Machine Learning Exercise 8

Shun-Lung Chang, Anna Ruby

In this report, we create features using principal components analysis (PCA hereafter) and k-means cluster analysis, use these features to train progressively more and more flexible classifiers, and analyze the optimal classifier size for classifying our data set while contrasting our feature creation methods.

PCA Feature Selection

Build a classifier with the first 20 principal components

We applied PCA to perform dimensions reduction on the data using the first 20 principal components. The transformed data with a bias vector is then used to build a linear regression classifier. The performance is shown below, where MSE stands for Mean Square Error and MISS for the misclassification rate.

$MSE_{train} = 0.0366$
$MISS_{train} = 0.067$
$MSE_{test} = 0.0373$
$MISS_{test} = 0.068$

Working with 20 PC's, we see only a small difference in the MSE_{train} and MSE_{test}, which is similarly seen in the misclassification rates of testing and training. As we discussed in class, the difference between these two curves will increase as we increase the number of parameters, m, here known to be PC's. Below we explore the optimal number of PC's for classification.

Build classifiers with different numbers (from 1 to 240) of principal components

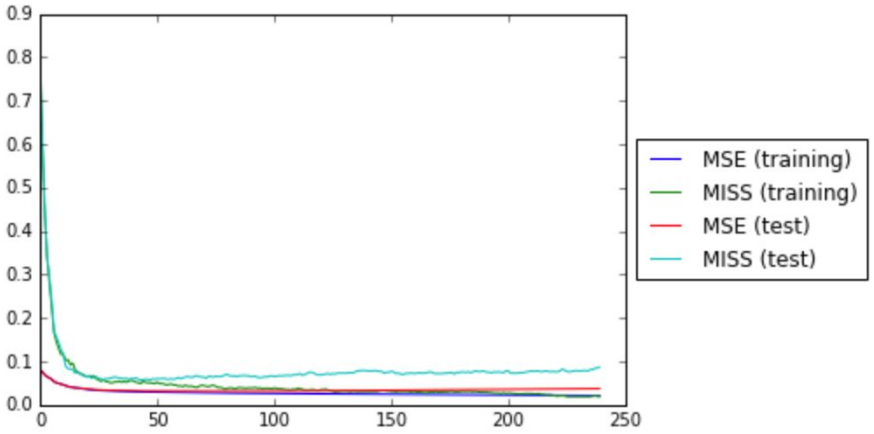

Figure 1: Error rates against the number of parameters in a given classifier.

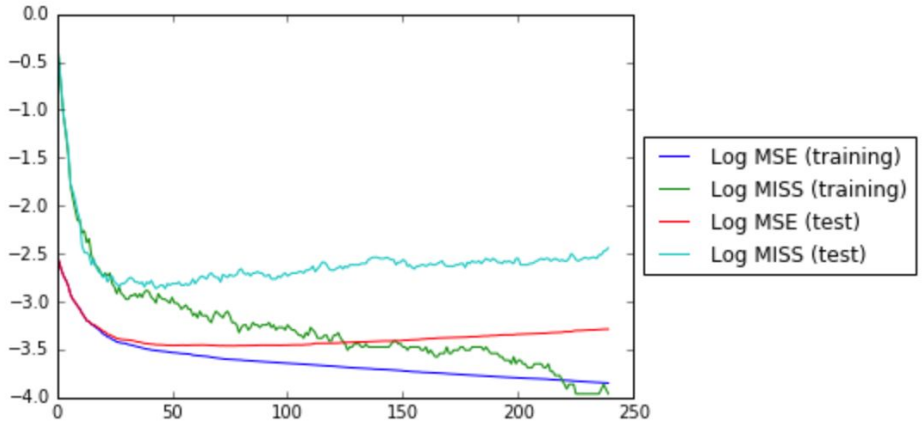

Figure 2: The logarithm of error rate against the number of parameters in a given classifier.

Both of the above figures show the increasing error on testing as we increase our parameter count, but the second, having taken the natural log of the MSE and MISS for testing and training, provides a clearer view of the separation between testing and training. As seen above, increasing the number of PC's used as features in our classification first causes a drastic decrease in both training and testing error (and, following from this, a decrease in the respective misclassification rates as well). As we continue to increase classifier flexibility we begin to see a separation in the MISS of training and testing, as we would expect with overfitting. There is a similar, but less drastic, separation in the MSE of training and testing.

Judging purely from figure 2, the optimal number of PC's for our classification is between 25 and 40.

K-Means Feature Selection

Build a classifier with 20 clusters

The features created by k-mean clustering analysis are based on the distance between each point and the codebook vector of each cluster. We have chosen to run our first training and testing using 20 clusters, to contrast with the above analysis using PCA.

$MSE_{train} = 0.033$
$MISS_{train} = 0.062$
$MSE_{test} = 0.034$
$MISS_{test} = 0.064$

As with the PCA feature selection method, we see a small, but notable, difference in the training and testing rates for both MSE and MISS. It is also worth noting that, with 20 features, the k-means clustering selection method has generated a classifier with a lower error rate, across the board, than the PCA selection method.

