Laboratory 2: Optimal Bayes Classifier

Kevin Mato - K5529

November 2019

1 Identification and removal of outliers

From a first plot of the training dataset over two dimensions it's noticeable that two points lay significantly far away from the main clusters. By means of simple descriptive statistics we recognise their influence. A first point can be discovered by searching the minimum in the dataset, and the second by searching the maximum. The indexes of the samples are 186, 642 or 641 .

2 Feature selection

After the removal of the outliers, we have better visualizations of the points. The plot of features 1 and 3 gives the best insight. This one because it displays well separated clusters without any overlap. Afterwards 3 Bayes classifiers were built (Table 1) with a priori of 0.125 and 3 different probability density functions.

Table 1: Results of the Bayes classifiers for different pdf functions and a priori=0,125.

pdf indep	**pdf multi**	**pdf parzen** $window = 0.001$
0.0263158	0.0049342	0.0241228

3 Quality evaluation

3.1 Based on training set dimension/reduction

An interesting part was evaluating the quality of the classifier in relation with the dimension of the training set. The 3 classifiers were tested with respectively 10%, 25%, 50% of the whole training set. The selection of the samples is randomic, hence due to this element in the experiment, the procedure is repeated 5 times. In the next tables for each classifier are reported in order: mean error, standard deviation, maximum value of error and minimal error.

The main result shown is that the standard deviation of the error rate is decreasing, while the mean has the same effect. This is a positive result for our classifier as it means the greater number of samples the better is quality of the classifier. The classifier with a Parzen pdf seems to be the one

Table 2: Results of the Bayes classifiers, 10% of training data, a priori=0,125

pdf indep	**pdf multi**	**pdf parzen** $window = 0.001$
0.0298246	0.0091009	0.0978070
0.0033929	0.0032803	0.0135828
0.0334430	0.0142544	0.1118421
0.0246711	0.0065789	0.0838816

Table 3: Results of the Bayes classifiers, 25% of training data, a priori=0,125

pdf indep	**pdf multi**	**pdf parzen** $window = 0.001$
0.0257675	0.0067982	0.0540570
0.0025421	0.0035657	0.0075630
0.0290570	0.0120614	0.0668860
0.0230263	0.0038377	0.0476974

Table 4: Results of the Bayes classifiers, 50% of training data, a priori=0,125

pdf indep	**pdf multi**	**pdf parzen** $window = 0.001$
0.0279605	0.0061404	0.0332237
0.0013429	0.0012502	0.0035019
0.0296053	0.0076754	0.0383772
0.0263158	0.0043860	0.0296053

with best improvement but the best classification is performed by the classifier using pdf indep. The decreasing of the error is not linear with the increasing number of samples but eventually it will reach a point of saturation. This is because for a good classification we don't really need the whole dataset, and after a certain number or samples the error will decrease of very little.

3.2 Based on different Parzen windows

On the first row the window widths, on the second the coefficient of error for that window.

Table 5: Bayes classiers with Parzen pdf, a priori=0,125 and different windows widths.

window 1	**window 2**	**window 3**	**window 4**	**window 5**
0.00010000	0.00050000	0.00100000	0.00500000	0.01000000
0.02850877	0.01699561	0.02412281	0.07949561	0.13925439

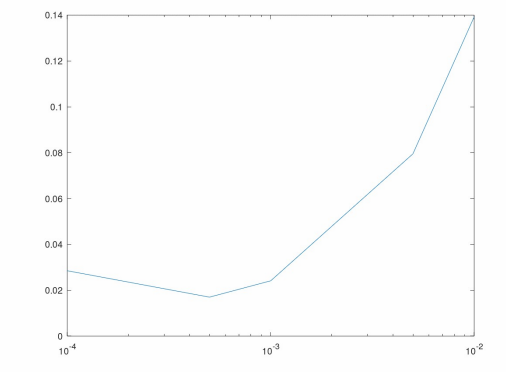

Figure 1: Parzen window widths and respective error

The results of the experiment show us that the best window width is the second as its related error coefficient is the lowest.

4 Change in the a priori probability

The experiment was run by setting a priori of the red suits two times higher than the black ones. The number of black suits was reduced of 50% in the testing set only and the experiment is repeated 5 times. Three different windows were used in the case of the parzen classifier. The next tables show the results.

Table 6: Bayes classifiers with different Parzen windows, 50% black suits

$window0.0005$	$window0.001$	$window0.005$
0.01432749	0.02017544	0.08508772
0.00065382	0.00151582	0.00240229
0.01461988	0.02119883	0.08918129
0.01315789	0.01754386	0.08333333

Table 7: Bayes classifiers with different pdf indep and pdf multi, 50% black suits

$pdfindep$	$pdfmulti$
0.03230994	0.00643275
0.00032691	0.00032691
0.03289474	0.00657895
0.03216374	0.00584795

There are not outstanding results. We can't detect a general behaviour in the results. Sometimes the changes in the error are minimal.

5 Comparison with 1NN

Before processing the data we check if they need normalization. The answer is positive as the standard deviations of the two features considered in the training set have a different order of magnitude (0.00884455 and 0.00095129). The error coefficient of 1NN is 0.0049342, and it's the same of the bayesian classifier with pdf multi in table 1. This is a great result because we can tell that 1NN is as good as a the bayesian classifier but with a higher complexity.