Exercise – 3: DS203-2024-S2

Submissions due by: February 16, 2025, 11:55 pm

This exercise is aimed at:

1. Getting introduced to and running various regression algorithms on a given data set and understanding their relative characteristics, performance, and advantages.
2. Calculating, effectively documenting, and understanding various regression metrics and developing an approach towards effectively using them.
3. Creating and consolidating multiple plots with the aim of comparing and contrasting the results of the regression algorithms.
4. Get introduced to the relevant ML functions of the Python library: ***sklearn***

Perform the following:

1. Review the Jupyter Notebook E3.ipynb (and data provided as E3-MLR3.xlsx) and:
    A. Create a summary of the code therein.
    B. Are there any learnings from this code that you wish to highlight?
2. Review the **sklearn** documentation for each **sklearn** function used in the Notebook (eg. PolyNomialFeatures, LinearRegression, mean_squared_error, etc.) and create a description of each to explain, to yourself, the functionality, the input parameters, and the outputs generated. Present this in the form of a two-column Table (Function name | Description).
3. You have been given 6 models to study : Tree based (Random Forest and XGBoost), non parametric (knn) and parametric (Linear Regression, SVR, and Neural Network). Each of these has its own advantages and disadvantages. We will study some aspects of the models one by one, keeping the other parameters fixed.
    A. Run the models with degree of the **PolynomialFeatures()** method fixed at 1,6 and 10, and show the regression fit line on the train set for the three settings. Which methods seem to be affected by this change the most, visually ? Support your answer by calculating the train MSE at degrees 1 and 10, and tabulate the difference between them (for the six models)?
    B. **Standardization** : This is a preprocessing step that rescales features to have zero mean and unit variance, ensuring that each feature contributes equally to the model:

$$z = \frac{x_i - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the feature.
Fix the degree back to 2. Comment the lines carrying out standardization of the data, and tabulate the MSEs (both train and test) of the six models, with and without standardization. Which models are affected the most, and which are not affected ?
    C. Fix the degree to 1. Run the Neural Network with 1, 2 and 3 hidden layers (each layer having 10 neurons) and document the train and test MSEs in this case. What do you observe ?
    D. Fix the degree to 6. Uncomment the line that adds outliers to the dataset. Now run the six models, and show the regression fit line on the test set (for the six models) before and after adding outliers. What models seem more robust to outliers (visually) ? What does this tell you about the choice of models in applications where noisy data is expected ?
    E. What are the limitations of the non-parametric methods ?
    F. Given the results, should LinearRegression be used at all? Why, when? Justify your answer.

4. In step ‘2’ you have already reviewed the important parameters and outputs related to the regression methods. Select 2 methods, vary the important parameters (visit the documentation of the functions to review function calls), and observe how the outputs change (eg. see the function calls for SVR and MLPRegressor for examples). Document the outcomes of your experiments.

5. Review sklearn documentation to understand and experiment with two more regression methods, in addition to the ones listed above. Fit these models on the data, and document the fitted lines and residuals (on Train and Test sets) as well as Train/Test metrics.

6. List your major learnings from this part of the exercise.

7. Create a single document by neatly capturing all the above analyses and comments in a well formatted document .

8. Convert the document into a PDF. Name of the PDF should be **E3-your-roll-number.pdf**. Upload it to the assignment submission point E3.

oooOOOooo