Department of Systems Design Engineering
University of Waterloo

Lab 1

Clusters and Classification Boundaries

Pattern Recognition
SYDE 372

Dan Hendry, 20207096.
Jenny Lu, 20172834

February 9, 2009

Introduction

This lab investigates the three related areas of calculating orthonormal transformations, creating decision boundaries, and assessing classification error. First, a set of training data for five classes are generated by applying orthonormal transformations to normally distributed clusters. Then, five classifiers – MED, GED, MAP, NN, and 5NN – are used to draw the contours of the decision boundaries using a numerical approach. Finally, another set of sample data are generated using the same distribution parameters, and the probability of error and confusion matrices for each of the five classifiers are calculated and assessed.

The lab is completed using MATLAB.

Discussion

Generating Clusters

Class A is generated according to the following multivariate Gaussian distribution parameters:

$$N_A = 200 \qquad \mu_A = \begin{bmatrix} 5 \\ 10 \end{bmatrix} \qquad \Sigma_A = \begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix}$$

Class B is generated according to the following multivariate Gaussian distribution parameters:

$$N_B = 200 \qquad \mu_B = \begin{bmatrix} 10 \\ 15 \end{bmatrix} \qquad \Sigma_B = \begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix}$$

To generate the required correlated data, normally distributed clusters are first created using the `randn` function. Then the transformation

$$V\Lambda^{\frac{1}{2}}$$

is applied to the normally distributed clusters, where V is the matrix whose columns are the eigenvectors of Σ, and Λ is the diagonal matrix with the eigenvalues of Σ on the main diagonal. The `eig` function is used to calculate matrices V and Λ for a given Σ. Finally, the transformed clusters are shifted by μ to obtain the clusters for classes A and B as required. The resulting clusters are plotted with their respective unit standard deviation contours in Figure 1.

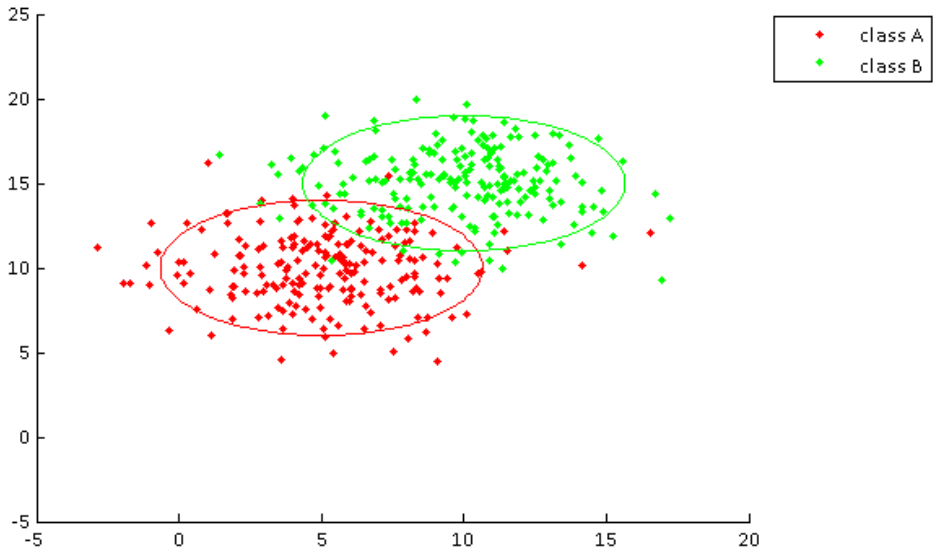

Figure 1: Clusters of classes A and B

Class C is generated according to the following multivariate Gaussian distribution parameters:

$$N_C = 100 \qquad \mu_C = \begin{bmatrix} 5 \\ 10 \end{bmatrix} \qquad \Sigma_C = \begin{bmatrix} 8 & 4 \\ 4 & 40 \end{bmatrix}$$

Class D is generated according to the following multivariate Gaussian distribution parameters:

$$N_D = 200 \qquad \mu_D = \begin{bmatrix} 15 \\ 10 \end{bmatrix} \qquad \Sigma_D = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$$

Class E is generated according to the following multivariate Gaussian distribution parameters:

$$N_E = 150 \qquad \mu_E = \begin{bmatrix} 10 \\ 5 \end{bmatrix} \qquad \Sigma_E = \begin{bmatrix} 10 & -5 \\ -5 & 20 \end{bmatrix}$$

To obtain the clusters for classes C, D, and E, the transformations described above are again applied to normally distributed clusters. The resulting clusters are plotted with their respective unit standard deviation contours in Figure 2.

