503 HW3 Problem 2

Katherine Wilkinson

2/18/2018

As previously, we will be looking at the automobile data, again focusing primarily on the non-categorical data. In this analysis however, we will be applying a number of classification methods to the auto data to obtain a classification rule to classify the cars into three different classes based on the number of cylinders. The classes will be labeled as Y = 1 if the number of cylinders is 5 or less, Y= 2 if the number of cylinders is 6, and Y = 3 otherwise. Additionally we will be looking at our data in three different ways; original data, standardized data, and PCA-preprocessed data based on the first two principal components. On these three versions of our data set, we will then apply LDA, QDA, logistic regression, and nearest neighbor classifier.

To begin, we will separate our data into a training set and a test set. Our training set consists of 75% of our data, with all 3 classes represented while our test set consists of the remaining 35% of the data. We can see in the below two plots that both data set have similar behaviors between each of the three variables and in regards to the separation by cylinders. We can also see that mpg and displacement have the best visible separation by cylinders. In our LDA analysis, we will use this fact to help visualize our data. In both sets the correlations between variables is similar and we can then assume that our test set is a good representation of our data. Thus we will use our train data to create our classification rule and test this on our test data set. We can then compare errors between both our training data set and our test data set.

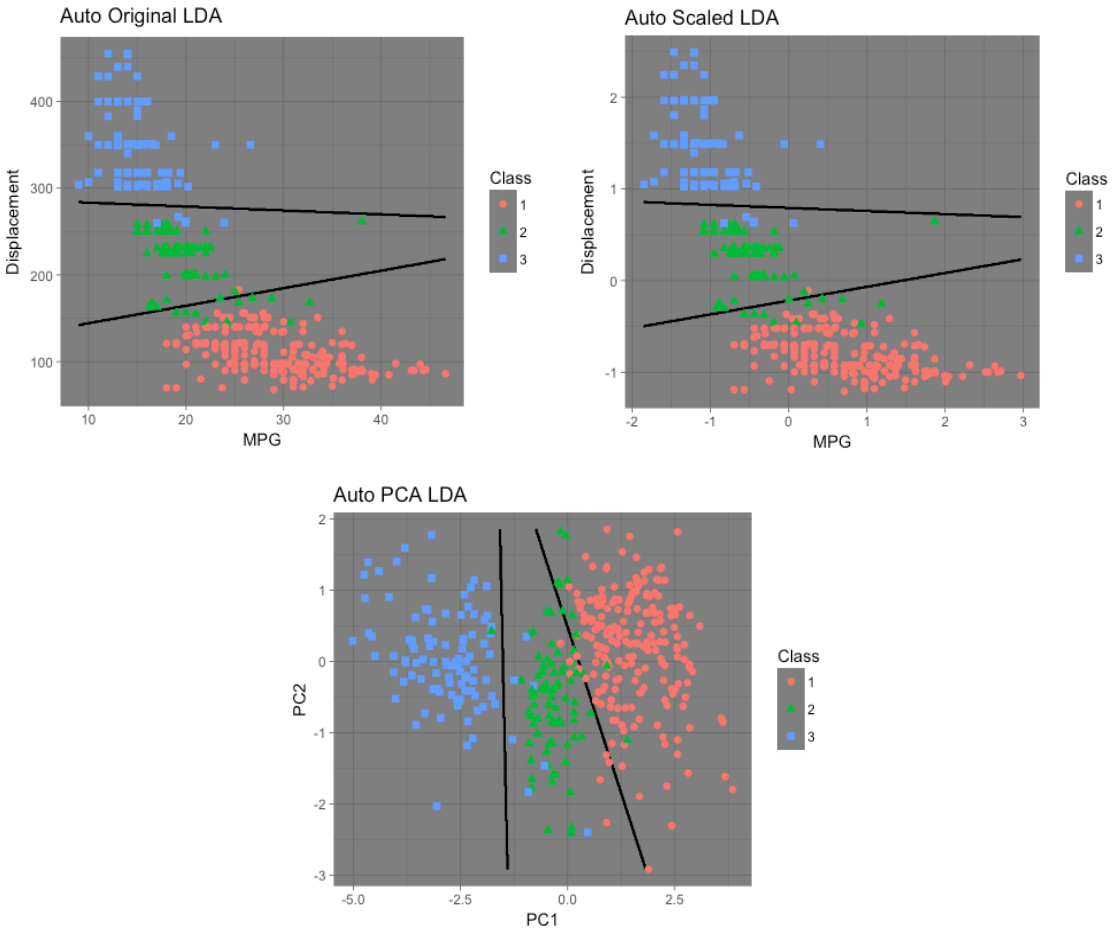
To start, we will use linear discriminant analysis (LDA) on our data with the lda function in R to find a classification method for the classes of cylinders. We can see in our table below that for each of our different preprocesses of our data, the test error is higher than our train error, as

expected. The original, untouched data and the standardized data appear to have very similar, if not identical, errors. In contrast, our PCA Preprocessed Data has a slightly higher error, for both Test and Train. However, the PCA Preprocessed Data also has a larger gab between the Test Error rate and the Train Error rate.

