University of Waterloo

SYDE 372: Pattern Recognition

Lab 1: Clusters and Classification Boundaries

**Team Members:**

Adrian Chan | 20548981

Devika Khosla | 20570142

Maria Cheng | 20567883

Michael Lim | 20567193

Vyshakh Sanjeev | 20554572

### Introduction

In pattern recognition, it is intuitive that if two data points have similar features, they should be assigned to the same class. In this lab, different methods of classification will be explored using clusters of random points generated from Gaussian distributions. This lab will cover three related areas in clusters and classification boundaries, which includes calculating orthonormal transformations, creating decision boundaries, and assessing classification error.

Two cases will be considered throughout the lab. The first case will consider two uncorrelated classes, with the same number of points, different means, and the same covariances. The second case will consider three correlated classes, with different sizes, means, and covariances. Distance and probability measure classifiers will be used to define the decision boundaries. The classifiers that will be investigated are minimum Euclidean distance (MED), generalized Euclidean distance (GED), maximum a posteriori (MAP), nearest neighbour (NN) and k-nearest neighbour (kNN), where k=5. Lastly, classifier performance will be observed by analyzing the error rates through experimental data and constructing confusion matrices.

### Generating Clusters

Two cases each with different sample clusters are randomly generated and plotted along with the unit standard deviator contour for each class. Figure 1 shows case 1 with two classes (A and B), and Figure 2 shows case 2 with 3 classes (C, D, and E).



