SYDE 372 Lab Report #1

Drew Gross - 20328775
Jake Nielsen - 20338042

Introduction

In this lab we examined 5 different types of classifiers, namely “Minimum Euclidean Distance” (MED), “Generalized Euclidean Distance” (GED), “Maximum A Posteriori” (MAP), “Nearest Neighbor” (NN), and k-Nearest Neighbor (KNN).

For each of the different classifiers two different sets of bivariate gaussian distributed class data were used. All of the 5 classifiers were trained on a sample of the class data, and then used to classify a new sample of the class data. The error rates and confusion matrices for the classifiers were generated experimentally and compared.

2. Generating Clusters

In this section we generate data points to use as our classes, based on supplied means and covariance matrixes.

Scatter plots of the classes, with unit standard deviation contours, are visible in figures 1 and 2.

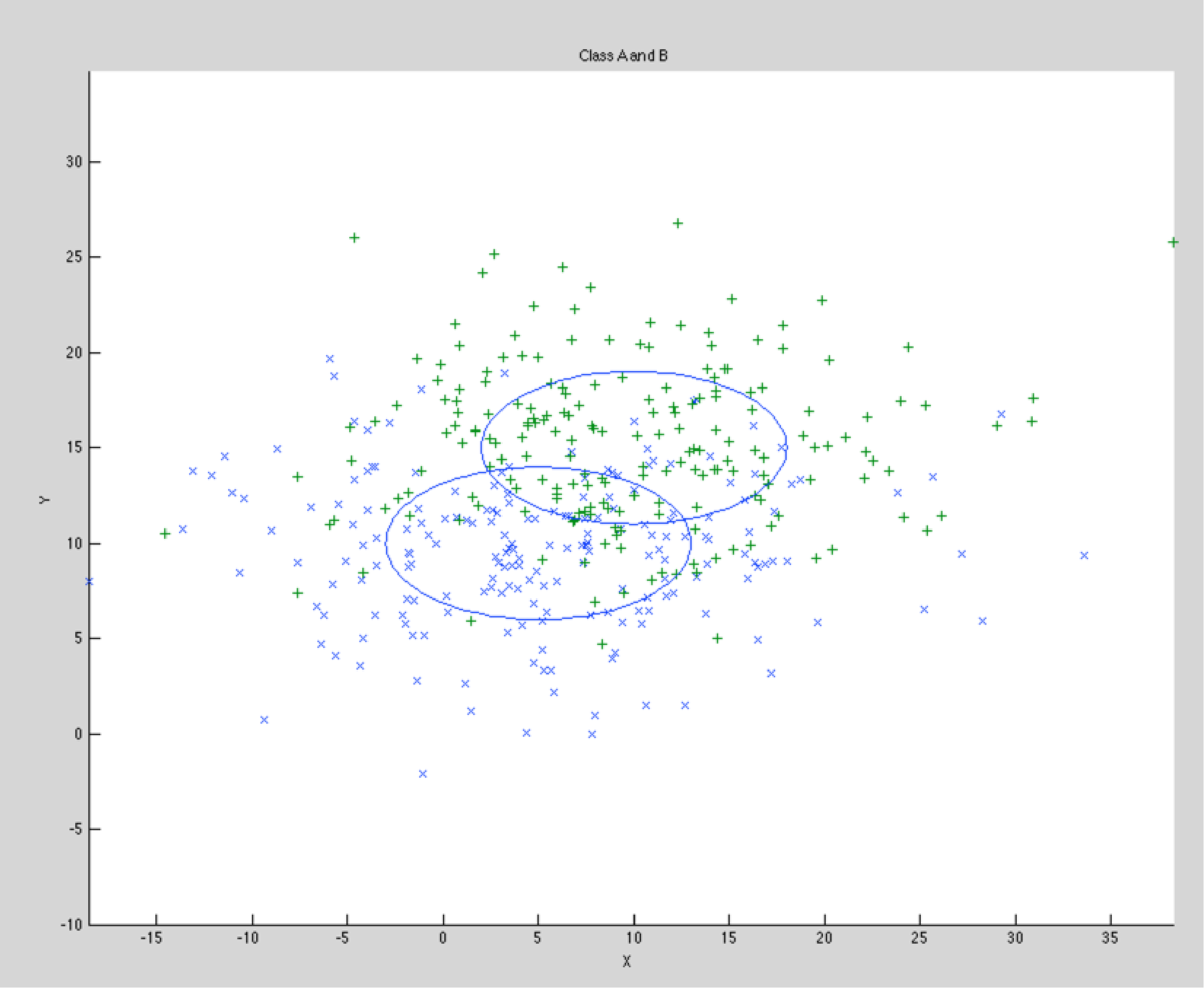

Figure 1: Plots of class A and B

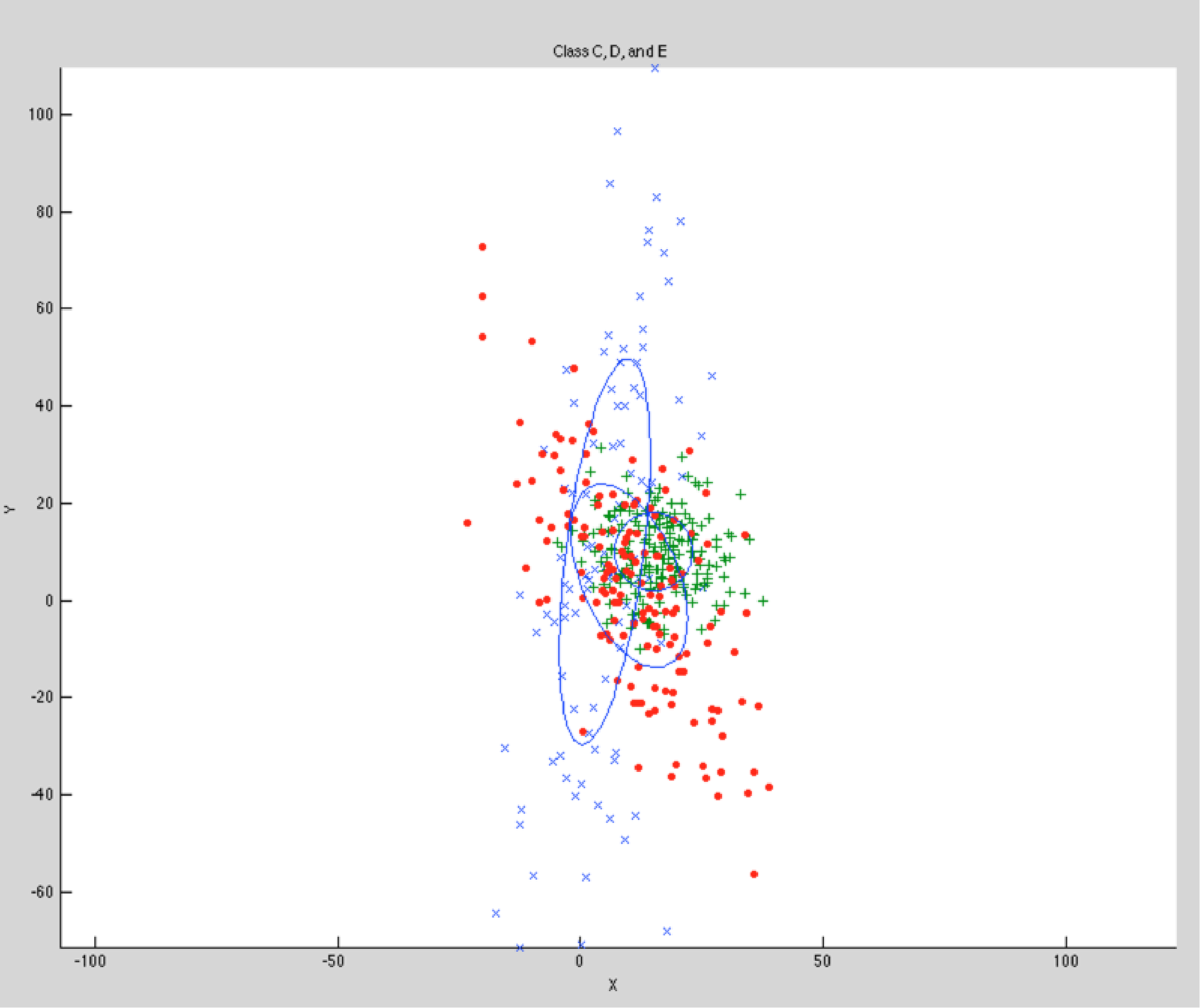

Figure 2: Plots of class C, D, and E

Visually, the unit standard deviation contour for each class captures the approximate shape and size of the distribution of points. The class distribution lies on the same axes as the contour, and the size of the ellipse is proportional to the spread of the class.

3. Classifiers

In this section we build 5 different types of classifiers using the data we generated in section 2.

First, we used an MED classifier to classify points. We found the decision boundary numerically by classifying every point in a grid and plotting where the decision changes. We then created a GED and MAP classifier and found the decision boundary in the same way. The boundaries for class A and B can be seen in figure 3, with the MED boundary in green. Because class A and B have the same number of data points, the boundary for GED and MAP is identical, and it is shown in blue. We did the same thing for classes C, D, and E, and the boundaries can be seen in figure 4. The boundary for MED is in red, GED in green, and MAP in blue.

