Machine Problem #3
COT 5610 - Machine Learning
Fall 2014
Due 10-16-2014

Edward Aymerich

1 Problem description

The MNIST data set consist of 60,000 training examples of handwritten digits, and 10,000 test examples. To reduce the dimensionality of this data set two algorithms were implemented and separately applied to it: Principal Component Analysis (PCA) and Fischer Discriminant Analysis (FDA). The resulting dimension-reduced data sets were tested using implementations of k-Nearest Neighbors (KNN) and Naive Gaussian Bayes (NGB) with Maximum Likelihood Estimation (MLE) classifiers from previous assignments.

In order to test the dimension-reduced data sets, three test protocols were used:

- Test Protocol 1: using all 60,000 training examples from dimension-reduced training set for training, and all 10,000 test examples from dimension-reduced test set for testing. To avoid division by zero, a small value $\epsilon = 0.000001$ was added to the estimated variance.

- Test Protocol 2: the training set was divided into a training subset (with 50,000 examples), and a validation set (with 10,000 examples). For KNN, the validation set was used to find the optimal value for k among the values {1,2,3,4,5,6,7,8,9,10}, and the optimal value was used to test the model against the test set. This test protocol was not used with NGB, because there is no parameter to tune in this algorithm when MLE is used.

- Test Protocol 3: only the training set was used in this protocol. KNN and NGB algorithms were used with 5-fold cross-validation and 10-fold cross-validation on the training set.

[1] http://eigen.tuxfamily.org

2 PCA and FDA implementation

The PCA algorithm was implemented on C++11, using the Eigen library[1] for linear algebra. The number of components used in the dimension-reduced data set is choose dynamically, using the calculated eigenvalues as indicators. When the sum of the d greatest eigenvalues accounts for 95% of the sum of all eigenvalues, then d is used as the number of dimensions in the reduced data set.

The FDA algorithm was implemented on C++11, again using the Eigen library for linear algebra. During implementation, it was found that the calculation of the within class scatter matrix (S_W) was by far the slowest part of the dimensionality reduction process, therefore additional efforts were made to accelerate this calculation. An OpenMP directive was used to distribute the calculation among all the cores in the system, and even with unbalanced distribution of work in general the parallel implementation on a 4 core system used 1/3 of the time used by the serial implementation on the same system.

3 Results

This section presents the error rates resulting from using the dimension reduced data set (obtained from the implemented PCA and FDA) on the KNN and NGB classifiers implemented on previous assignments. The implemented PCA and LDA algorithm were compiled using GCC 4.9.1 on a Windows 7 64 bit platform, and the experiments were executed using a Intel Core i5 3470 processor running at 3.2GHz.

3.1 PCA reduced data set

The original MNIST data set was transformed into a lower dimensionality data set using the implemented PCA algorithm. The implementation automatically chooses the amount of dimensions of the resulting

data set based on the calculated eigenvalues. Table 1 shows the accumulated value for the largest eigenvalues. For the MNIST data set, using the 154 largest eigenvalues accounts for 95.02% of the total eigenvalues, so the implemented PCA algorithm used only the corresponding 154 eigenvectors to build up the space of the dimension reduced data set.

Principal Components	% accumulated eigenvalue
1	9.70
2	16.80
3	22.97
5	33.23
7	40.81
9	46.46
11	50.92
17	60.74
26	70.02
44	80.33
87	90.01
154	95.02
565	99.99
784	100

Table 1: Accumulated eigenvalue according to the number of principal components used.

With the number of dimensions known, a transformation matrix W is build using the eigenvectors corresponding to the largest 154 eigenvalues, and both the train data set and test data set are transformed to the new space using W. KNN and NGB were used to test the obtain data set.

3.1.1 KNN results for PCA

Table 2 summarizes the results obtained for Test Protocol 1 when using PCA with KNN, and compares them with the results obtained in Machine Problem 1 on the original MNIST data set. For all tested k values, the error rate slightly improved, and the best result is found when $k = 3$, going from 2.95% to 2.78%. As for execution time, it improved dramatically, going from 143.7 seconds on average to 25.18 seconds on average.