LDA Error Rates

	Original Data	Standardized Data	PCA Preprocessed Data
Test Error	0.0677966	0.0677966	0.1101695
Train Error	0.0437956	0.0437956	0.0547445

We can also look at our boundary plot of each of our data sets to see how they differ.



Again, we can see that the Original data and the Scaled Data are quite similar, while our PCA data looks much different, even if you do not take the rotation into account. All three sets do not have perfect classification methods, which is reflected in the errors in the above table. We can see though that the original data and the scaled data have fewer misclassified points.

Regardless, all three data sets do have fairly good classifications with relatively lower error rates for both the training and the test data.

We can then compare our LDA classification to the quadratic discriminant analysis (QDA). Our test error rates are again larger than our train data error rates, with the exception of the PCA Preprocessed Data. However, the error rates for the PCA Preprocessed data are fairly similar at 0.0508 and 0.0511. Similar to LDA, the PCA data has larger error rates for both train and test sets. Unlike LDA however, our Original data error rates and our standardized data error rates are different, with the original error rate slightly worse for the Test data set. From this analysis and LDA, we would probably want to choose our the classification method for either the standardized data or the original data, as these errors are the smallest and seem to have the fewest misclassification. While QDA does have the smaller errors, it should be noted that this classification method would be a more complicated model and may be harder to accurately reproduce.

QDA Error Rates

	Original Data	Standardized Data	PCA Preprocessed Data
Test Error	0.0423729	0.0338983	0.0508475
Train Error	0.0145985	0.0255474	0.0510949

After doing LDA and QDA, we can also do a logistic regression on our data. Once again, we will use train and test data for each of our different data (original, standardized, and PCA). Looking at the regression equations found for the classification method (using 1 as the base comparison) we can see some differences now between the original and the standardized data (the PCA Data is more difficult to compare as we are using the first two Principal Components).

```
 Coefficients (Original Data):
##   (Intercept)        mpg displacement  horsepower     weight acceleration
## 2   -39.67848 -0.3213432    0.1965534 -0.03849161 0.01307534    -1.002804
## 3  -178.44706  0.4250007    0.6696404 -0.15181766 0.01979470    -1.538959
```

```
 Coefficients (Standardized Data):
##   (Intercept)       mpg displacement   horsepower    weight acceleration
## 2    10.82486 -1.337807      22.3289  0.003331877  9.687856    -1.909144
## 3   -48.46640 14.323985     124.3028 -1.529711319 20.917274    -3.052090
```

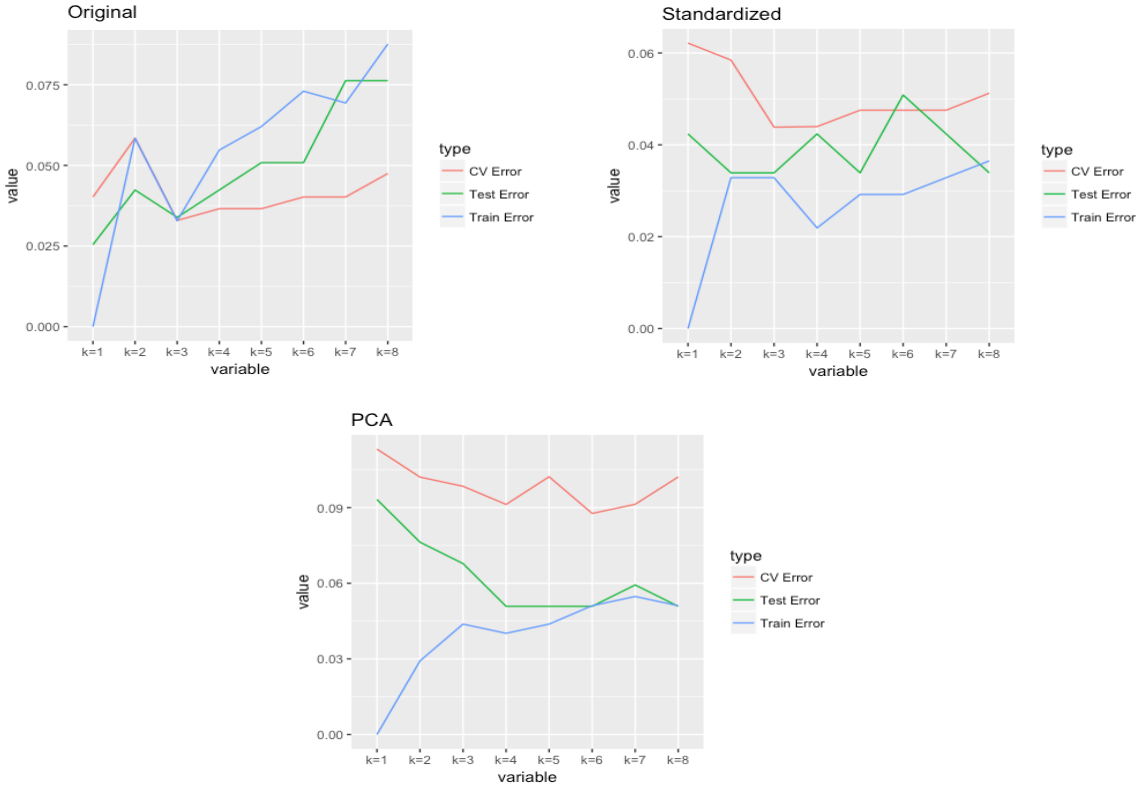
```
 Coefficients (PCA Preprocessed):
##   (Intercept)        PC1       PC2
## 2    1.733864  -5.870556 -1.388504
## 3   -3.978951 -10.367824 -2.669242
```

