CS 4320: Machine Learning

Assignment : Complex Linear Regression

In this assignment, you will use linear regression to fit a model to a collection of data. Your goal is to minimize the MSE on a set of test data. The data is more complex that the last assignment. You will need to use a data processing pipeline.

Use your personal data set available on Canvas in the `regression-2` folder.

Explore and analyze this data as you did in previous assignments. Include the plots and analysis in your report.

Design and use a data processing pipeline that will scale the data into a better range, and will add derived data features.

Fit a linear regression model to the data. Note this means find the parameters.

It is expected that you will use the `sklearn.linear_model.SGDRegressor` to find the best model.

You will need to record the MSE found on the training data, and the MSE found on the testing data.

Required Steps

- Download your data.
- Explore and analyze your data.
- Split the data 80%/20%, for training/testing.
- Write (or modify) a Python program using sklearn to process and fit the training data to a SGDRegressor model.
- Report the MSE loss obtained for your best model on the training data.
- Report the MSE loss obtained for your best model on the testing data.
- Report the linear model coefficients found.
- Report *your* model function.
- Include your analysis of the quality of your model's fit to the data.
- Commit and push your code in the git repository.
- Submit the report (as PDF) to Canvas.