Bayesian Classification Model

Shreyas Bapat, Bhavya Bhatt, GaganDeep Tomar

September 16, 2018

Abstract

The following discussion revolves around the modelling of data classifier in accordance with Bayesian Decision Theory. We then apply the model on three diffrent kinds of dataset which are precisly - linearly separable, non-linearly separable and actual data which can be of random nature. The goal is to compare the performance of the model with the above datasets under different conditions on covariance matrix.

1 Introduction

The Bayesian Decision Theory is a probablistic theory for classifying the data points on the basics of pre-known prior and class conditional probabilities(which in real scenario is not known in any closed form expression). We give below the basics of Bayes rule in the context of pattern recognition.

1.1 Bayes Rule

The bayes rule states that if we have prior probabilities of a class and we know class conditional probability for each class then given a sample data point we can find the probability that it belongs to some particular class i as follows

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{\sum_{j=1}^{m} P(x|C_j)P(C_j)} \tag{1}$$

Where $P(C_i)$ is prior probablity of class i and $P(x|C_i)$ is class conditional probablity. Then we can find to which class the data point belong to as follows

$$C = \{C_i, max\{P(C_i|x), i = 1 \dots m\}\} \tag{2}$$

where m is the number of classes.

1.2 Assumptions

In the above equation 1 we have to have $P(C_i)$ and $P(x|C_i)$ for the classification. We can obtain prior probablity by observing the training dataset. The ad-hoc assumption in our model is that class conditional probablity is considered to be normal distributed with parameters μ_i and σ_i^2 which are mean and variance of the dataset belonging to that particular class i. Note that the above parameters are scalar in univariate case(dimension of data point is one) but they would be a vector (mean vector)and a matrix(covariance matrix) respectively in bivariate and multivariate case. The expression for multidimensional gaussian distribution is

$$P(x|C_i) = \frac{1}{\sqrt{det(2\pi\mathbf{\Sigma_i})}} exp\{\frac{-1}{2}(x-\mu_i)^\intercal \mathbf{\Sigma_i}^{-1}(x-\mu_i)\} \tag{3}$$

Where μ_i is mean vector or class i and $\mathbf{\Sigma_i}$ is covariance matrix of class i.

2 Coutour Curves and Covariance Matrix

In this section[1] we discuss the relation between the shape of the cross section produced by slicing the bivariate gaussian distribution with a hyperplane parallel to the 2D-feature plane and covariance matrix. The bivariate gaussian distribution is a follows

$$P(x|C_i) = \frac{1}{\sqrt{det(2\pi\mathbf{\Sigma_i})}} exp\{\frac{-1}{2}(x-\mu_i)^\intercal \mathbf{\Sigma_i}^{-1}(x-\mu_i)\} \tag{4}$$

with mu_i be a 2×1 mean column vector and Σ_i be 2×2 covariance matrix. So we assume the covariance matrix in its expanded form as

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$$

where diagonal terms are variance of the features and off-diagonal terms are covariance bewteen feature-1 and feature-2. The above matrix is symmetric precisly due to the fact that $cov(x_i, x_j) = cov(x_j, x_i)$. Now we set the above distribution function to some constant k and find the resultant curve projected on the feature space which is called constant contour curve.

$$\frac{1}{\sqrt{det(2\pi\mathbf{\Sigma_i})}} exp\{\frac{-1}{2}(x-\mu_i)^\intercal \mathbf{\Sigma_i}^{-1}(x-\mu_i)\} = k$$

after some manupilation and taking log both sides we get

$$(x-\mu)^\intercal \mathbf{\Sigma}^{-1}(x-\mu) = -2\ln(\sqrt{2\pi\,|\mathbf{\Sigma}|}k)$$

where we have dropped the index i for simplicity and the whole analysis can be done without the loss of generality. Now writing the matrix in full and evaluating the required operation on the column vector we get finally

$$\Sigma_{22}X_1^2 + \Sigma_{11}X_2^2 - 2\Sigma_{12}X_1X_2 + 2\ln(\sqrt{2\pi\,|\mathbf{\Sigma}|}k) = 0 \tag{5}$$

where $x = [x_1 x_2]^\intercal$, $\mu = [\mu_1 \mu_2]^\intercal$, $X_1 = x_1 - \mu_1$ and $X_2 = x_2 - \mu_2$. This equation is in the form of general equation for conic section

$$ax^2 + by^2 + cxy + d = 0 \tag{6}$$

Now in our case the coefficients a and b are Σ_{22} and Σ_{11} respectively. The above equation thus represents an ellipse in our case as variances are always positive values. Now from the elementary analysis of conics we know that the coefficient of xy represent the extend to which the ellipse is titled w.r.t to the axis. Also the coefficients of x^2 and y^2 represents the length of major and minor axis respectively. Now we consider $\Sigma_{12} = 0$ (covariance matrix is diagonal) then the we recover the familiar equation of ellipse with major and minor axis parallel to the x-y axis. The equation is

$$\frac{X_1^2}{\left(\frac{-2\ln(\sqrt{2\pi|\mathbf{\Sigma}|}k)}{\Sigma_{22}}\right)} + \frac{X_2^2}{\left(\frac{-2\ln(\sqrt{2\pi|\mathbf{\Sigma}|}k)}{\Sigma_{11}}\right)} = 1 \tag{7}$$

which is of the form

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} = 1$$

the above is the equation of countour curve projected in the feature space. Now we consider following three cases

$$\Sigma_{22} < \Sigma_{11} \rightarrow A > B \quad ellipse$$
$$\Sigma_{22} = \Sigma_{11} \rightarrow A = B \quad circle$$
$$\Sigma_{22} > \Sigma_{11} \rightarrow A < B \quad ellipse$$

[1] for a complete discussion refer to the appendix

3 Training and Testing

In the presented work we would be taking test size of 0.25. We would use training data points to train our classifier and testing data points to analyse the performance of the trained model(by performance we mean the ability of the model to correctly classify the unseen data points *i.e* training examples). Training the model has different meaning for different models, here training means calculating the covariance matrix using training dataset and obtaining the required decision boundary.

3.1 Discriminant function

The decision boundary is represented mathematically using discriminant function. For Bayes Classifier, the discriminant function is as follows

$$f(x) = g_i(x) - g_j(x) \tag{8}$$

where $g_i(x)$ is

$$g_i(x) = \ln(P(C_i|x)) \tag{9}$$

We observe that $f(x) = 0$ is the equation for decision boundary. And $f(x) > 0$ the data point belongs to class i and $f(x) < 0$ the data point belongs to class j. Now we know that $P(C_i|x) = P(x|C_i)P(C_i)$ and taking logarithm simplfies the expression for the discriminant function[2]. We have omitted the total probability term in the denominator because it is same for all the classes and does not contribute to the classifion process.

