CSCI 183
Fall 2024
Exam 1
10/25/2024
Time Limit: 65 Minutes

Name: Arman Mini
SCU ID: _____

This exam contains 6 pages (including this cover page) and 3 questions.
Total of points is 25.

Grade Table (for teacher use only)

Question	Points	Score
1	5	
2	8	
3	12	
Total:	25	

1. (5 points) **Data Visualization, Feature Selection and Encoding**

   (a) (1 point) Ross selected a feature with $r = 0.94$ and built a univariate linear regression model and recorded the error (e_1). He added some more features for training with $r\_list = [0.87, 0.23, 0.04, -0.3, -0.01]$ and trained a linear regression model on the same dataset and noticed the new error(e_2) is less that e_1. Is this possible?

   By adding new values the model either god more accurate or the new data fit it well thus lowering error since less outliers, or the new data caused the line to adjust changing the inclusion lowering error

   (b) (2 points) For the given scenario answer the following: Chandler is plotting a scatter plot using *annual income* (feature) and *loan approved* (target).

   - He uses red and blue circles to represent *Approved* and *Not Approved* classes.
   - He aims to check if the data is linearly separable.
   - He is struggling to distinguish between the two classes in the plot.

   i. What could be the reasons for Chandler's difficulty in distinguishing the classes in the scatter plot?

   ii. How can Chandler overcome this challenge to visualize the classes better to decide a classification algorithm?

   what if not linearly seperable?

   i. the difficulty could be a result of overlap in points → more factors than annual income thus 2 people same income can have different results

   ii. he can either adjust plot to space out better to find the gaps and zones for the line or he can assign the values to two scatter plots rather than one to see a comparison → maybe not linearly seperable

   (c) (2 points) Assume the following scenario:

   - Rachel's dataset has *PurchaseAmount* and *ProductCategory* [Electronics, Clothing, Home].
   - She splits the data into train and test sets and applies one-hot encoding (get_dummies) on *ProductCategory* in each of the subsets.
   - The number of columns in the train set differs from the test set.

   Is this possible? Explain your answer.

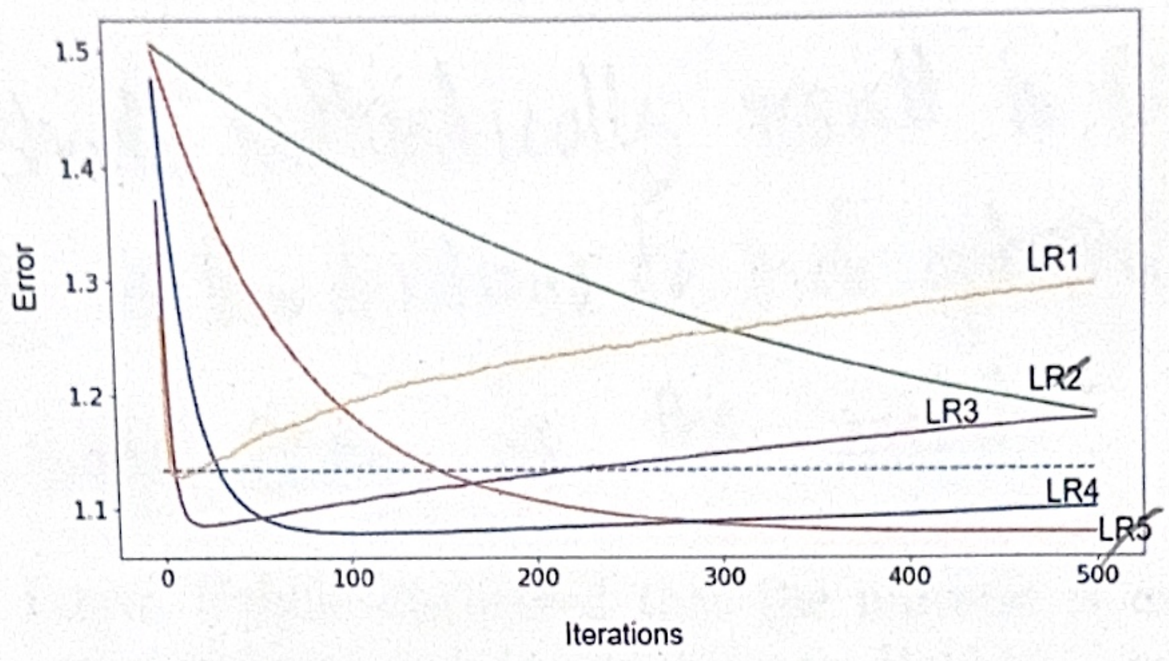
   yes since the split can result in test data and training data having diff items thus more/less colums because of one hot encoding training can have feature A, B, C, D, F while test only has B, C, F

2. (8 points) **Linear Regression**

Answer the following questions (part a and part b are independent of each other):

(a) (2 points) Joey was experimenting with five values of learning rates: LR1 (yellow), LR2 (green), LR3 (purple), LR4 (blue) and LR5 (red) and plot the graph below.

