**Homework 2: Linear Regression** **CSCE 633**
Due: 11:59pm on February 16, 2024

---

**Instructions for homework submission**
a) There are two sections in this homework. Write the solution to the first section in Latex. For the programming questions, explain your thought process, results, and observations in a markdown cell after the code cells in a Jupyter Notebook.
b) You need to submit a zip file:

- The name of the .zip file: FirstName_LastName_HW2.zip
- The zipped folder includes the following files:
  - A pdf (generated using latex) for the math section: FirstName_LastName_HW2.pdf
  - A .ipynb jupyter notebook file for the programming section: FirstName_LastName_HW2.ipynb
  - A csv file with 2 main columns. One for the linear regression predictions with the name pred_linear and one for the logistic regression predictions with the name pred_logistic. file name: FirstName_LastName_preds.csv

c) Start early :)
d) Total: 100 points

**Math Questions**

---

**Problem 1: Gradient Calculation (8 points)**
**NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.**

In this question you are required to calculate gradients for 2 scalar functions.

(1) Calculate the gradient of the function $f(x, y) = x^2 + ln(y) + xy + y^3$. What is the gradient value for $(x, y) = (10, -10)$?

(2) Calculate the gradient of the function $f(x, y, z) = tanh(x^3 y^3) + sin(z^2)$. What is the gradient value for $(x, y, z) = (-1, 0, \pi/2)$?

**Problem 2: Matrix Multiplication (8 points)**
**NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this problem. Show your work.**

In this question you are required to perform matrix multiplication.

(1)

$$\begin{bmatrix} 10 \\ -5 \\ 2 \\ 8 \end{bmatrix} \begin{bmatrix} 0 & 3 & 0 & 1 \end{bmatrix} = ?$$

(2)

$$\begin{bmatrix} 1 & -1 & 6 & 7 \\ 9 & 0 & 8 & 1 \\ -8 & 1 & 2 & 3 \\ 10 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 6 & 2 & 0 \\ 0 & -1 & 1 \\ -3 & 0 & 4 \\ 3 & 4 & 7 \end{bmatrix} =?$$

**Programming Questions**

The goal of this section is to create a regression-based models to assess air quality. The data for this homework is uploaded on CANVAS (data_train.csv) which includes the training data (features and labels) and (data_test.csv) which includes the test data. Remember, the test data only contains the features we use for the training, but not the labels. For each row in the test data, you need to use the trained model to predict the corresponding labels. Each row of the data corresponds to a sample and the columns include the following information:

1. NMHC(GT): hourly averaged overall Non Metanic HydroCarbons concentration in microg/ $m^3$
2. C6H6(GT): hourly averaged Benzene concentration in microg/$m^3$
3. C6H6(GT): hourly averaged Benzene concentration in microg/$m^3$
4. PT08.S2(NMHC): hourly averaged sensor response to NMHC
5. NOx(GT): hourly averaged NOx concentration in ppb
6. PT08.S3(NOx): hourly averaged sensor response for NOx
7. NO2(GT): hourly averaged NO2 concentration in microg/$m^3$
8. PT08.S4(NO2): hourly averaged sensor response for NO2
9. PT08.S5(O3): hourly averaged sensor response for O3
10. T: Temperature in C
11. RH: Relative Humidity
12. AH: Absolute Humidity
13. PT08.S1(CO): TARGET VARIABLE - hourly averaged sensor response for CO

**(a) Data Processing (4 points)**

1. Download and read the data. For Python, you may use *pandas* library and use *read csv* function.
2. Print the first 5 rows of the data. You may use *head()* function in *pandas* library. Print the shape of the training dataframe. Write a short description of the data.

3. Does the data have any missing values? How many are missing? Return the number of missing values. (In *pandas*, check out *isnul()* and *isnul()*.sum())

4. Drop all the rows with any missing data. (In *pandas*, check out *dropna()*. *dropna()* accepts an argument *inplace*, check out what it does and when it comes in handy.)

5. Extract the features and the label. The label is *PT08.S1(CO)*.

**(b) Exploratory Data Analysis (20 points)**

1. Plot the histograms of all the features in the data. Do all the features have a normal distribution? Do you see any outlier values? Do you need to apply any normalization technique to these values? If so, you can transform your data in this step and explain your thought process in the corresponding markdown cell.

2. Pick 2 features and create a scatter plot to illustrate the correlation between these two features. Is there a high correlation between these features?

3. Compute the Pearson's correlation between all pairs of variables 1-12. Assign the resulting correlation values in a 12x12 matrix C, whose (i; j) element represents the correlation value between variables i and j, i.e., C(i; j) = corr(i; j). Visualize the resulting matrix C with a heatmap and discuss potential associations between the considered variables. Note: You can use the 'heatmap' function from 'seaborn'.

**(c) (30 points) Linear Regression Implementation** Implement a linear regression model **from scratch** to regress the target variable, Carbon monoxide (CO). (Remember: The only library you may use for this question is the *NumPy* library.)

**(d) (20 points) Result Analysis** Perform a 5-fold cross validation. Compute RMSE for each validation set across 5 folds. Report average and standard deviation of RMSE values. Do you see a big change across different folds? How can you use the coefficient of this model to find the most informative features?

**(e) (10 points) Inference**

1. Use the trained linear regression model and predict the PT08.S1(CO) value for the test data.

2. Save the predictions in a csv file with a single column. The column will contian the linear regression predictions with the name pred_linear.

3. Add this csv file to your submission.