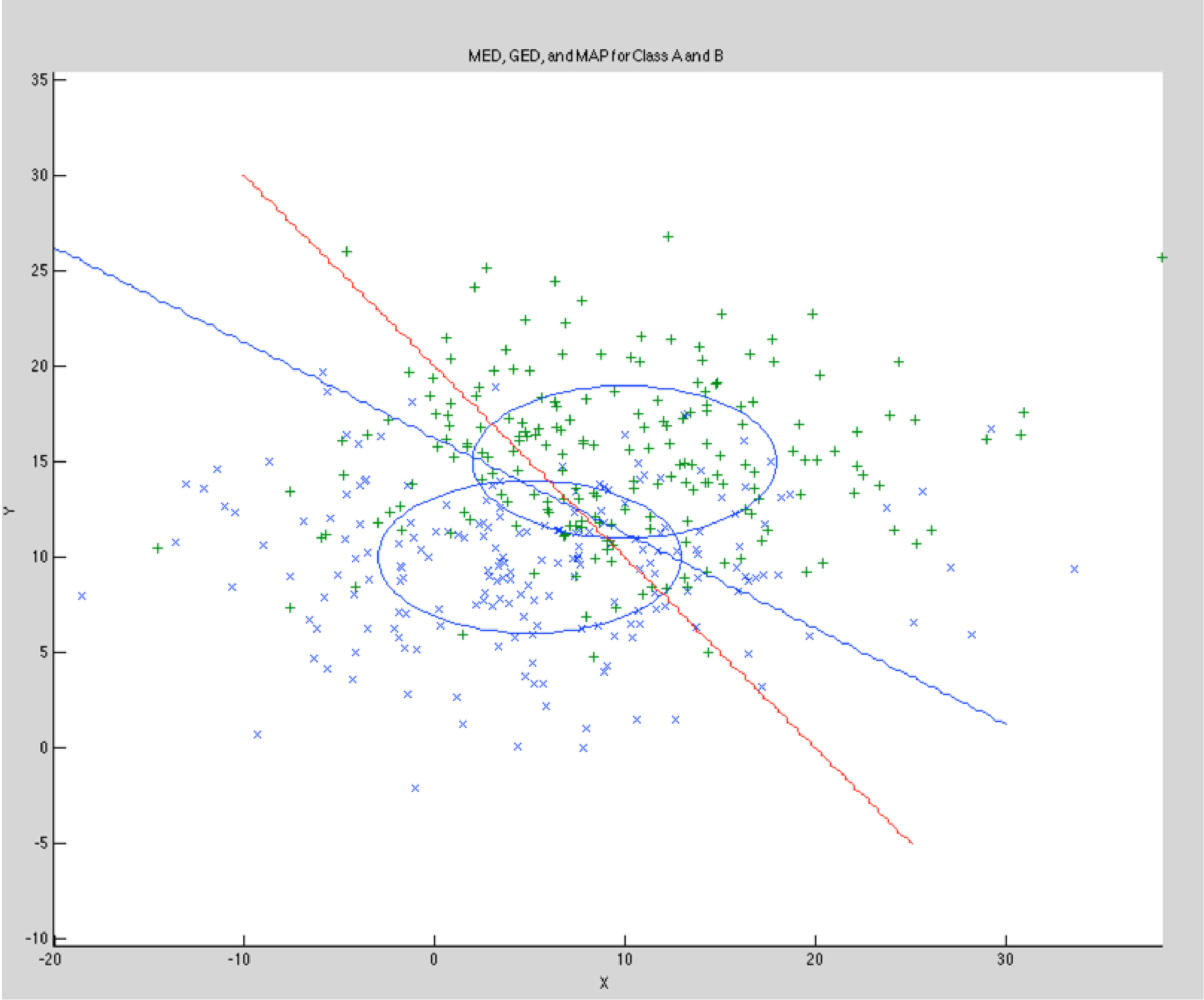

Figure 3: MED, GED, and MAP decision boundary for class A and B

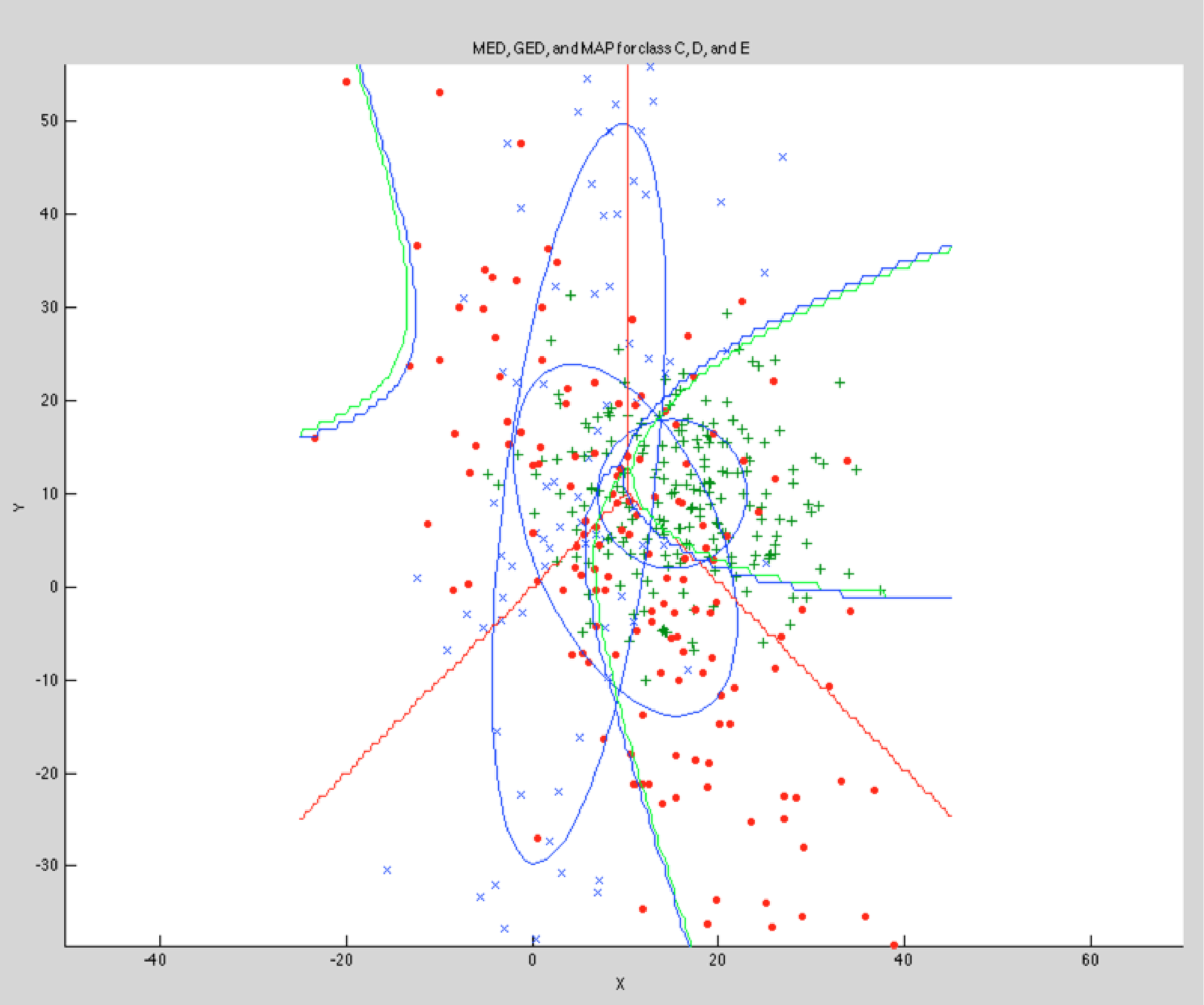

Figure 4: MED, GED, and MAP decision boundary for classes C, D and E

Next, we used NN and kNN with k=5 classifiers to classify points, and found the decision boundary in the same way as we did previously. The resulting decision boundaries can be seen in figure 5 for class A and B, with the boundary for NN in red and the boundary for kNN in blue, and in figure 6 for class C, D, and E, with the boundaries in the same colours.