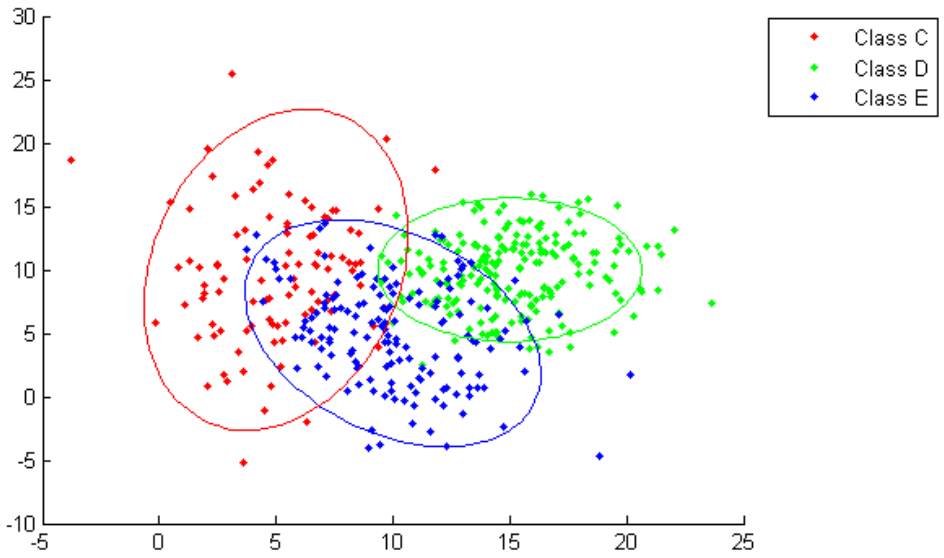

Figure 2: Clusters of classes C, D, and E

As shown in Figure 1 and Figure 2, most of the samples lie within the unit standard deviation contour. Visually, the unit standard deviation contour forms a rough outline of the shape of each cluster of sample data.

Classifiers

Parametric Classifiers

The MED classifies a point x as c_i by minimizing the following distance function to the prototype z_i for classes $i = 1 \dots n$:

$$d_E^2 = (\vec{x} - \vec{z_i})^T (\vec{x} - \vec{z_i})$$

The GED classifies a point x as c_i by minimizing the following distance function to the prototype z_i for classes $i = 1 \dots n$:

$$d_{GED}^2 = (\vec{x} - \vec{z_i})^T \Sigma_i^{-1} (\vec{x} - \vec{z_i})$$

The GED classifies a point x as c_i by maximizing the following probability function for classes $i = 1 \dots n$ with Gaussian distributions:

$$\mathrm{P}(\vec{x}) = \mathrm{P}(\vec{x}|c_i)\mathrm{P}(c_i)$$

$$= \frac{1}{2\pi|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{z_i})^T \Sigma_i^{-1}(\vec{x} - \vec{z_i})\right) \frac{N_i}{\sum_{i=1}^{n} N_i}$$

Non-Parametric Classifiers

NN classifies a point x as c_i if c_i contains the sample data point that is closest to x by the Euclidean distance. 5NN classifies a point x as c_i if c_i contains the sample data point that is the 5th closest to x by the Euclidean distance.

Parametric Classifiers for Classes A and B

The boundaries of parametric classifiers MED, GED, and MAP for classes A and B are shown in Figure 3.

