Derek Yadgaroff
LNU ID: dy222ax
Derek.chase84@gmail.com
April 22, 2020
2DV516 – Intro to Machine Learning, Distance

Assignment 2, Exercise 7

Report

Table of Contents

Summary **3**

A look at the data **3**

Numerical *3*

Categorical *3*

Detailed Steps Taken **4**

Load data *4*

Categorize data *4*

Models and Method *4*

Linear Regression 4

Lasso and Ridge 5

Elastic Net Regression 5

Performance **6**

Summary

In this exercise, we are given a data set and asked to use the knowledge that we have accumulate throughout this course to process and analyze the data and report on it:

1. Find a good linear regression model by evaluating each of these variants:
    - Standard linear regression
    - Lasso regression
    - Ridge regression
    - Elastic net regression
2. Extend the model through the following optimizations and refinements:
    - optimizing for the regularization hyperparameter λ (this is called alpha in sklearn)
    - add transformed features (higher degree polynomial)
    - combine transformed features and regularization
3. Several variables are categorical and need to be transformed
4. Determine how to evaluate each model and decide which is best
5. Present findings and arguments

A look at the data

We have been provided 8 variables. The first 7 are data points collected about the person, and the 8th variable is the amount of charges for that person.

Numerical

The following variables are given as numerical data and will be treated as numerical data:

- Age
- BMI
- Children
- Shoe Size
- Charges

Categorical

The following variables are categorical given as strings and need to be preprocessed.

- Smoker – given as ‘yes’ or ‘no’. We will encode this as 1 and 0 respectively (binary encoding)
- Sex – given as one of two labels, ‘male’ or ‘female’. We will encode this as 1 and 0 respectively

- Region – given as one of four labels, ‘southeast’, ‘southwest’, ‘northeast’, ‘northwest’. We will use the LabelEncoder from scikit-learn to transform these to int variables 0, 1, 2, 3, 4. OneHotEncoder was also considered and tested but did not show any big difference so I kept the LabelEncoder.

Detailed Steps Taken

I have taken the following steps as you can see in the code.

Load data

The CSV file did not read in correctly using numpy loadtxt. I therefore used the built in CSV class to read in the CSV. I ignored the header, and since the data has extra lines in between rows, I ignored empty lines.

Categorize data

As mentioned earlier, I converted categorical data into int representations. I then created a new data object containing just the finalized X and y data

Models and Method

I am using the the SKLearn models and methods:

Linear Regression

This is the standard Linear Regression model. I ran two tests:

<u>Using train test split</u>

- I used train_test_split to separate the X and y into test and train data.
- I used a test size of .2
- I tested this model with X and with a higher polynomial of X**2

The MSE for both of these was much higher than the models that used cross validation.
X degree 1: MSE=31874408.807573073
X degree 2: MSE=31727735.240174975

<u>Using Cross Validation</u>

- I used cross validation with K values 3, 4, 5
- I also tested these with X and a higher polynomial of X**2

X degree 1
K=3, MSE= 6107.278617124394
K=4, MSE= 6067.31440225263
K=5, MSE= 6066.272489453042

X degree 2:

K=3, MSE=6088.121744780007
K=4, MSE=6029.402770763166
K=5, MSE=6031.827946008691

Lasso and Ridge

These are two models that use regularization to produce better fit curves and reduce overfitting.

- I was able to test X and X**2 polynomial with these
- I used GridSearchCV to find the best value of alpha that produces the lowest MSE
- I tested alphas in the range from .001 to 100, with a step of 100. This was necessary to be able to compute the X**2 polynomial. A higher step did not produce very different results.
- I also iterated over multiple values of k

X degree 1

Lasso Regression
K=3, Optimal MSE=6102.147759969593, Optimal Alpha=73.73763636363637
K=4, Optimal MSE=6067.314521793873, Optimal Alpha=0.001
K=5, Optimal MSE=6066.263709496919, Optimal Alpha=2.021181818181818

Ridge Regression
K=3, Optimal MSE=6107.2786804955795, Optimal Alpha=0.001
K=4, Optimal MSE=6067.314383207093, Optimal Alpha=0.001
K=5, Optimal MSE=6066.272547357652, Optimal Alpha=0.001

X degree 2

Lasso Regression
K=3, Optimal MSE=6084.606565325385, Optimal Alpha=44.445
K=4, Optimal MSE=6028.77549213139, Optimal Alpha=22.223000000000003
K=5, Optimal MSE=6031.785940504825, Optimal Alpha=2.021181818181818

Ridge Regression
K=3, Optimal MSE=6088.121624750279, Optimal Alpha=0.001
K=4, Optimal MSE=6029.341676057231, Optimal Alpha=1.011090909090909
K=5, Optimal MSE=6031.670197653992, Optimal Alpha=1.011090909090909

Elastic Net Regression

This method is a combination of ridge and lasso. It is computationally more expensive so I was not able to test X**2 degree in a stable way and I will not use those results here.

- I used GridSearchCV to iterate over alphas and also to iterate over the l1_ratio parameter
- I also iterated over multiple values of 7

K=3, Optimal MSE=6107.279220687361, Optimal Alpha=0.001, Optimal l1_ratio=0.99
K=4, Optimal MSE=6067.3142127659685, Optimal Alpha=0.001, Optimal l1_ratio=0.963030303
K=5, Optimal MSE=6066.273139272747, Optimal Alpha=0.001, Optimal l1_ratio=0.99}

Performance

The following are the average MSE values for each test:

Average MSE for Standard Regression with cross validation: 6065.036328
Average MSE for Lasso: 6063.482332
Average MSE for Ridge: 6064.999852
Average MSE for Elastic Net: 6080.288858

We can see that the Lasso Regression using cross validation produced the lowest MSE. However, all methods were relatively close in their minimal MSE value. The Elastic Net could not compute a higher degree polynomial for K=5, so I omitted those results.

I see that finetuning the parameters such as alpha, K, and l1_ratio are necessary for finding an optimal model. I also see that higher order polynomials can also improve the MSE value. I will look into further code optimizations or environments so that I am able to test with higher polynomials in the future.

For this exercise, I will say that because Lasso has the lowest average MSE over the test iterations, that it is the best model to use for this data set. However, I understand that each method has strengths and weaknesses and a similar selection process should be used for future data analysis.