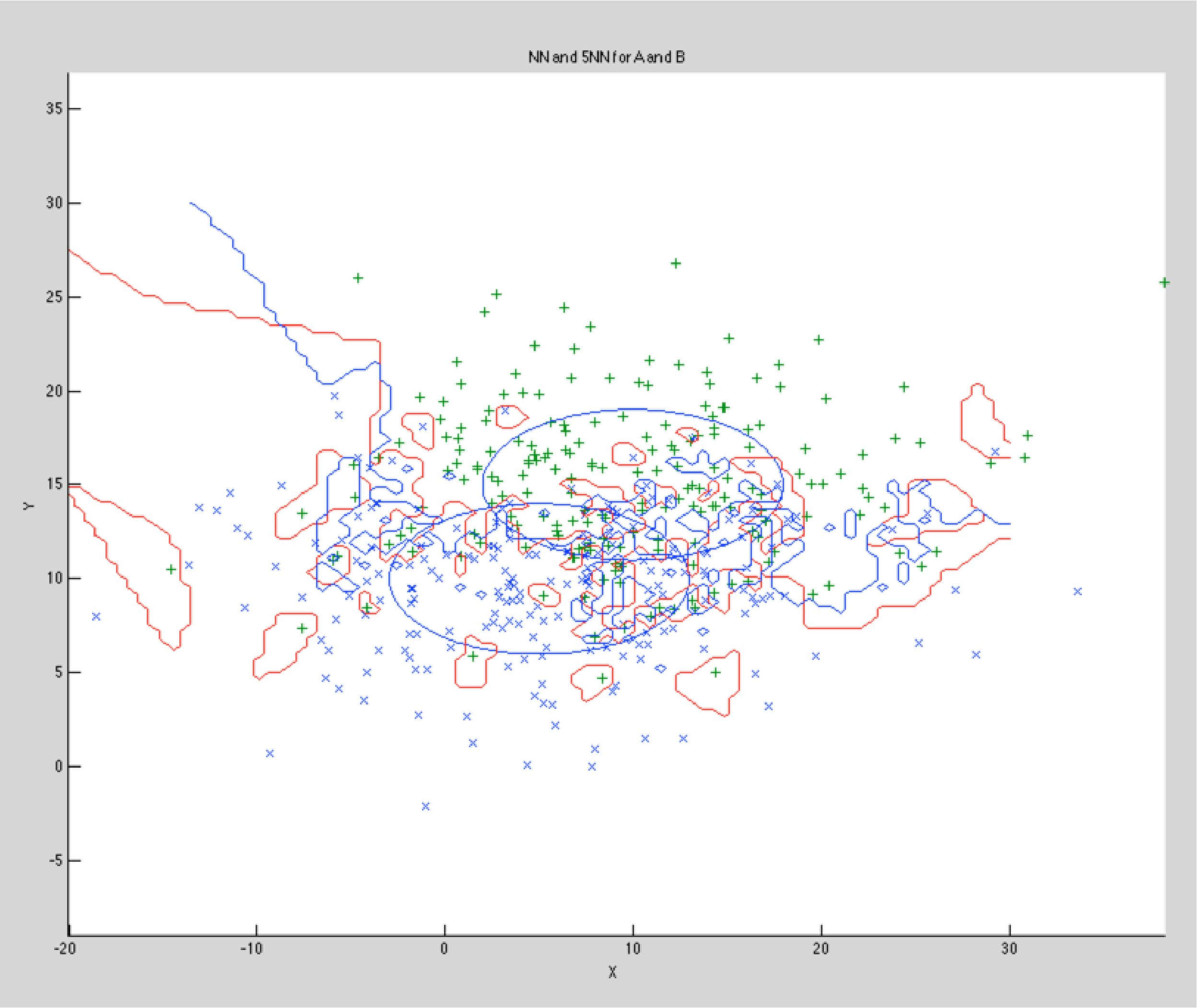

Figure 5: NN and kNN boundaries for class A and B

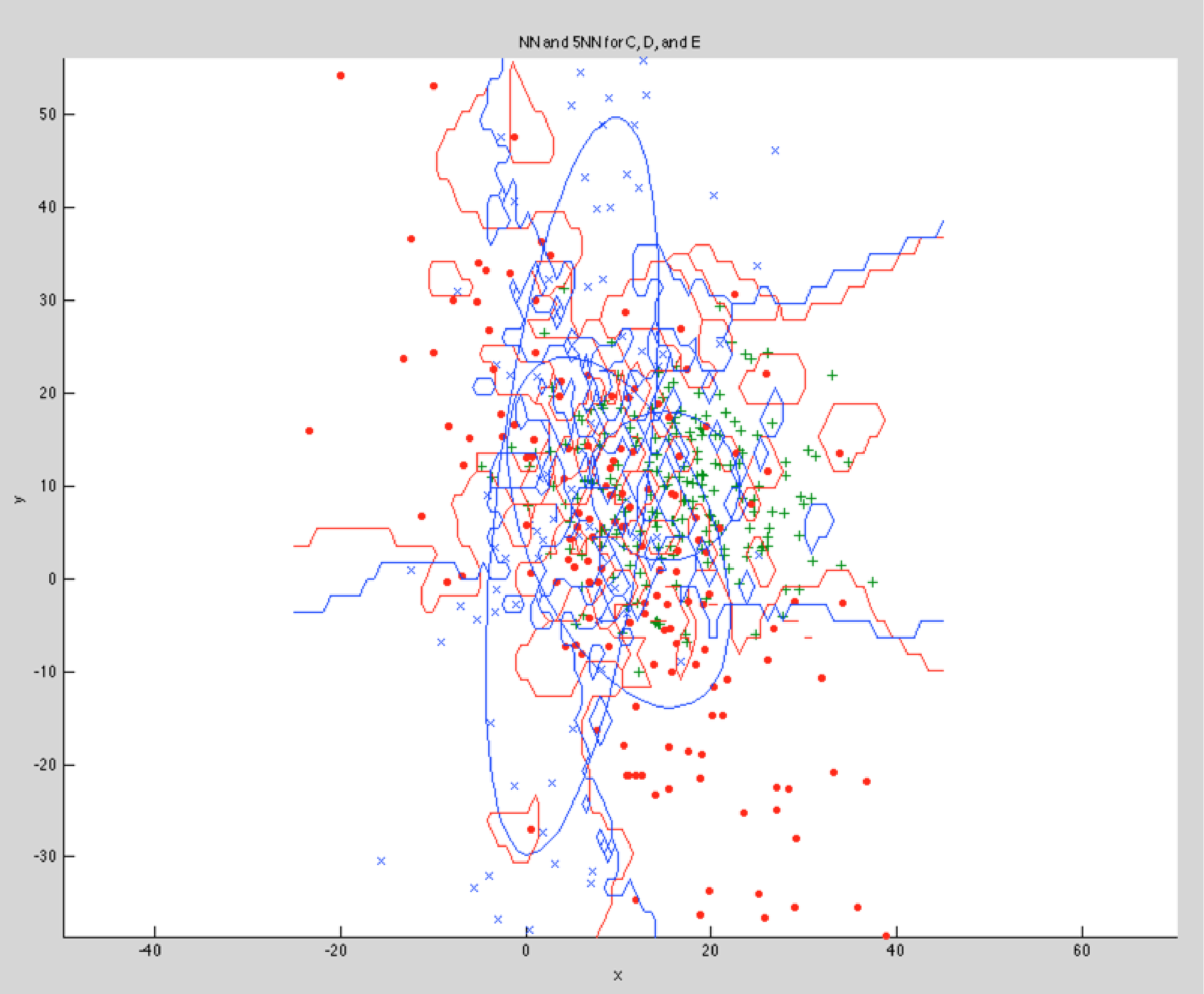

Figure 6: NN and kNN boundaries for class C, D, and E

The different classification boundaries are very different. The MED decision boundary forms a hyperplane, equidistance to the means of the classes it is distinguishing between. GED and MAP are very similar, and in class A and B they are the same because A and B have the same prior probability. NN and kNN are radically different from MED, GED, and MAP. NN has many small pockets of regions with a different classification, due to the outliers in the original distribution. kNN has smaller pockets, however every set of 5 outliers will still have a small pocket at their center where they will influence classification. Unfortunately it is difficult to see these pockets in the class C, D, and E plot, but they can be seen in the class A and B plot.

4. Error Analysis

MED Classifier

The MED classifier was the most error-prone. Its error rate and confusion matrix for the A and B classes were:

$P(\varepsilon) = 0.2925$

143	57
60	140

For the C, D, and E classes they were:
