DS 310 Machine Learning
Fall 2021 / Amulya Yadav
HW #2: 50 points

Please submit two files on Canvas:
- *your answer to this homework;*
- *the code for problem 3 (named "yourlastname_HW2.ipynb").*

Remember to submit the two files by clicking on "add another file" in Canvas, instead of submitting one zipped file of the aforementioned two files.

Please reach out to TA Hangzhi Guo (email: hangz@psu.edu) if there is any question about the assignment.

1. (5 points) What objective function do we use in logistic regression?
   A. Least Square Error
   B. Maximum Likelihood
   C. Both A and B

2. (5 points) Which of the following evaluation metrics does not make sense if applied to logistic regression output to compare with target?
   A. Accuracy
   B. Log loss
   C. Mean Squared Error

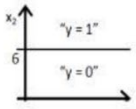
3. (5 points) Suppose you train a logistic regression classifier and your hypothesis function h is

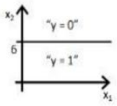
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \text{ where } \theta_0 = 6,\ \theta_1 = 0,\ \theta_2 = -1.$$

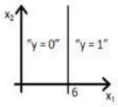
$6 - x_2 \geq 0$

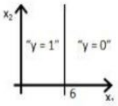
$6 \geq x_2$

Which of the following figure will represent the decision boundary as given by above classifier?

A) 

B) 

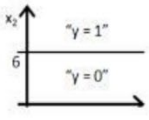
C) 

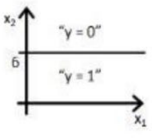
D) 

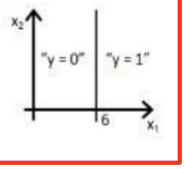
4. (5 points) If you replace coefficient of x1 with x2 what would be the output figure?

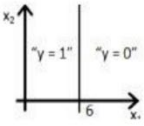
$6 - x_1 \geq 0$

$6 \geq x_1$

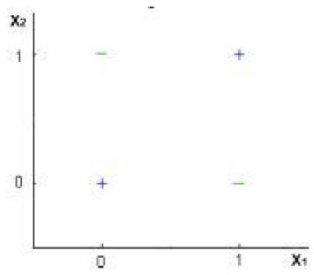
A) 

B) 

C) 

D) 

5. (5 points) Can a Logistic Regression classifier do a perfect classification on the data shown below? (*Note: You can use only X1 and X2 variables where X1 and X2 can take only two binary values (0,1)*)



A. True
B. False
C. Can't say
D. None of these

6. (25 points) Consider the Adult Income dataset. The ML model is aimed to predict whether the income will exceed 50k or not. You need to first preprocess the dataset. Next, you need to build a machine learning model to predict individual's income will >=50k (positive) or <50k (negative). Please follow the instruction on `hw2.ipynb`.

   a. (5 points) Transform the categorical features into the one-hot-encoding format.
   b. (5 points) Normalize the continuous features.
   c. (5 points) Split the dataset into *train_X, train_y, test_X, test_y.*
   d. (10 points) Build a ML model to predict income in the testing set.
      You should report the accuracy score in the testing set. To obtain full credit, you need to achieve accuracy score higher than 82.5% on the testing set. Please briefly describe your approach. *(5 bonus point if accuracy score is higher than 83.5% or the highest accuracy if no one reaches 83.5%)*

Implement all the questions and upload your code named `{yourlastname}_hw2.ipynb` to Canvas. Your code need to be reproducible for your results.