Assignment 1

Bayesian Classifier is minimum error rate classifier for classifying an n-dimensional data. The minimum error rate classification can be achieved by use of the discriminant functions.

$$g_i(x) = \ln p\left(\frac{x}{C_i}\right) + \ln P(C_i) \qquad (1)$$

$p\left(\frac{x}{C_i}\right)$ is the conditional probability or likelihood of x in C_i , $P(C_i)$ is the prior probability.

As the data given follows normal distribution, that is conditional probabilities $p\left(\frac{x}{C_i}\right) \sim N(\mu_i, \sigma_i)$, therefore the multivariate normal density in equation 2 becomes to equation 3 to find the discriminant functions between two regions.

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-(x-\mu)^t \Sigma^{-1}(x-\mu)] \qquad (2)$$

$$g_i(x) = \frac{-(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)}{2} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| + \ln P(C_i) \qquad (3)$$

In this assignment, the task is to determine functionality of Bayes Classifier for 3 different types of data, under 4 different cases as given below.

3 Types of data are:

1) Linearly Separable 2) Non Linearly Separable 3) Real World data.

4 different cases are:

Case 1: Classifier for which covariance matrices of all 3 classes is same and is equal to $\sigma^2 I$

Case 2: Classifier for which full covariance matrices of all 3 classes are equal.

Case 3: Classifier for which covariance matrix is diagonal and is different for each class.

Case 4: Classifier for which each class has full individual covariance matrices.

All data points are assumed to follow Gaussian distribution.

Each type of data contains 3 classes {C1, C2 and C3} having two dimensions x and y values. Set of *discriminant functions* $g_i(x)$ is computed for each class i = 1, 2, 3.

The classifier is said to assign the given point to Class i using $argmax_i\ (g_i(x))$. As per the equation 1, $g_i(x)$ is the maximum discriminant function corresponds to the maximum posterior probability. $g_i(x)$ can be any function as long as classification remains unchanged. So, we choose a function 'f' such that $f(g_i(x))$ is a monotonically increasing function. One such function is logarithmic function which converts all the products terms to sum terms and makes the computation easy.

1. Case1: Covariance matrix for all the classes is the same and is $\sigma^2 I$

"If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyper plane of d-1 dimensions, perpendicular to the line separating the means" [Duda, Hart, Stork]. In this case, the boundary which separates two distribution is a line orthogonal to the line joining the mean of the two distributions. In the Figure 1, we can observe that the discriminant boundary is linear and orthogonal to the mean of the distributions. Figure 1

Figure 1 below shows the data points plotted with their corresponding constant density contours. Since the variance is forced to be same for all classes, we observe circular contours, which means, points are equidistant from the mean μ and are spread-out in a way that they form a symmetric Gaussian bell curve. Concentric circles show that the probability density of innermost circle is higher than the outer circle. Points closer to the mean are highly probable to belong to that class than the points that are farther from the mean. From the plot we can see that the data points of any two classes can be separated by a straight line.

The appropriate covariance matrices are achieved by averaging the covariance matrices of all classes, making the non-diagonal elements 0 and assigning the resultant matrix to the covariance matrix of each of the class. Also, the classifier assumes that the feature vectors are independent of each other.

1.1. Classification of Linearly Separable Data

In this type, the classes C1, C2 and C3 are disjoint sets and do not overlap each other. It can be classified using the linear discriminant boundary i.e. line when the covariance of the classes are equal and using quadratic discriminant boundary when the covariance are unequal. In this assignment, we analyze the nature of the discriminant boundary under four different cases.

Confusion Matrix

Confusion matrix shows the visualization of the performance of an algorithm.

	Predicted Class1	**Predicted Class2**
Actual Class 1	True Positive (TP)	False Negative (FN)
Actual Class 2	False Positive (FP)	True Negative (TN)

True Positive: In the total number of test examples, the actual is class 1(positive) and predicted is also class 1(positive).

True Negative: In the total number of test examples, the actual is class 2(negative) and predicted is also class 2(negative).

False Negative: In the total number of test examples, the actual is class 1, but predicted is class 2(negative).

False Positive: In the total number of test examples, the actual is class2, but predicted is class 1(positive).

Accuracy:

This metrics give the proportion of correct classification of test vectors given the trained model. It is given by,

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN})} \qquad (\mathbf{4})$$

Precision:

The fraction of sum of True Positives to the sum of predicted positives are known as Precision.

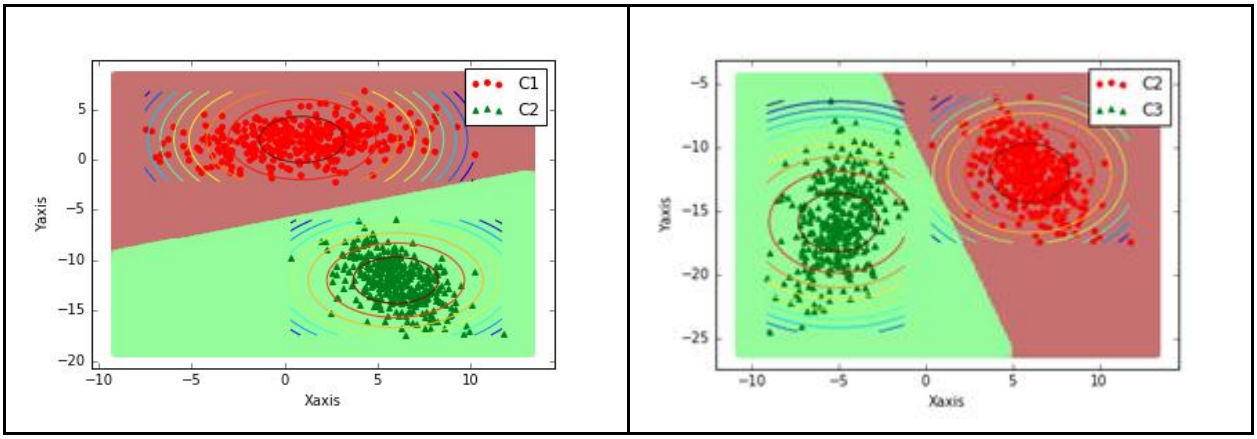
$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \qquad (\mathbf{5})$$

Recall:

