SYDE 675
Assignment 2
Shala Chen
Feb 22nd 2017

1 Analytical Error Rates

1.1

Table 1: Analytical Probability of Error

	case 1		case 2		case 3	
	a1	b1	a2	b2	a3	b3
MED	0.0668	0.0668	0.3085	0.3085	0.3085	0.2854
	avg:0.0668		avg:0.3085		avg:0.2970	
GED	0.0668	0.0668	0.2248	0.2248		
	avg:0.0668		avg:0.2248			

I calculated error bounds and used them as limits in multivariate normal cumulative distribution function to get those probability of errors.

Since MED classifier only used sample mean, so the error rates on case 2 and case 3 are much higher than on case 1. GED classifier used sigma of the data as well so it's more accurate than MED classifier in case 2.

1.2

As for probability of error for case 3 using GED, I will try the bound. When its Gaussian distribution with more than 1 feature and 2 classes with unequal covariance matrix, we can use Kolmogoroff/Bhattacharyya bounds to estimate probability of error. To get upper bound and lower bound of the error, let

$$\begin{aligned} K &= P_{GED}(correct) - P_{GED}(error) \\ &= 1 - 2P_{GED}(error) \end{aligned}$$

$$\rho = \int_{\Omega} [P(x/c1) * P(x/c2)]^{1/2} dx$$

and get

$$K \geqslant 1 - 2 * \sqrt{P(c1) * P(c2)} * \rho$$

where lower bound for K = upper bound for $P_{GED}(error)$
and upper bound for K = lower bound for $P_{GED}(error)$.

2 Experimental Error Rates

2.1 use 5 samples to learn the classifier and 195 samples to test for 40 times

Table 2: probability of error using 5 samples

	MED		GED		1NN		3NN		5NN	
	$\mu_{P_{(\varepsilon)}}$	$\sigma_{P_{(\varepsilon)}}$	$\mu_{P_{(\varepsilon)}}$	$\sigma_{P_{(\varepsilon)}}$	$\mu_{P_{(\varepsilon)}}$	$\sigma_{P_{(\varepsilon)}}$	$\mu_{P_{(\varepsilon)}}$	$\sigma_{P_{(\varepsilon)}}$	$\mu_{P_{(\varepsilon)}}$	$\sigma_{P_{(\varepsilon)}}$
case1	0.0783	0.0153	0.1512	0.0881	0.1058	0.0444	0.0906	0.0321	0.0919	0.0302
case2	0.3390	0.0637	0.3447	0.0724	0.3556	0.0553	0.3567	0.0668	0.3628	0.0747
case3	0.2858	0.0649	0.3131	0.0616	0.2967	0.0539	0.2896	0.0548	0.3052	0.0611
case4	0.2188	0.0188	0.2295	0.0801	0.1617	0.0685	0.2029	0.0508	0.2275	0.0401

I select 5 samples as training set and the other 195 as testing set, and then collected the sum of wrong cases and divided them by total test cases. I put all of 40 errors in one matrix and get the overall average and standard deviation in the end.

All classifiers had low error rates for case 1 since it had fewer over lapping. They all had a high error rate for case1 2 and case 3 due to the same reason. GED and KNN classifiers could be curves so they had a much lower error rates for case 4. Since we used only 5 samples to train the classifier, 1nn had the best performance among knn classifiers while 5nn had the worst.

2.2 use 199 samples to learn the classifier and 1 sample to test for 200 times

Table 3: probability of error using 199 samples

	$\mu_{P_{(\varepsilon)}}$ for MED	$\mu_{P_{(\varepsilon)}}$ for GED	$\mu_{P_{(\varepsilon)}}$ for 1NN	$\mu_{P_{(\varepsilon)}}$ for 3NN	$\mu_{P_{(\varepsilon)}}$ for 5NN
case1	0.0600	0.1575	0.1050	0.0725	0.0700
case2	0.3325	0.2275	0.3025	0.2525	0.2800
case3	0.2650	0.2300	0.2775	0.2675	0.2300
case4	0.2050	0.1425	0.0950	0.0825	0.0750

I selected 1 sample as test case and used all the rest as training sets. and then iterated 200 times until all the cases have been used as test case. I summed up all the wrong cases and divided it by 400 to get the overall overage.

Since there were more training data for knn classifiers, their performance for 1nn, 3nn and 5nn intended to be similar. The error rates for case 2 and case 3 were still high due to the overlapping while they were low for case 1 and 4 due to the same reason.

2.3 Analytical and Experimental Error Rates Comparison

Table 4: Analytical and Experimental Error Rates Comparison

	MED Analytical	MED - 5	MED - 199	GED Analytical	GED - 5	GED - 199
case1	0.0668	0.0783	0.0600	0.0668	0.1512	0.1575
case2	0.3085	0.3390	0.3325	0.2248	0.3447	0.2275
case3	0.2970	0.2858	0.2650		0.3131	0.2300

3 Discussion

When using Analytical method to get MED and GED error rates, they were identical(0.0668) because they used the same average. When it came to Experimental methods, they were different because it involved actual distances.

The KNN classifier for 5 sample experimental methods were always high because it had too few training sets and the patterns for 2 classes were hugely overlapped.

When using 199 samples to train the KNN classifier, the error rates were high for case 2 and case 3 due to the overlapping, however, the error rates dropped a lot for case 4 because of the abundant training sets.

The classifiers trained with 199 samples usually had lower error rates compared with 5 sample cases. Because all the samples were used 200 times, which provided us with more data each time to train the classifier better.