4 Performance of the model

In this section[3] we start our actual task of performance check on three different datasets. The analysis includes for each dataset we have considered four cases on covariance matrix which are as follows

- Covariance matrix for all the classes is the same and is $\sigma^2\mathbf{I}$.
- Full Covariance matrix for all the classes is the same and is $\mathbf{\Sigma}$.
- Covariance matric is diagonal and is different for each class
- Full Covariance matrix for each class is different

and for each such case we calculate

- Classification accuracy, precision for every class, mean precision, recall for every class, mean recall, F-measure for every class and mean F-measure on test data
- Confusion matrix[4] based on the performance for test data
- Constant density contour plot for all the classes together with the training data superposed
- Decision region plot for every pair of classes together with the training data superposed
- Decision region plot for all the classes together with the training data superposed

[2]for more insight you can refer to the textbook, R.O.Duda, D.G.Stork Pattern Classification-wiley

[3]Note that in the further discussion class-1 red, class-2 green, class-3 blue

[4]Note that entry C_{ij} of confusion matrix represents number of points known to be in class i but classified as class j

4.1 Dataset-1

This dataset contains two types of two dimensional data points - linearly separable and non-linearly separable. Both have classification among three classes.

4.1.1 Linearly Separable

First we show the scatter plots of the data points in training dataset.

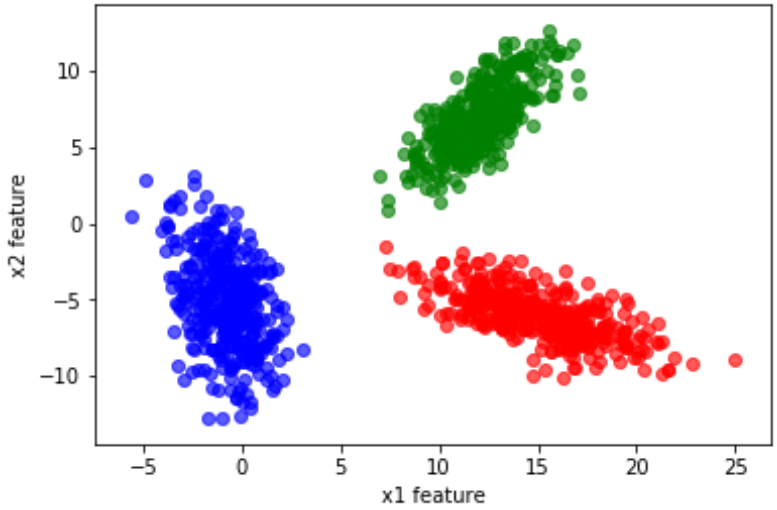


Figure 1: Dataset-1: Linearly Separable

As we can see from the figure 1 the data points are linearly separable. Now we consider the four stated cases on this dataset.

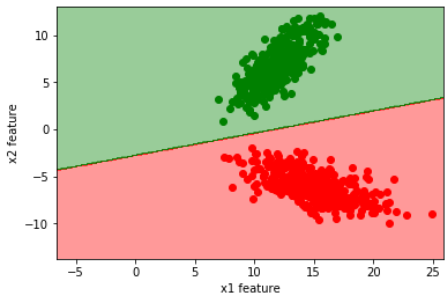
Case-1: $\Sigma_i = \sigma^2 \mathbf{1}$ According to the theory considered for this model we get a linear[5] discriminant function of the form

$$\mathbf{w}^\intercal (x - x_0) = 0 \tag{10}$$

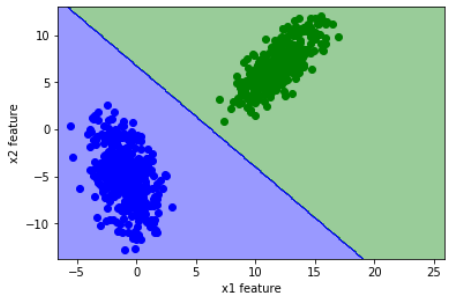
$$\mathbf{w} = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{|| \mu_i - \mu_j ||^2} \ln \frac{P(C_i)}{P(C_j)} (\mu_i - \mu_j)$$

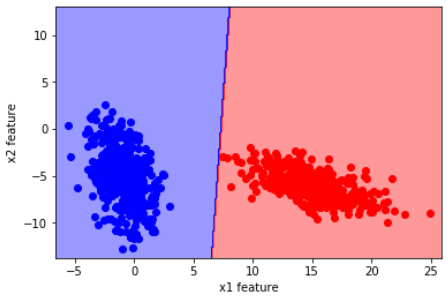
where $\mathbf{w}$ is weight matrix and x_0 through which the decision boundary(linear) passes. The decision surface between each pair of three classes are shown in figure 2. Now we state the performance parameters obtained by testing the classifier on the test dataset(0.25).



(a) case-1: class-1 class-2



(b) case-1: class-2 class-3



(c) case-1: class-3 class-1

Figure 2: Decision Boundary and Training Data points for Linearly Seperable Dataset-1, Case-1

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 84 & 0 & 41 \\ 21 & 90 & 14 \\ 107 & 97 & 46 \end{bmatrix}$$

We observe here is that even though the diagonal terms (correctly classified) are appreciable but the cross diagonal terms suggest that this case of covariance matrix cannot be a pratical approximation for the dataset.

Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.6639344262	0.3875598086	0.4894259819
class-2	0.72	0.4812834225	0.5769230769
class-3	0.184	0.5187165775	0.271642518

Table 1: Performance parameters for Linearly Seperable Dataset-1, Case -1

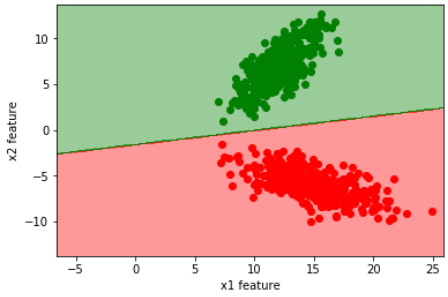
Accuracy 43.66%, **Mean – Precision** 0.4625, **Mean – Recall** 0.522.

[5]linear means that if feature is two dimensional then it is a line, if three dimensional it is a hyperplane and similarly for higher dimensions

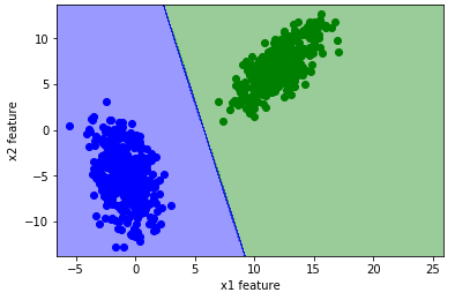
Case-2: $\Sigma_i = \Sigma$ According to the theory considered for this model we get a linear[6] discriminant function of the form

$$\mathbf{w}^\intercal(x - x_0) = 0 \tag{11}$$
$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_i - \mu_j)$$
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln(P(C_i)) - \ln(P(C_j))}{(\mu_i - \mu_j)^\intercal \mathbf{\Sigma}^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

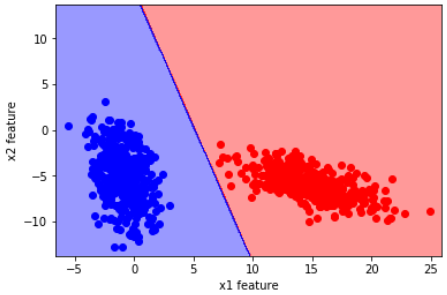
where $\mathbf{w}$ is weight matrix and x_0 through which the decision boundary(linear) passes. The decision surface between each pair of three classes are shown in figure 3. Now we state the performance



(a) case-2: class-1 class-2



(b) case-2: class-2 class-3



(c) case-2: class-3 class-1

Figure 3: Decision Boundary and Training Data points for Linearly Seperable Dataset-1, Case-2

parameters obtained by testing the classifier on the test dataset(0.25).