Figure 1. Case 1 distributions with unit contours

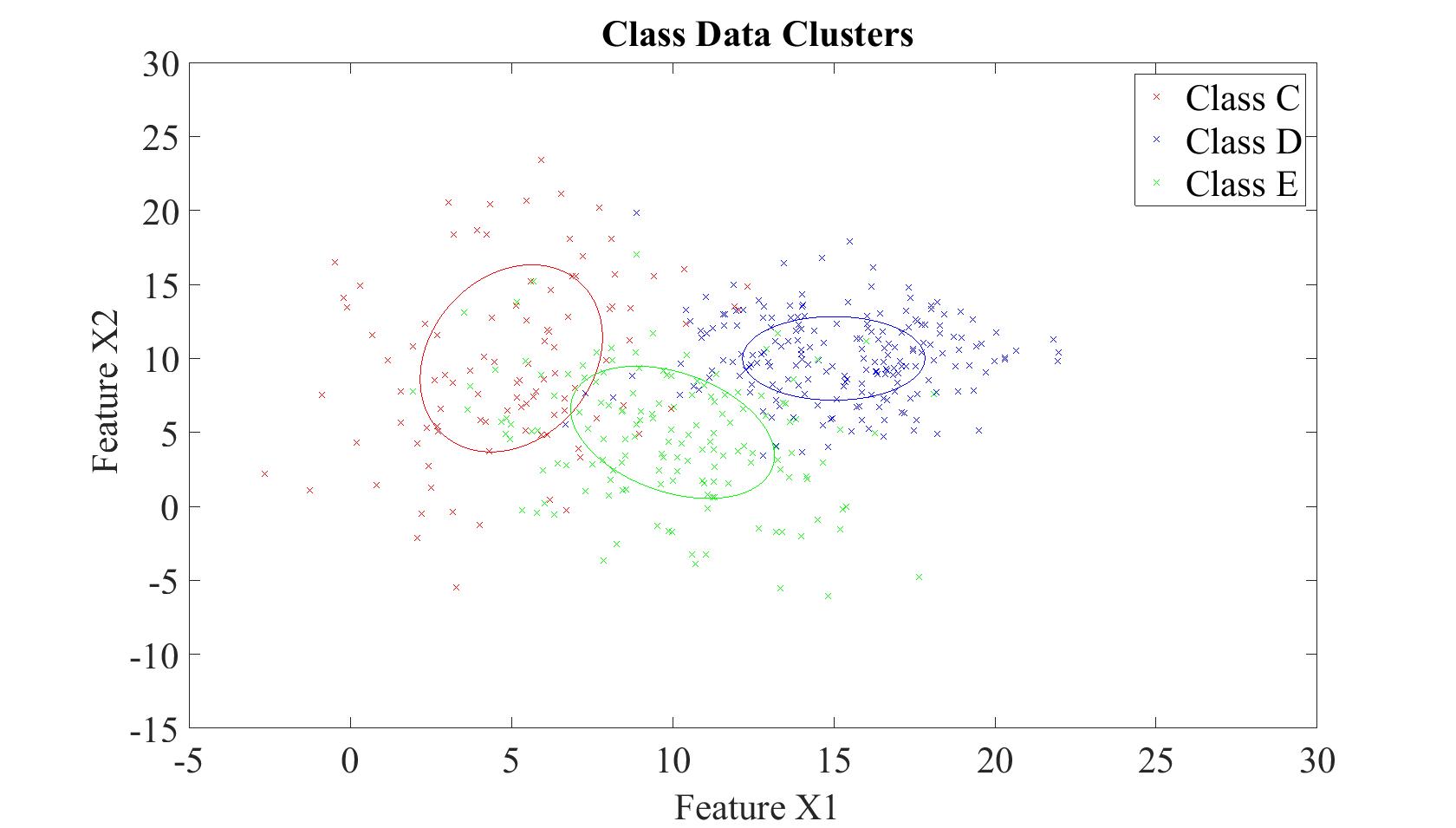


Figure 2. Case 2 distributions with unit contours

The unit contours represent the covariance and mean within the data of a class. Visually, the contours reflect the mean and spread of the distributions as well as the correlation between the variables within the different classes.

### Classifiers

Intuitively, patterns with similar features should be assigned to the same class. A simple method to quantify similarity would be to measure the Euclidean distance between the pattern and the prototype of each class and assign the pattern to the class with the minimum Euclidean distance – the MED classifier, where the prototype of each class is the mean point of that class. This would mean that the decision boundary would be a hyperplane that bisects the class prototypes. A given pattern *x* belongs to class A if the criteria below is met, where μ is the mean of the class distribution.

Ideally, we want to maximize the distance between classes and minimize the distances within classes (intra-class distance). Thus, we opt for the Minimum Intra-class Distance (MICD) classifier, also known as the generalized Euclidean distance (GED) classifier. This is an extension of the MED classifier, where the covariance of the data is also accounted for. Thus, in addition to the mean of each class, the spread in each class is also taken into consideration in calculating the decision boundary. This is done by identifying the locations of the unit standard deviations of each class and then defining a hyperplane that goes through the intersection point of every unit standard deviations of the given classes. Thus, the unit of measure becomes the number of standard deviations and the pattern is assigned to a class to which it is closest to in units of standard deviations. A given pattern *x* belongs to class A if the criteria below is met, where μ is the mean and S is the covariance of the class distribution.

Further extending the MICD classifier to incorporate the shape of the distribution (in addition to the mean and covariance), we have the Maximum a Posteriori (MAP) classifier. This becomes particularly useful for non-Gaussian distributions. By incorporating the conditional probability density distributions, we can better predict which class a given pattern belongs.

In most applications however, the posterior distribution is difficult to measure. The likelihood however is much simpler to calculate from the data. Thus, by using Bayes’ rule, we can restructure this classifier to use the likelihood function and the prior (from the data at hand) to redefine the MAP classifier to be:

For a simple Gaussian distribution, the above equation can be simplified to be,

Where μ is the mean and is the standard deviation and P(A) and P(B) are the priori distributions that are proportional to the number of elements within each class.

Figure 3 and Figure 4 show plots of the case 1 and 2 class samples, their unit standard deviation contours, and the MED, MICD, and MAP decision boundaries.

