CSE 4622 - Machine Learning Lab
Lab - 02
Topics: Simple and Multiple Linear Regression

[The tasks have been designed in such a way that it improves your overall understanding of the topics and makes you ready for research. If any part of the task is not clear, please post a comment in the submission thread. I'll provide the necessary clarification. Please try to be honest and avoid plagiarism. Any unfair means will be strictly handled.**]**

In this assignment, we will be working with the **House Price Prediction** dataset and implementing the **Linear Regression** algorithm for predicting house prices based on related features.

- Use **Google Colab** to write all the codes in a single notebook file (**.ipynb** format). For each task, show the output for different input in individual code blocks (in ipynb) before submitting for the ease of evaluation

Tasks:

1. Load the dataset and print the total number of samples, column names, null values in each column, and data types of the columns. Remove the samples that have null values.
2. Identify the feature columns and the value you want to predict (In this case, house prices).
3. Split the dataset into training and test sets. The ratio should be 80-20.
4. Apply gradient descent algorithm to calculate the coefficients of individual features on this dataset.
    a. If the error difference between any two consecutive epochs is less than 0.5, stop the training.
    b. Plot Epoch VS Loss graph and identify the stopping epoch.

    c. Show the effect of different step sizes on both training and testing.
5. Implement Linear Regression using Scikit-Learn.
    a. At first, use all the features to predict the hyperplane, thus, calculating the coefficients. Then, calculate the MSE (Mean Squared Error).
    b. Generate Pearson Correlation Matrix from the training set. Instead of using all features to train the models, select features that are highly correlated with house prices, but not correlated with each other. After selecting hand-picked features, implement Linear Regression again and show if the MSE value decreases or not.
    The following materials might be helpful to understand what Pearson Correlation Matrix is.
        i. Analytics Vidhya
        ii. Real Python
        iii. Kaggle Notebook

Submission:

1. Submit a PDF and IPYNB formatted version of your notebook where the output of each cell is visible.
2. Write a report on your experimental findings from the above-mentioned tasks (1~5). Try to reflect your deep understanding of the topics. You have to include some analysis on your solution approach, the problems you faced during experimenting, how you overcame those issues, and so on. You can use the necessary figures to enrich your explanation (will carry bonus marks). Your report should include the following:
    a. A brief description of your dataset preprocessing
    b. A brief description of your model training
    c. Feature selection mechanism
    d. Justification of obtained result
    e. Conclusion

3. Submit the notebook files and report in a single zipped file. The naming convention should be ‘StudentID_Lab02.zip’.

Your submitted report will be checked for plagiarism. Do not copy from somewhere or someone else. Try to write only what you have understood.