SYDE 675
Pattern Recognition
Assignment 2

Xin Chen

20705271

FEB 11

Analytical Error Rates

Our goal is to consider deriving analytical error rates for the first three cases and derive the exact probability of error P(ϵ) for GED and MED for each of cases 1 and 2. But for case 3, we can only derive the P(ε) for MED and discusses how to figure out P(ε) for GED.

We should use this formula:

$P(\varepsilon)=\int_{R1} P(x|C_2)P(C_2)dx+\int_{R2} P(x|C_1)P(C_1)dx$

For case 1 and 2, $P(C_1)= P(C_2)=0.5$.

The probability of error for each situation:

CASE	**1**	**2**	**3**
MED	**0.067**	**0.309**	**0.239**
GED	**0.067**	**0.225**	**N/A**

Table: Analytical Error Rates

For case1, the P (ε) were also equal because the decision boundaries of MED and GED were same. As for case2, MED do not need to consider the probability of error, so P (ε) for GED is lower than MED. While for case3, it is difficult to find the P(ε) for GED. Therefore, for addressing this problem, I would like to apply an error bound estimates like using upper bounds and lower bounds rather than the exact value.

Classifier variability given very limited data

	Case 1		Case 2		Case 3		Case 4	
	μ	σ	μ	σ	μ	σ	μ	σ
MED	0.0783	0.0153	0.3390	0.0637	0.2858	0.0649	0.2188	0.0188
GED	0.1512	0.0881	0.3447	0.0724	0.3131	0.0616	0.2295	0.0801
NN	0.1058	0.0444	0.3556	0.0553	0.2967	0.0539	0.1617	0.0685
3NN	0.0906	0.0321	0.3567	0.0668	0.2896	0.0548	0.2029	0.0508
5NN	0.0919	0.0302	0.3628	0.0747	0.3052	0.0611	0.2275	0.0401

It is clear that the MED classifiers nearly had the lowest accuracy. And the data samples for Case1 overlapped least, so Case1 had the highest accuracy of all kinds of classifiers. As for the kNN classifier, when K increased, the mean of error probability also increased. But the covariance of error probability is decreasing, and the accuracy is decreasing and more focused when k increased.

Classifier accuracy and assessment using all of the data as training

	Case1	Case2	Case3	Case4
MED	0.060	0.333	0.265	0.205
GED	0.058	0.228	0.230	0.143
NN	0.105	0.303	0.278	0.095
3NN	0.073	0.253	0.268	0.083
5NN	0.070	0.280	0.230	0.075

From the table above, we could see that when K increased, the error rates decreased. It is mainly because decision boundaries of KNN blurred as K increased. Comparing Case 2 and Case 3 between Case 1 and Case 4, we will find that the error rates for Case 2 and Case 3 were higher because of the overlapping.

Experimental vs. Analytical Errors

	MED			GED		
	Analytical	Experimental value(Limited)	Experimental value(Jackknife)	Analytical	Experimental (limited)	Experimental (Jackknife)
Case1	0.067	0.078	0.060	0.067	0.151	0.058
Case2	0.309	0.339	0.333	0.225	0.345	0.228
Case3	0.239	0.286	0.265	N/A	0.313	0.230

It is clear that for Case1 and Case2, the MED and GED is equal because of the same average. Comparing the difference between limited data and jackknife in experimental value, jackknife normally had lower error rates, which is because that we use samples 200 times and get more data to train the classifier.