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

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.

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 a train/test split with test size .2, and a cross_val_score. I was able to test this with degrees 1, 2, 3 and 4

Lasso and Ridge

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

I was able to test these with degrees 1, 2, and 3. I used Grid Search with a K fold (cv) value of 5.

Elastic Net Regression

This method is a combination of ridge and lasso. I was able to test this with degrees 1, 2 and 3. I used Grid Search with a K fold (cv) value of 5.

Performance

The following are the findings ordered by MSE values, ascending:

1	method	algorithm	degree	score	alpha	l1
2	Grid Search	Lasso Regression	2	6030.91768	31.6330204	
3	Grid Search	Ridge Regression	2	6031.31037	15.3068163	
4	Cross Validation	Linear Regression	2	6031.83483		
5	Grid Search	Lasso Regression	3	6037.6035	50	
6	Cross Validation	Linear Regression	4	6043.53828		
7	Grid Search	Ridge Regression	3	6049.811	23.4699184	
8	Cross Validation	Linear Regression	3	6056.41828		
9	Grid Search	Lasso Regression	1	6066.23891	4.08255102	
10	Grid Search	Ridge Regression	1	6066.2533	2.04177551	
11	Grid Search	ElasticNet Regression	1	6066.2548	0.001	0.1
12	Grid Search	ElasticNet Regression	2	6066.2548	0.001	0.1
13	Grid Search	ElasticNet Regression	3	6066.2548	0.001	0.1
14	Cross Validation	Linear Regression	1	6066.27249		
15	Train/Test Split	Linear Regression	3	31609217.8		
16	Train/Test Split	Linear Regression	2	31727725.1		
17	Train/Test Split	Linear Regression	4	31833472.3		
18	Train/Test Split	Linear Regression	1	31874408.8		

- We can see that the Lasso regression with a polynomial of 2 degrees performed the best and had the smallest MSE value of 6030.917
- The top performing algorithms were all of polynomial 2
- The elasticNet algorithm was the second worst performer
- The Linear Regrssion without cross validation was the worst performer

We can see that the Lasso Regression using cross validation produced the lowest MSE. However, the top three performing methods were relatively close in their minimal MSE value, and, all of the models that used Cross validation were also relatively close.

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 do not always provide a better MSE.

I was not able to compute ElasticNet at a degree of 4 without further increasing the max iterations, so I omitted it. I see that high degree polynomials will obviously become more expensive to compute.

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.