Computational Machine Learning COSC2793

Assignment 1 - Report

INTRODUCTION

This report present the use of multiple regression model to predict the life expectancy of a newborn based on several features related to the region which he/she was born, in other words making an ultimate judgement of how to approximate the unknown target function. In terms of principle of Ockham's Razor the best approach is to find the simplest hypothesis that is consistent with the experience.

EXAMINING THE DATASET

Both the given train.csv and test.csv data is loaded into the data frame. Using head function we determine a sample of the data. We see the presence of ID attribute which will not facilitate to application of suitable regression models this we drop that column. Now to identify obvious skews, to obtain a better understanding of what is going on and to identify if the data satisfies the assumptions required for machine learning, we draw a histogram of each attribute. Furthermore a correlation plot to analyse the relation of attribute on target is draw in form of heat map to determine the correlation of other attributes on the target life expectancy. The light colours represent less correlation thus attributes like adult mortality and thinness 1-19 years has less correlation to target variable.

EVALUATION METRICS

The various metrics used to evaluate the results of the prediction are :

- Mean Squared Error(MSE) - The MSE catches even a small error which leads to over-estimation and help us find how bad the model is
- Mean-Absolute-Error(MAE) -The MAE is more vigorous to outliers and does not effect the errors as extremely as MSE. MAE is a linear score which implies all the individual contrasts are weighted similarly. Not suitable where outliers are predominant.
- R^2 - While R-squared provides an estimate of the strength of the relationship between your model and the response variable, it does not provide a formal hypothesis test for this relationship.
- Coefficient of Determination - it is a statistical measure in a regression model that decides the extent of variance in the dependent variable that can be explained by the independent variable

METHODOLOGY

Linear Regression - Multi-Variate Regression

The goal for this methodology was to find the line that pass through all the value point but its too good to be true so target is line of best fit. A line of best fit will have the minimum value of the loss function. In gradient descent we will start from the point , find the slope

and then decide whether to increase or decrease the $\theta 1$ value. If we make the alpha α value too small it will take forever to converge to the end point and if we make it too large it will just keep jumping around. Here we use the Multi Variate Regression since we are searching for multiple factors that affect the life expectancy of a new born.

The X data frame contains all attributes other than the target and using linear regression we find the effect of these features on the life expectancy which is kept in Y. For application of machine learning algorithms the data needs to be splitter into training and testing data. The train_test_split() function is used to split the data into 80% training data and 20% test data for each of the data frame. We cannot directly compare the hypothesis with target function however we can use some unseen data to measure the performance of given hypothesis. Now we train a linear regression model and fit the theta parameters. The training data is fixed and the theta zero and theta one parameters are calculated. Theta zero is the intercept value and theta one is the coefficient .

The evaluation matrices are used and we get the following values .

```
Mean squared error  19.912327082753734
Mean absolute error  3.4281613277841707
Root Mean squared error  4.462323058985502
Coefficient of Determination  0.7675548028524057
```

Regularisation

In regularisation we automatically find which hypothesis space best suits our problem. How we do it is, we come up with a way to let some of the un import things go to zero , it's the perfect answer to over fitting. Feature selection happens implicitly.
Regularisation of loss function has two parts, the smoothness part pushes towards simpler model and the data part combined with high degree of freedom pushes towards over fitting, this integration is balanced with the parameter lambda λ . If lambda λ value is too large it will ignore the data part and will end up giving a straight line parallel to x axis. In hypothesis what we can control is the θ which is the alpha value.

Polynomial Regression

The hypothesis space in polynomial regression takes both linear and non linear data. On increasing the degree the feature space becomes so large that we cannot run algorithms on it anymore. Put polynomial features in code then run linear fit function.
With this many features given a polynomial function with degree 3 will have too many coefficients.

Polynomial regression of degree 2

Using K Fold cross validation - Evaluate then repeat, with different partitions as test set. Choose a model which gives an error which is similar to the average error.
We use the following regularisation weights to evaluate which weight is best for the polynomial regression model. Here the best fit line is a curve unlike the straight line in linear regression. We predict the values for predY using the model we have already trained on trainX. In ridge regression when lama value is set to 0, then the equation becomes like a normal linear regression.