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \qquad (\mathbf{6})$$

F- Measure:

$$F - score = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (\mathbf{7})$$



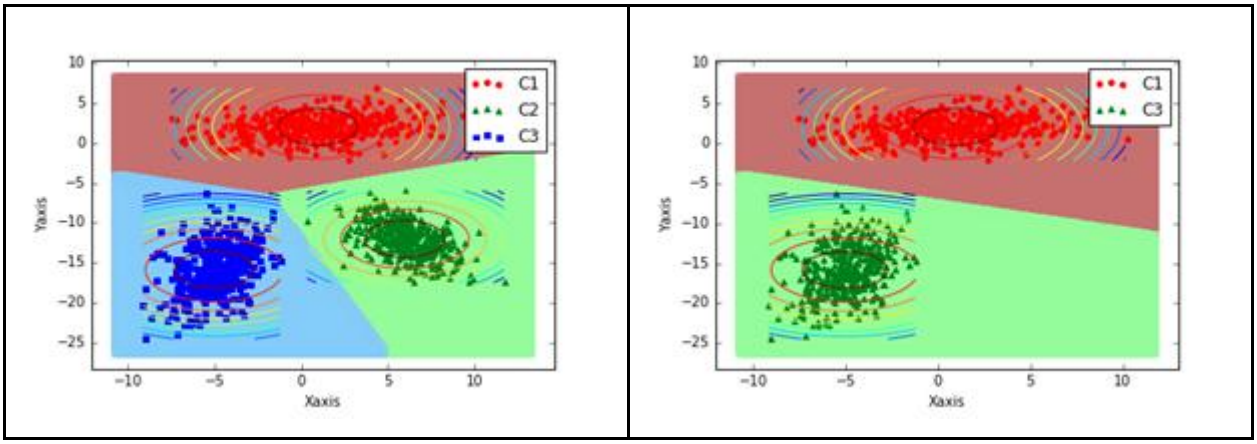


Figure 1: Type 1, Case 1: Region plot and contour plots considering Case 1 covariance matrices for Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

In fig1, pic 1, looks like P(C1) == P(C2) because the decision boundary is cutting the line joining the means at the midpoint.

Table 1.1.1 The Confusion Matrix for the linear separable, case 1

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*125*	*0*	*0*
	Class 2	*0*	*125*	*0*
	Class 3	*2*	*0*	*123*

Table 1.1.2 The Performance Matrix for the linear separable, case 1

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*98.43*	*100*	*99.21*
Class 2	*100.0*	*100.0*	*100.0*
Class 3	*100.0*	*98.4*	*99.19*
Mean Value	*99.48*	*99.47*	*99.47*

Classification Accuracy (%): 99.4

The given points correspond to the linear data. So, they can be ideally separable by a line but we are forcing the covariance matrices to be diagonal and same for all classes which is why, the test data is mis-classified by a very small percentage.

Non-Linearly Separable Data

Using Case 1 covariance matrices, the non-linear data has been classified as shown in the Figure 2, the decision boundary is linear perpendicular to the mean of the corresponding classes.

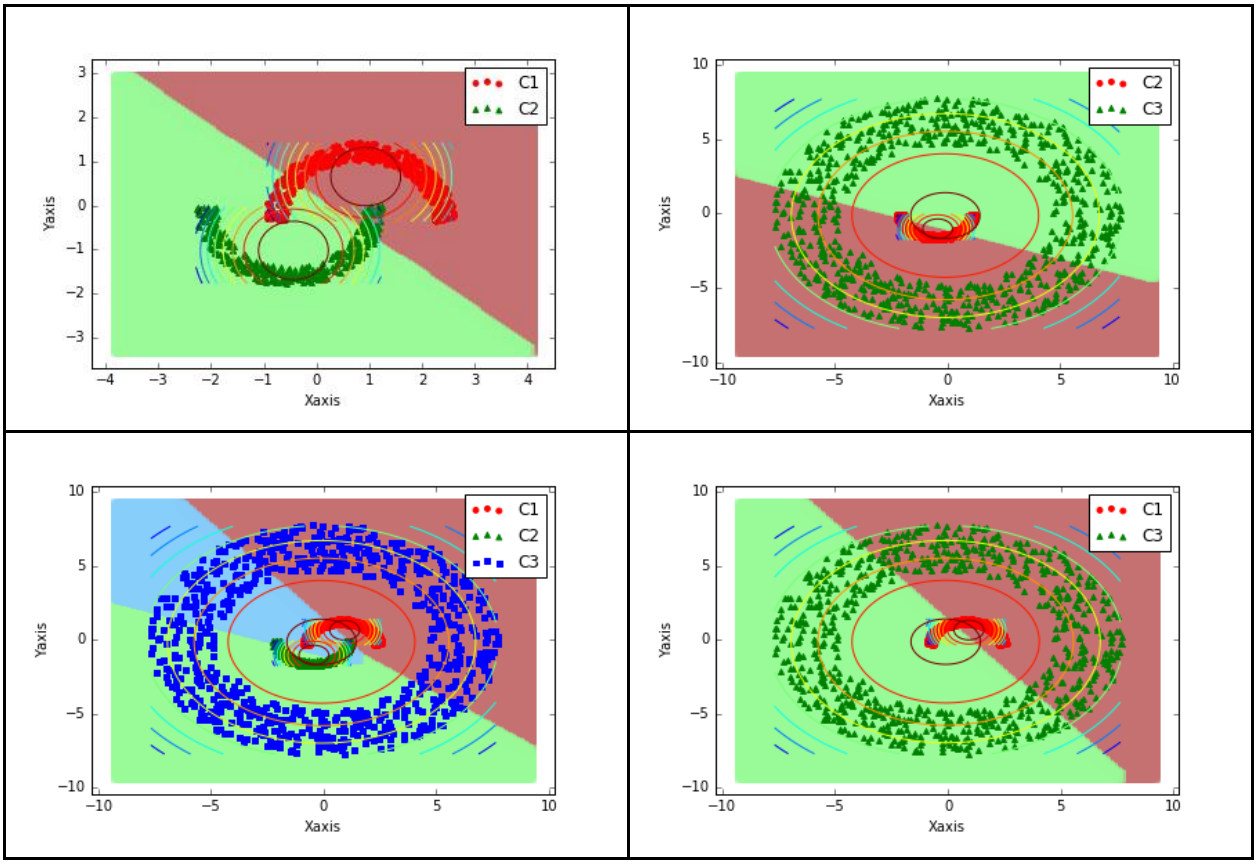


Figure 2: Type 2, Case1: Region plot and contour plots considering Case 1 covariance matrices for Non-Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

The data points are nonlinear. We have forced the classifier to believe that the covariance matrices of all classes are same. We have calculated the mean of all the data points in a class independently. So the Gaussian distribution forms concentric circles around the means of their respective classes. Also, the decision boundary is not accurately discriminating between the two classes but is perpendicular to the line joining the two means.

Table 2.1.1 The Confusion Matrix for the non-linear separable, case 1

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*95*	*0*	*30*
	Class 2	*16*	*100*	*9*
	Class 3	*126*	*98*	*26*

Table 2.1.2 The Performance Matrix for the non-linear separable, case 1

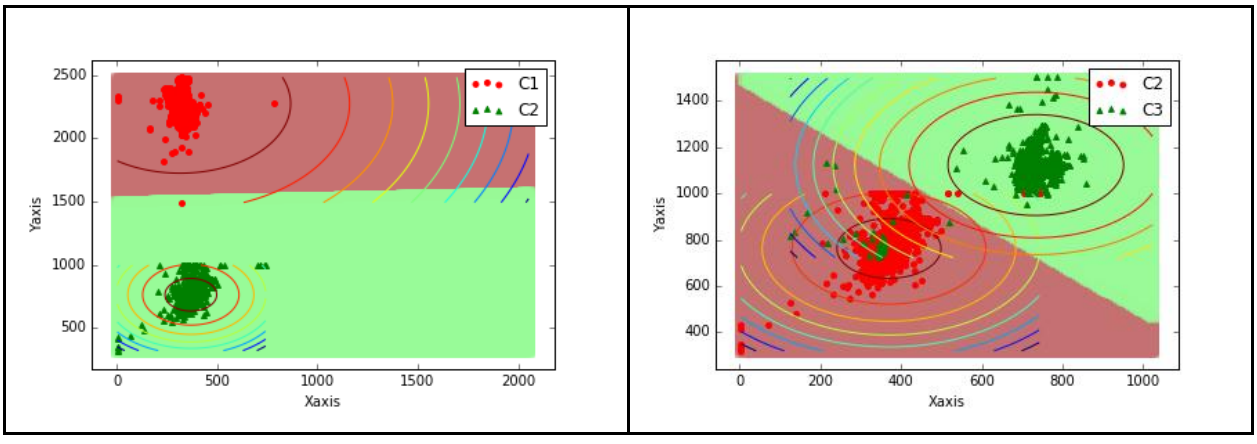
	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*40.08*	*76.0*	*52.49*
Class 2	*50.51*	*80.0*	*61.92*
Class 3	*40.0*	*10.4*	*16.51*
Mean Value	*43.53*	*55.47*	*43.641*