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 0 & 0 & 125 \\ 0 & 0 & 125 \\ 39 & 7 & 204 \end{bmatrix}$$

We can observe that this is performing even worst as it has two diagonal terms equal to zero which means it does not correctly classify any point in class 1 and class 2. So this also is not a realisitic approximation of the covariance matrix.

Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0	0	NaN
class-2	0	0	NaN
class-3	0.816	0.449339207	0.5795454545

Table 2: Performance parameters for Linearly Seperable Dataset-1, Case -2

Accuracy 40.8%, **Mean – Precision** 0.149, **Mean – Recall** 0.272.

[6]linear means that if feature is two dimensional then it is a line, if three dimensional it is a hyperplane and similarly for higher dimensions

Case-3: Σ_i is diagonal and different According to the theory considered for this model we get a non-linear discriminant function of the form

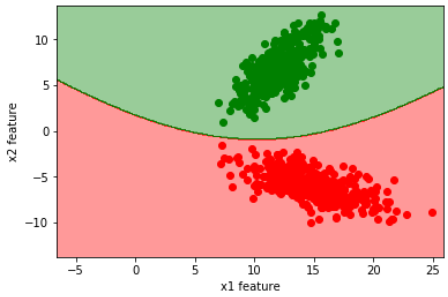
$$g_i(x) = x^\intercal \mathbf{W_i} x + w_i^\intercal x + \omega_{i0} \tag{12}$$

$$\mathbf{W_i} = \frac{-1}{2}\mathbf{\Sigma_i}^\intercal$$

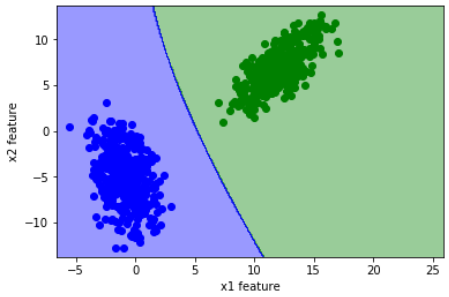
$$w_i = \mathbf{\Sigma_i}^{-1}\mu_i$$

$$\omega_{i0} = \frac{-1}{2}\mu_i^\intercal \mathbf{\Sigma_i}^{-1}\mu_i - \frac{1}{2}\ln(|\mathbf{\Sigma_i}|) + \ln(P(C_i))$$

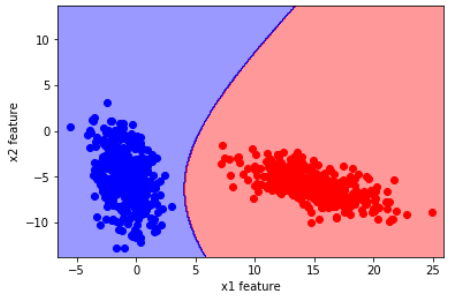
. From equation the decision surface between each pair of three classes would be non-linear as shown in figure 4



(a) case-3: class-1 class-2



(b) case-3: class-2 class-3



(c) case-3: class-3 class-1

Figure 4: Decision Boundary and Training Data points for Linearly Seperable Dataset-1, Case-3

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 120 & 5 & 0 \\ 7 & 117 & 1 \\ 3 & 0 & 247 \end{bmatrix}$$

This shows that off diagonal terms are appreciably decreased which makes this approximation to be reasonable one for this dataset.

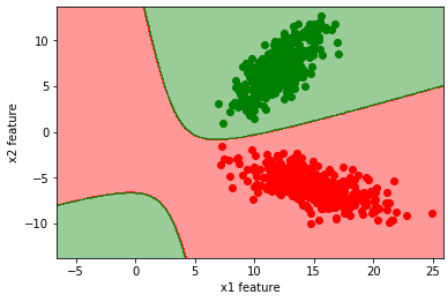
Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.96	0.9230769231	0.9411764706
class-2	0.936	0.9590163934	0.9473684211
class-3	0.9959677419	0.9959677419	0.9959677419

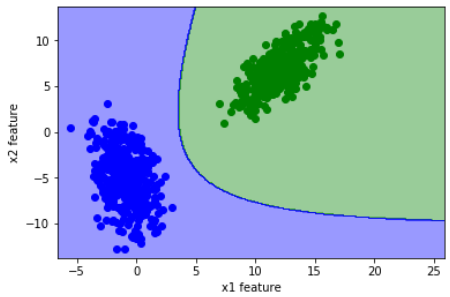
Table 3: Performance parameters for Linearly Seperable Dataset-1, Case -3

Accuracy 96.8%, **Mean – Precision** 0.959, **Mean – Recall** 0.963.

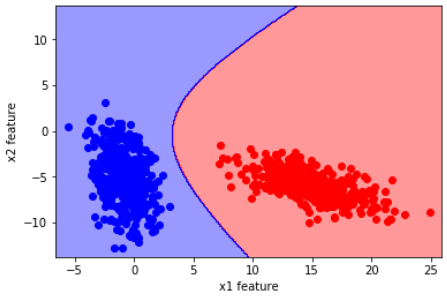
Case-4: Σ_i is arbitary In this case also we can use the equation 12. The decision surface between each pair of three classes would again be non-linear as shown in figure 5



(a) case-4: class-1 class-2



(b) case-4: class-2 class-3



(c) case-4: class-3 class-1

Figure 5: Decision Boundary and Training Data points for Linearly Seperable Dataset-1, Case-4

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 119 & 6 & 0 \\ 7 & 117 & 1 \\ 3 & 0 & 247 \end{bmatrix}$$

This also performs similar to that of the previous case.