In order to run a loop to find the optimised lambda value, lRegPara value was set to produce 20 values between 0.001 and 0.5 . This array was fed in and k fold iteration of data produced 10 individual unique set of training and testing data and the polynomial regression model with a adjusted values from loop runs over each unique data set of train and test data. We get the training and validation data in an array form and set the variables. The train data is fit in and validation dataset is fed into other variable , we run the algorithm and calculate the mean square error which shows to be 16.846955 and is quite less that the linear regression thus showing the curved line to be a better fit to minimise the loss function and cover larger ground with the hypothesis.

Weight of a model with lasso.

As shown in figure 3 the lasso tool helps us to find the highest weighted feature. It can perform both variable selection and regularisation. Providing optimum feature selection will make the model much easier to interpret and reduces the effect of overfitting. In this case where three re 22 dimension the knowledge about right feature is important. As shown income composition of resource has the highest precedence, status and schooling also seem to effect the target life expectancy .

Polynomial regression of degree 3

We try to increase the degree of polynomial to 3 to check for a better fitting model but the number of coefficients are too large and the Figure shows the end result of it. The hypothesis space is so complex that Machine learning algorithms cannot perform on it. The alpha value provides a straight dip and thus causing a unambiguous choice of lambda variable. Here we can see a high variance thus it will failed for the test data under consideration ,

RESULTS AND CONCLUSION

After taking all the metrics into consideration, the final selected approach decided upon is the ridge polynomial regression of degree 2 and alpha value of 0.03. As seen from Figure 2 the model chose is neither under fitting or over fitting the data thus providing low bias and low variance among all the models we looked into. It is more efficient because it can take both linear and non linear data and the MSE observed is 16.846 which is less than the 19.912 offered by linear regression and thus providing a better model which shows the derived hypothesis (model) generalise well to new data. It fits a wide range of curvature

For parameter settings an appropriate value of alpha is required. To achieve that a loop to put several alpha value is run to check which suits the best. The Figure 1 shows the consistency at first jump, that is between 0.02726316 and 0.05352632. Tuning to values in between them that is 0.02, 0.03 and 0.04 gave us different mean square error. Note this is for 10 iteration, thus the mean of the MSE was taken for the combined 10 iteration . As shown in table 1 the value was least for alpha/ lambda value of 0.03, thus proving it to be best suitable lambda value to hold up against over fitting or under fitting.