Classification Accuracy (%): 44.2

**The reason for the low accuracy is very evident. Classifier is trying to discriminate nonlinear data points using a line. So, more than 50% of the test data points are misclassified. Classifier assumes that the covariance matrices for all classes are same, while in reality they are not.

Real-World Data:

Using Case 1 covariance matrices, the Real world data has been classified as shown in the Figure 2, the decision boundary is linear perpendicular to the mean of the corresponding classes. The contour plot in this case is a circle illustrates the spread of the data in the two dimensional space.



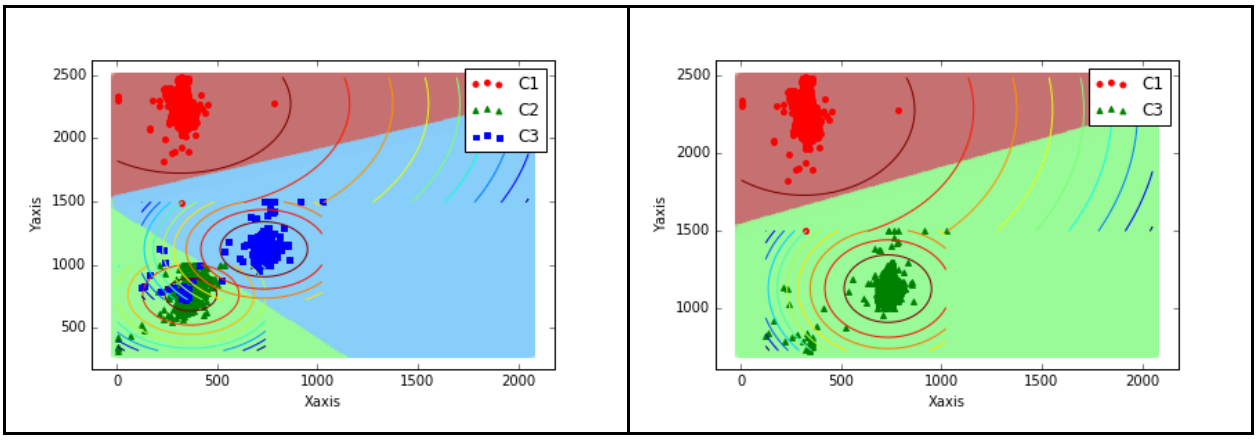


Figure 3: Type 3, Case1: Region plot and contour plots considering Case 1 covariance matrices for Real World Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 3.1.1 The Confusion Matrix for the real world, case 1

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*589*	*3*	*4*
	Class 2	*0*	*613*	*9*
	Class 3	*0*	*27*	*514*

Table 3.1.2 The Performance Matrix for the real world, case 1

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*100.0*	*98.33*	*99.41*
Class 2	*95.33*	*98.55*	*96.92*
Class 3	*97.53*	*95.01*	*96.25*
Mean Value	*97.62*	*97.46*	*97.53*

Classification Accuracy (%): 97.55

2. Case 2: Classifier for which full covariance matrices of all 3 classes are equal.

We are forcing the covariance matrices of all classes to be equal and we are achieving this by averaging covariance matrices of all classes and assigning to each of them.

This will result in linear discriminant function.

Linearly Separable Data

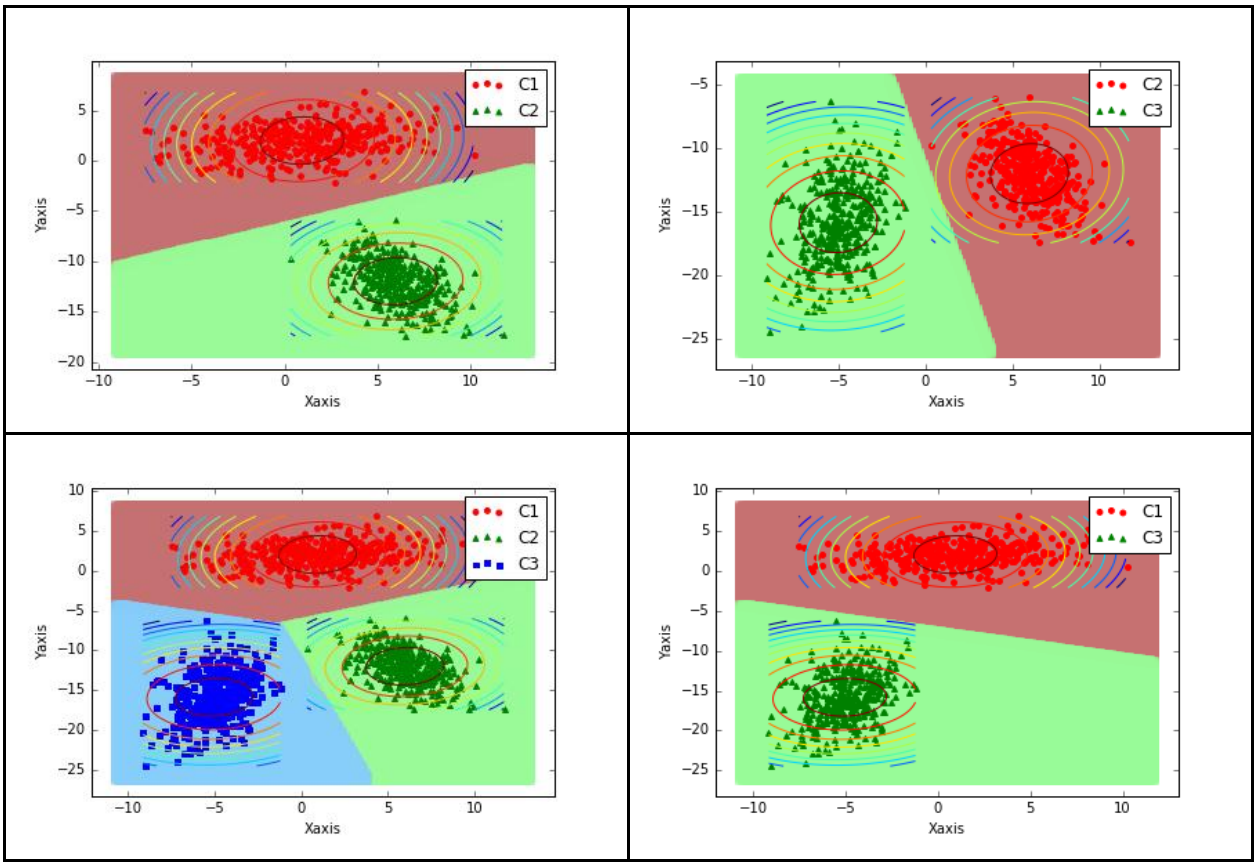


Figure 4: Type 1, Case2: Region plot and contour plots considering Case 2 covariance matrices for Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 1.2.1 The Confusion Matrix for the linear separable, case 2