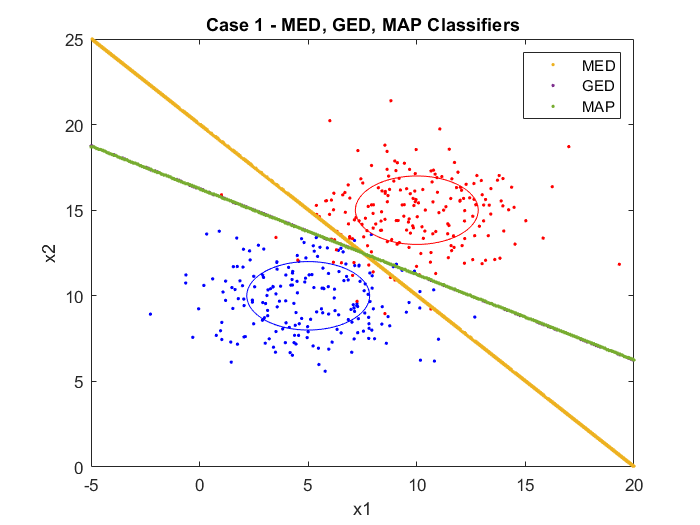


Figure 3. Case 1 with MED, MICD, and MAP decision boundaries.

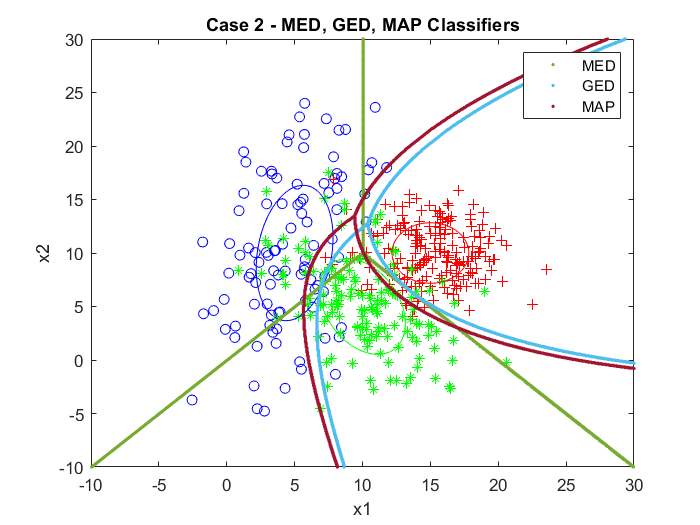
Recall in case 1, there are two sets of random, normally distributed points, with different means and the same covariances. The MED classifier forms a boundary between the two classes that is equidistant and perpendicular to a line drawn between the means of each class. The MICD and MAP classification boundaries reside on top of each other. Although the decision boundary generated using the MICD classifier is in general not linear, in Case 1, both distributions have the same covariance. As a result, the decision boundary is linear. The slope is different to that generated using MED as the spread in x and y is different - since there is more spread along x than y, the line becomes less steep. Since the a priori class probabilities for MAP are proportional to the number of samples in each class, there is equal a priori probabilities. Furthermore, as previously mentioned, the covariances of the two classes are also identical. As a result, the MAP classifier simplifies to the MICD classifier and the decision boundaries formed are the same.

Figure 4. Case 2 with MED, MICD, and MAP decision boundaries.

Case 2 consists of classes with different number of sample points, means, and covariances. The MED classifier plots boundaries that still remain equidistant from the means of each class and is not influenced at all by the distributions of the class. MICD considers these distributions and by multiplying them by the inverse covariance matrix, normalizes them to give uncorrelated features with unit variance for comparable distances. As a result, the decision bounds are represented by parabolas and wraps more around the points. The MAP classifier takes into account the a priori probabilities proportional to the number of points in each sample. The MAP bound appears similar to the MICD classifier shifted left, to account for higher number of points for class D (red).

NN and kNN classifiers also rely on distance measures to classify points; however, the prototype that defines the class differs and is not represented by the class mean. Rather, for NN the prototype is the sample point closest to the point to be classified.

NN:

kNN is a variation of NN in which the prototype is the mean of the k NNs closest to the point to be classified. In this lab, k = 5, in which the mean of the 5 NN to the test point is chosen as the prototype.

Figure 5 and Figure 6 show plots of the case 1 and 2 class samples, their unit standard deviation contours, and the NN and kNN decision boundaries.

