Assignment-2: Linear and Ridge Regression

Problem Statement:

A trading company X needs to predict the purchasing power of consumers in a given city, so as to decide the company's investment in that city. The company has some historical data that has different attributes of various cities and corresponding purchasing power of consumers (more details in Dataset section). Develop linear and ridge regression methods to help the company in predicting purchasing powers of the customers.

Implementation: [2+4+2+2=10]

- Exploratory data analysis and Feature scaling
- Implementation of closed form solution approach towards linear regression [LIN_MODEL_CLOSED]
- Implementation of gradient descent approach towards linear regression [LIN_MODEL_GRAD] with Minibatch
- Implementation of gradient descent approach towards linear regression with regularization (ridge regression) with Minibatch [LIN_MODEL_RIDGE]

**__Implement [LIN_MODEL_CLOSED] from scratch__. You may make use of the numpy library to perform matrix operations.

** For implementation of [LIN_MODEL_GRAD] and [LIN_MODEL_RIDGE], you may use the scikit-learn library.

**In general, you may use libraries to process and handle data.

**Experiment 3, 4 and 5 should be done with feature scaled data.

**For training [LIN_MODEL_GRAD] and [LIN_MODEL_RIDGE], use minibatch size of 256 and total no of epochs 50.

**Performance Metric to be used for evaluating the models is Mean Square Error (MSE)

Experiments: [2+2+3+3+3=13]

The dataset will be split into Train:Validation:Test with 60:20:20 ratio.

1. **Experiment 1:** EDA: Show the distribution of the features and their pair-wise correlation

2. **Experiment 2:** Report and compare performance (MSE) of [LIN_MODEL_CLOSED] with and without feature scaling. Performance should be computed with 20% held out test data.

3. **Experiment 3:** Report the effect of varying learning rate in [LIN_MODEL_GRAD] on validation data. Choose learning rate values from [1e-5, 1e-4, 1e-3, 1e-2, 0.05, 0.1]. Plot MSE vs learning rate. Find the best value of the hyperparameter learning rate.

4. **Experiment 4:** Report the effect of varying ridge regression hyperparameter (alpha) in [LIN_MODEL_RIDGE] on validation data. Use the learning rate found in Experiment 1. Choose alpha values from the range 0.0 to 1.0 with an increment of 0.1. Plot MSE vs alpha. Find the best value of the hyperparameter alpha.

5. **Experiment 5:**
    a. Derive performance of the three models [LIN_MODEL_CLOSED], [LIN_MODEL_GRAD] and [LIN_MODEL_RIDGE] with the optimal hyperparameters found in the earlier experiments. Use held out 20% data (test) for estimating the performance measured with MSE.
    b. Report your observations with appropriate explanations.

Datasets:

This dataset comprises sales transactions captured at a retail store.You can find the dataset here.

Data Overview:

- User_ID: Unique ID of the user
- Product_ID: Unique ID of the product.
- Gender: indicates the gender of the person making the transaction.
- Age: indicates the age group of the person making the transaction.
- Occupation: shows the occupation of the user, already labelled with numbers 0 to 20 (e.g., 0 may correspond to banker, 1 may correspond to doctor, etc)
- City_Category: User's living city category. Cities are categorised into 3 different categories 'A', 'B' and 'C'.
- Stay_In_Current_City_Years: Indicates how long the user has lived in this city.
- Marital_Status: is 0 if the user is not married and 1 otherwise.
- Product_Category_1 to _3: Category of the product. All 3 are already labelled with numbers.
- Purchase: Purchase amount.

Handling NaN (Missing) Values:

1. Features with >50% NaN values can be dropped
2. For numerical features, the NaN values can be imputed with the following methods:
    a. Mean
    b. Median
    c. Forward fill (fill missing value with the previous data point)
    d. Backward fill (fill missing value with the next data point)
3. For categorical features:
    a. Mode
    b. Forward fill
    c. Backward fill

Problem: Predict purchase amount

Submission:

A .zip file containing the python source code and a PDF report file. The final name should follow the template: <Assign-No>_<Your Roll No>.zip. For example, if your roll no is 15CE30021, the filename for Assignment 2 will be: `Assign-2_15ce30021.zip`

1. A single python code (.py) containing the implementations of the models and experiments with comments at function level. The first two lines should contain your name and roll no.

2. A report [PDF] containing **[2 points]**
    a. Experiment 1: EDA Plots and correlation heatmap
    b. Experiment 2: Performance values for models with and without feature scaling.
    c. Experiment 3: MSE vs learning rate plot. Also the best choice for the learning rate value.
    d. Experiment 4: MSE vs alpha plot. Also the best choice for the alpha value.
    e. Experiment 5: Table capturing comparison among [LIN_MODEL_CLOSED], [LIN_MODEL_GRAD] and [LIN_MODEL_RIDGE] followed by your observations

Responsible TAs:

Please write to the following TAs for any doubt or clarification regarding Assignment 2
Aditya Chawla: adityachawla700@gmail.com
Ramishetti Sai sreeja: ramishetti@iitkgp.ac.in

Deadline:

The deadline for submission is **26th August (Saturday), 11:55PM, IST**. Irrespective of the time in your device, once submission in moodle is closed, no request for submission post-deadline will be entertained. No email submission will be considered. So, it is suggested that you start submitting the solution at least one hour before the deadline.