CS F320 Foundations of Data Science

Assignment-2

Submission Time & Date: 1PM, 10th Nov 2022

Instructions

- This assignment is a coding project and is expected to be done in groups. Each group can contain at most three members. Make sure that all members in the group are registered to this course.
- This assignment is expected to be done in Python using standard libraries like NumPy, Pandas and Matplotlib/Seaborn only. Try writing a modularized and vectorized code, avoid hardcoding. Jupyter Notebook can be used. Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.
- All deliverable items (ex. .py files, .ipynb files, reports, images) should be put together in a single .zip file. Rename this file as A1_<id-of-first-member>_ <id-of-second-member>_ <id-of-third-member> before submission.
- Submit the zip file on CMS on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions/exemptions will be given. The demos for this assignment will be held later, the date of which shall be conveyed to you. All group members are expected to be present during the demo.

Problem Statement

The given dataset is used to predict application energy usage (in Wh) (in the Application column of the dataset) using various attributes such as temperatures, relative humidity various rooms in the house, windspeed, visibility, etc (in the columns: T1,RH_1,T2,RH_2,,, Visibility,rv1,rv2 of the dataset).

Dataset: https://drive.google.com/file/d/1hkc5d67JY1eHQ6EsidRMe3rPS3v3LeGs/view?usp=share_link

You are required to build a regression model to predict the application energy usage. You can perform data preprocessing like standardization or min-max scaling, and do a 80:20/90:10 train-test split.

2-A Correlation coefficients and Principal Component Analysis

i) Select a set of 1,2,3,...26 features by checking which of these features show a maximum linear relationship with the target attribute by using Pearson correlation coefficient. Use these features to build the regression model and find the training and testing error for each set of features.

ii) Perform Principal component analysis to select a set of 1,2,3,...26 principal components. Use these features to build the regression model and find the training and testing error for each set of features. Also find the percentage of variance captured by each of the feature sets.

2-B Greedy Forward and Backward Feature Selection

Perform

i) greedy forward feature selection and

ii) greedy backward feature selection

to find the subset of features that make the optimal regression model. Find the minimum training and testing error of the optimal model (using 1,2,3,...26 features).

2-C Comparative Analysis

Perform comparative analysis of the best models obtained using each of the feature selection techniques (from 2-A and 2-B) and the performance with the regression model using all the features.

What needs to be documented

- Give a brief description of the model, and how you implemented each of the feature selection techniques.
- Tabulate the training and testing error values when 1,2,....26 features are used (obtained using each of the feature selection technique).
- The best model obtained using each of the feature selection technique
- The results of the best models, the overall best model and the features/eigen values [whichever applicable] being used by each of these models.