Precision, Recall and F-measure [7]

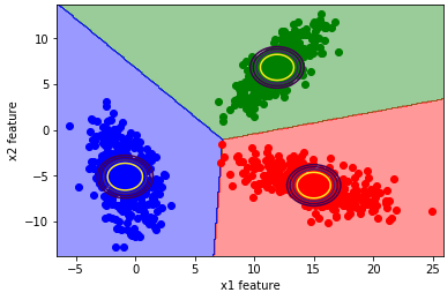
class	Recall	Precision	F-Measure
class-1	0.952	0.9224806202	0.937007874
class-2	0.936	0.9512195122	0.9435483871
class-3	0.9959677419	0.9959677419	0.9959677419

Table 4: Performance parameters for Linearly Seperable Dataset-1, Case -4

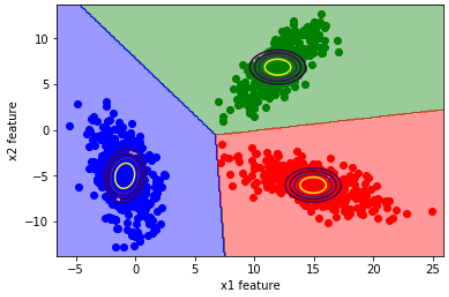
Accuracy 96.6%, **Mean – Precision** 0.956, **Mean – Recall** 0.961.

[7] The discriminant functions discussed above for all the different cases would be same for all the subsequent datasets

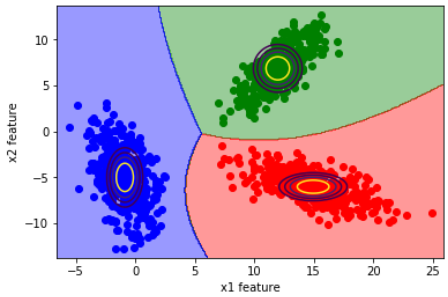
Conclusion The best performance is given by case-3 and worst performance is given by case-2. Now we shown the final decision boundary simultaneously for all the three classes and for all the four cases along with their contour plots superimposed on them in figure 6



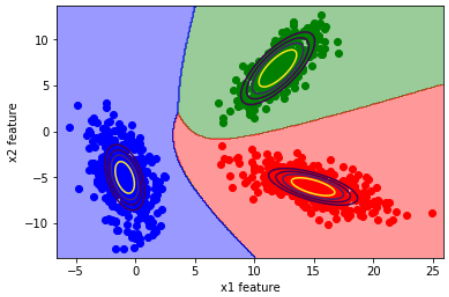
(a) case-1: interclass decision surface and contour plot



(b) case-2: interclass decision surface and contour plot



(c) case-3: interclass decision surface and contour plot



(d) case-4: interclass decision surface and contour plot

Figure 6: decision boundary for all three classes and contour plots

4.1.2 Non-Linearly separable

First we show the scatter plots of the data points in training dataset. As we can see from the

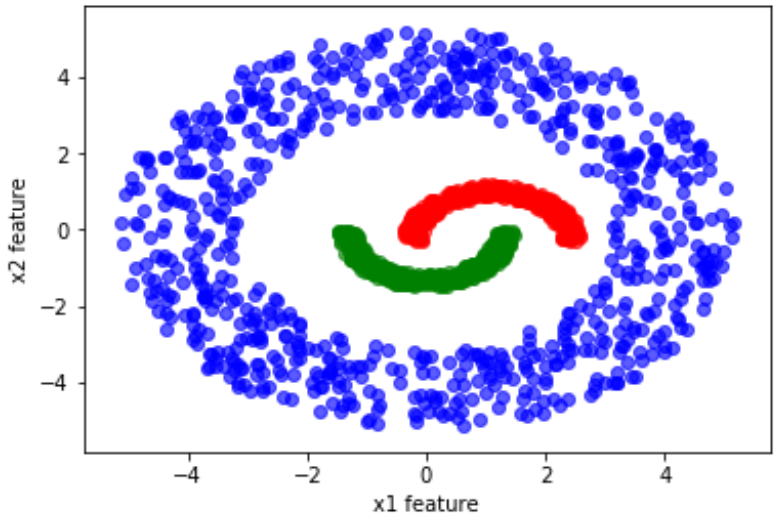
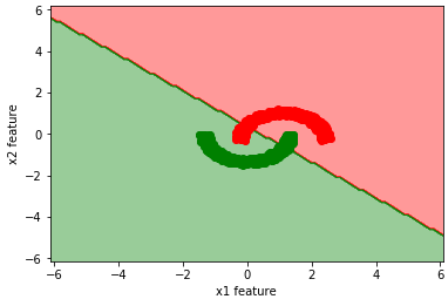


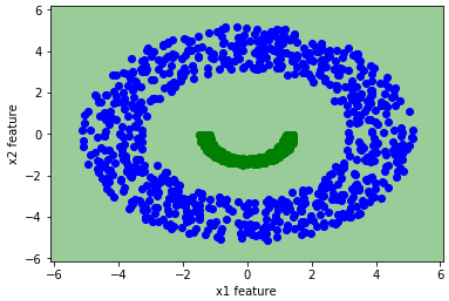
Figure 7: Dataset-1: Non-Linearly Separable

figure 7 the data points are not linearly separable. Now we consider the four stated cases on this dataset.

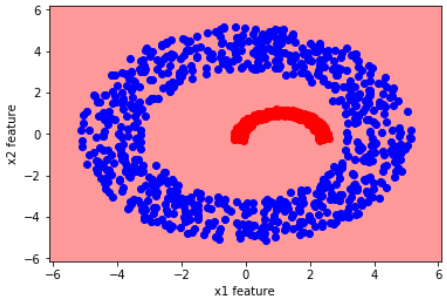
Case-1: $\mathbf{\Sigma_i = \sigma^2 1}$ The decision boudary between each pair of classes is shown in figure 8



(a) case-1: class-1 class-2



(b) case-1: class-2 class-3



(c) case-1: class-3 class-1

Figure 8: Decision Boundary and Training Data points for Non-Linearly Seperable Dataset-1, Case-1

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 85 & 0 & 40 \\ 22 & 96 & 7 \\ 88 & 114 & 48 \end{bmatrix}$$

We can observe here that cross-diagonal terms are exceptionally greater than diagonal terms so this is not a good approximation for the dataset.

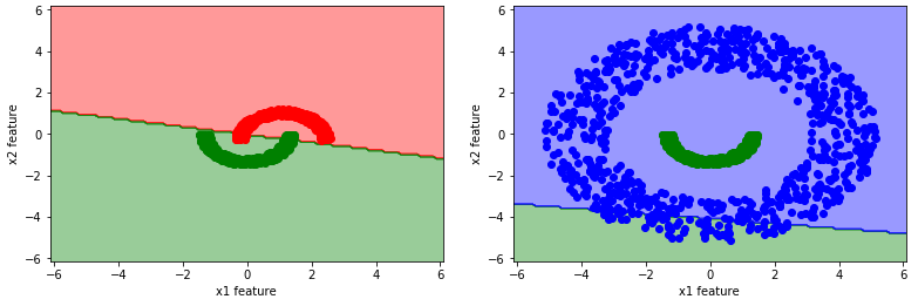
Precision, Recall and F-measure :

class	Recall	Precision	F-Measure
class-1	0.6639344262	0.4358974359	0.526275559
class-2	0.768	0.4571428571	0.5731343284
class-3	0.192	0.5052631579	0.2782608696

Table 5: Performance parameters for Non-Linearly Seperable Dataset-1, Case -1

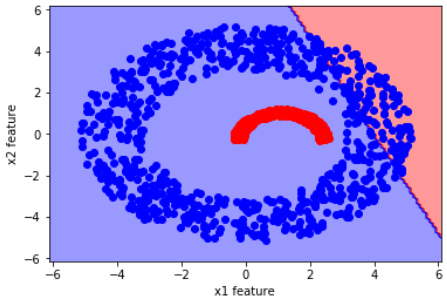
Accuracy 45.8%, **Mean – Precision** 0.466, **Mean – Recall** 0.541.

