282566 *282566*

Student Name: Carlos Valle-Diaz

Room: Virtual

MOSES CENTER: CONFIRMED EXAM

PROFESSOR: Christopher Paul Musco

PROFESSOR CONTACT: cpm303@nyu.edu, +1 646 997 3346

COURSE: Intro to Machine Learning 4563 A Lecture

IN-CLASS DURATION: 1 hour; 30 minutes

SUPPLEMENT MATERIALS:

ARE QUESTIONS ALLOWED:

SCRAP PAPER:

SCANTRON:

DATE: 03-09-2020

START TIME: 9:00 AM

STUDENT NAME: Carlos Valle-Diaz

END TIME: 11:15 AM

ACCOMODATION(S):

- Extra Time: Extended time (1.5x) on in-class timed exams, in-class timed assignments and out of class timed exams
- Smaller proctored testing environment

EXAM RECEIPT:
Email/Upload
Professor delivery

EXAM RETURN:
Scan/Email to:
Professor pickup: Signature: ______________________ Date: ______________

Semester: Spring Session 2020

282566 *282566*

Student Name: Carlos Valle-Diaz

Room: Virtual

MOSES CENTER:
STUDENT ACCOMODATED EXAM INFO

PROFESSOR: Christopher Paul Musco

COURSE: Intro to Machine Learning 4563 A Lecture

9:13

DATE: 03-09-2020

START TIME: 9:00 AM

STUDENT NAME: Carlos Valle-Diaz

END TIME: 11:15 AM

RETURN THIS SHEET, EXAM SHEETS, & ALL SCRAP PAPER TO THE FRONT DESK

NYU TANDON SCHOOL OF ENGINEERING Carlos valles-Diaz

New York University Tandon School of Engineering
Computer Science and Engineering

CS-UY 4563: Midterm Exam 1.
Monday, Mar. 9th, 2020, 9:00 - 10:15pm
50 Total Points

Directions

- Show all of your work to receive full (and partial) credit.
- If more space is required, you may use extra sheets of paper clearly marked with your name, netid, and the problem you are working on.

1. Always, Sometimes, Never. (12pts – 3pts each)

Indicate whether each of the following statements is ALWAYS true, SOMETIMES true, or NEVER true. **No justification is necessary to receive full credit for a correct answer.** To earn partial credit if you are wrong, you may provide a short justification or example to explain your choice.

(a) The empirical risk of a model is lower than the population risk.
ALWAYS SOMETIMES NEVER

(b) You train a multiple linear regression model with varying levels of ℓ_2 regularization. Let $\vec{\beta}^{(1)} = \arg\min_{\vec{\beta}} \|X\vec{\beta} - \vec{y}\|_2^2 + \lambda_1\|\vec{\beta}\|_2^2$ and let $\vec{\beta}^{(2)} = \arg\min_{\vec{\beta}} \|X\vec{\beta} - \vec{y}\|_2^2 + \lambda_2\|\vec{\beta}\|_2^2$.
If $\lambda_1 > \lambda_2$, is $\|X\vec{\beta}^{(1)} - \vec{y}\|_2^2 < \|X\vec{\beta}^{(2)} - \vec{y}\|_2^2$?
ALWAYS SOMETIMES NEVER

They are equal unless Sub stub skips are off

(c) The linear classifier found by logistic regression minimizes error rate (0-1 loss) on the training data.
ALWAYS SOMETIMES NEVER

(d) Consider a multiple linear regression problem where each data example has the form $(\vec{x}, y) = ([x_1, x_2], y)$. Transform the predictor variables by adding quadratic terms, so each new data example has the form $(\vec{x}_{trans}, y) = ([x_1, x_2, x_1^2, x_2^2, x_1x_2], y)$. Let L^* be the minimum training loss for the original problem and let L^*_{trans} be the minimum training loss for the transformed problem. Is $L^*_{trans} \le L^*$?
ALWAYS SOMETIMES NEVER

Question which is more accurate data?, original or transformed? This case original