Build classifiers with different numbers (from 1 to 800) of clusters

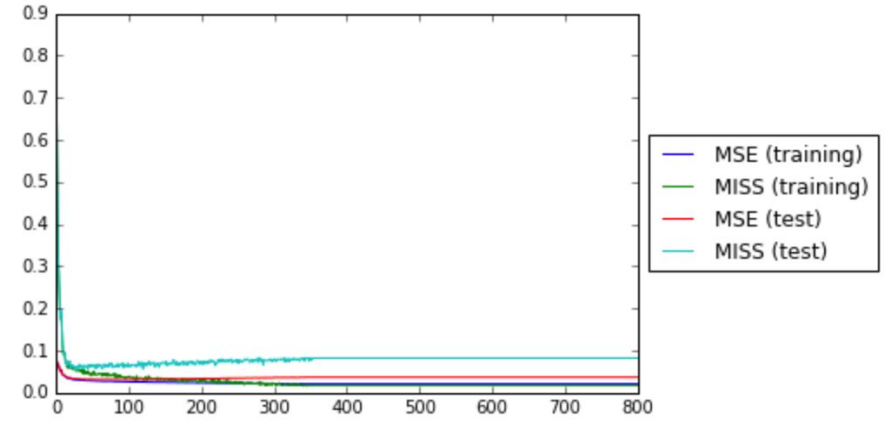

Figure 3: Error rates against the number of parameters in a given classifier.

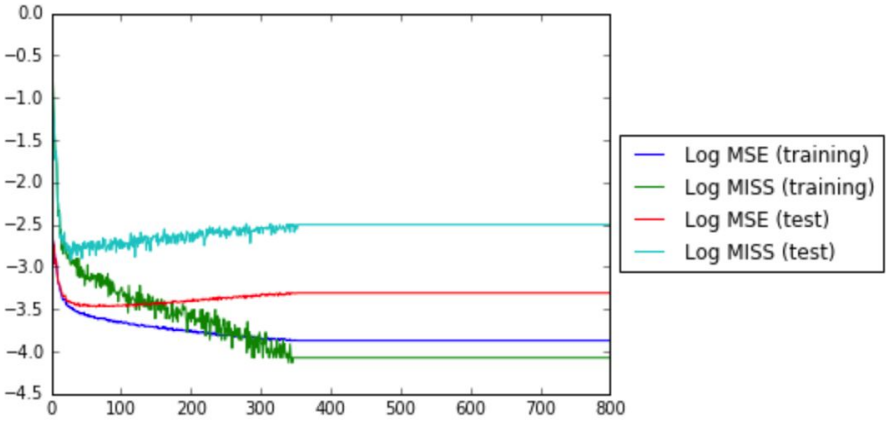

Figure 4: Logarithmic transform on the error rates against the number of parameters in a given classifier.

Again, we have above to graphs showing the error rate of our classifier, considering both misclassification and MSE, as we increase the number of parameters considered in the classifier. As in figure 2, figure 4 shows the natural logarithmic transform on the error rates against the parameter count. In both figures, the training and test error become constant when the number of clusters is larger than 350, as most clusters are empty. As with our PCA chosen features, we see first a drastic decrease in the error rate for both MSE and MISS as we begin to use more and more cluster features. the MISS train rate quickly drops far below the MISS train error rate, signalling that, as we increase our classifier flexibility, we again experience overfitting.

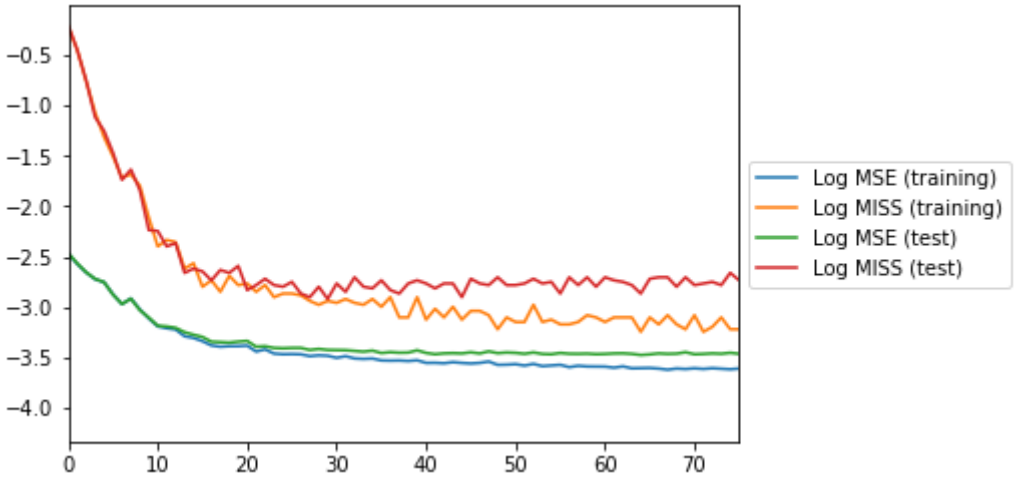

Figure 5: Zooming in on the first 75 classifiers represented in figure 4.

When we zoom in on the first 75 classifiers tested, we can estimate that the optimal amount of cluster features is between 20 and 40 clusters. This may seem somewhat counterintuitive to some, as one may have assumed that, given that there are 10 different numbers represented within our data set, the logical optimal number of clusters would be 10. It is though, frequently the case that additional subdivision of clusters corresponding to individual classes can improve classification overall, which is why we experience a continued decrease in our error rates, across the board, for classifiers with more than 10 parameters.