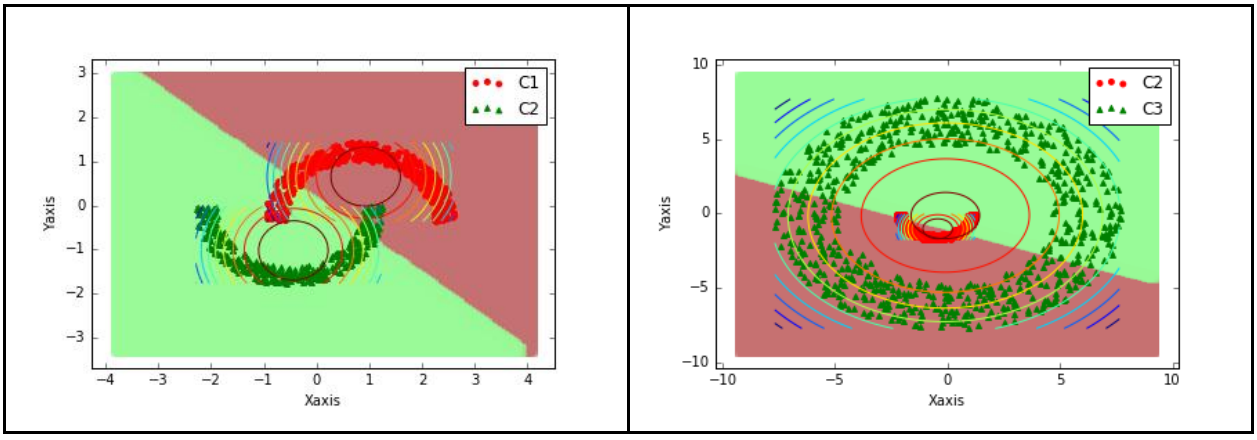
	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*125*	*0*	*0*
	Class 2	*0*	*125*	*0*
	Class 3	*2*	*0*	*123*

Table 1.2.2 The Performance Matrix for the linear separable, case 2

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*98.43*	*100.0*	*99.21*
Class 2	*100.0*	*100.0*	*100.0*
Class 3	*100.0*	*98.4*	*99.1*
Mean Value	*99.48*	*99.47*	*99.47*

Classification Accuracy: 99.46

Non Linear Data:



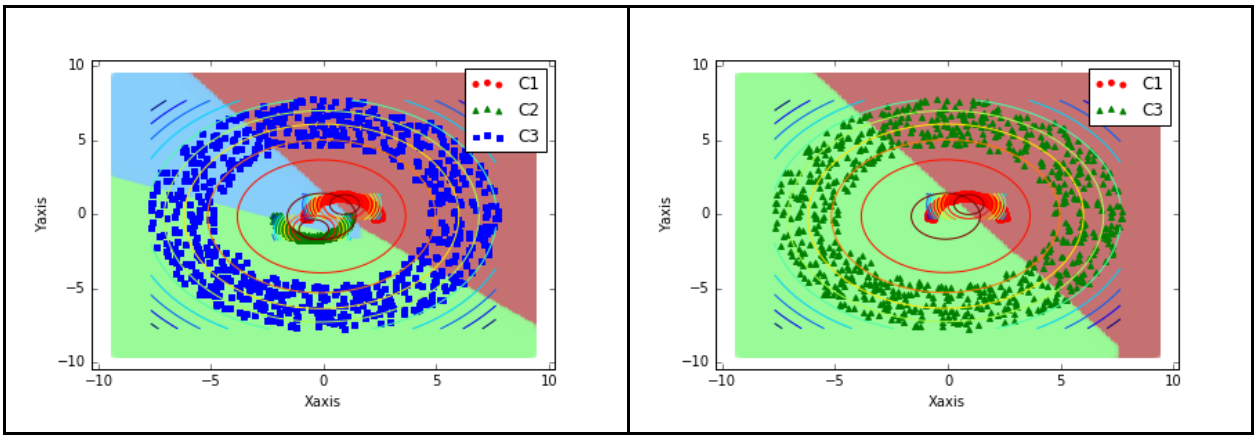


Figure 5: Type 2, Case2: Region plot and contour plots considering Case 2 covariance matrices for Non-Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

In Fig 5, pic 1, we can see that decision boundary is cutting the line joining the two means in the min point because the P (C1) == P (C2).

Table 2.2.1 The Confusion Matrix for the nonlinear separable, case 2

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*95*	*0*	*30*
	Class 2	*16*	*100*	*9*
	Class 3	*125*	*98*	*27*

Table 2.2.2 The Performance Matrix for the nonlinear separable, case 2

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*40.25*	*76.0*	*52.63*
Class 2	*50.51*	*80.0*	*61.92*
Class 3	*40.91*	*10.8*	*17.09*
Mean Value	*43.89*	*55.6*	*43.88*

Classification Accuracy (%): 44.4

**The reason for the low accuracy is very evident. Classifier is trying to discriminate nonlinear data points using a line. So, more than 50% of the test data points are misclassified. Classifier assumes that the covariance matrices for all classes are same, while in reality they are not.

