Lab Assignment 3 - Regression - Part 1

Build a **Linear Regression Model** from scratch for predicting Behaviour of the urban traffic of the city of Sao Paulo in Brazil.

Steps

1. **Dataset:** Download the dataset from the link https://tinyurl.com/braziltraffic. The dataset contains 17 features and 1 target (Slowness in traffic (%) which is last column).
2. **Encoding:** Load the dataset into the code for pre-processing. 1st feature 'Hour' can be discretized into labels such as [morning, noon, afternoon, evening, night], which can be further codes using one-hot encoding, where morning can be represented as [0,0,0,0,1], noon can be [0,0,0,1,0] and so on. This results single feature "Hour" to be represented using five features of binary values. This now makes the dataset to have four extra columns. Choice of discretization is up to you. You can have like [day, night] also.
3. **Normalization:** Since the features are in different ranges, each column can be normalized into 0 to 1 using different methods such as scaling, standardizing etc. This part you have to explore. Note: Normalization should not be done for the target feature.
4. **Data Splitting:** After the range normalization, its time to split the data into training and testing. Dataset contain 135 entries (5 days data, each day 27 entries), so keep the last 27 rows of the original dataset (data of last 1 day) for testing, and rest of them for training.
5. **Linear Regression Training:** Now we have to train this model based on gradient descent algorithm. Automatically initialize a linear regression model with n+1 parameters (n features and 1 bias) with small random float values. Read the training samples one by one. Compute the error for each training sample, update the n+1 parameters. Updation should be done simultaneously for all the parameters before iterating for the next sample. Train the model for many iteration and epochs in the training dataset. Plot the error of the model as a line graph for training iterations.
6. **Testing:** Test the model with the test data and compute the mean squared error (MSE) for test data. Explore different preprocessing strategies on how to improve the performance of the model in testing set.
7. **Playing with the Model:** You can try different strategies to see whether testing error comes down or not. Strategies can be different 1. Initialization of parameters 2. Encoding of features, 3. removal of some features, 4. normalization methods, 5. Shuffling of training samples. Check the model error for the testing data for each setup.

Reference: Algorithm can be studied from https://machinelearningmastery.com/linear-regression-tutorial-using-gradient-descent-for-machine-learning/. One hot encoding https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as Numpy, Pandas

Submission: Submit your files in **Single ipython Notebook** in LMS before **Sunday 11th Aug, 11.59 pm.**

Marking: Marking is based on both **performance during the lab hours** as well as **complete submission in LMS**.

Important Note: Please feel free to think out of the box by exploring all the possibilities in the web. Objective of any assignment is only to improve your learning experience, not just about getting output!