Homework 3, STAT365/665

Girish Sastry

February 27, 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