Case-2: $\mathbf{\Sigma_i = \Sigma}$ The decision boundary between each pair of classes is shown in figure 9



(a) case-2: class-1 class-2

(b) case-2: class-2 class-3



(c) case-2: class-3 class-1

Figure 9: Decision Boundary and Training Data points for Non-Linearly Seperable Dataset-1, Case-2

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 0 & 0 & 125 \\ 0 & 0 & 125 \\ 25 & 26 & 199 \end{bmatrix}$$

This is performing worst[8].

Precision, Recall and F-measure

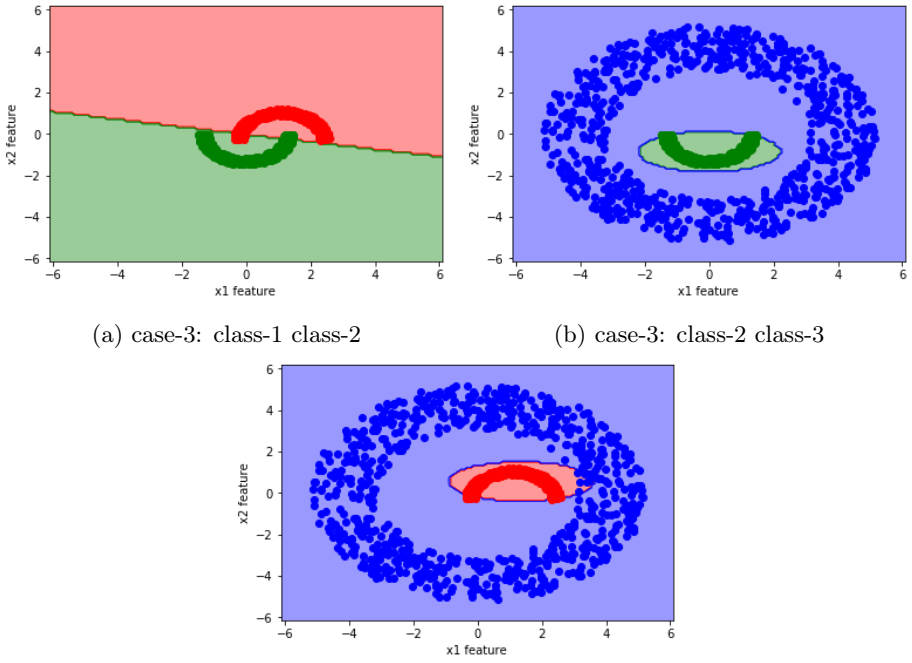
class	Recall	Precision	F-Measure
class-1	0	0	NaN
class-2	0	0	NaN
class-3	0.796	0.4432071269	0.5693848355

Table 6: Performance parameters for Non-Linearly Seperable Dataset-1, Case -2

Accuracy 39.8%, **Mean – Precision** 0.147, **Mean – Recall** 0.265.

[8]same reason as given in section 4.1.1

Case-3: Σ_i is diagonal and different The decision boundary is shown in figure 10



(a) case-3: class-1 class-2

(b) case-3: class-2 class-3

(c) case-3: class-3 class-1

Figure 10: Decision Boundary and Training Data points for Non-Linearly Seperable Dataset-1, Case-3

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 109 & 9 & 7 \\ 5 & 117 & 3 \\ 2 & 0 & 248 \end{bmatrix}$$

This approximation on covariance matrix is performing fairly good.

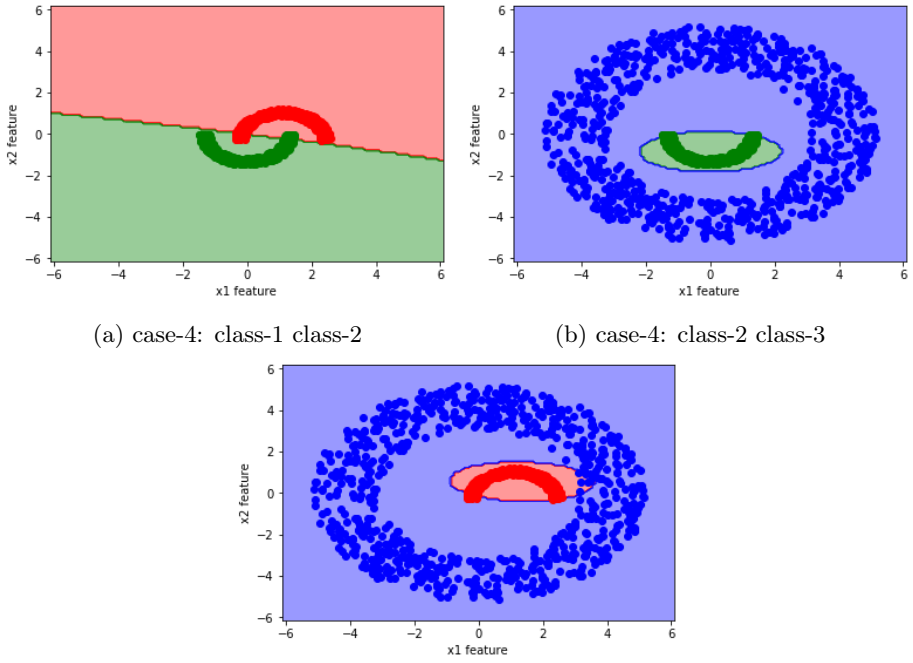
Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.872	0.9396551724	0.9045643154
class-2	0.936	0.9285714286	0.9322709163
class-3	0.992	0.9612403101	0.9763779528

Table 7: Performance parameters for Non-Linearly Seperable Dataset-1, Case -3

Accuracy 94.8%, **Mean – Precision** 0.943, **Mean – Recall** 0.933.

Case-4: Σ_i is arbitary The decision boundary is shown in figure 11



(a) case-4: class-1 class-2

(b) case-4: class-2 class-3

(c) case-4: class-3 class-1

Figure 11: case-4 Decision Boundary and Training Data points for Non-Linearly Seperable Dataset-1, Case-4

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 113 & 9 & 3 \\ 5 & 117 & 3 \\ 2 & 0 & 248 \end{bmatrix}$$

This also peforms fairly good similar to the above case.