Real-world Data

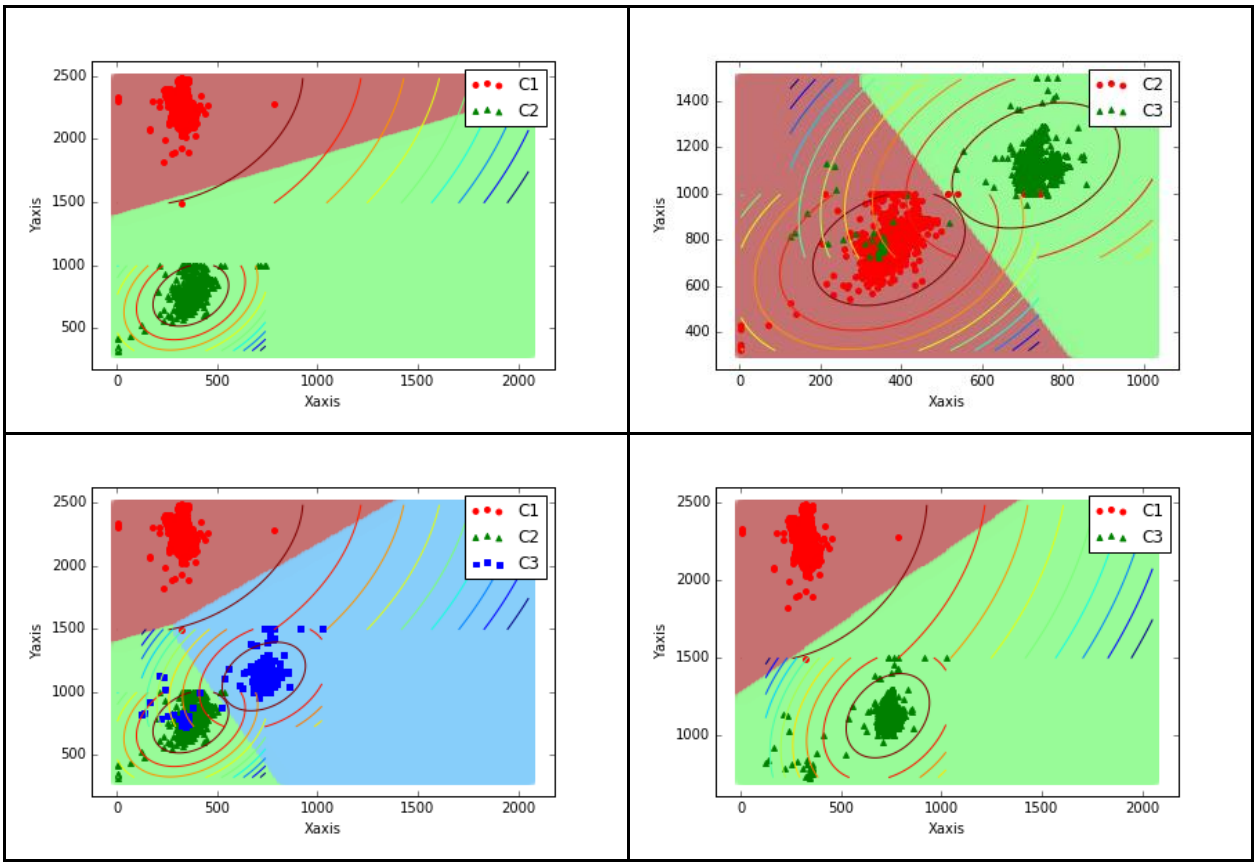


Figure 5: Type 3, Case2: Region plot and contour plots considering Case 2 covariance matrices for Real-World Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 3.2.1 The Confusion Matrix for the real world, case 2

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*579*	*3*	*14*
	Class 2	*0*	*613*	*9*
	Class 3	*1*	*28*	*512*

Table 3.2.2 The Performance Matrix for the real world, case 2

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*99.83*	*97.15*	*98.47*
Class 2	*95.19*	*98.55*	*96.84*
Class 3	*95.7*	*94.64*	*95.17*
Mean Value	*96.9*	*96.78*	*96.83*

Classification Accuracy (%): 96.87

Case 3: Classifier for which covariance matrix is diagonal and is different for each class.

In this case, covariance matrices are different but are forced to be diagonal. This can be achieved by simply converting all the non-diagonal elements to 0. Also, the classifier assumes that the feature vectors are independent of each other.

Case 3 is the sub class of Class 4.

We can also notice that the contour plots are ellipsoidal in nature unlike in case 1 and case 2.

Linear Data

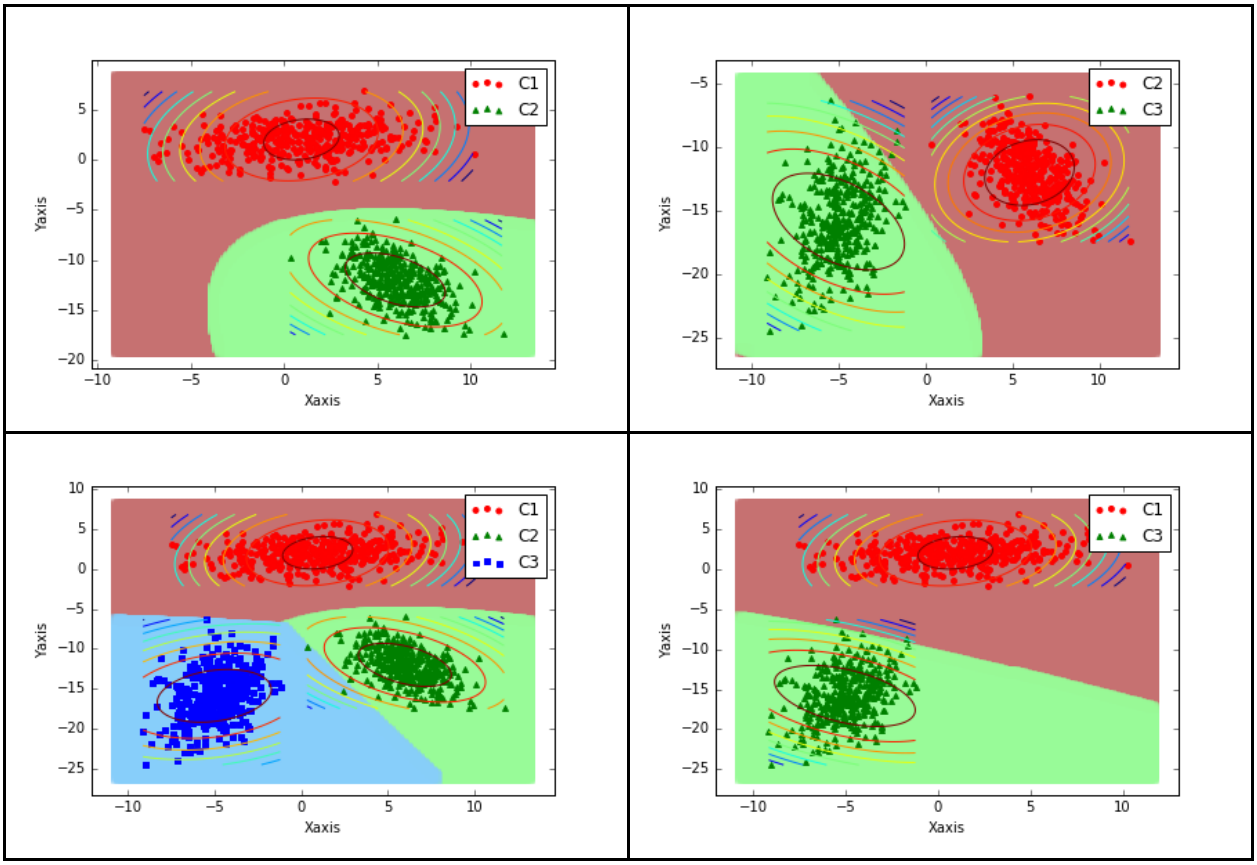


Figure 7: Type 1, Case3: Region plot and contour plots considering Case 3 covariance matrices for Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 1.3.1 The Confusion Matrix for the linearly separable, case 3

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*125*	*0*	*0*
	Class 2	*0*	*125*	*0*
	Class 3	*0*	*0*	*125*

Table 1.3.2 The Performance Matrix for the linear separable, case 3

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*100.0*	*100.0*	*100.0*
Class 2	*100.0*	*100.0*	*100.0*
Class 3	*100.0*	*100.0*	*100.0*
Mean Value	*100.0*	*100.0*	*100.0*

Classification Accuracy: 100.0

Ideally, line is enough to discriminate between the classes with linear data points. So, quadratic decision boundary will do a better job any way, which is evident from the performance matrix.