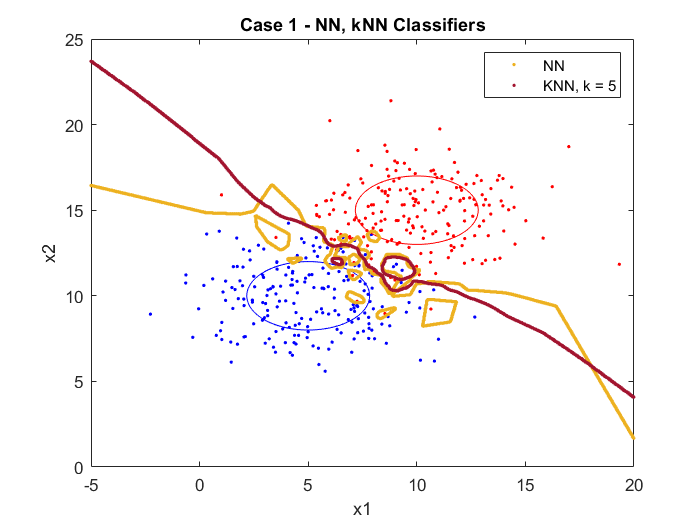


Figure 5. Case 1 with NN and kNN decision boundaries.

In the NN case, the decision boundary curves and fits the data points. This is an example of overfitting as blue points that are clearly in the red region result in decision boundaries forming around them. However, this only pertains to this particular set of points. The kNN classifier generalizes better than NN and smooths out the boundaries, being less sensitive to noise and outliers. However, a mini decision boundary was formed by a small cluster of noise in the red region. One way to mitigate this would be to increase the value of k points needed to determine the prototype.

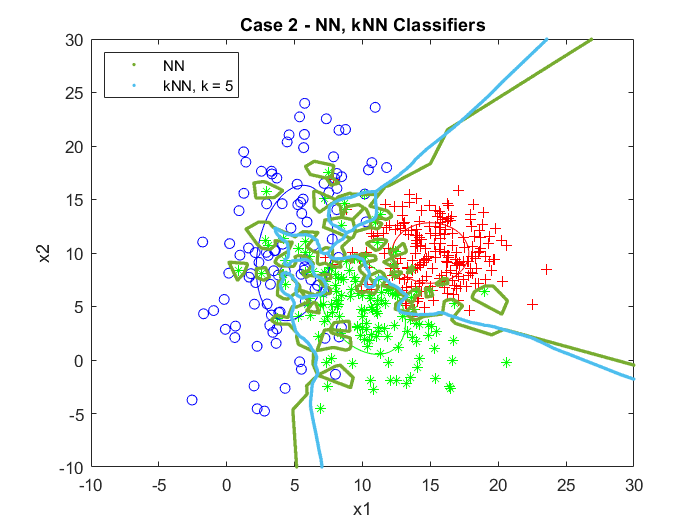


Figure 6. Case 2 with NN and kNN decision boundaries.

Similar decision boundary formation is represented in Figure 6 with case 2 samples. NN is extremely sensitive to outliers and results in multiple formations of decision boundaries. kNN results in more generalized boundaries forming around each class.

### Error Analysis

For case 1 the MICD and MAP classifiers generate the same result because both classes have the same covariance matrices and the same a priori class probability. The classifiers with the lowest errors in case 1 were 5NN, with an error rate of 4.25%, followed by MICD/MAP with an error rate of 5.75%.

For case 2, the classifier with the smallest error was MAP, at 14.89%, followed by MICD, at 18.22%. From the confusion matrices, it appears unlikely that any points from C will be confused with D and vice versa. It is more likely to misclassify points between classes D and E, or between classes C and E. Looking at the clusters, this makes sense - spatially, class E is in the center, between classes C and D.

The confusion matrices were generated using the decision boundary grids created for the previous section (Classifiers). For each data point in the cluster, the code finds the nearest point in the decision boundary grid. Then it compares the label at that point to the class that the data point belongs to and adds 1 to the appropriate category. For example, for case 1 if the data point were from class A, and the label was A, it would add 1 to the top left corner of the matrix (predicted A + actual A). If the label was B, it would add 1 to the top right corner (predicted B + actual A). The error rates were calculated by taking the total number of errors and dividing it by the total number of points: .

The confusion matrices and error rates are shown below.

