DTU

Analysis of Parkinson's Data Set

Machine Learning Project 2

Authors

Maximilian Stalzer, Akira-Miranda Adeniran-Lowe - s215141, s215170

Section	Responsible
Regression A	s215170
Regression B	s215170
Classification	s215141
Question 1	s215170
Question 2	s215141
Question 3	s215141
Question 4	s215141

Table 1: Table of responsibilities

Contents

1 Regression Part A

1.1 Feature Transformation

Commonly, Regression is a method in Machine Learning used to predict continuous values. Our aim during this regression was to predict the likelihood that a patient will be positive for Parkinson's Disease based on the variables described in assignment 1. The Data set used for this report did not contain any missing values so no cleaning up was necessary.

Before applying the regression algorithm to the data we had to first standardise the data. This was done by subtracting the mean and then dividing by the standard deviation. This produces a new standardised data matrix of the form below:

$$\hat{X}_{ij} = \frac{X_{ij} - \mu_j}{\hat{s}_j}$$

This leaves each column of the data with a mean of 0 and standard deviation of 1.

1.2 Regularised Linear Regression Model

We will now introduce a regularisation parameter λ which will control the complexity of the regression model. The range of the parameter λ that we have chosen to be reasonable is the series ranging from $\lambda = 10^{-3}$ to $\lambda = 10^{2}$. We then used $K = 10$ Fold Cross-Validation to estimate the generalisation error and plot a graph of this as a function of the Regularisation Error λ.

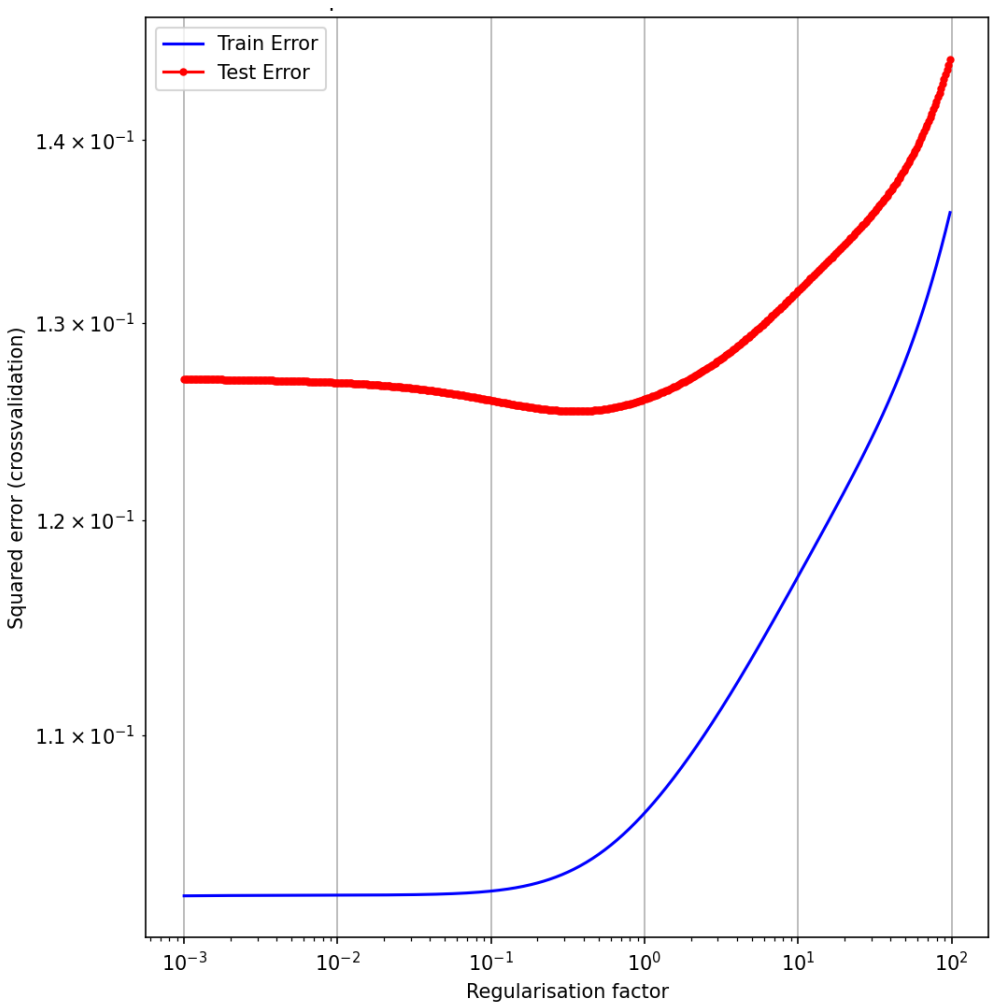

Figure 1.1: Estimated Generalisation Error vs λ

The optimal value for λ can be seen on the above figure as the point with the lowest estimated Generalisation error. This is can be seen on the graph to be $\lambda = 1 \cdot 10^{-0.45}$