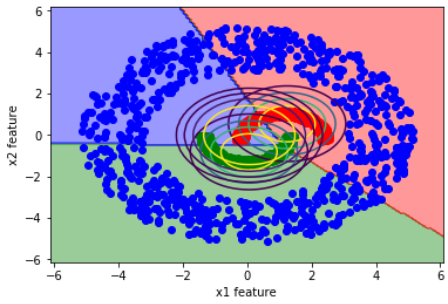
Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.904	0.9416666667	0.9224489796
class-2	0.936	0.9285714286	0.9322709163
class-3	0.992	0.9763779528	0.9841269841

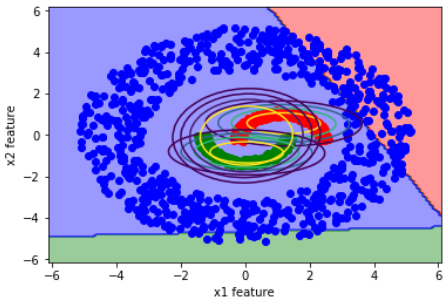
Table 8: Performance parameters for Non-Linearly Seperable Dataset-1, Case -4

Accuracy 95.6%, **Mean – Precision** 0.948, **Mean – Recall** 0.944.

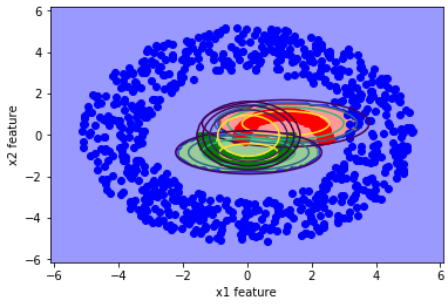
Conclusion The best performance is given by case-3 and worst performance is given by case-2. Now we shown the final decision boundary simultaneously for all the three classes and for all the four cases along with their contour plots superimposed on them in figure 12



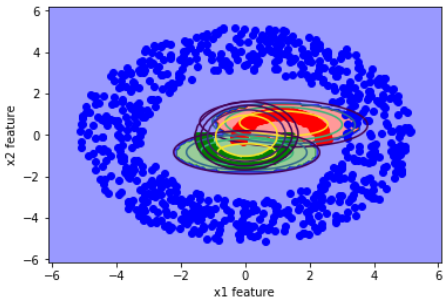
(a) case-1: interclass decision surface and contour plot



(b) case-2: interclass decision surface and contour plot



(c) case-3: interclass decision surface and contour plot



(d) case-4: interclass decision surface and contour plot

Figure 12: decision boundary for all three classes and contour plots

4.2 Dataset-2

This dataset is real dataset and is much more random[9] than the dataset-1 considered above. We carry out similar analysis for this dataset also. First we show the scatter plots of the data points in training dataset in figure 13

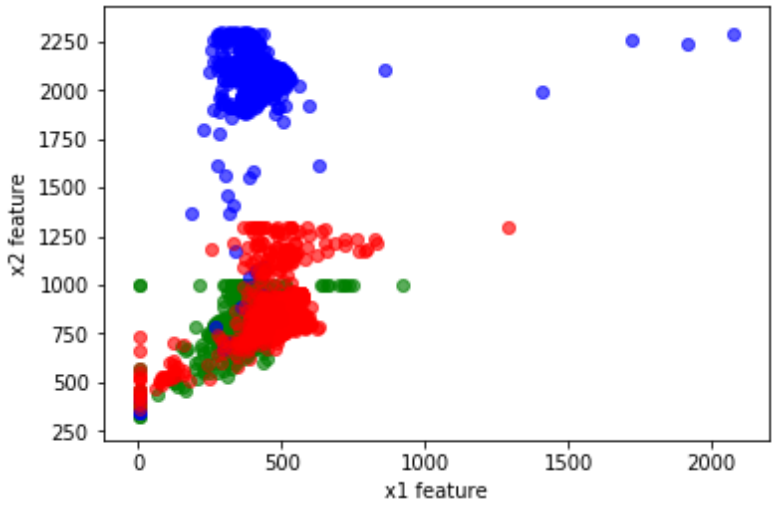
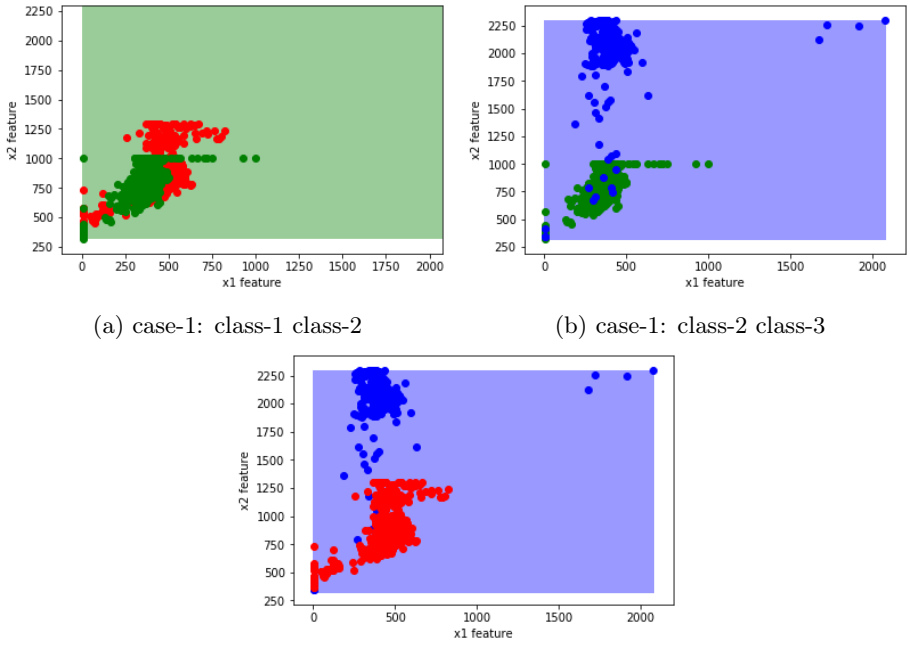


Figure 13: Dataset-2: real dataset

[9]randomness here means that classes can be overlapping also

Case-1: $\mathbf{\Sigma_i = \sigma^2 1}$ The decision boundary is shown in figure 14



(a) case-1: class-1 class-2

(b) case-1: class-2 class-3

(c) case-1: class-3 class-1

Figure 14: Decision Boundary and Training Data points for Dataset-2, Case-1

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 535 & 79 & 0 \\ 89 & 533 & 0 \\ 2 & 3 & 568 \end{bmatrix}$$

We observe here that it is performing good with the approximated covariance matrix.