5,2,4,3,1



slop 2>5

slope 1>3>4

What would be the **ascending** order of the value of the learning rates and why? Explain your answer.

5,2,4,3,1 this is since 5, 2 converge the fastest so ib

those first but the following 3 diverge thus

I ranked in order of slope greatest to least

5,2,4,3,1

gh low

(b) (2 points) Joey and Chandler are performing Linear Regression on the same set of features and target variable. They are using the same loss function and the same threshold for the epsilon method. After their algorithm **converged**, they both get significantly different error values. Is this possible? If so, why? If not, why not?

Assuming gradient descent

yes this is possible, if they were to have different learning rates it will affect the convergence thus will result in different values one must've had a much smaller or larger one than the other, After x iteration it will always converge learning rate -> speed

(c) (1 point) Monica was training univariate linear regression model with the least square regression as the loss function on a dataset for predicting the sales of her restaurant which has a target column called 'sales'. She initialized the model parameters to be 0. In the first iteration she observed that the loss is 0. Is this possible? Explain your answer.

Assuming a hypothesis of $\hat{y} = \theta_1 x + \theta_2$ if she initializes at 0 it should mathmatically result in 0, this would require a few iterations to find optimal θ_1, θ_2 values to make sense of the model.

(d) (3 points) Edwin Hubble discovered that the universe is expanding. If v is a galaxy's recession velocity and d is its distance, Hubble's law states that v = Hd, where H is known as Hubble's constant. The following are distance (in millions of lightyears) and velocity (thousands of miles per second) measurements made on five galactic clusters.

Assume: **Hypothesis function: $\hat{y} = Hd + c$**

distance	68	137	315	405	700
velocity	2.4	4.7	12.0	14.4	26.0

Answer the following questions:

(a) (2 points) Find the updated parameters H and c after the first iteration of Gradient Descent algorithm. The following is given: Predict new x_1

- Initially, $H = 1, c = 0$ → θ_1, θ_2 init'd?
- Learning rate is 0.001.
- Gradients of $J(H, c)$ with respect to H is 149604.78
- Gradients of $J(H, c)$ with respect to c is 313.1

Possible these θ_1, θ_2 for x_1?

Show all your steps, work and calculation.

(b) (1 point) After 1000 iterations you find H = 0.0368 and c = -0.002. What will be the predicted velocity for a distance of 900?

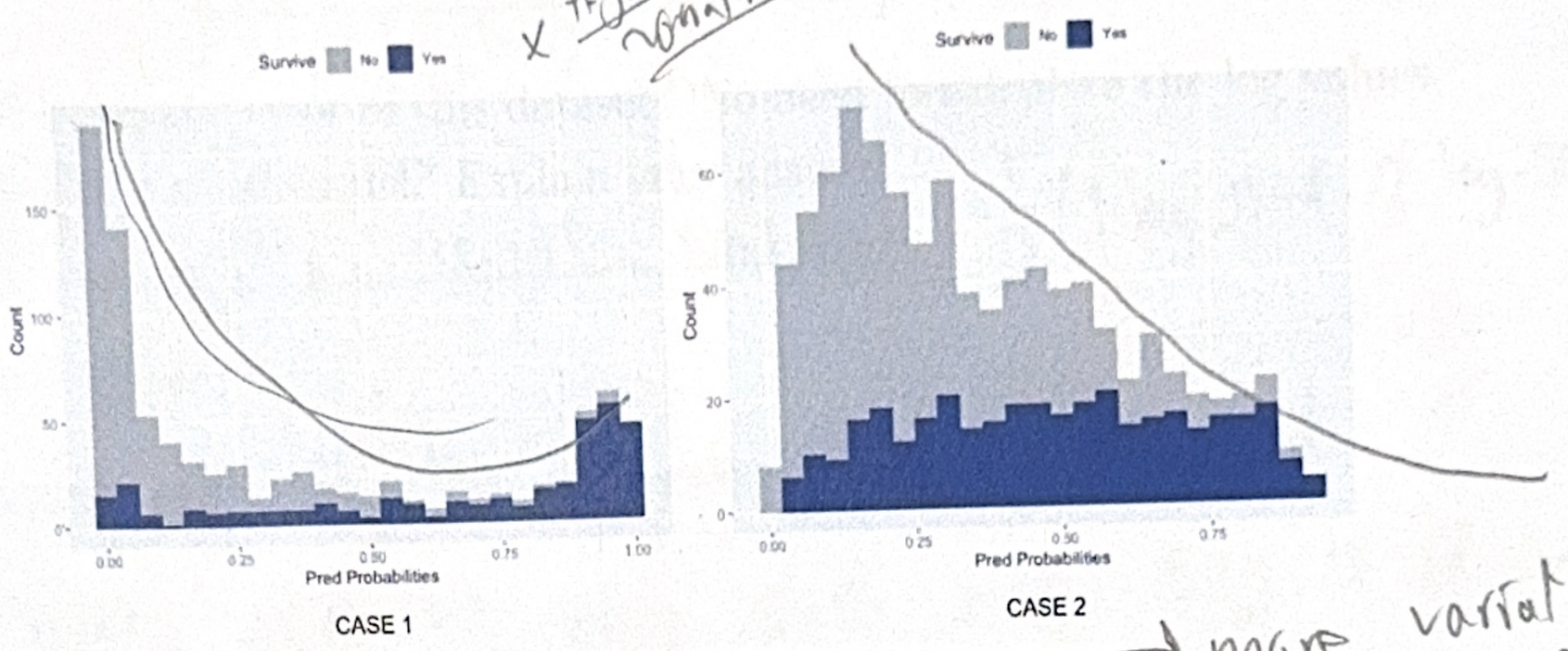
b) $\hat{y} = 0.0368(900) + (-0.002) \rightarrow =$