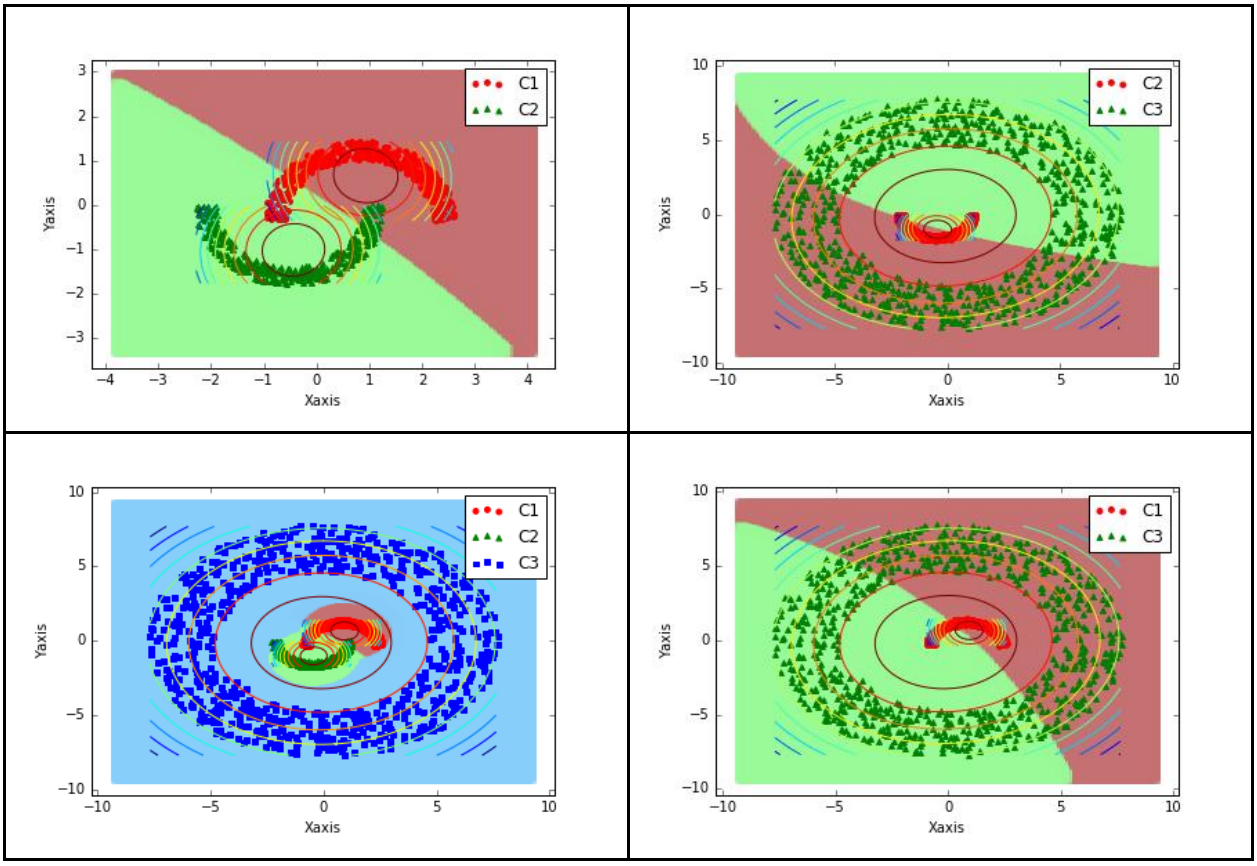


Figure 8: Type 2, Case3: Region plot and contour plots considering Case 3 covariance matrices for Non-Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

From fig 8, many more points are classified appropriately compared to the linear decision boundary. This is because the decision boundary employed here is quadratic in nature.
We can clearly this in Fig 8, pic 3.

Table 2.3.1 The Confusion Matrix for the nonlinear separable, case 3

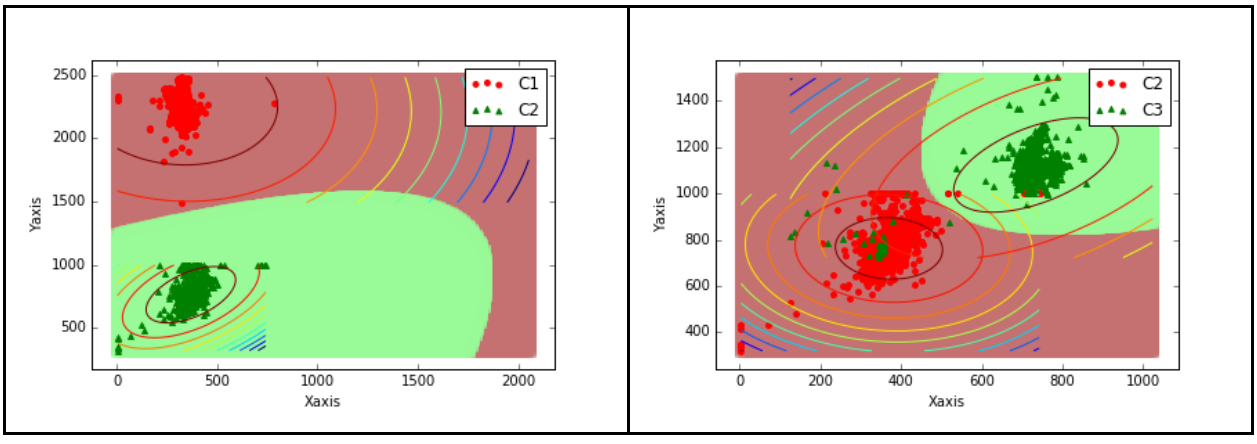
	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*121*	*4*	*0*
	Class 2	*6*	*119*	*0*
	Class 3	*0*	*0*	*125*

Table 2.3.2 The Performance Matrix for the nonlinear separable, case 3

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*95.28*	*96.8*	*96.03*
Class 2	*96.75*	*95.2*	*95.97*
Class 3	*100.00*	*100.00*	*100.00*
Mean Value	*97.34*	*97.33*	*97.33*

Classification Accuracy (%): 98.0

Here, we can see great improvement in the classification using the quadratic decision boundary than linear decision boundary in particular with non-linear data.