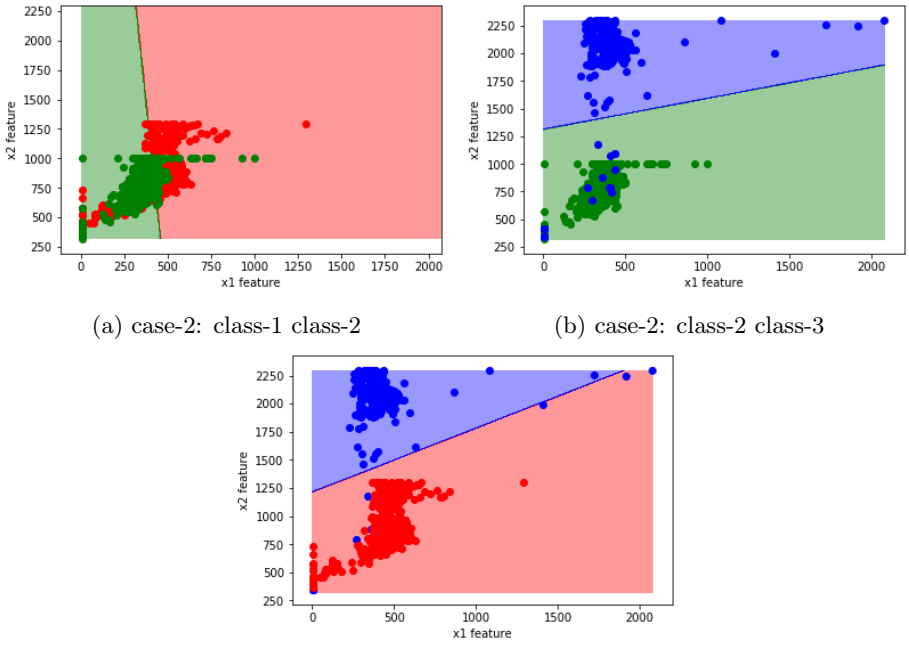
Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.8713355049	0.8546325879	0.8629032258
class-2	0.8569131833	0.8666666667	0.8617623282
class-3	0.9912739965	1	0.9956178791

Table 9: Performance parameters for Dataset-2, Case-1

Accuracy 90.4%, **Mean – Precision** 0.907, **Mean – Recall** 0.906.

Case-2: $\mathbf{\Sigma_i = \Sigma}$ The decision boundary is shown in figure 15



(a) case-2: class-1 class-2

(b) case-2: class-2 class-3

(c) case-2: class-3 class-1

Figure 15: Decision Boundary and Training Data points for Dataset-2, Case-2

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 533 & 81 & 0 \\ 48 & 574 & 0 \\ 3 & 3 & 567 \end{bmatrix}$$

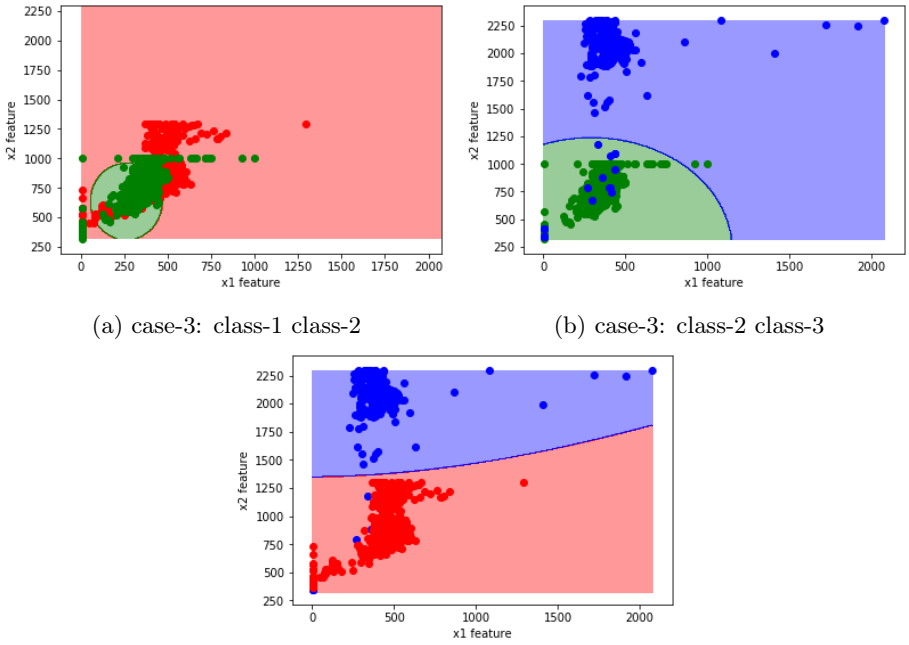
Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.8680781759	0.9126712329	0.8898163606
class-2	0.922829582	0.896875	0.9096671949
class-3	0.9895287958	1	0.9947368421

Table 10: Performance parameters for Dataset-2, Case-2

Accuracy 92.5%, **Mean – Precision** 0.936, **Mean – Recall** 0.926.

Case-3: Σ_i is diagonal and different The decision boundary is shown in figure 16



(a) case-3: class-1 class-2

(b) case-3: class-2 class-3

(c) case-3: class-3 class-1

Figure 16: Decision Boundary and Training Data points for Dataset-2, Case-3

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 530 & 84 & 0 \\ 68 & 554 & 0 \\ 2 & 2 & 569 \end{bmatrix}$$

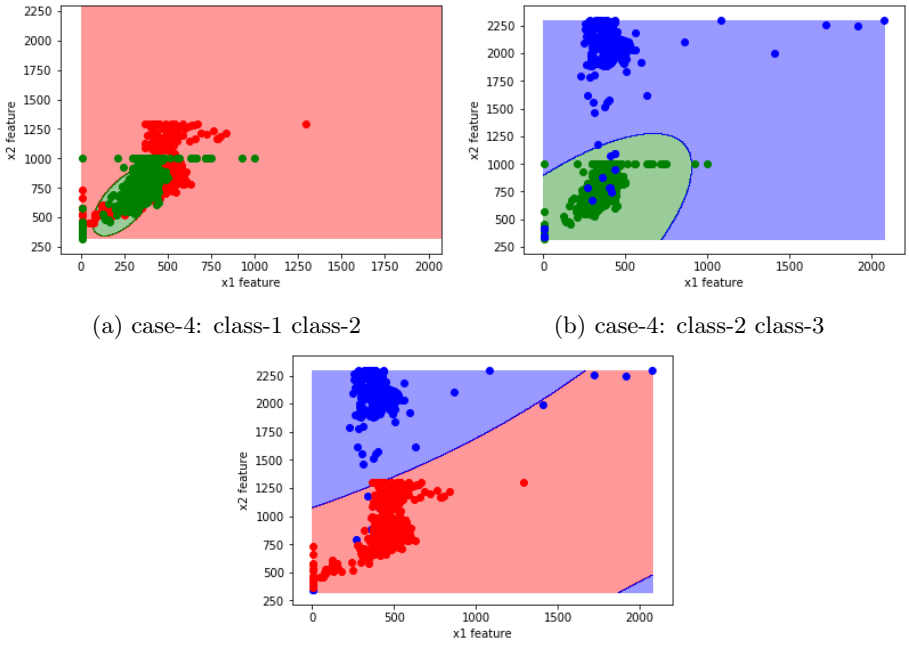
Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.8631921824	0.8833333333	0.8731466227
class-2	0.8906752412	0.865625	0.8779714739
class-3	0.9930191972	1	0.996497373

Table 11: Performance parameters for Dataset-2, Case-3

Accuracy 91.3%, **Mean – Precision** 0.916, **Mean – Recall** 0.915.