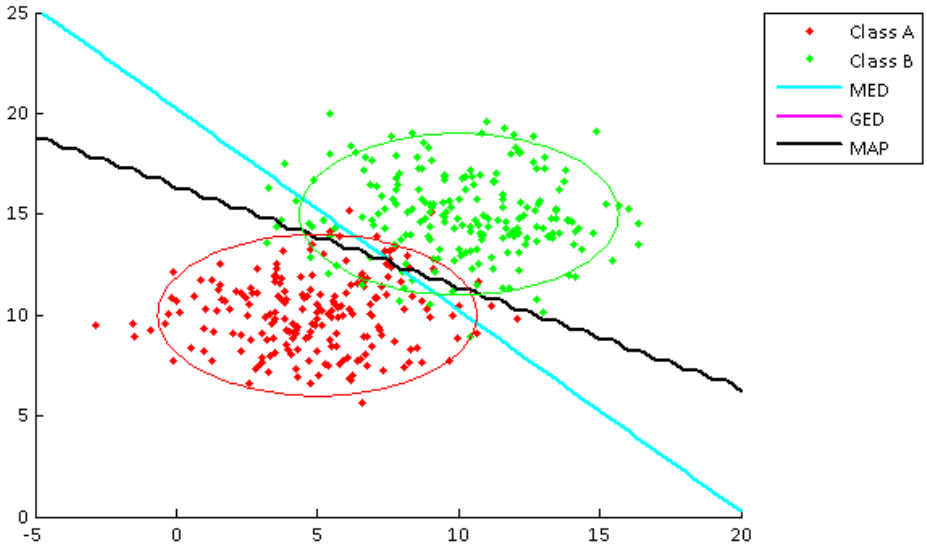

Figure 3: MED, GED, and MAP boundaries for classes A and B

The boundary of the MED is the perpendicular bisector between the prototypes of classes A and B as expected. It does not take into account the scaling of the distributions along the axes.

Since the distribution for both classes is Gaussian, the MAP is expected to be similar to the GED, but with a bias towards more compact and more likely classes. In this case, since $\Sigma_A = \Sigma_B$, the classes have unit standard deviation contours with equal areas, implying that the two classes are equally compact. Since $N_A = N_B$, the number of samples in classes A and B are equal, implying that the two classes are equally likely. Hence, the GED and MAP for classes A and B are exactly the same, and their boundaries are overlapping as shown in Figure 3. As expected, the MAP and GED are better classifiers than the MED because they both take into account the scaling of the distributions along the axes.

In all three cases the boundaries of the parametric classifiers are dependent only on the distribution parameters and are visually related to the unit standard deviation contours. They are not influenced by the specific set of training samples used.

Non-Parametric Classifiers for Classes A and B

The boundaries of non-parametric classifiers NN and 5NN for classes A and B are shown in Figure 4.

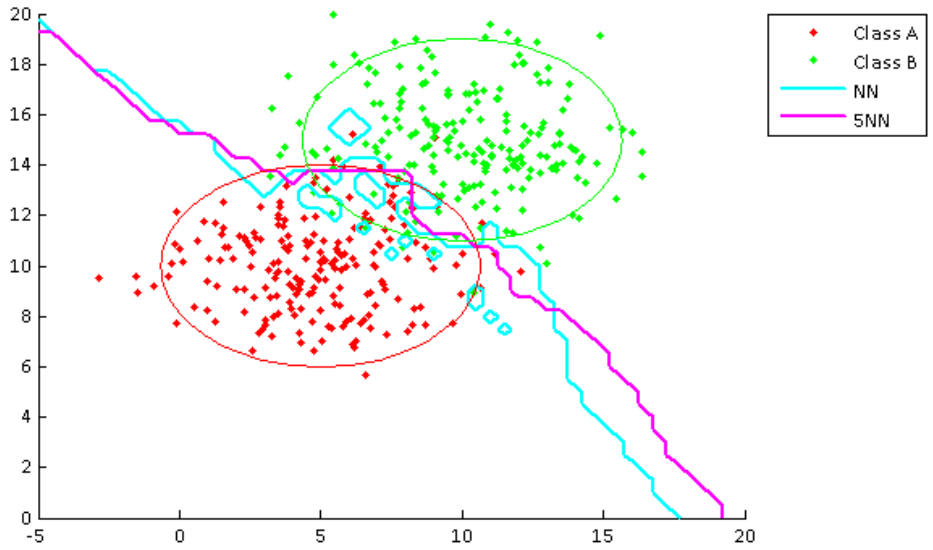

Figure 4: NN and 5NN boundaries for classes A and B

The boundary for NN shows that the classifier is sensitive to outliers as expected, as single outlier samples influence the classifier and create boundaries of small areas around the single outlier samples that should not exist. The boundary for 5NN shows that the classifier is significantly less outlier sensitive, as the extraneous boundaries around single outlier samples no longer exist.

In both cases the boundaries of the non-parametric classifiers are dependent on and are visually related to the specific set of training samples used.

Parametric Classifiers for Classes C, D, and E

The boundaries of parametric classifiers MED, GED, and MAP for classes C, D, and E are shown in Figure 5.

