BUAN 6341
APPLIED MACHINE LEARNING
ASSIGNMENT 1
Due date: February 11, 11:59 pm

In this assignment, we will be using linear and logistic regression on a given dataset. In addition, we will experiment with design and feature choices.

We will be using Online News Sharing dataset available for download at
https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#

Goal:
You are required to implement the following two algorithms:

1. Linear regression
2. Logistic regression

You **can** use **any** publicly available R or Python library/package.

Tasks:

1. Divide the dataset into train and test sets sampling randomly. Use only predictive attributes and the target variable (do not use non-predictive attributes).
2. Use linear regression to predict the number of shares. Report and compare your train and test error/accuracy metrics. You can pick any metrics you like (e.g. mean squared error, mean absolute error, etc.)
3. Convert this problem into a binary classification problem. The target variable should have two values (large or small number of shares).
4. Implement logistic regression to carry out classification on this data set. Report accuracy/error metrics for train and test sets.

Experimentation:

1. Experiment with various model parameters for both linear and logistic regression and report on your findings as how the error varies for train and test sets with varying these parameters. Plot the results. Report your best parameters. Examples of these parameters can be learning rate for gradient descent, convergence threshold, etc.
2. Pick ten features randomly and retrain your model only on these ten features. Compare train and test error results for the case of using all features to using ten random features. Report which ten features did you select randomly.
3. Now pick ten features that you think are best suited to predict the output, and retrain your model using these ten features. Compare to the case of using all features and to random features case. Did your choice of features provide better results than picking random features? Why? Did your choice of features provide better results than using all features? Why?

Deliverables:
You are required to turn in your code and a report. We should be able to run the code as is and get the results and plots that you have included in the report. You should include and describe results for all the

experiments above. You should also mention how you constructed the classes for the classification problem (value of threshold and why you picked it). You can be creative and include other plots/results too. However, the report should not exceed 6 pages. Also describe your interpretation of the results. What do you think matters the most for predicting the number and class of shares? What other steps you could have taken with regards to modeling to get better results?

Grading:
Total weightage: 10% of final grade

Breakdown:

Code: 20 points (Code should execute and produce the results presented in the report with minimum effort).

Report: 80 points
Points will be awarded based not only on how good your results are, but also on how well you describe them as well as underlying experimentation.