CS 622 – Homework 3

Shen Chan Huang

March 24 2023

1 Python Version and Libraries Used

- Python Version: 3.8.13
- Pandas
- Numpy

2 Datasets

The dataset we are working on is the 'auto-mpg.data'. Since it is not a csv file, we use pd.read_fwf. I used the following to load the dataset with the given column names.

col_names = ['mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'model year', 'origin', 'car name']

X = pd.read_fwf(filename, names = col_names)

Since the dataset had missing values, data imputation of some sort had to be used. I chose to impute data using the mean of the feature was used for filling the missing values. Of course, before imputing, I had to split the dataset into training and testing set.

3 Description

To run the script, go to cmd and use "python -u Huang_HW3_CS622.py"

In this assignment, I implemented linear regression from scratch and used 10-fold cross validation.

The problem is to model the correlation the dependent variable 'mpg' and the 7 independent variables:

$$Y = X \cdot \vec{b}, \tag{1}$$

where $\vec{b}$ is the unknown vector of coefficients. To optimize it, we use the formula

$$\vec{b} = (X'X)^{-1}X'Y. \tag{2}$$

To calculate the RMSE, the professor's formula is

$$RMSE = \sqrt{\sum_{i=1}^{N}(y_i - \mathbf{X}_i \cdot \mathbf{b})^2}. \tag{3}$$

First, I loaded the data set and shuffled it. I wrote functions to perform the following:

1. Split the full dataset into training and testing set given an N for N-fold cross validation.
2. Impute data by filling in the mean.
3. Linear Regression.
4. Normalize data and normalize test data.

5. Compute the RMSE.

6. Compute total error.

To ensure randomization, I first shuffled the dataset. Then pass it to (1.) to split and obtain a training and testing set.

Next, since there are missing values, I imputed the missing values with the mean, within the training and testing set separately.

I then normalized the training set and obtained the normalizing parameters (mean and standard deviation, to normalize the test data later).

Then I appended a column of ones for setting up the matrix to perform linear regression.

I calculated the coefficient of the linear regression, the RMSE, and the R^2 value each time during cross validation.

4 Experimental Results

	cylinders	displacement	horsepower	weight	acceleration	model year	origin	RMSE
Fold 1	−0.0673	0.2077	−0.0498	−0.7083	0.0344	0.3618	0.1524	2.6198
Fold 2	−0.0908	0.294	−0.0733	−0.735	0.0474	0.3652	0.1417	2.418
Fold 3	−0.1362	0.2816	−0.0432	−0.7501	0.054	0.3476	0.112	3.1042
Fold 4	−0.0732	0.2216	−0.0404	−0.7319	0.0362	0.3616	0.1426	2.8051
Fold 5	−0.1091	0.3113	−0.0685	−0.7433	0.0293	0.3672	0.1506	2.0094
Fold 6	−0.0496	0.1368	−0.0642	−0.6664	0.0086	0.3583	0.1486	3.1907
Fold 7	−0.1101	0.263	−0.0266	−0.7481	0.038	0.3577	0.1547	3.3534
Fold 8	−0.0727	0.2286	−0.0512	−0.7238	0.0393	0.3619	0.1427	2.1895
Fold 9	−0.1141	0.2971	−0.0872	−0.7292	0.0303	0.3483	0.1554	3.1671
Fold 10	−0.0901	0.2822	−0.0579	−0.7413	0.0422	0.3494	0.1539	2.1603

Table 1: Coefficients of seven independent variables and RMSE.

	R^2
Fold 1	$R^2 = 0.8134309367892417$
Fold 2	$R^2 = 0.7497888351588635$
Fold 3	$R^2 = 0.7990822959353385$
Fold 4	$R^2 = 0.7516948400135615$
Fold 5	$R^2 = 0.9103504917061611$
Fold 6	$R^2 = 0.8025174489758431$
Fold 7	$R^2 = 0.7569253566054717$
Fold 8	$R^2 = 0.8568706158111572$
Fold 9	$R^2 = 0.7723414863150336$
Fold 10	$R^2 = 0.8644754761165729$

Table 2: The R^2 value of each fold.