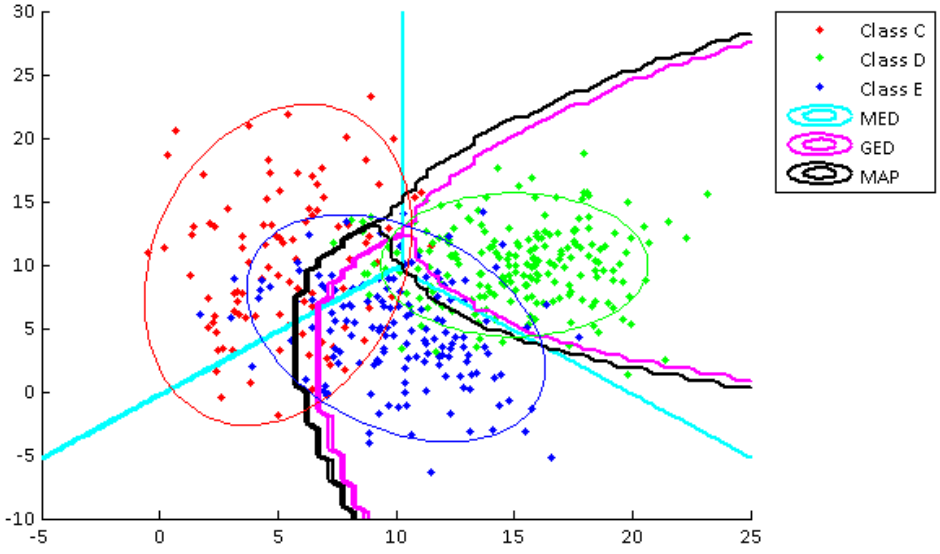

Figure 5: MED, GED, and MAP boundaries for classes C, D, and E

The boundaries of the MED are the perpendicular bisectors between the prototypes of classes C, D, and E as expected. It does not take into account the scaling of the distributions along the axes nor the rotation of the distributions from the directions of the axes.

Since the distribution for both classes is Gaussian, the MAP is expected to be similar to the GED, but with a bias towards more compact and more likely classes. In this case, since $|\Sigma_{\mathrm{D}}| < |\Sigma_{\mathrm{E}}| < |\Sigma_{\mathrm{C}}|$, the areas of the unit standard deviation contours increase in order of classes, D, E, and C, implying that the compactness of the classes are decreasing in the same order. Since $N_D > N_E > N_C$, the number of samples in are decreasing again in order of classes D, E, and C, implying that the likelihood of the classes are decreasing in the same order. Hence, compared to the GED, MAP has a bias toward class D followed by class E, and the boundaries of the MAP are shifted away from classes D and E and towards class C as shown in Figure 5. As expected, the MAP and GED are better classifiers than the MED because they both take into account the scaling of the distributions along the axes and the rotation of the distributions from the directions of the axes. Furthermore, the MAP is a better classifier than the GED because it also takes into account the compactness and likelihood of the classes.

In all three cases the boundaries of the parametric classifiers are dependent only on the distribution parameters and are visually related to the unit standard deviation contours. They are not influenced by the specific set of training samples used.

Non-Parametric Classifiers for Classes C, D, and E

The boundaries of non-parametric classifiers NN and 5NN for classes C, D, and E are shown in Figure 6.

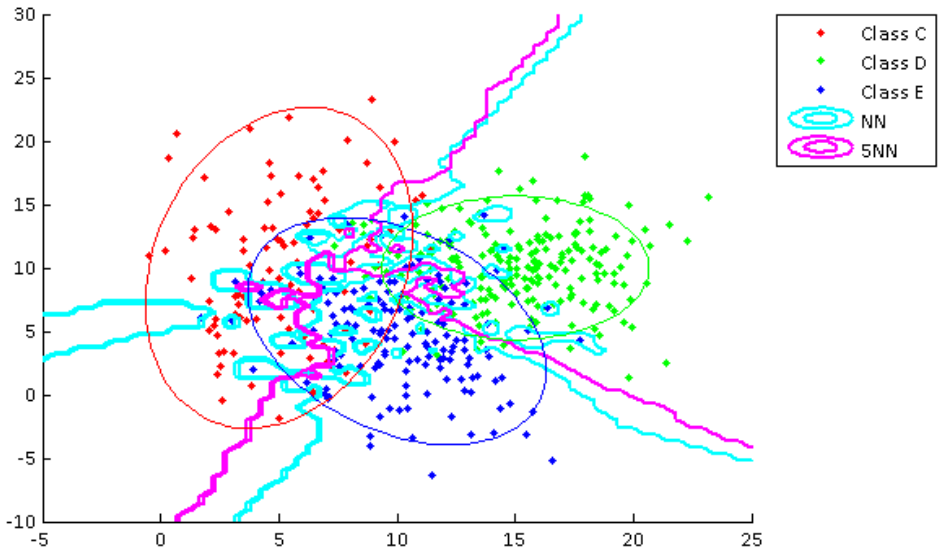

Figure 6: NN and 5NN boundaries for classes C, D, and E

The boundary for NN shows that the classifier is sensitive to outliers as expected, as single outlier samples influence the classifier and create boundaries of small areas around the single outlier samples that should not exist. The boundary for 5NN shows that the classifier is significantly less outlier sensitive, as the extraneous boundaries around single outlier samples are mostly removed.

