DATA 220 Mathematical Methods for Data Analytics – Homework – 5

Deadline – 11.59 PM – 12/08/2023

20 points

For the following tasks use the superconductivity_data (uploaded in the Canvas) that contains 81 features extracted from 21263 superconductors along with the critical temperature in the 82nd column (target). Details of the data, you can find it here.

Split the dataset into training and test sets of 80:20 ratio (use **random_seed = 2023**) and test_size = 0.20. You must train the Multiple Linear Regression model using the training data and compute R^2 and MSE using the test dataset.

Problem 1 (Coding): Apply Multiple Linear Regression (MLR) using normal (least square solution). You must not use any direct or in-built package for MLR.

a. **(5 pts)** Check the five assumptions (mentioned in the classroom) of MLR (use – training dataset) and proper interpretation – why the assumptions are met or not
b. **(2 Pts)** Derive the normal equation for linear regression.
c. **(2 Pts)** Apply the standardization technique to all 81 extracted features to ensure that all features have a consistent scale. Utilize **'fit_transform'** for the training data and **'transform'** for the test data to prevent data leakage.
d. **(6 pts)** Find optimal values of intercept and coefficients using the normal equation of the linear regression ($\boldsymbol{b_{opt}} = \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T y}$) using the training data. To avoid inverse matrix error, you may use pseudo inverse (np.ling.pinv)
e. (**2 pts**) Find ($\hat{y}$) (predict for each datapoints of x_test) – show in dataframe – making two columns: y_actual & $\hat{y}$_predict
f. (**2 +1 = 3 pts**) Finally, for the test dataset:
    a. Calculate coefficient of determination (R^2) and interpret the result
    b. Find MSE (mean of sum of squares of error (residual)

You are required to submit:

1. An MS/PDF/Scanned document:
    a. Include all the steps of your calculations.
    b. Attach screenshots of the code output.
2. Source code:
    a. Python (Jupyter Notebook)
    b. Ensure it is well-organized with comments and proper indentation.

- Failure to submit the source code will result in a deduction of 5 points.

- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.