Yiling Chen

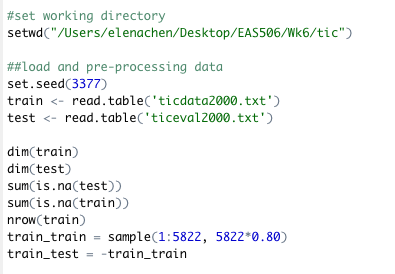
ID: yilingch

Hw3 Write-Up

**Question 1**

The data provided is the insurance company benchmark information on customers. There are 86 variables on product-usage data and socio-demographic data derived from zip codes. The data also has both training (5822) and test (4000) set. The goal is to predict who will be interested in buying a caravan insurance policy using OLS estimates, forward selection, backward selection, Lass regression, as well as Ridge regression.

My first step is to load the data and pre-process the data by checking if there is any missing data, splitting the training data into two parts for forward and backward selection to use later.



Next, I use linear regression method to see whether it is a good predictor to identify who will be interested in buying a caravan insurance policy (v86). The result I got back from using linear regression shows the MSE is 0.05. Therefore, the error rate is relatively high for OLS. But base on the summary report, v59(contribution fire policies) and v49(Contribution motorcycle/scooter policies) are highly correlated with the choice of whether or not to buy a caravan insurance policy. And base on OLS, the prediction of people interested in buying a caravan insurance policy is 0.

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated

I then performed forward and backward selection on the data. Base on the summary of both forward and backward selections, v47(contribution car policies) and v82 (number of boat policies) has the top relationship with caravan insurance policy buyers.

Text

Description automatically generated

Forward subset selection taken into account of variables one at a time from the top, and backward subset selection taken into account of variables one at a time from the back. I used plotting and try to visualize the minimum error which is the lowest point in both Cp and BIC. Forward selection: Cp(29) BIC(8). Backward selection: Cp(28) BIC(8).

Engineering drawing

Description automatically generated Shape, engineering drawing

Description automatically generated

The next step is to look at the MSE and predict how many people are likely to buy caravan insurance policy. By setting the prediction model, I put if people are likely to buy the insurance as 1 in b for forward selection and c for backward selection. The result I got back shows there are 72 people likely to buy the insurance. However, I then wanted to find the error rate of both forward and backward selection and the results are 5% for forward and 6% for backwards. Hence, so far the forward selection is giving us the lower rate of error but it is still very high.

Graphical user interface, text

Description automatically generated

Finally, I use lasso and ridge regression the determine how many people are interested in purchasing the caravan insurance as well as the error rate. The result I got back are, 71 people will purchase the insurance if using lasso, and 70 people if using ridge. The test error for lasso and ridge are shown below.

Graphical user interface, text, application, email

Description automatically generated

In conclusion, contribution of car policies and boat policies are likely purchasers of the caravan insurance policy, and it seems like around 70people will be likely to purchase base on our model. The reason for contribution of car policies and boat policies owner might because some of the car policies owners owns a caravan, and the boat policies owners are likely to have enough income to purchase a caravan and purchases the insurance. However, the error rate is not optimal for the models and this is because we are using linear regression predicting therefore, we encounter the issue of masking. Masking in classification is when the linear regression model is too rigid that masked the classes in the middle. In our case, we have too many variables, therefore linear regression is not the best model to predict the outcome.

**Question 2**

The purpose of this problem is to compare the error rate between the training and testing data. The first step is to generate a data set that contains 1000 observations with 800 in training and 200 in testing where the features are 20.

I first generate the dataset by identifying the p,n,y, and e base on 𝑌 = 𝑋𝛽 + 𝜀 , as well as beta has five elements that are 0. Setting the beta to zero will help us to see the error rate. I then, divide the data into test and train where test contain 200 observations and train contain 800 observations according to the requirement.

Graphical user interface, text, application

Description automatically generated

The next step is to perform best subset selection on the training set and get the lowest MSE point. I use the predict model of training using best subset selection and put it into every value of p and the result use the mean square error equation is that the lowest error rate in training data is when p=20, therefore it is when p is at its greatest. I also plot the graph by using ggplot to get a better visual of this.

Graphical user interface, text, application, email

Description automatically generatedChart, line chart

Description automatically generated

I then did the same with our testing set. I use the predict model of training using best subset selection and put it into every value of p and the result use the mean square error equation is that the lowest error rate in test data is when p=17. I also plot the graph by using ggplot to get a better visual of this. Therefore, when P=17, is when the data has its lowest error rate. The reason for the error rate will increase in testing is because the more features the data contains, it will cause overfitting in which the error rate will increase in the testing set.

Graphical user interface, text, application, email

Description automatically generatedChart, line chart

Description automatically generated

Finally, the true model where beta has 5 components equal to 0, so the true min is 15. However, our test set min of error is 17, this happened because during selection we took into the consideration of coefficient. Even though, we set the betta but the model will follow the data and get new beta(coefficients). Also the variable depend on the dependent variable’s significant rate rather than beta.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

**Question3**

Iris is a data frame with 150 rows and 5 columns, it gives the measurements of detail variables for 50 flowers from each of 3 species of iris. The three species are Setosa, versicolor and virginica. The first goal is to perform kNN on a range of k values and plot the error rate, then report the confusion matrix for optimal model. The second goal is to perform kNN on the first two principal components and plot the error rate, then report the confusion matrix.

My first step is to load and pre-process the data. I checked to see if there is any missing data, which it doesn’t have. Then I look at the names and details of the dataset to get a broad idea of the iris dataset. Then, I split the dataset into 80% train and 20% test for KNN testing.

Graphical user interface, text

Description automatically generated

Next, I created a loop to look at kNN for different values of k. For k from 1 to 20, the minimum error will be at when k = 1. This make sense because as k increase, the error will increase. I then, plot the kNN error rate using ggplot as shown below.

Chart, histogram

Description automatically generatedGraphical user interface, text, application

Description automatically generated

As required by the question, I report the confusion matrix to get a closer look at our predicted table. We can see the table more closely, when k=20 the accuracy is .9. The kNN error rate is lowest at k=1, as shown in the matrix down below, it has an accuracy rate of 97%. This shows us that when k=1 the error rate of prediction is at it’s lowest.

Graphical user interface, text, application

Description automatically generated

A picture containing table

Description automatically generated