In both cases the boundaries of the non-parametric classifiers are dependent on and are visually related to the specific set of training samples used.

Error Analysis

A new set of testing data is generated following the steps used to generate the set of training data.

Experimental Error Rate

The experimental error rate is calculated as

$$\mathrm{P}(\varepsilon) = \frac{\text{\# of misclassified samples}}{\text{total \# of samples}}$$

Table 1 lists the experimental error rates for each of the classifiers in each case.

Table 1: Experimental Error Rate

	MED	GED	MAP	NN	5NN
c_A, c_B	0.0850	0.0600	0.0600	0.0975	0.0800
c_C, c_D, c_E	0.1978	0.1689	0.1533	0.2356	0.1667

As expected, MAP gives the minimum experimental error rate in both cases. In the first case, since the GED and MAP are exactly the same, the experimental error rate for the two classifiers is also the same, as expected. In the second case, MAP gives a smaller experimental error rate than GED, which gives a smaller experimental error rate than MED, as expected.

In the non-parametric case, in both cases, 5NN gives a smaller experimental error rate than NN, as expected. Furthermore, the experimental error rate of 5NN approaches the minimum experimental error rate of MAP due to the relatively large set of training data used.

Confusion Matrices

Table 2, Table 3, Table 4, Table 5, and Table 6 show the confusion matrices of MED, GED, MAP, NN, and 5NN respectively. The counts shown by the confusion matrices are consistent with the experimental error rates determined above.

The confusion matrices for GED and MAP are identical, as expected.

Table 2: MED confusion matrix for classes A and B

		Classified As	
		c_A	c_B
True Class	c_A	189	11
	c_B	23	177

Table 3: GED confusion matrix for classes A and B

		Classified As	
		c_A	c_B
True Class	c_A	192	8
	c_B	16	184

Table 4: MAP confusion matrix for classes A and B

True Class \ Classified As	c_A	c_B
c_A	192	8
c_B	16	184

Table 5: NN confusion matrix for classes A and B

True Class \ Classified As	c_A	c_B
c_A	181	19
c_B	20	180

Table 6: 5NN confusion matrix for classes A and B

True Class \ Classified As	c_A	c_B
c_A	189	11
c_B	21	179

Table 7, Table 8, Table 9, Table 10, and Table 11 show the confusion matrices of MED, GED, MAP, NN, and 5NN respectively. The counts shown by the confusion matrices are consistent with the experimental error rates determined above.

For all classifiers, the number of misclassified samples in class E and the number of samples misclassified as class E is far larger than the number of samples misclassified between classes C and D. This is consistent with the distribution of the clusters, as class E overlaps considerably with both classes C and D.

Table 7: MED confusion matrix for classes C, D, and E

True Class \ Classified As	c_C	c_D	c_E
c_C	73	3	24
c_D	1	184	15
c_E	28	18	104

Table 8: GED confusion matrix for classes C, D, and E

True Class \ Classified As	c_C	c_D	c_E
c_C	91	1	8
c_D	2	184	20
c_E	31	14	105

Table 9: MAP confusion matrix for classes C, D, and E

True Class \ Classified As	c_C	c_D	c_E
c_C	83	1	16
c_D	0	190	10
c_E	20	22	108

Table 10: NN confusion matrix for classes C, D, and E

True Class \ Classified As	c_C	c_D	c_E
c_C	72	4	24
c_D	1	175	24
c_E	22	31	97

Table 11: 5NN confusion matrix for classes C, D, and E

True Class \ Classified As	c_C	c_D	c_E
c_C	78	3	19
c_D	1	189	10
c_E	20	22	108

Conclusions

The results from this lab are consistent with the theoretical behavior of the five classifiers. In both cases, MAP proved to be the best classifier, both in terms of the visual relationship between the boundaries and training data set and minimizing classification error. For classes of equal standard deviation and probability, such as in the first case, the GED and MAP are identical, and thus give the same boundaries and classification error. In the non-parametric case, 5NN proved to be a better classifier than NN, both in terms of the visual relationship between the boundaries and training data set and minimizing classification error, since it effectively minimizes sensitivity to outliers. In both cases, the classification error rate of 5NN approaches the minimum classification error rate of MAP due to the relatively large set of training data used.