Homework 3, STAT365/665

Girish Sastry

March 1, 2011

1 Classification of Zipcode Data

We use k-NN with $k = 1, 7, 15$. We carry out LDA using the built in package in R. For QDA, the process is a bit more complicated: because some of the covariance matrices are singular, we must do some sort of dimensionality reduction before running QDA. We use PCA to do this. Note that the prediction errors are essentially the same in both the 50 and 100 component cases, so for computational efficiency we use the first 50 principal components. For logistic regression we make use of the multinom package. For reduced rank LDA, we first project our data onto a subspace of dimension 10. Essentially, we define $\hat{Y} = X(X^TX)^{-1}X^TY = X\hat{B}$. Next we compute the eigen-decomposition of $\hat{Y}Y$ with PCA and we crossvalidate on the training set. PCA allows us to choose the number of components each iteration to measure the error. The optimal number of components will minimize the error, and is the set of components that we will use for reduced rank LDA. The optimal number of components is 9 from crossvalidation. We obtain the following test and training errors for the various methods:

Method	Test Error	Training Error
1-NN	2.47%	0.00%
7-NN	3.02%	0.58%
15-NN	3.85%	0.93%
LDA	6.38%	6.20%
QDA	6.52%	1.77%
Reduced Rank LDA	6.38%	6.20%
Logistic Regression	9.49%	0.00%

We see that in general, nearest neighbor methods once again perform better than the other methods. Just as was the case for linear regression, this is probably due to the nature of the problem, and perhaps due to our error metric for cross validation. Zipcode classification is a discrete classification problem and thus we use misclassification error instead of mean squared error (as in regression). Thus, we note that linear methods generally do not perform as well as nearest neighbors for the zipcode problem. Also, according to *The Elements of Statistical Learning* website, a 2.5% test error is "excellent". So the best method is probably much more complicated than linear models.

Furthermore, there may be a Gaussian assumption in the linear methods used, which could explain the worse results vs nearest neighbors. Looking at LDA vs Reduced Rank LDA, we see that there is no difference in training and test errors. We used 9 out of the 10 components for reduced rank LDA, so it is

likely that the unused component had such a small proportion of the variance that it was negligible to the model.

Next, note that QDA performs better than the linear methods, probably due to the Gaussian assumption mentioned earlier. QDA allows for variation within each class and thus is looser than the linear models. This seems more valid than a single variance across all classes.

2 Classification of Phoneme Data

For both logistic regression and QDA, we choose to split the data in half. The first half is the training set and the second half is the test set. Looking at a plot of the data, it seems that there is no significant difference in the distribution of data in the test and training sets.

2.1 Raw and Regularized Logistic Regression

First, we carry out raw and regularized logistic regression on a subset of the data. This subset corresponds to the "aa" and "ao" phonemes. Regularized regression was done using 12 knots, as in *The Elements of Statistical Learning*. The results are as follows:

Method	Test Error	Training Error
Raw Logistic Regression	24.71%	7.92%
Regularized Logistic Regression	18.97%	20.08%

These results represent the same trend seen in *The Elements of Statistical Learning*. Using a natural cubic spline fit causes our test error to decrease but also causes our training error to increase a lot. Just as in the text, smoothing the estimator probably causes this. The high error rate is most likely due to the linear logistic regression model. Many of the features are highly correlated and the model doesn't necessarily take this into account. When extending the analysis to the full data set, we obtain the following results:

Method	Test Error	Training Error
Raw Logistic Regression	15.04%	1.82%
Regularized Logistic Regression	14.62%	12.90%

In both the raw and regularized cases, we notice a lower error rate than in the subset of "aa" and "ao" phoneme data, probably because we simply have more data to train our model on. The raw logistic regression model is more accurate than the regularized logistic regression model.

2.2 Raw and Regularized QDA

Next, we run raw and regularized QDA on the phoneme data set, obtaining the following results:

Method	Test Error	Training Error
Raw QDA	29.02%	14.46%
Regularized QDA	17.66%	15.21%

When regularizing QDA, we choose our knots to be of size 5, 6, 7, 8, and 9. Then we use 10-fold cross-validation and find that 8 knots is optimal. Thus, we train and test using 8 knots. We observe that our training error is a little higher in the regularized model but our test error is lower: this, just as before,

is likely due to the smoothing effect in the raw case. When comparing QDA and logistic regression, it seems that logistic regression gives us better results. This is probably due to the nature of the data set: the data are quantitative values of sound, and so there is likely some kind of relationship between the data points, which QDA does not take into account. To find a better model for this data, we would need something that takes this inter-correlation into account.

3 Appendix

This section contains the R code for the analysis.