Gouthamrajan Nadarajan

Homework 3

**Problem 1**

I generated a pseudo dataset with 20 features and 1000 observations using a normal distribution, I also generated a random beta vector of coefficients using a uniform distribution, and I also generated an error (noise) vector using an normal distribution.

I then set a random 5 beta coefficents equal to zero before calculating the response Y matrix as

Y = X \* Beta + error.

After splitting the data into a 90:10 train:test split, I performed forward subset selection and report the results below in Figure 1 where the x axis is the number of variables and the Y axis is the MSE.

(d)



Figure 1: Forward subset selection of simulated dataset

(e)

The individual MSE for each variable count is shown below where I found that the model with 15 variables performed the best. We also notice from the graph that there is a definitive convergence of the overall error after about 10 variables.

Table

Description automatically generated

(f)

For this best model trained with 15 variables, I show the coefficients below compared to the original ground truth coefficients for comparison. Where the coefficients variable is the coefficients of the trained model and beta is the ground truth.

Text, email

Description automatically generated

Interestingly enough, we see that the model learned coefficients are very similar to the ground truth coefficients and that the model learns to ignore the variables whose coefficents are zero thus telling us that this model is effectively able to ignore variables that carry no information.

Note that this makes sense that this is the best model because there are 20 total variables and we selected 5 of the beta coefficients to go to zero so the model is in fact using all the relevant data for classification.

(g)

I then created a plot depicting the number of variables on the x axis and

On the y axis. This term is essentially an error term that represents the total overall error of the difference in the predicted and actual values of the coefficients.



Figure 2: Error of Coefficients.

Here in Figure 2, we see a similar trend as we saw in Figure 1 as the error in coefficients reaches a minimum in the 15 variable case. I again attribute this to the model using all the relevant information while for the cases where the irrelevant information is being used (where beta =0), the overall error increases.

**Problem 2**

(a)

Looking at the weekly dataset, I first show a pair plot of how the variables change together in figure 3. We can tell from the pair plot that each variable follows a normal distribution and the variables are not very well correlated with each other (covariances close to zero).

A picture containing diagram

Description automatically generated

Figure 3: pair plot of Weekly data

A summary of the numerical data in the weekly dataset is shown below. One thing to note is that the lag variables have very similar mean and quartile data.

Text, table

Description automatically generated

(b)

Using the full dataset to perform logistic regression and taking a look at the individual coefficients, we see that lag2 (X2) appears to be a statistically significant variable. This tells us that the percentage return for 2 previous weeks has the most impact on the Direction. This is interesting because lag3, lag4, and lag5 do not appear to be statistically significant which tells us that overally extensive historical data may not help the model and neither does too recent historical data as lag1 also does not appear to be statistically significant.

Table

Description automatically generated

(c)

Now Looking at the confusion matrix results for the logistic regression (using all the variables), we see that the model has very poor performance where it is unable to differentiate between both classes. Here looking at the accuracy we see an accuracy of 0.5629 but this is poor as we only have two classes and thus the baseline itself is 50% accuracy. Instead the kappa score tells us a more comprehensive story of the model output with an output of 0.0377

Text

Description automatically generated with medium confidence

(d)

Now looking at the confusion matrix using only the lag2 variable, we see that the model performs significantly better (kappa of 0.1414 as opposed to a kappa of 0.0377). This reaffirms my original conclusion that Lag2 is the ideal amount of historical information for the model to make classifications.

Table

Description automatically generated

(e)

Repeating this using LDA, we found almost identical results shown below which is not surprising as LDA and logistic regression have similar mathematical formulations.

Table

Description automatically generated

(f)

Now, performing the experiments using KNN with k=1, I report the following results below. Interestingly, we see that the model performs even worse using KNN which may be due to the fact that k=1 which could lead to significant overfitting.

(g)

From these results, we see that the LDA and the logistic regression perform better than the KNN method which is to be expected as I again suspect the KNN method is overfitting with k=1.

Table

Description automatically generated

(h)

I first changed the value of k to k=3 and repeated the same KNN experiment and found that this significantly improved model performance which is shown below. Comparing this to the experiment with k=1, kappa metric increased from -0.0033(k=1) to 0.0116(k=3)

Table

Description automatically generated

I also wanted to explore using a second variable for LDA and logistic regression. I specifically wanted to work with lag1 in addition to using lag2. My justification is that keeping the most recent weekly data as its own variable may improve classification.

Repeating those experiments, first for logistic regression, I show the results below using the two variables. Here we see that this actually performed worse than using only 1 variable. This could be attributed to the lag1 being “junk data” that is not relevant for the classification. I also repeated the same experiment using LDA but did not find anything different than the logistic regression case.

Table

Description automatically generated

**Problem 3**

(a)

The pairs plot of the variables is shown below in Figure 3. We can tell that most of the variables (telling by the scatterplots) appear to follow a normal distribution except for the SSPG variable which seems to have some sort of bimodal component to the data.

Looking at the covariance of the variables, we see that (glufast vs. glutest) and (glufast vs sspg) and (glutest vs. sspg) have a significant positive covariance (increase in one leads to increase in the other) whereas the other variable combinations do not show as significant of a correlation.

Diagram

Description automatically generated with low confidence

Figure 3: Pair plot showing variable correlations for diabetes dataset

(b)

Applying LDA using all the data, I show the following confusion matrix below:

Table

Description automatically generated

Below I show the confusion matrix for a model training using QDA:

Table

Description automatically generated

When comparing LDA with QDA, we see that the overall accuracy and the distribution of the confusion matrices are similar (0.95 for QDA and 0.90 for LDA), but LDA struggles more with the Normal class classifications and QDA struggles more with the Chemical Diabetic class classifications.

(c)

Looking now at the new data for the new individual, using LDA, I found that the individual was classified as Chemical diabetic, but using QDA I found that the individual was classified as normal.

I would probably agree with the QDA model’s classification of Normal in this case, because the LDA model’s confusion matrix tells us that the model tends to misclassify Normal as being Chemical Diabetic (The positive predictive value for the chemical diabetic class is only 0.79).