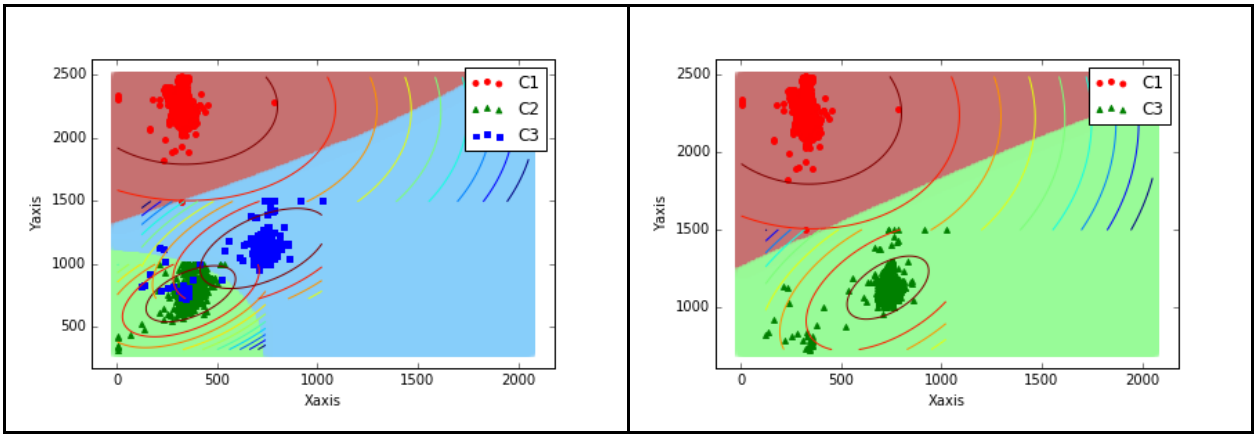


Figure 9: Type 3, Case3: Region plot and contour plots considering Case 3 covariance matrices for Real World Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 3.3.1 The Confusion Matrix for the real world, case 3

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*583*	*3*	*10*
	Class 2	*0*	*613*	*9*
	Class 3	*1*	*27*	*513*

Table 3.3.2 The Performance Matrix for the real world, case 3

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*99.83*	*97.82*	*98.81*
Class 2	*95.33*	*98.55*	*96.92*
Class 3	*96.43*	*94.82*	*95.62*
Mean Value	*97.2*	*97.07*	*97.12*

Classification Accuracy (%): 97.15

Case 4: Classifier for which each class has full individual covariance matrices.

Classifier considers covariance matrices of all classes as is. No constraints imposed.

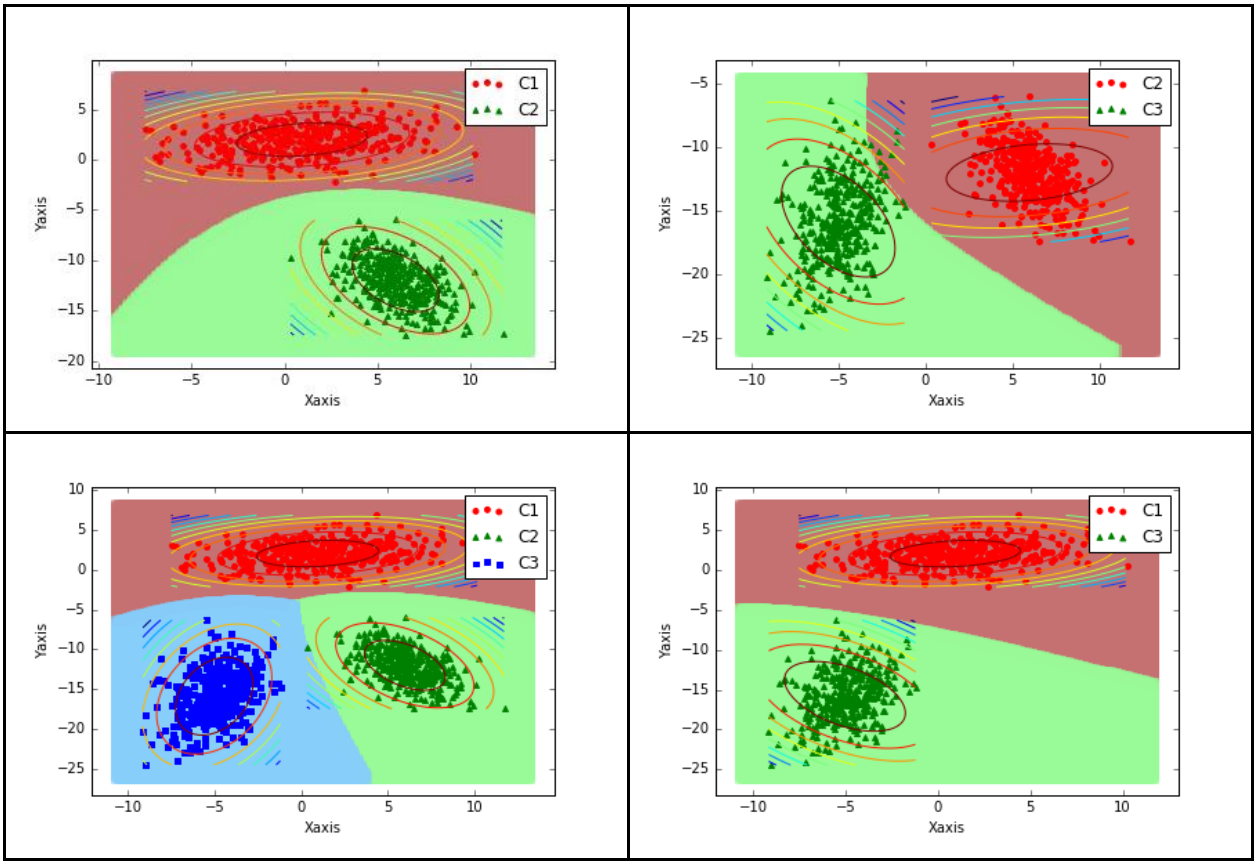


Figure 10: Type 1, Case4: Region plot and contour plots considering Case 4 covariance matrices for Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 1.4.1 The Confusion Matrix for the linear separable, case 4

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*125*	*0*	*0*
	Class 2	*0*	*125*	*0*
	Class 3	*0*	*0*	*125*

Table 1.4.2 The Performance Matrix for the linear separable, case 4

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*100.0*	*100.0*	*100.0*
Class 2	*100.0*	*100.0*	*100.0*
Class 3	*100.0*	*100.0*	*100.0*
Mean Value	*100.0*	*100.0*	*100.0*

Classification Accuracy (%): 100

Resultant $g_i(x)$is a quadratic equation which does a better job on the linear data points than a linear decision boundary.