1.3 Explanation and Conclusion

Weights in the final fold	
Attribute	Weight
MDVP: Fo(Hz)	0.0
MDVP:Fi(Hz)	-0.04
MDVP:Flo(Hz)	-0.15
MDVP:Jitter(%)	-0.87
MDVP:Jitter(Abs)	0.35
MDVP:RAP	0.28
MDVP:PPQ	-0.22
Jitter:DDP	0.29
MDVP:Shimmer	0.12
MDVP:Shimmer(dB)	0.23
Shimmer:APQ3	-0.06
Shimmer:APQ5	-0.28
MDVP:APQ	0.16
Shimmer:DDA	-0.06
NHR	-0.01
HNR	-0.04
RPDE	-0.08
DFA	0.12
spread1	0.27
spread2	0.11
D2	-0.03
PPE	-0.0

Table 1.1

The effect that each attribute has on the model can be seen by the weight associated with it in Table 1.1. An attribute with a large magnitude will have a large effect on the prediction outcome. An example of an attribute with a large magnitude would be MDVP:Jitter(%). An attribute with a weight of a very small magnitude (for example MDVP:Fo(Hz)) will have little/no effect on the outcome of the prediction.

A weight that is negative in value will result in a prediction for the status that is lower in value. In the context of our data set, this would mean that the patient is less likely to have Parkinson's Disease.

From this information, we can conclude that a MDVP: Jitter(%) and MDVP:Jitter(Abs) have a larger effect on the predictions made by this linear regression model. As The Weight for MDVP:Jitter(%) is largely negative, the model is more likely to predict that the corresponding patient is less likely to have Parkinson's Disease.

2 Regression Part B

2.1 Implementation of Two Level Cross-Validation

We now want to evaluate the efficacy of different regression models. We will in this section test three different types of models; a regularised Linear Regression, an Artificial Neural Network and a Baseline model. And compare them using Two Level Cross-validation

The Regularised Linear Regression model will be the same as in regression part A. We will also be finding the optimal λ value for each fold.

For the Artificial Neural Network the controlling parameter will be the number of hidden layers. Through trial and error we have come to the range [1,3, 5, 7, 10, 20]

For the baseline model there will be no complexity adjustment. The baseline model was made to output the mean of the training set. This can be written as the equation below:

$$f(X)_{baseline} = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} y_i^{train}$$

The baseline model will then be evaluated on the test set as any other model.

2.2 Two Level Cross-Validation Table

Outer Fold	**ANN**		**Linear Regression**		**Baseline**
Index	$h*_i$	E_i^{test}	λ_i^{test}	E_i^{test}	E_i^{test}
1	7	0.23	1e-0.22	0.10	0.24
2	5	0.33	1e-0.12	0.16	0.19
3	3	0.17	1e-0.21	0.16	0.16
4	3	0.15	1e0.21	0.11	0.17
5	1	0.18	1e-2.4	0.13	0.24
6	10	012	1e-0.35	0.11	0.096
7	10	0.43	1e-3.0	0.063	0.12
8	5	0.10	1e-1.13	0.14	0.25
9	3	0.22	1e-0.31	0.13	0.19
10	5	0.11	1e-0.30	0.087	0.19

Table 2.1: Table of responsibilities

From the table it can be seen that the optimal value for $h_i^* = 5$ as it has the lowest corresponding test error. It can also be seen that the optimal value for lambda is 1e-3 as this is the value with the lowest corresponding test error.

2.3 Statistical Evaluation of Model Performance

In order to evaluate the performance of the models, we will use a paired t-test.

To compare the models, we will again use K = 10 fold cross validation to get the test errors for each model. From this the Estimated difference in generalisation error can be computed to be: $Z_i = z_i^A - z_i^B$. Using scipys confidence interval and cdf functions, the method in box 11.3.4 can be followed. The previously stated values have been calulated and written in the table below.