Table 3 summarizes the results obtained for Test Protocol 2 when using PCA with KNN, and compares those results with the ones obtained in Machine Problem 1 on the original MNIST data set. In all cases for the validation set there was a slight improvement on the obtained error rate, but the most interesting result is that the tuned value of k is different when the reduced dimension data set was used. On the original MNIST data set, the lower error rate on the validation set wast attained when $k = 4$, but on the lower dimension data set the lower error rate on the validation set occurs when $k = 3$. This change on the tuned k together with the use of the lower dimension data set results in the error rate for the test set to go from 3.42% to 2.95%, an improvement of 0.47%.

k	MNIST error rate (%)	PDA error rate (%)	MNIST exe. time (s)	PDA exe. time (s)
1	3.09	3.06	144.96	25.09
2	3.73	3.46	143.98	25.75
3	2.95	2.78	143.46	25.23
4	3.18	2.94	143.81	25.08
5	3.12	2.88	143.61	25.08
6	3.23	3.00	143.68	24.83
7	3.06	2.90	143.56	25.27
8	3.30	3.07	143.67	25.55
9	3.41	3.04	143.49	25.13
10	3.35	3.16	143.47	25.34
20	3.75	3.63	143.36	24.56
30	4.04	3.88	143.63	25.96
50	4.66	4.34	143.41	24.61
100	5.60	5.35	143.71	25.07

Table 2: Results obtained from Test Protocol 1 for KNN with the original MNIST data set and the PDA dimension reduced data set.

set	k	MNIST error rate (%)	PCA error rate (%)	MNIST exe. time (s)	PCA exe. time (s)
val.	1	2.88	2.63	119.50	20.15
	2	3.33	3.19	119.55	19.93
	3	2.80	2.56	119.67	21.21
	4	2.74	2.56	119.47	20.28
	5	2.82	2.62	119.78	21.89
	6	2.90	2.76	119.38	20.42
	7	2.92	2.69	119.42	20.28
	8	2.92	2.75	119.63	20.31
	9	2.95	2.71	119.41	19.97
	10	2.99	2.88	119.59	20.08
test	4/3	3.42	2.95	119.59	20.08

Table 3: Results obtained from Test Protocol 2 for KNN with the original MNIST data set and the PDA dimension reduced data set.

Table 4 summarizes the results obtained for Test Protocol 3 when using PCA with KNN, and compares those results with the ones obtained in Machine Problem 1 on the original MNIST data set.

test	k	MNIST avg. error rate (%)	PCA avg. error rate (%)	MNIST exe. time (s)	PCA exe. time (s)
5-fold	1	3.06	2.86	690.11	115.20
	3	2.99	2.75	690.62	115.69
	5	3.07	2.90	689.26	115.54
	7	3.22	3.03	689.44	135.22
	9	3.38	3.03	690.38	154.91
10-fold	1	2.95	2.74	782.17	175.41
	3	2.88	2.64	777.63	162.17
	5	2.99	2.81	776.88	161.30
	7	3.09	2.94	775.51	179.67
	9	3.28	3.07	778.56	180.92

Table 4: Results obtained from Test Protocol 3 for KNN with the original MNIST data set and the PDA dimension reduced data set.

As before, error rate improves slightly in all cases, and the execution time is considerably reduced.

3.1.2 NGB results for PCA

For Test Protocol 1 when using PCA with NGB, the resulting error rate was 22.52%, which is worse than the result obtained in Machine Problem 2 when the original MNIST data set was used (20.88%).

As it was mentioned before, Test Protocol 2 doesn't apply in this case, because NGB using MLE doesn't have any parameters to tune.

Regarding Test Protocol 3, the data set generated by PCA obtains an average error rate of 24.04% for the 5-fold case, which is a worst result than the one obtained by NGB using the original data set (22.58%). For the 10-fold case, the data set generated by PCA obtains an average error rate of 23.98%, which again is a worst result that the one obtained using the original data set (22.54%).

3.2 FDA reduced data set

The original MNIST data set was transformed into a lower dimensionality data set using the implemented FDA algorithm. In this case, the number of dimension in the resulting data set was selected as the highest value permitted by the FDA method, this is, number of classes minus one, or 9 for MNIST.