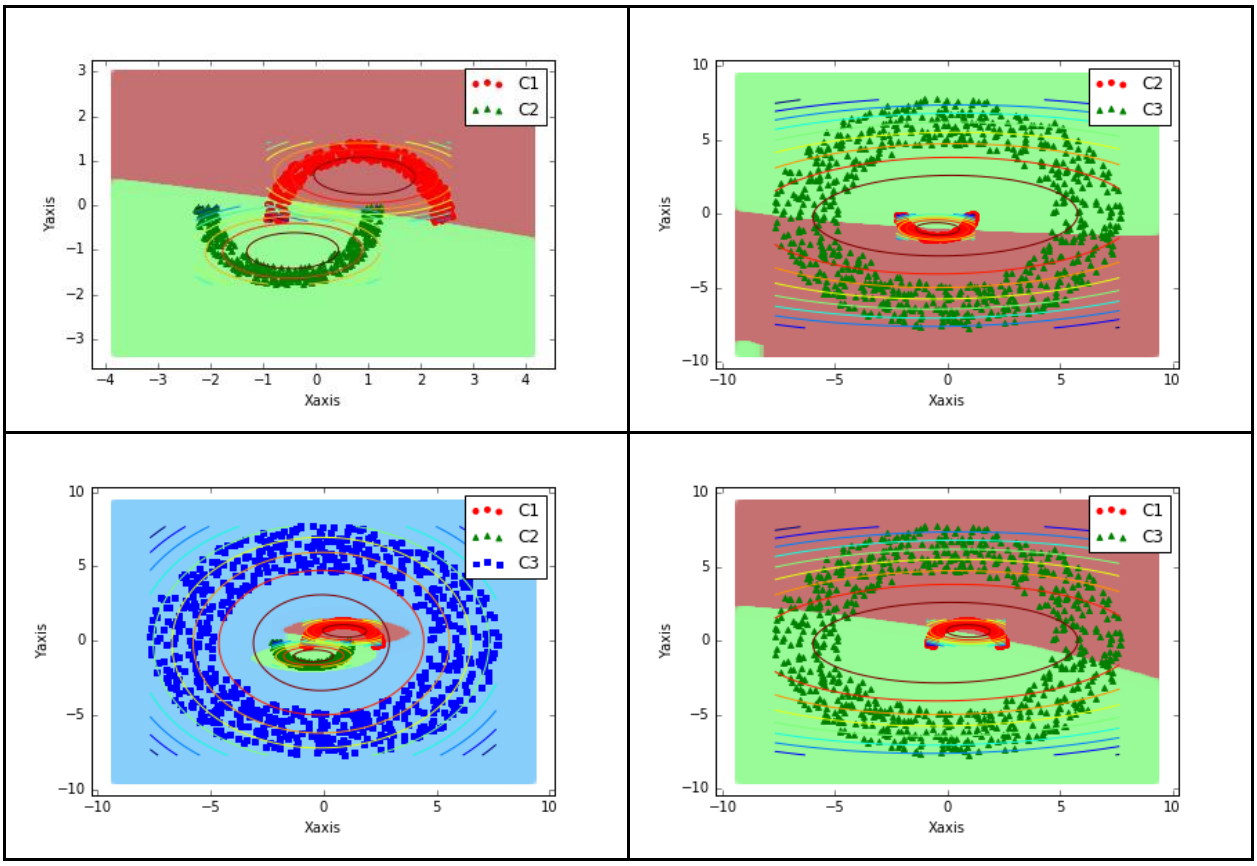


Figure 11: Type 2, Case4: Region plot and contour plots considering Case 4 covariance matrices for Non-Linearly Separable Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 2.4.1 The Confusion Matrix for the nonlinear separable, case 4

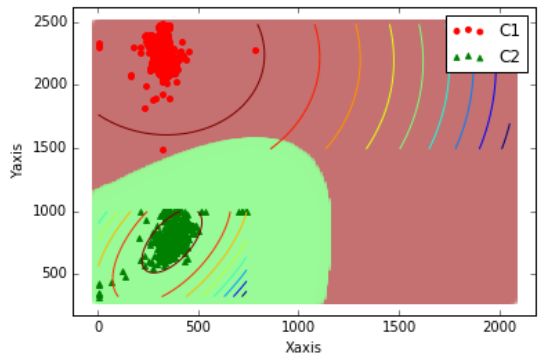
	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*120*	*5*	*0*
	Class 2	*4*	*121*	*0*
	Class 3	*0*	*0*	*250*

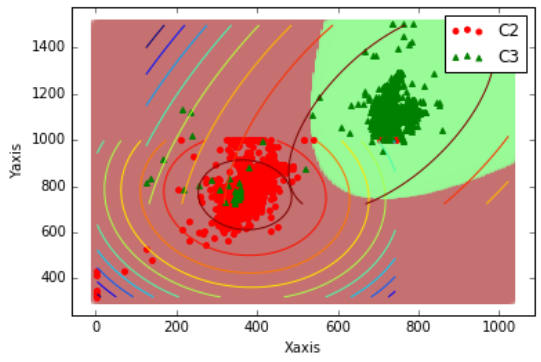
Table 2.4.2 The Performance Matrix for the nonlinear separable, case 4

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*96.77,*	*96.0*	*96.39*
Class 2	*96.03*	*96.8*	*96.41*
Class 3	*100.0*	*100.0*	*100.0*
Mean Value	*97.6*	*97.6*	*97.6*

Classification Accuracy (%): 98.2

Case 4 has slightly better performance than Case 3 on non-linear data.





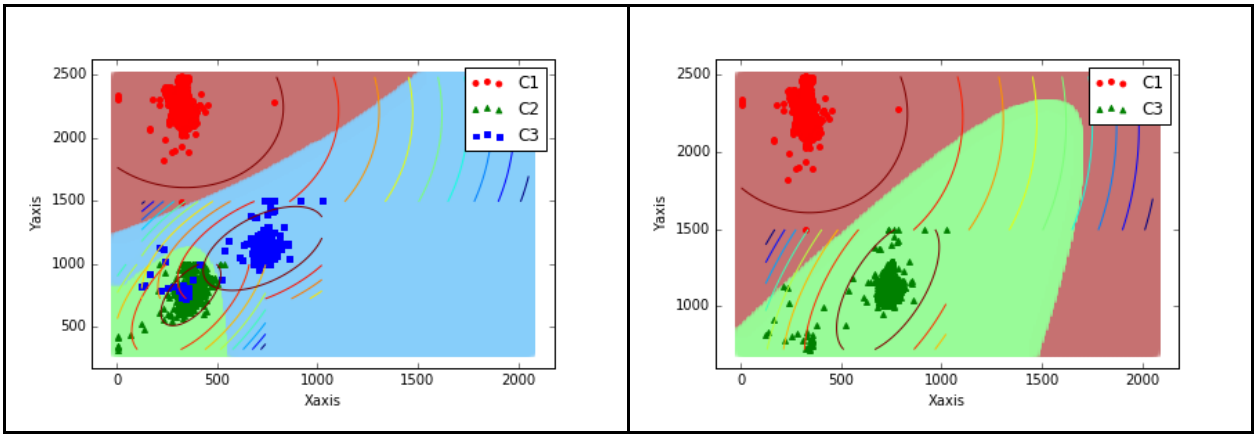


Figure 12: Type 3, Case4: Region plot and contour plots considering Case 4 covariance matrices for Real World Data. From clockwise, the plots for Class 1 and Class 2, Class 2 and Class 3, Class 1 and Class 3, Class1 and Class2 and Class3.

Table 3.4.1 The Confusion Matrix for the real world, case 4

	Class assigned by the Classifier			
Actual Values		*Class 1*	*Class 2*	*Class 3*
	Class 1	*581*	*3*	*12*
	Class 2	*0*	*611*	*11*
	Class 3	*1*	*128*	*512*

Table 3.4.2 The Performance Matrix for the real world, case 4

	Precision (%)	*Recall Rate (%)*	*F Score (%)*
Class 1	*99.83*	*97.48*	*98.64*
Class 2	*95.17*	*98.23*	*96.68*
Class 3	*95.7*	*94.64*	*95.17*
Mean Value	*96.9*	*96.78*	*96.83*

Classification Accuracy (%): 96.87

Conclusion:

The take away from using Bayesian classifier on the different types of data points is

Linear data which can be discriminated with a line, can do almost as good with quadratic decision boundaries also.

Real world data that we are dealing with, are spaced such that, linear boundary or quadratic boundary is not making much of a difference, as we can see from the performance matrix.

The key observation is with nonlinear data. Quadratic decision boundaries are performing more than 2 times better than linear decision boundary on nonlinear data.