Hyperparameter Tuning and Cross Validation

DS4400: Machine Learning I
Bilal Ahmed

Ridge Regression – Assignment 1 (Part-II)

Due By: 7/17/2023 11:59 PM EST

Introduction

In this assignment, we will implement hyperparameter tuning using grid search. Additionally, we will use random k-fold cross validation to estimate the error for our ridge regression implementation. You [illegible] submit your solutions as python programs [illegible] submit a Jupyter notebook. If using Ju[illegible] f.

Datasets

- ***Concrete***: The [illegible] ets repository and has nine num[illegible] ne compressive strength of di[illegible] column is named 'strength' whi[illegible] to ridge regression.

Instructions

1. Read the conc[illegible] a dataframe using pandas [illegible]
2. Read the sciki[illegible] Pipeline (here), and Ri[illegible]
3. Create a pipel[illegible] ict / train a ridge regressor.
4. ***Hyperparameter Tuning:*** To estimate the best value of alpha (lambda in the course notes) for our ridge regression model, we will use grid search. Scikit-learn has a built-in method for doing grid search called GridSearchCV (doc). Additionally, we also need to implement cross validation for estimating model performance at each grid point. To this end, we will use k-fold cross validation that is implemented in scikit-learn as KFold (doc).
    a. Create a KFold object with ***k***=5 (for five fold cross validation), setting random_state=44 and shuffle=True. What do these parameters signify and what is their importance for estimating model performance? (*5 points*)

b. Perform grid search using the k-fold object in the previous step optimizing mean squared error (MSE).
    i. Use a grid with alpha values = [0, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 50.0]
    ii. Report the best value of the alpha parameter and the best score for the concrete dataset. (Note: Scikit treats higher value of scores as better model performance and to minimize the error the negative value of the error should be used for scoring (see here)) (*15 points*)

5. Estimating MSE for the dataset: Using the optimal value of alpha that we obtained using grid search, we will now estimate the MSE of our ridge regressor on the concrete dataset.
    a. For k = [5, 10] set up a KFold object similar to the settings in part 4a. (These
    b. Use sc e model perfor for both k=5 and k=

6. Using only a s nates and a standard way sing different partitions (rar implementati lement 5a and 5b using t te the same and set the n_
    a. Impler
    b. Why w *points*)
    c. How w the dataset both ir of input featur