CS669 : Pattern Recognition

Bayes Classifier

Assignment 1

Group 9

16th September 2018

Team members

Name	Roll Number
Lokesh Kumar	B16061
Aditya Singh	B16085
Anand Ramrakhyani	B16124

1 Introduction

Pattern Recognition involves looking for any kind of similarity (what we call a pattern) in the provided data, and then to use the nature of that pattern to classify the data into different classes. In this assignment, we have built a simple classifier, known as Bayes Classifier, on three different datasets :

(a) Linearly Separable Dataset,
(b) Non Linearly Separable Dataset,
(c) Real World Dataset

The data was randomly split into two parts -
(i) 75% of it was used as training data,
(ii) 25% of it was used as testing data.

2 Basic Terminology

1. Covariance Matrix : In probability theory and statistics, a covariance matrix is a matrix whose element in the (i,j) position is the covariance between the i-th and j-th elements of a random vector, where a random vector is a random variable with multiple dimensions. [1]

$$\begin{pmatrix} Cov(x_1,x_1) & Cov(x_1,x_2) & \cdots & Cov(x_1,x_n) \\ Cov(x_2,x_1) & Cov(x_2,x_2) & \cdots & Cov(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_n,x_1) & Cov(x_n,x_2) & \cdots & Cov(x_n,x_n) \end{pmatrix}$$

2. Accuracy : Accuracy measures the total number of correctly classified data points with respect to the total number of data points tested. Mathematically, in a given confusion matrix,

True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \ .$$

3. Precision : For a given class, precision is the ratio of number of data points identified as belonging to that class which actually belong to that class and the total number of data points which are identified as belonging to that class, i.e. for the positive class,

$$\text{Precision} = \frac{TP}{TP+FP} \ .$$

4. Recall : For a given class, recall is the ratio of number of data points identified as belonging to that class which actually belong to that class and the total number of data points which actually belong to that class, i.e. for the positive class,

$$\text{Recall} = \frac{TP}{TP+FN} \ .$$

5. F-Measure : It is the harmonic mean of precision and recall, i.e.

$$\text{F-Measure} = \frac{2*Precison*Recall}{Precision+Recall} \ .$$

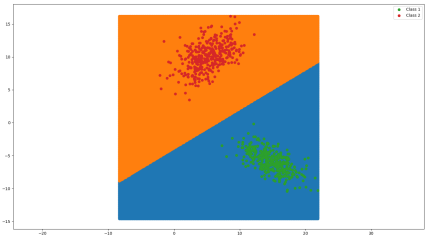
3 Observations

3.1 Covariance matrix for all the classes is the same and is σ^2I.

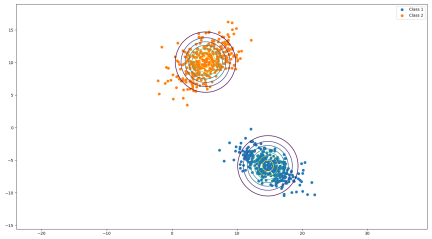
In this case, we have obtained the same Covariance matrix for all the classes by taking the average of Covariance matrices of all the classes and have obtained the same variance by averaging all the variences.

3.1.1 Linearly Separable Data

3.1.1.1 For Class 1 and Class 2

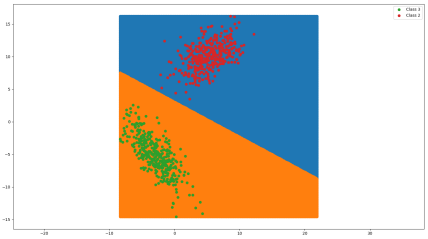


(a) Decision region plot.

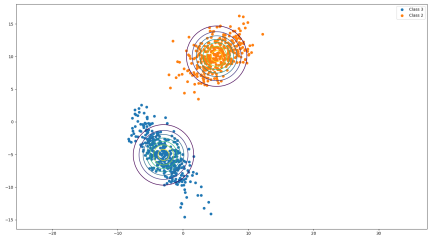


(b) Constant density contour plot.

3.1.1.2 For Class 2 and Class 3

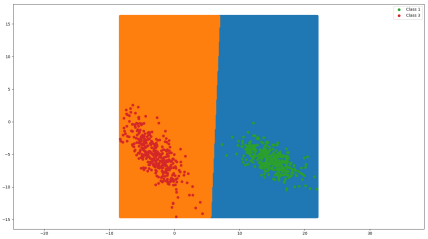


(c) Decision region plot.

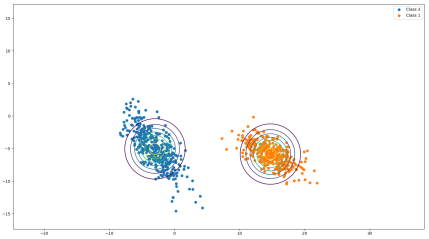


(d) Constant density contour plot.

3.1.1.3 For Class 3 and Class 1

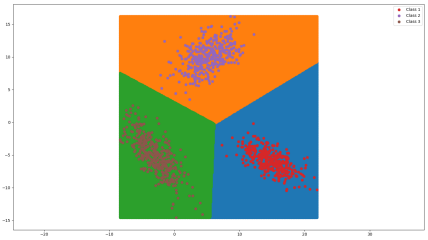


(e) Decision region plot.