a) $\theta_1 := 1 - (0.001) \sum_{i=1}^{5} 2(1 \cdot x^{(i)} + 0 - y^{(i)}) \cdot x^{(i)}$

$\theta_2 := 0 - (0.001) \sum_{i=1}^{5} 2(1 \cdot x^{(i)} + 0 - y^{(i)}) \rightarrow =$

3. (12 points) **Linear Classification:**

(a) Assume you have the following figure that describes the predicted probability of data points in Logistic Regression for a target variable 'Survive'. The y-axis represents the count of predicted probabilities of data points belonging to each bin in the histogram. The grey color represents 'No' (Class 0) and blue color represents 'Yes' (Class 1):



i. (2 points) In which case in the figure above, will the log loss value be more and why?

ii. (2 points) Based on the answer of i) which do you think is a better model and why?

i. ~~Case 2 as it follows the log loss line of converging where case 1 seems to increase again with the exponential in a way.~~

ii. case 2 since it follows the log loss trend which is optimal for O,1 output problems, however it using something like exponential then case 1 may be better, it also has less variation in case 2 so better

(b) (3 points) We use an AND gate to generate a dataset where the two inputs to the AND gate make up two boolean features of my dataset and output of the AND gate makes up the target variable.Given this data, which linear classification algorithm (discussed in class) would you choose to perform classification on the above dataset? Justify your answer.

the sigmoid function, since we want a nonlinear, 0,1 determining function for the given problem at hand. this will result of only 0,1 which is log since there is no halt alive, Also determines most likely output which we want in this prediction

(c) (2 points) You are applying Logistic Regression to predict whether a user will click on an advertisement. You obtain the following table:

Table 1: Ad Click Prediction Results

User ID	True Label	Predicted Probability
456	1	0.15
789	0	0.92
321	1	0.45

a) What is the error of this dataset? No need to calculate the log values.

b) Is this a good model? Explain your answer.

a) $(.15-1)^2 + (.92-0)^2 + (.45-1)^2 =$ total loss

$-\frac{1}{3}(1\log .15 - (0\log .85) + (0\log .92 - 1\log .08) + (1\log .45 - 1\log .55)$

b) very bad example as it has a super high loss and almost gets every prediction over 50% wrong say .15 it will happen and it happens when .92 it will and it doesn't

(d) (3 points) Assume there is a dataset with a categorical target variable having values {cat, dog and rabbit}. Phoebe wants to perform linear classification on this dataset using multiple **logistic regression** models. As a data scientist how would you help Phoebe with this scenario? Explain your answer with an example.

As a data scientist my advice would be to use a combined model as opposed to multiple log regression models as it will result in the least error. this happened in our hw 2 problem 6 where we saw our combined model for the different features resulted in much lower error than individual ones. As well as with a classification task like this it may make it less demanding computationally like with a DTE where initial steps can be shared like a yes/no whiskers helping cut the overall number of steps and computational power need. should help reduce loss especially since still smaller the log loss function by balancing it out