Indian Institute of Technology Madras

Web MTech Industrial AI

ID5002W: Industrial AI Lab

End semester examination

Date: 08/04/2023 Time: 3.5 hours Total Points: 100

Instructions

1. The students can use their own codes and notes. Taking help from others (live beings) through any means will be considered malpractice. It will be reported to the authority.
2. The late submissions will be fined as per the rule.
3. All the codes must be uploaded as a single .zip file with your roll number as the file name.

Problem

Q1 A team of researchers has collected data based on superconductor materials to predict critical temperatures. The file shared 'EndsemDataset.zip' has the required data. Students must work on the folder that is named after their respective roll numbers. All datasets must be split into train and test sets for training and validation respectively with train size = 80% of all the samples and random state = 42.

(a) Preprocess the dataset by checking for NaN values and standardizing the dataset. [Marks: 5]

(b) Fit a linear regression model to the training dataset. Perform regression diagnostics and report on the validity of ordinary least-squares assumptions. Generate necessary plots as a part of diagnostics. [Marks: 10]

(c) Fit Principal Components Regression (PCR) and Partial Least-Squares (PLS) model to the training dataset. Report on the optimal number of components in both models based on R^2. Generate necessary plots to compare the performance of different models with the components. [Marks: 10]

(d) Fit a LASSO and Ridge regressor to the training dataset. Report on the optimal value of regularization term α based on R^2. [Marks: 10]

(e) Fit a support vector regressor (SVR) using the polynomial kernel and report on the optimal value for the degree and C based on R^2. [Marks: 20]

(f) Implement a neural network regression model in PyTorch, containing two hidden layers with ReLU activation and a dropout layer between them. Train the model for 10 epochs. [Marks: 20]

(g) Visualize the performance of all the models in terms of R^2 and $RMSE$ values for the validation dataset. [Marks: 10]

(h) Choose the two best linear regression models and the SVR and Neural Network nonlinear regression models based on the previous subsection (g). [Mark:5]

(i) Use 5-fold cross-validation for these four models and compare these models using the $RMSE$ values. Suggest the best model for this data set [Marks: 10]