Slide 1

Good evening. We are Team Bengio.

Our project is to compare six 10-classs classifiers using the NotMNIST data set.

Slide 2

The NotMNIST set is similar to the MNIST data set that used in our homework except that it consists of 10 letters instead of 10 numbers.

Similar to the MNIST data set, each sample consists of a picture with 28x28 pixels.

Our training set has 60000 samples divided into 10 letters. Each letter has 6000 samples.

The test set has 10000 samples with 1000 samples per letter.

In our experiments, we reduce the dimension of each training/testing sample with N principle components and feed them into a classifier to evaluate the accuracy of the test data and the elapsed CPU time of training and testing as we change the number of Principle Components from 1 to 50.

Slide 3

To determine the trade-off of accuracy, elapsed CPU time, and the number of principle components, we change the number of PC from 1 to 50.

For each number n of PC, we reduced the dimension of the data into n-dimension using principle components analysis.

To do training, the reduced dimensional data with the training label are feed to a classifier of interest to obtain the model paraments of the classifier.

To do testing, the trained model parameters are inserted back to the classifier together with the testing data to get a prediction of the class label of the testing data.

In our experiments, we tested the Bayesian Classifier, the linear classifier, and the linear classifier using non-linear combinations of features.

The linear classifier using non-linear combinations of features were discussed in our class notes and we’ll talk more about the detail in the next slide.

These three multi-class classifiers were formulated using two formulations. One formulation is the same as our home works. It is called One-Vs-Rest multiclass classification.

The other formulation is called One-Vs-One Multiclass classification and we will explain more in detail later.

Therefore, we had a total of six classifiers.

Slide 4

In a regular linear classifier, a feature vector x is converted into an augmented feature vector xa by pre-pending a one in the feature vector before further processing by the linear classifier.

Slide 5

Linear classifier with non-linear combination of features is similar to the linear classifier in the previous slide, except that the augmented by one operation is replaced by augment by non-linear combinations.

In our implementation, the augment by non-linear combinations of features is implemented in three steps.

In the first step, the data vector x is augmented by one to form xa.

In the second step, the xa is used to form a matrix xaT\*xa which consists of 1, xi, the square terms xi^2 and the cross term xi\*xj.

In the third step, the elements in the upper triangle of the matrix are collected to form the non-linear feature vector xnc.

Slide 6

Two approaches of multiclass classification were implemented.

The more familiar approach is One-vs-Rest.

This is the approach that we used in our homework assignments.

For the One-vs-Rest approach, we have our linear classifier with Kesler construction. We have the Bayesian classifier. And we have the linear classifier with non-linear feature combination.

Slide 7

The second approach of multiclass classification that we implemented is called One-vs-One.

In One-vs-One, the K class classifications is performed at one pair of classes at a time.

In this case, we have (K+1)\*k/2 binary classifiers as shown in the matrix on the right.

In particular the (C,E) is a classifier for classification the input data vector as the letter C or the letter E.

For each data vector, it has a total of (K+1)\*K/2 class labels. The most frequent label is selected as the predicted label of the data vector.

In effect, we let the binary classifiers vote against each other. And the winner is the one with the most votes.

As in the One-vs-Rest approach, we also have similar three classifiers.

So we have a total of six classifiers for comparison.

Slide 8

Here is a comparison of the accuracy of the 6 classifiers as a function of the number of principle components.

In particular, accuracy of a classifier is the percentage of correct classifications.

In this graph, the horizontal axis is the number of PC, the vertical axis is the accuracy.

The legend of the graphs is ordered by the accuracy at PC=50. So, the graph with the highest accuracy is on the top.

Here are the three One-Vs-One accuracy results.

And here are the three One-vs-Rest accuracy results.

As shown here for PC<15, the accuracy of One-Vs-one is better than One-Vs-Rest.

When the number of PC increases, the results are less clear.

Let us look at the results with 16 PCs.

Slide 9

We are going to compare the accuracy, PPV, Sensitivity, and Specificity of the six classifiers using boxplot.

Boxplot shows the max, min, first, second, third quartile of the metric among the 10 letters.

Here is our comparison of the six classifiers.

We also added the random classifier as a benchmark.

At 16 PCs, the linear classifiers with non-linear combination of features are the best or near the best accuracy.

For PPV, the linear classifier with non-linear combination and one-vs-one approach is distinctive better than the others.

All of them are significantly better than the random classifier.

Slide 10

Here is a comparison of elapsed CPU time for training plus testing and the accuracy for different number of PC. Each marked data point corresponds to certain number of PC.

Please notice that the CPU time is in log scale.

Let us zoom in more to find out what happened for accuracy above 80%.

Slide 11

Let us look in more detail for accuracy above 80%.

In particular, the Bayes classifier is order or magnitude more computation than the others.

OveVsOne is slower than OneVsRest.

And Linear classifiers are fasters.

Linear classifiers with non-linear combination of feature have the best accuracy.

Slide 12

In conclusion, ….

Slide 13

Thank you very much.

Slide 14

Slide 15

Slide 16