3.2.1 KNN results for FDA

Table 5 summarizes the results obtained for Test Protocol 1 when using FDA with KNN, and compares them with the results obtained in Machine Problem 1 on the original MNIST data set.

k	MNIST error rate (%)	FDA error rate (%)	MNIST exe. time (s)	FDA exe. time (s)
1	3.09	72.57	144.96	1.94
2	3.73	72.65	143.98	2.10
3	2.95	71.02	143.46	2.21
4	3.18	70.09	143.81	2.36
5	3.12	68.99	143.61	2.32
6	3.23	68.36	143.68	2.65
7	3.06	68.56	143.56	2.30
8	3.30	68.27	143.67	2.69
9	3.41	67.36	143.49	2.45
10	3.35	67.27	143.47	2.56
20	3.75	66.10	143.36	2.53
30	4.04	65.96	143.63	2.18
50	4.66	66.61	143.41	2.07
100	5.60	67.21	143.71	2.18

Table 5: Results obtained from Test Protocol 1 for KNN with the original MNIST data set and the FDA dimension reduced data set.

The error rate increases for all values of k when using the dimension reduced FDA data set. Even worse, the error rates jump from just 1 digit percent values to around 70% error rates.

Table 6 summarizes the results obtained for Test Protocol 2 when using FDA with KNN, and compares those results with the ones obtained in Machine Problem 2 on the original MNIST data set.

set	k	MNIST error rate (%)	FDA error rate (%)	MNIST exe. time (s)	FDA exe. time (s)
val.	1	2.88	72.52	119.50	1.59
	2	3.33	72.74	119.55	1.60
	3	2.80	71.09	119.67	1.60
	4	2.74	69.75	119.47	1.60
	5	2.82	68.77	119.78	1.60
	6	2.90	68.51	119.38	1.60
	7	2.92	68.19	119.42	1.60
	8	2.92	67.51	119.63	1.60
	9	2.95	67.78	119.41	1.60
	10	2.99	67.47	119.59	1.60
test	4/10	3.42	67.59	119.59	1.60

Table 6: Results obtained from Test Protocol 2 for KNN with the original MNIST data set and the FDA dimension reduced data set.

Again, the results obtained by using the FDA reduced data set are worse than the ones obtained by using the original MNIST data set. And once again, the error rate skyrocketed in comparison.

As for Test Protocol 3, Table 7 summarizes the results obtained when using FDA with KNN, and compares those results with the ones obtained in Machine Problem 1 on the original MNIST data set.

test	k	MNIST avg. error rate (%)	FDA avg. error rate (%)	MNIST exe. time (s)	FDA exe. time (s)
5-fold	1	3.06	72.68	690.11	9.20
	3	2.99	71.20	690.62	9.20
	5	3.07	68.87	689.26	9.21
	7	3.22	67.99	689.44	9.21
	9	3.38	67.47	690.38	9.27
10-fold	1	2.95	72.50	782.17	10.41
	3	2.88	71.05	777.63	10.43
	5	2.99	68.63	776.88	10.67
	7	3.09	67.79	775.51	10.41
	9	3.28	67.18	778.56	10.42

Table 7: Results obtained from Test Protocol 3 for KNN with the original MNIST data set and the FDA dimension reduced data set

3.2.2 NGB results for FDA

The FDA reduced data set keeps performing bad even when the classifier gets changed to NGB. For Test Protocol 1, the reduced data set gets an error rate of 87.29%, compared to an 20.88% when the original MNIST data set was used.

Test Protocol 2 was not used in this case, because when NGB is used with MLE there are not parameters to be tuned.

Regarding Test Protocol 3, the FDA reduced data set gets an average error rate of 86.22% for the 5-fold case, which a lot higher than the one obtained by NGB using the original data set (22.58%). For the 10-fold case, the data set generated by PCA obtains an average error rate of 86.29%, which again is a worst result that the one obtained using the original data set (22.54%).

3.3 Conclusions

The PCA reduced data set performs very well when used with KNN. It reduces substantially the execution time, and also lowers the error rate. However, when this same data set was used with a NGB classifier, the results were worse than the results obtained by using the original MNIST data set.

Regarding the FDA reduced data set, the results are terrible. In all test cases the error rate went up, and by an order of magnitude. Given that FDA reduces the dimensionality to just 9 dimensions, the execution times are much faster than using the original MNIST data set, but the error rates simply won't allow any practical use of the FDA reduced data set.

The result of the FDA reduced data set may be attributed to the lower dimensionality of the data set. It is possible that 9 dimensions aren't sufficient to capture enough information to make a good discrimination of the classes. The eigenvalues obtained by the PCA algorithm (Table 1) indicate that 9 dimensions just gets 46.46% of the total eigenvalues, which may be a clue that 9 dimension aren't sufficient.