BUAN 6341
APPLIED MACHINE LEARNING
ASSIGNMENT 1
Due date: February 10, 11:59 pm

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

We will be using the Facebook Comment Volume dataset available for download at
https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset
You can use any variant of the dataset.

Goal:
Implement a linear regression model on the dataset to predict the total number of comments a post will receive in next H hours. You are not allowed to use any available implementation of the regression model. You should implement the gradient descent algorithm with batch update (all training examples used at once). Use the sum of squared error normalized by 2*number of samples [$J(\beta_0, \beta_1) = (1/2m)[\sum(y^(i) - y(i))^2]$ as your cost and error measures, where m is number of samples. You can use any number of features greater than 10. You can decide on which features to use using some experimentation and exploratory analysis.

Tasks:
Part 1: Download the dataset and partition it randomly into train and test set using a 70/30 split.

Part 2: Design a linear regression model to model the number of comments a post will receive in next H hours. Include your regression model equation in the report.

Part 3: Implement the gradient descent algorithm with batch update rule. Use the same cost function as in the class (sum of squared error). Report your initial parameter values.

Experimentation:
1. Experiment with various values of learning rate $\propto$ and report on your findings as how the error varies for train and test sets with varying $\propto$. Plot the results. Report your best $\propto$ and why you picked it.
2. Experiment with various thresholds for convergence. Plot error results for train and test sets as a function of threshold and describe how varying the threshold affects error. Pick your best threshold and plot train and test error (in one figure) as a function of number of gradient descent iterations.
3. Pick five features randomly and retrain your model only on these five features. Compare train and test error results for the case of using your original set of features (greater than 10) and five random features. Report which five features did you select randomly.
4. Now pick five features that you think are best suited to predict the output, and retrain your model using these five features. Compare to the case of using your original set of 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. In your report, include the final equation of your model (with all parameters). You should include and describe results for all the four experiments above. In addition, you should also include your final error values for train and test sets. You can be creative and include other plots/results too. However, the report should not exceed 8 pages total.

Grading:
Total weightage: 10% of final grade

Breakdown:

Report: 100 points
If your code doesn't run or doesn't produce the same results then you get zero 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.

Experiment 1: 20 points
Experiment 2: 20 points
Experiment 3: 20 points
Experiment 4: 20 points

Discussion: 20 points
Describe your interpretation of the results. What do you think matters the most for predicting the number of comments? What other steps you could have taken with regards to modeling to get better results?