Exercise – 3: DS203-2023-Sem2

Submissions due by: Feb 11, 2024, 11:55pm

This exercise is aimed at:

- Getting introduced to and running various regression algorithms on a given data set and understanding their relative characteristics, performance, and advantages.
- Calculating, effectively documenting, and understanding various regression metrics and developing an approach towards effectively using them.
- Creating and consolidating multiple plots with the aim to compare and contrast the results of regression algorithms
- Get introduced to the relevant functions of the Python library: **sklearn**
- (Optional) Using LLM tools like ChatGPT to generate code

Perform the following:

1. Review the Jupyter Notebook E3.ipynb 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. Generate outputs by setting **degree=1, degree=3, degree=6, degree=10,** in the **PolynomialFeatures** function used in E3.ipynb and analyze and record your observations and conclusions:
    a. Review the **augmented_data.csv** file generated in each case and document your observations.
    b. Create an overall qualitative summary based on a review and analysis of the Figures generated.
    c. Summarize and explain the variations in the metrics across regression methods for a given **degree** (ie. a given set of polynomial features). Cover both, train and test, metrics, and compare them.
    d. Summarize and explain the variations in the metrics across **degrees** for a given regression method. Cover both, train, and test metrics, and compare them.
    e. When **degree = 1** which method(s) result in acceptable regression models? Why?
    f. When **degree = 6** which method(s) result in acceptable regression models? Why?
    g. As the value of **degree** is increased to 10 which regression methods show the most impact? Why?
    h. Why do Non-parametric methods like KNN / Tree based methods generate good results even without feature engineering?
    i. What are the limitations of the non-parametric methods?
    j. 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 2-3 methods, vary the important parameters, and observe how the outputs change (eg. see the function calls for SVR and MLPRegressor). Document the outcomes of your experiments.
5. Review sklearn documentation to understand and experiment with a few more (2-3) regression methods and document the outcomes of your experiments.
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.

Optional:

- Most of the code in E3.ipynb was generated using ChatGPT. Manual intervention was required to correct wrongly generated portions. Create an appropriate set of prompts to re-create the code. Debug the code to ensure that it works. Include the prompts as well the generated code and its output in your report.

oooOOOooo