However, when we look at the Mean Error Rates for the Logistic Regression, we can see again that the original data and the standardized data are relatively close together, especially when compared to the PCA data. Again, the PCA data has higher error rates than the other two data set types. For the Original data and the PCA data the error rates for the logistic regression are smaller than those found in LDA and QDA. In all cases for the logistic regression however, the test error rate is smaller than the train error rate. Interestingly, the standardized data and the PCA data have the same or very similar mean error rate for the test data. Overall, in comparison to LDA and QDA, logistic regression does have smaller errors and may give a slightly better classification method.

Logistic Regression: Mean Error Rate

	Original Data	Standardized Data	PCA-preprocessed data
Test	0.0169492	0.0423729	0.0423729
Train	0.0182482	0.0109489	0.0437956

Finally, we can also look at the nearest neighbor classifier using the Euclidean distance metric.. In this case, we will look at a number of nearest neighbor values (k) and use cross-validation to check which value of k will give us the smallest error for each type of data. Using a 10-fold cross-validation on just the training portion for each of our different data sets, we find that k=3 gives us the smallest error rates for the original data (0.0329) and the standardized data (0.0438), while k = 6 gives the smallest error rate for PCA data (0.0877). Once again, it appears that our PCA data has the highest error rate. This can also be seen in the graphs of the different error rate for each of the data sets.



For the original data we can also see that as the number of neighbors increases, the error rate increases for all three types of error. In general, this is the case for the other two data sets, but not quite as drastically. While k = 6 seems to be the best choice for the PCA data, we will look further into the error rates for k = 3 to compare the differences on the different data sets. This is also a less complicated model choice and thus may be easier to recreate the classification method for the number of cylinders and will be less likely to overfit the data.

From our cross-validation and plots of the training, cross-validated, and test errors as a function of k, we will then look more closely at the specific errors for k = 3. As noted above, the PCA data again has the highest error rate for the train error, test error, and the cross-validated error. We can also see that the error rates for train and test for both the original data and the standardized data are very similar. This is reflected in the graphs above as well.

Nearest Neighbor Error: k =3

	Train	Test	CV
Original Data	0.0328467	0.0338983	0.0329293
Standardized Data	0.0328467	0.0338983	0.0438384
PCA Preprocessed Data	0.0437956	0.0677966	0.0984512

When looking at all 4 of our classification methods, we can see some consistent patterns, mainly that the PCA data generally has the higher error rates. This may be do to the processing of the data, as our PCA Preprocessing changes the data by projecting it onto a smaller plane. Standardizing on the other hand, keeps the same number of dimensions and merely shifts the data slightly. Overall the logistic regression method has the smallest error rates while LDA has the largest error rates, that is LDA tends to have more misclassification of the number of cylinders to the car. It should be noted however, that when looking at all the error rates, the complexity of the model should also be taken into account. Our LDA and QDA methods uses the assumption that the density of our data is Gaussian which may not necessarily be the case. Logistic regression and nearest neighbor on the other hand have fewer assumptions about the data. If our underlying data is truly Gaussian, then LDA may be the most suitable. Overall, the number of assumptions and the complexity of the resulting model as well as the error rates of missclassification must be taken into account when choosing which classification method should be used.