**Case 1:**

Overall experimental error rate = 6.55%

MED:

Error rate = 7.75%

|  |  |  |
| --- | --- | --- |
|  | Predicted A | Predicted B |
| Actual A | 187 | 13 |
| Actual B | 18 | 182 |

GED/MICD:

Error rate = 5.75%

|  |  |  |
| --- | --- | --- |
|  | Predicted A | Predicted B |
| Actual A | 191 | 9 |
| Actual B | 14 | 186 |

MAP:

Error rate = 5.75%

|  |  |  |
| --- | --- | --- |
|  | Predicted A | Predicted B |
| Actual A | 191 | 9 |
| Actual B | 14 | 186 |

NN:

Error rate = 9.25 %

|  |  |  |
| --- | --- | --- |
|  | Predicted A | Predicted B |
| Actual A | 180 | 20 |
| Actual B | 17 | 183 |

5NN

Error rate = 4.25%

|  |  |  |
| --- | --- | --- |
|  | Predicted A | Predicted B |
| Actual A | 186 | 14 |
| Actual B | 3 | 197 |

**Case 2:**

Overall experimental error rate = 20.27%

MED:

Error rate = 20.22%

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted C | Predicted D | Predicted E |
| Actual C | 74 | 7 | 19 |
| Actual D | 1 | 178 | 20 |
| Actual E | 30 | 13 | 107 |

GED/MICD:

Error rate = 18.22%

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted C | Predicted D | Predicted E |
| Actual C | 89 | 2 | 9 |
| Actual D | 4 | 168 | 28 |
| Actual E | 29 | 10 | 111 |

MAP:

Error rate = 14.89%

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted C | Predicted D | Predicted E |
| Actual C | 84 | 2 | 14 |
| Actual D | 1 | 182 | 17 |
| Actual E | 17 | 16 | 117 |

NN:

Error rate = 27.11%

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted C | Predicted D | Predicted E |
| Actual C | 64 | 2 | 34 |
| Actual D | 3 | 170 | 27 |
| Actual E | 29 | 27 | 94 |

5NN:

Error rate = 21.11%

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted C | Predicted D | Predicted E |
| Actual C | 62 | 3 | 35 |
| Actual D | 1 | 181 | 18 |
| Actual E | 15 | 23 | 112 |

### Conclusions

In summary, we compared the use of 5 different classifiers with two different cases to study the effects of the classifier on each case, and to study how the decision boundaries are affected by the class parameters (number of sample points, mean, and covariance).

By comparing the MED, MICD and MAP boundaries in case 1, we can see that the MICD boundary (which accounts for the covariance in addition to the mean) has a different slope to MED (which is only dependent on the mean) as the spread in the classes are different along the x and y axes. The MICD boundary is linear as the two classes have the same covariance. The MAP and MICD boundaries are identical and, in this case, where the two classes have the same number of sample points and covariance, the MAP classifier simplifies to the MICD classifier. In case 2 however, where the covariances are different, we see that the MICD boundary is quadratic. Since the number of sample points are different (the a priori distribution is different), the MAP classifier is offset from the MICD classifier to account for this difference. The MED classifier is still linear as it is only dependent on the position of the means. Thus, the MED and MICD classifiers can be thought of as specific case uses of the MAP classifier.

We also studied the effects of using the NN and kNN classifiers on the two cases. As discussed above, the NN overfits the data and perfectly envelopes all the sample points - this is observed in both cases. Since the kNN classifier takes the average point of k nearest neighbours, it is less susceptible to noise. However, to better improve its performance against noise, we can increase the value of k. Note however that as the k value increases, we are approaching the MED classifier. Setting k = n, where n is the total number of sample points, we have the MED classifier and by setting k = 1, we have the NN classifier.

Overall, for both cases the MAP classifier had the lowest error. Intuitively this makes sense, because the MAP classifier considers the most information. MAP considers the whole probability distribution, whereas NN only considers the points, MED the mean, and MICD the mean and covariance. Moreover, in case 2 the confusion matrices showed that the least errors occurred when classifying between C and D, which shows that classes C and D overlap less than C and E, or D and E.