APPENDICES

Figure 1

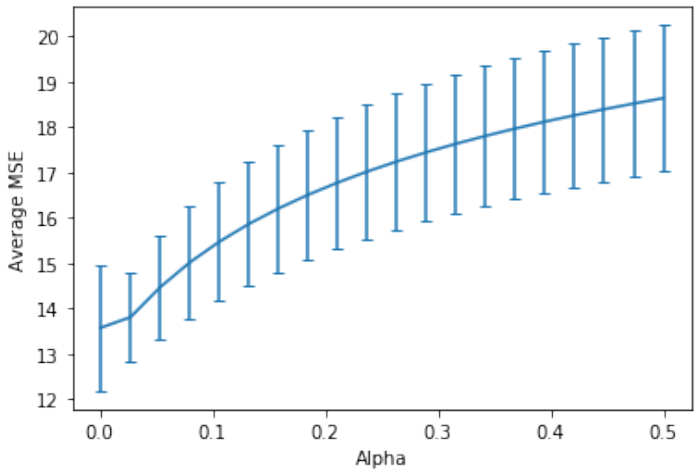

Figure 2

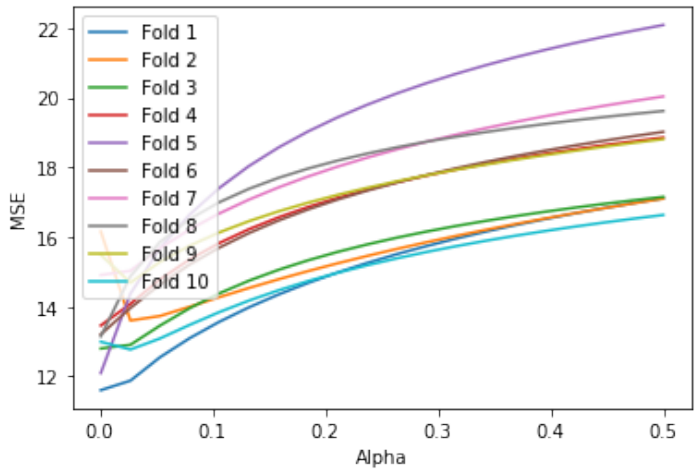

Figure 3

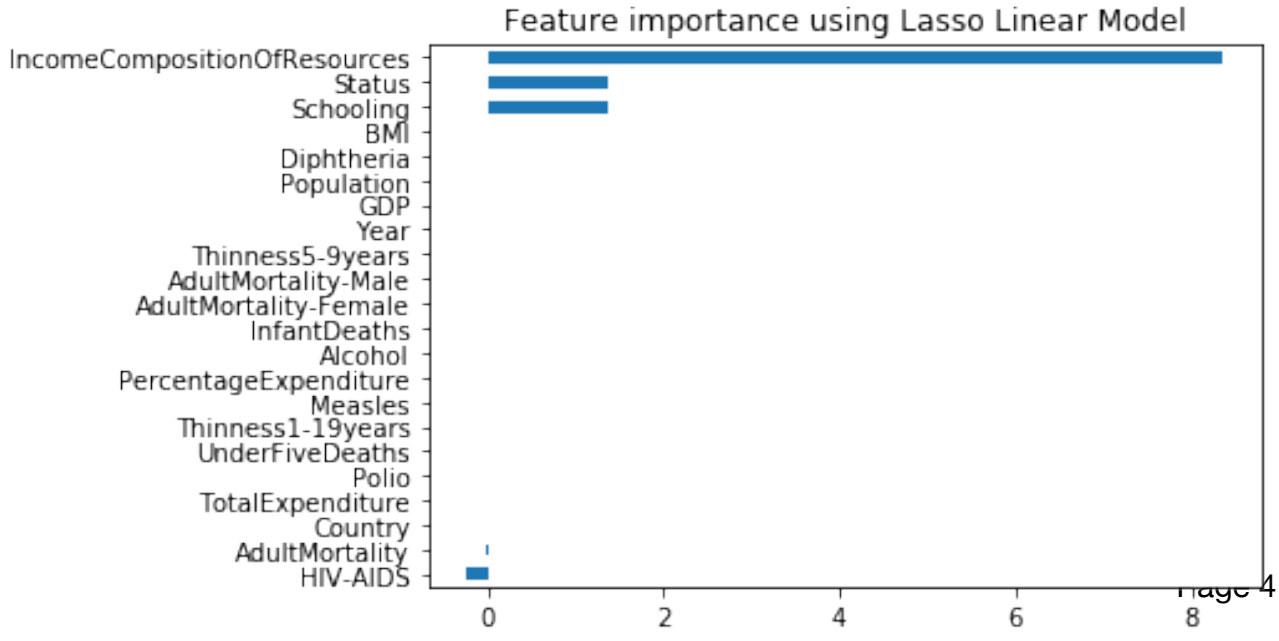

Figure 4

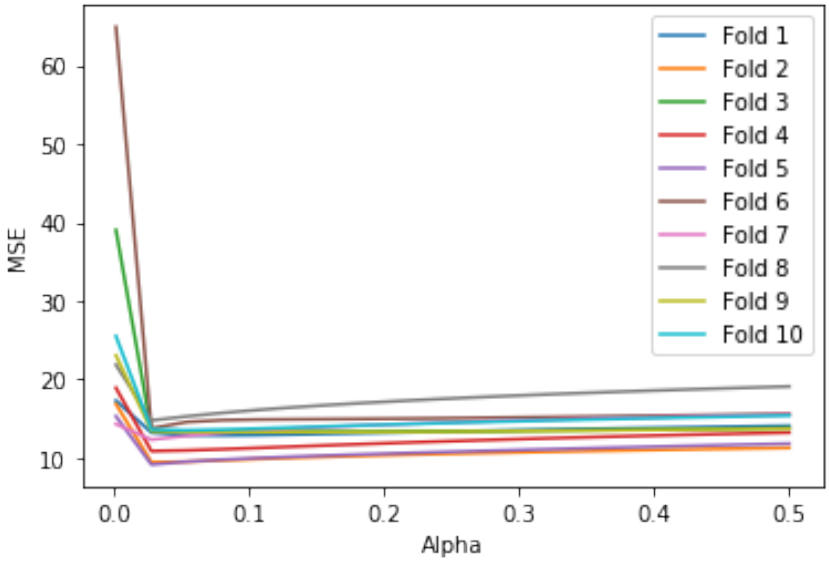

Table 1

	Mean of MSE (iteration)
Alpha = 0.04	14.165
Alpha = 0.03	13.958
Alpha = 0.02	14.231