i > \log \alpha$

monotonic transformations of the features do not change the decision tree.

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

Not covered this year.

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.

Not covered this year.

$\hat{y}_i = h_1(x_i) + h_2(x_i)$

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.

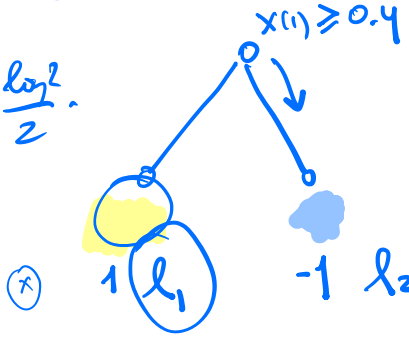
Table 2: Data

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

Solution: Call leaf 1 ℓ_1, 2 ℓ_2.
We must choose ℓ_1 & ℓ_2 to minimize the exp loss:
ℓ_1: min: $2e^{-\ell_1} + e^{\ell_1}$ // $\ell_1 = \frac{1}{2}\log 2$
ℓ_2: min: $e^{-\ell_2} + 2e^{\ell_2}$ // $\ell_2 = -\frac{1}{2}\log 2$
Computed by taking deriv & setting $=0$.

ℓ_2:
$\exp\{-1\cdot\ell_2\} + 2\exp\{1\cdot\ell_2\}$
$-e^{-\ell_2} + 2e^{\ell_2} = 0$
$2e^{\ell_2} = e^{-\ell_2}$
$\log 2 + \ell_2 = -\ell_2$
$\ell_2 = -\frac{\log 2}{2}$.

(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.



Find leaf values that minimize sum of exp losses:
$\exp\{-1\cdot\ell_1\} + \exp\{-1\cdot\ell_1\} + \exp\{-(-1)\cdot\ell_1\}$
$= 2\exp\{-\ell_1\} + \exp\{\ell_1\}$
$= -2e^{-\ell_1} + e^{\ell_1} = 0$ / $e^{\ell_1} = 2e^{-\ell_1}$
$\ell_1 = \log 2 - \ell_1$
$\ell_1 = \frac{1}{2}\log 2$

(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)"

Q: Why $h_1 + h_2$?

$$\min: \sum_{i=1}^{6} \exp\{-y_i \cdot [h_1(x_i) + h_2(x_i)]\}$$
$$= \sum_{i=1}^{6} \exp\{-y_i h_1(x_i)\} \cdot \exp\{-y_i(\beta_1 x_i(1) + \beta_2 x_i(2))\}$$

What is AdaBoost?

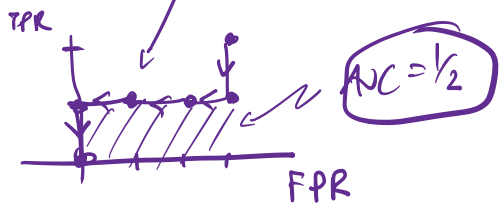
Boosting: Produce a sol of form $h_1 + h_2 + \cdots + h_k$

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.

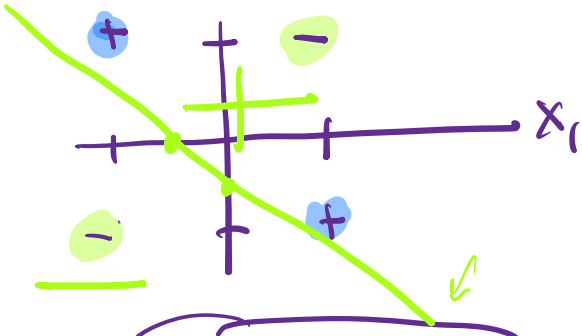
Truth	1	0	0	0	1
model	0.3	0.4	0.5	0.7	0.9

θ	FPR	TPR
$0 \le \theta \le 0.3$	3/3	2/2
$0.3 < \theta \le 0.4$	3/3	1/2
$0.4 \le \theta < 0.5$	2/3	1/2
$0.5 \le \theta \le 0.8$	1/3	1/2
$0.8 \le \theta \le 0.9$	0/3	1/2
$0.9 \le \theta < 1$	0/3	0/2



Part B: Consider this dataset

x_1	x_2	y
1	1	−1
−1	−1	−1
1	−1	+1
−1	1	+1



(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).

Plugging in, the rule is: $\hat{y} = \text{sign}\left(\frac{1}{2} + x_1 + x_2\right)$

hence the boundary is given by: $\frac{1}{2} + x_1 + x_2 = 0$.

We can plug in any point one one side, say, $\begin{pmatrix}0\\0\end{pmatrix}$, to see which side gets labeled + and which −.

In this case: $\text{sgn}\left(\frac{1}{2} + 0 + 0\right) = +1$, so upper right is +.

(b) Is it possible that such a linear classifier can correctly classify all the examples in this dataset ? NO.

Max Likelihood :

Coin : want to estimate $P(C = \text{heads})$.

Flip n times, k heads

$P(C = \text{heads}) = \frac{1}{2}$

ML principle: choose $\hat{p}$ that makes the outcome we saw as likely as possible.

If $P(H) = p$ what is $P(k \text{ heads out of } n \text{ flips})$

$$\binom{n}{k} p^k (1-p)^{n-k}$$

likelihood of what we saw if truth is $P(H) = p$.

Take deriv, set $= 0$;

$$\frac{d}{dp}\left[\binom{n}{k} p^k (1-p)^{n-k}\right] = \binom{n}{k}\left[(n-k)(1-p)^{n-k-1}(-1)\cdot p^k + k p^{k-1}(1-p)^{n-k}\right] = 0$$

$n-k-1$

$$k \cdot p^{k-1}(1-p)^{n-k} - (n-k)\,p^{k}(1-p) = 0$$

$$p^{k-1}(1-p)^{n-k-1}\left[k(1-p) - (n-k)p\right] = 0$$

$$k - kp - np + kp = 0$$

$$k = np$$

$$\boxed{p = k/n}$$

When to boost more?

A: Boosting reduce training error, but it might increase overfitting.

Same answer for a diff strategy.

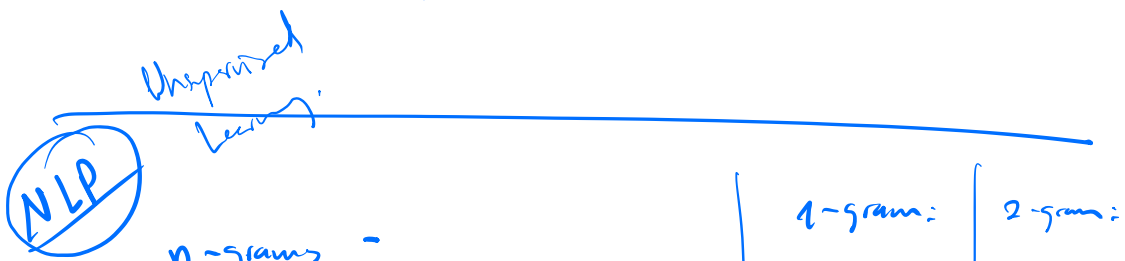
Stacking. [X]

If taking on-line:

1. Camera on

2. scan & upload 1 file w/ solns by

~~Be~~ 1:50



n-grams -

freq. of all n-tuples.

1-gram:	2-gram:
a	ab
b	ac
c	ad
	ae
	⋮