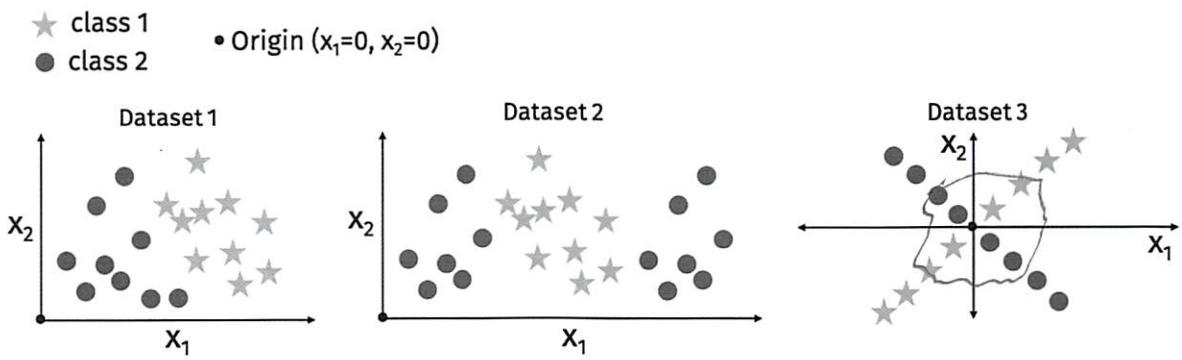
L^* min of training loss of orig prob

L^* trans the min training loss tranformed proble.

Carlos Valle-Diaz

3. Model Diagnosis 2 (10pts)

Consider the following scatter plots of data for three binary classification problems. x_1 and x_2 are the independent variables and class labels are indicated by points with a different shape and shade.



(a) (4pts) Indicate which of the three clustering problems could be solved to high accuracy (small error rate) using a logistic regression model with no regularization and no feature transformations.

One V.S Rest could be used to solve all three models although slow it will be able to seperate and cluster each individual group based on the point of its shape if its circle or star.

(b) (6pts) For any of the problems that you believe are not *directly solvable* with logistic regression, suggest a possible feature transformation which *would make it possible* to obtain a high accuracy solution with logistic regression. For each problem, your solution should be a set of new features $\phi_1(x_1, x_2), \phi_2(x_1, x_2), \ldots, \phi_q(x_1, x_2)$ that depend on the original features x_1 and x_2. You may use as large a q as you need.

For dataset 3 we can use a pnorm w/ a high p to group them and have a better solution then split off from there rather then do a one vs rest solution.

NYU TANDON SCHOOL OF ENGINEERING Carlos Valle-Diaz

2. Model Diagnosis Short Answer (8pts)

You are trying to solve a prediction problem using a multiple linear regression model with ℓ_2 loss. You first split the data set into a train set (80%) and a test set (20%). You then train the model on the train set to obtain a parameter vector $\vec{\beta}$. Using $\vec{\beta}$, you evaluate the average squared loss of the regression model on the train and test set, separately.

For each of the following scenarios, circle all answers that apply. **No justification is necessary to receive full credit for a correct answer.** To earn partial credit if you are wrong, you may provide a short justification.

(a) (4pts) The average squared loss on the train set is 1.5 and the average squared loss on the test set is 12.6. **Which of the following techniques is likely to improve your average test loss?**

REGULARIZATION FEATURE SELECTION FEATURE TRANSFORM DATA SCALING

if we regulate the data we can minimize the loss same goes for choosing the specific features,

(b) (4pts) The average squared loss on the train set is 10.2 and the average squared loss on the test set is 9.9. **Which of the following techniques is likely to improve your average test loss?**

REGULARIZATION FEATURE SELECTION FEATURE TRANSFORM DATA SCALING

- Same as above for F.S and Regulization
- Data Scaling we can scale the data inorder to reduce the amount of unecessary data needed or garbadge data to reduce LOSS.
- Feature Transform - Can reduce the amount of features and ∴ reduce loss by reducing the # of entries and points and error b/n points that may occur

NYU TANDON SCHOOL OF ENGINEERING

Carlos Valle-Diaz

4. Loss Minimization. (10pts)

For data with one predictor and one target: $(x_1, y_1), \ldots, (x_n, y_n)$, consider a linear regression model:

$$f_{\beta_0,\beta_1}(x) = \beta_0 + \beta_1 x$$

$u = (y_i - f_{0,\beta_i}(x_i))^2$

$du = +2x(y_i - f_{0\beta_i\beta_i}(x_i))\frac{1}{\beta_0+\beta_i}$

with *exponential loss*:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} e^{(y_i - f_{\beta_0,\beta_1}(x_i))^2}$$

(a) (5pts) Write down an expression for the gradient of the loss L.