The Null Hypothesis we are testing is: M_A and M_B have equal performance. Where our value for alpha in this case is $\alpha = 0.05$

RLR vs Baseline			ANN vs Baseline			RLR vs ANN		
$\hat{z}$	CI	p-val	$\hat{z}$	CI	p-val	$\hat{z}$	CI	p-val
-0.0690	(-0.105,-0.0326)	0.00203	0.0172	(-0.0751, 0.109)	0.684	0.0862	(0.00347, 0.169)	0.0428

Table 2.2: Model comparison

From table 2.2 it can be concluded that the Regularised Linear Regression does not perform much better than the base line. As well as this the second set of model comparisons shows that the Artificial Neural Network does not perform as well as the baseline. As for the third part of the test, this does not show much about the performance between the Regularised Linear Regression Model and the Artificial Neural Network Model. Due to this we can not reject our null hypothesis.

3 Classification ANN

Explanation of classification problem

Here we will use an Artificial Neural Network which will predict if a person has PD or not based on the other data attributes. This will be a binary-class problem since we are trying to predict to which out of only two classes a person belongs. In the data set the 17th column labelled ”status” identifies whether a person has PD (1) or not (0).

Comparison of Models

For the ANN, we will again be using the range [1, 3, 5, 10] for the number of hidden layers. For the logistic regression we will be testing a range of lambda from 10^-8 to 10^2. The baseline model in this case will simply choose the most common class in the training set. As we became limited on time we decided to use a value of K=5 for cross validation instead of 10.

The error measure in this case can be the misclassified observations over the total number of observations in that test set.

Outer Fold	**ANN**		**Logistic Regression**		**Baseline**
Index	$h*_i$	E_i^{test}	λ_i^{test}	E_i^{test}	E_i^{test}
1	1	0.21	1e1.84	0.18	0.18
2	3	0.18	1e1.75	0.13	0.23
3	10	0.077	1e1.53	0.13	0.23
4	5	0.025	1e1.90	0.15	0.23
5	10	0.18	1e1.97	0.13	0.36

Table 3.1: Table of responsibilities

From the table we see that for the ANN the lowest test error was seen when $h*_i$ was 5 and the best value for lambda was $1e-3.0$.

Statistical Evaluation of Model Performance

We again use a paired t-test to statistically evaluate the performance of the models in relation to each other.

Logistic Regression vs Baseline			ANN vs Baseline			Logistic Regression vs ANN		
$\hat{z}$	CI	p-val	$\hat{z}$	CI	p-val	$\hat{z}$	CI	p-val
-0.102	(-0.206,0.00757)	0.0508	-0.113	(-0.233, 0.00757)	0.0599	-0.0102	(-0.107, 0.0868)	0.784

Table 3.2: Classification Model comparison

From the p-values we find that the Logisitic Regression model performs marginally better than Baseline. The ANN however outperforms the baseline while comparing linear regression to the ANN directly shows the linear regression as better suited.

4 Discussion

Regression

Using regression in this report, we have tried to predict the likelihood that a patient is positive for Parkinson's Disorder based on 22 attributes of recordings of their speech. Three different models were used in this report including; regularised linear regression, an Artificial Neural Network and a Baseline. From the results of the evaluation of the models, it seems that regression was not an appropriate way to predict the status of the Patients. []

Classification

When comparing with the baseline we see that our Linear Regression model only marginally outperformed it. The Artificial Neural Network also outperforms a baseline model that simply chooses the mode of a test set. We would have liked to explore more of what attributes in the logistic regression were most significant in classifying the status but became limited on time.

Previous analysis

The data set used for this report was found on the UCI machine learning repository site, it was previously used in a paper titled 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection' [1]. Whilst there is previous analysis of our dataset, the majority of this analysis is mathematically proving non-linear models rather than validation. This makes it quite difficult assess our models performance against it.

5 Exam Problems

5.1 Question 1: C

We build the confusion matrices for different thresholds of prediction, and calculate the true and false positive rates for each. We can then build an ROC curve by plotting the true positive rate against the false positive rate and we can see that option C corresponds to the ROC curve given.

5.2 Question 2: B

We compute the fractions of how many observations fall into each congestion level based on if we are looking at first all observations, then just $x_7 = 2$ and lastly just $x_7 \neq 2$. Out of these three we choose the largest fractions (most observations in one group).

5.3 Question 3: C

There are 7 input units, each input unit points to each of the 10 hidden layer units so we have 70 parameters between the input and the hidden layer. From the hidden layer to the 4 output units we have a further 40. Adding these together we get a total of 110.

5.4 Question 4: D

The first split at A must be so that the false split following it, B must split into congestion level 1 and 2. The only option that does this is when A: $b_1 \geq -0.76$ followed by B: $b_2 \geq 0.03$.

Bibliography

[1] Little MA et al. *Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection*. 2007. URL: https://archive.ics.uci.edu/ml/datasets/Parkinsons.

Technical
University of
Denmark

DTU Lyngby
2800 Kgs. Lyngby
Tlf. 4525 1700