$P(\varepsilon) = 0.5244$

30	28	42
39	113	48
63	16	71

GED Classifier

The GED classifier was second-worst. The error rate and confusion matrix for the A and B classes were:
$P(\varepsilon) = 0.2375$

149	51
44	156

For the C, D, and E classes they were:
$P(\varepsilon) = 0.4333$

91	3	6
51	98	51
76	8	66

MAP Classifier

The MAP classifier’s performance was better than MED and GED, but worse than NN and KNN . The error rate and confusion matrix for the A and B classes were:
$P(\varepsilon) = 0.2375$

149	51
44	156

For the C, D, and E classes they were:
$P(\varepsilon) = 0.4156$

89	4	7
43	106	51
73	9	68

NN Classifier

The NN classifier's performance was second-best. The error rate and confusion matrix for the A and B classes were:
$P(\varepsilon) = 0.3275$

132	68
63	137

For the C, D, and E classes they were:
$P(\varepsilon) = 0.3533$

62	12	26
8	140	52
19	42	89

KNN Classifier

The KNN classifier's performance was the best. The error rate and confusion matrix for the A and B classes were:
$P(\varepsilon) = 0.3125$

135	65
60	140

For the C, D, and E classes they were:
$P(\varepsilon) = 0.3222$

61	11	28
7	141	52

17	30	103

Summative Statements

In the second case (which was the most complicated case), KNN performed the best. The error rate when the experiment was run multiple times varied by as much as 8%, but KNN was always in the lead in the CDE case. KNN was also the most computationally expensive. The majority of the time taken in running the experiment was spent in the KNN section of the code.

In the confusion matrices a trade-off seems to have been made by the NN and KNN classifiers as compared to the MED, GED, and MAP. In particular, NN and KNN are more likely to misclassify C as D or E, but less likely to misclassify D as a C. That trade-off seems to have been made to the overall advantage of the system as a whole, although at the cost of the accuracy of classifying things of type C.

Conclusion

In conclusion, in more complex systems (even ones that are completely gaussian in nature) NN and KNN algorithms seem to perform significantly better than MED, GED, or MAP. Unfortunately NN and KNN are also much more computationally expensive than MED, GED, or MAP. In more simple systems, MED, GED, and MAP seem to perform significantly better than NN, and KNN.