$\nabla L(\beta_0,\beta_1) = 2x_i(y_i) - f_{\beta_0\beta_1}(x_i)))\frac{1}{\beta_0+\beta_i} \cdot e^{(y_i - f_{\beta_0\beta_1}(x_i))^2}$

(b) (2pts) Name two algorithms/methods which could be used to minimize L.

* Brute force
* Log Gradient

(c) (3pts) In general, is this exponential loss more or less robust to outliers when compared to ℓ_2 loss? How about when compared to ℓ_∞ loss?

Exponetianal loss is more roblast than l_2 loss when compare to l_2 loss on l_∞ loss if drawn correctly the graph looks somewhat like:

where exponetial capture more outliers that the l_2 loss as it is able to expand more freely

 Carlos Valle-Díaz

5. Bayesian Crab Classification (**10pts**)

A biologist is collecting specimens from two species of crabs, species S_0 and S_1. These species live in the same habitat and look similar to the human eye. To accelerate crab sorting by species, the biologist wants to develop a simple classification rule based on body measurements. She observes that the ratio of *forehead breadth* to overall *body length* differs between crabs in species S_0 and S_1. The biologist proposes to measure this ratio (denoted by R) and use it as a single predictor variable for classification.

The biologist assumes that the crab data comes from a "mixture of Gaussians" probabilistic model. In particular, she assumes that for each species, R follows a normal (Gaussian) probability distribution, with different parameters for each species. The biologist makes the following concrete observations:

- 35% of all crabs collected belong to S_0 and the remaining 65% belong to S_1.
- For crabs in S_0, the average value of R is .5. For crabs in S_1, the average value of R is .4.
- For both species, the standard deviation of R is .1.

(a) (6pts) Suppose we collect a new crab with forehead breadth to body length ratio R_{new}. The biologist would like to assign this crab to S_0 or S_1 using a maximum a posterior (MAP) classification rule. Denote this rule by $f : \mathbb{R} \to \{S_0, S_1\}$. The rule takes as input the ratio R_{new} and outputs S_0 or S_1.

Write down all mathematical expressions that would need to be evaluated to compute f for a given input R_{new}. Your expressions do not need to be simplified, but they should not involve unknown variables besides R_{new}. **Hint:** Use Bayes rule.

$S_0 = .35$

$S_1 = .65$

$\overline{S_0} = .5$

$\overline{S_1} = .4$

new Crab?

$$\Pr(\beta \mid (x, y)) = \frac{\Pr((x, y) \mid \beta^{new}) \Pr(\beta^{ne})}{\Pr((x, y))}$$

$$\Pr(R_{new} \mid S_0/S_1)_{f: R\{S_0, S_1\}} = \frac{\Pr((R_{S_0}, R_{S_1}) \mid R_{new}) \Pr(R_{new})}{\Pr((R_{S_0}, R_{S_1}))}$$

(b) (4pts) Show that, for this problem, the classification rule f has the following form:

$$f(R_{new}) = \begin{cases} S_0 \text{ if } R_{new} \geq \lambda \\ S_1 \text{ if } R_{new} < \lambda, \end{cases}$$

for some fixed threshold parameter λ (you do not need to explicitly compute λ).

There is some fixed threshhold as the crabs have a destinct size of .4 and .5 ∴ there must be a threshhold size where this new crab is.

(c) (3pts – extra credit) ~~Given the~~ biologist's data above, will the threshold λ for the MAP classification rule be EQUAL TO, LARGER, or SMALLER than .45? Justify your answer in a sentence or two. This problem can be solved without a calculator.

Smaller as the crabs sit b/n .4 & .5 and most likely fit in the range above the .4 range as 65% of them are .5 size.

Question 1, b if it is a typo ~~that~~ is it suppose to be β_1 and β_2 or β^1 and β^2?