Case-4: Σ_i is arbitary The decision boundary is shown in figure 17



(a) case-4: class-1 class-2

(b) case-4: class-2 class-3

(c) case-4: class-3 class-1

Figure 17: Decision Boundary and Training Data points for Dataset-2, Case-4

Confusion Matrix The confusion matrix obtained is

$$\mathbf{\Sigma} = \begin{bmatrix} 518 & 94 & 2 \\ 61 & 561 & 0 \\ 4 & 2 & 567 \end{bmatrix}$$

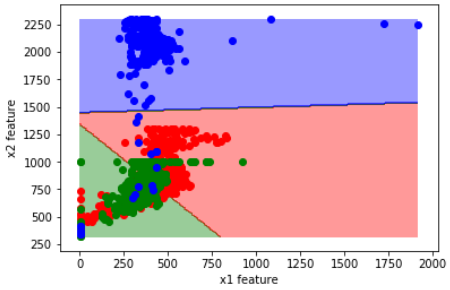
Precision, Recall and F-measure

class	Recall	Precision	F-Measure
class-1	0.8575667656	0.898911353	0.8777524677
class-2	0.9019292605	0.8538812785	0.8772478499
class-3	0.9895287958	0.9964850615	0.9929947461

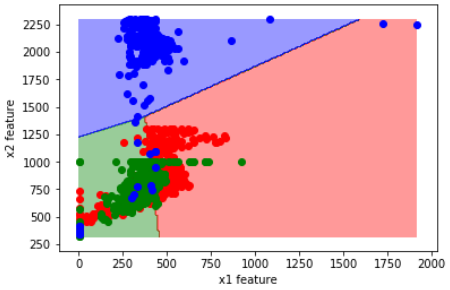
Table 12: Performance parameters for Dataset-2, Case-4

Accuracy 94.3%, **Mean – Precision** 0.916, **Mean – Recall** 0.916.

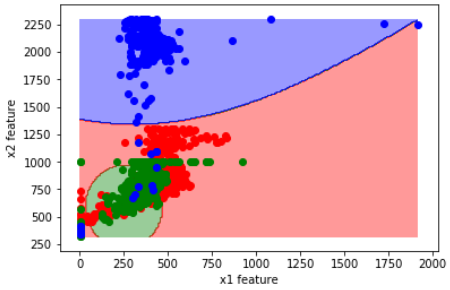
Conclusion We can see that all of the above cases on covariance matrix are good approximation for this dataset and any of the approximate form can be used for classification. Now we shown the final decision boundary simultaneously for all the three classes and for all the four cases along with their contour plots superimposed on them in figure 18



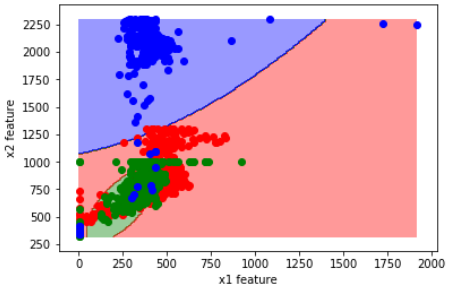
(a) case-1: interclass decision surface and contour plot



(b) case-2: interclass decision surface and contour plot



(c) case-3: interclass decision surface and contour plot



(d) case-4: interclass decision surface and contour plot

Figure 18: decision boundary for all three classes and contour plots

5 Appendix

This section includes some of the derivations relating the angle with which the ellipse(constant contour curves) is titled with the axis and the elements of the covariance matrix. This section can be skiped without any loss. The derivation includes all the points discussed in section 2. We first assume general form of elliptical curve in two dimensions as follows[10]

$$ax^2 + by^2 + cxy + d = 0 \tag{13}$$

Now we rotate the axis(active rotation) using rotation matrix for two dimensional space as follows

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Now we find x and y in terms of primed x' and y'. After simple algebraic calculations and manipulations we obtain following equation in x'-y' coordinate system

$$\begin{aligned}(a\cos^2\theta + b\sin^2\theta - c\cos\theta\sin\theta)x'^2 + (a\sin^2\theta + b\cos^2\theta + c\cos\theta\sin\theta)y'^2 \\ +(2a\cos\theta\sin\theta - 2b\sin\theta\cos\theta + c(\cos^2\theta - \sin^2\theta))x'y' + d = 0\end{aligned} \tag{14}$$

Here θ is the angle with which the coordinate axis are rotated(which we further state that is the angle with which the major and minor axis of the tilted ellipse is making with x-y axis). Now we impose the condition that in this coordinate system the above ellipse should be parallel to the axis and is of the standard form

$$\frac{x'^2}{A^2} + \frac{y'^2}{B^2} = 1 \tag{15}$$

for this we should have coefficient of $x'y'$ in the equation 14 to be zero which is equivalent as

$$\begin{aligned}2a\cos\theta\sin\theta - 2b\sin\theta\cos\theta + c(\cos^2\theta - \sin^2\theta) &= 0 \\ (a-b)\sin 2\theta + c\cos 2\theta &= 0 \\ \tan 2\theta = \frac{c}{b-a}\end{aligned} \tag{16}$$

This precisely gives the dependence of angle on the coefficients in equation 13. This can be easily proved that if $c = 0$ (angle would then be zero acc. to 16)then the ellipse given by equation 13 would be parallel to the x-y axis. Now for this ellipse to be a circle we equate the coefficients of x'^2 and y'^2 in equation 14

$$\begin{aligned}a\cos^2\theta + b\sin^2\theta - c\cos\theta\sin\theta &= a\sin^2\theta + b\cos^2\theta + c\cos\theta\sin\theta \\ a\cos^2\theta - a\sin^2\theta + b\sin^2\theta - b\cos^2\theta &= 2c\cos\theta\sin\theta \\ (a-b)\cos 2\theta &= c\sin 2\theta \\ \tan 2\theta &= \frac{a-b}{c}\end{aligned}$$

but from equation 16 we can finally write the above condition as

$$-(a-b)^2 = c^2 \tag{17}$$

which is true only when $a = b$ and $c = 0$(this condition is not ad-hoc and makes much more sense as there is no such thing as rotated circle so we can safely consider $c = 0$ without loss of generality). So for ellipse to be a circle the coefficients of x^2 and y^2 should be equal and you recover the standard equation of circle

$$x^2 + y^2 = r^2 \tag{18}$$

Now this whole analysis can be used to relate the coefficients of equation 13 to that of the section 2 which on comparision with equation 5 gives us

$$\begin{aligned}a &= \Sigma_{22} \\ b &= \Sigma_{11} \\ c &= -2\Sigma_{12}\end{aligned}$$

and now you can correlate the conclusions given in section 2 with the above calculations.

[10]Note that we are not considering the terms which are because of shift of origin because it would not affect the shape and tilted angle