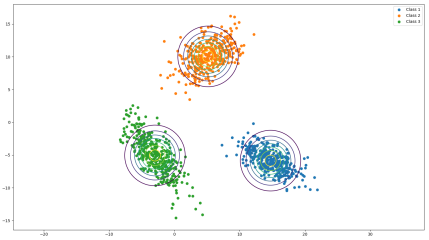


(f) Constant density contour plot.

3.1.1.4 For All Three Classes



(g) Decision region plot.



(h) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	125	0	0
Class 2	0	125	0
Class 3	0	0	125

Statistical Analysis

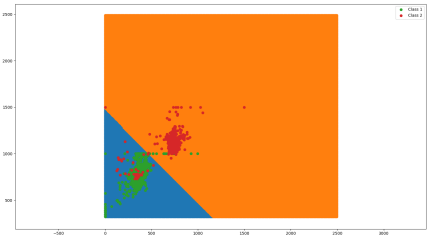
	Class 1	**Class 2**	**Class 3**
Precision	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F-Measure	1.00	1.00	1.00

A. Classification Accuracy = 100%

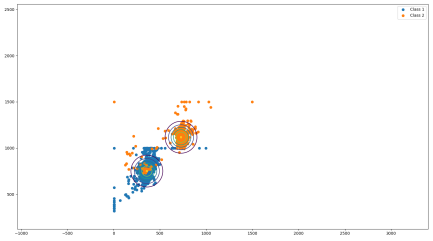
B. Mean Precision = 1.00
C. Mean Recall = 1.00
D. Mean F-Measure = 1.00

3.1.2 Non-Linearly Separable Data

3.1.2.1 For Class 1 and Class 2

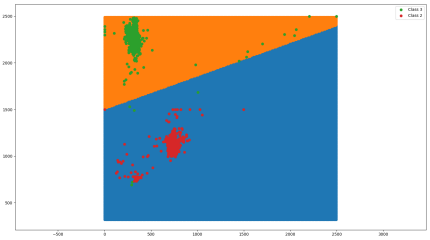


(i) Decision region plot.

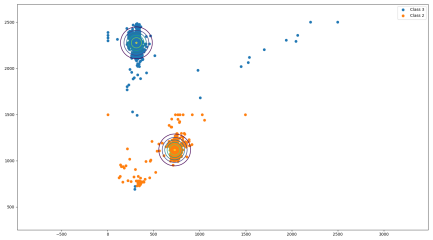


(j) Constant density contour plot.

3.1.2.2 For Class 2 and Class 3

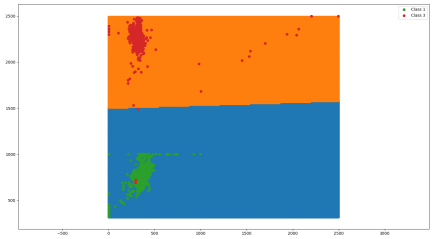


(k) Decision region plot.

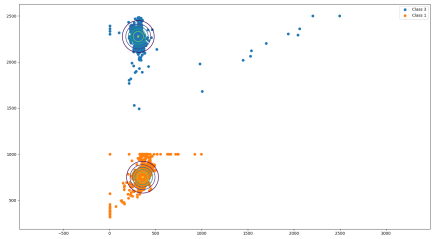


(l) Constant density contour plot.

3.1.2.3 For Class 3 and Class 1

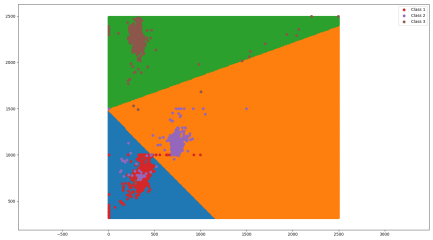


(m) Decision region plot.

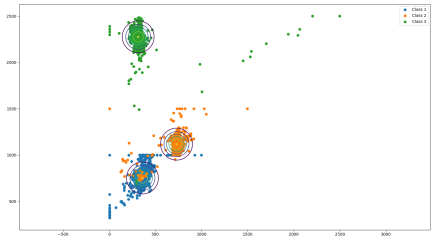


(n) Constant density contour plot.

3.1.2.4 For All Three Classes



(o) Decision region plot.



(p) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	617	5	0
Class 2	14	527	0
Class 3	1	1	595

Statistical Analysis

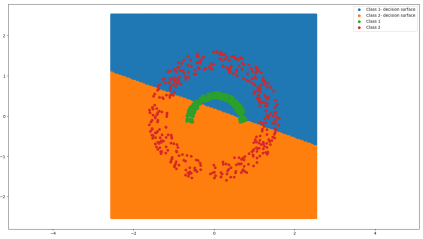
	Class 1	**Class 2**	**Class 3**
Precision	0.991	0.974	0.996
Recall	0.976	0.988	1.00
F-Measure	0.984	0.981	0.998

A. Classification Accuracy = 98.80%

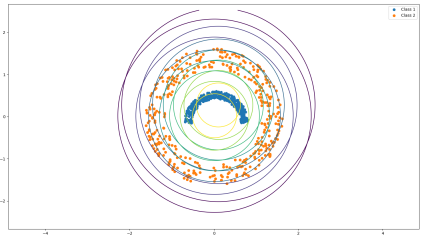
B. Mean Precision = 0.9875
C. Mean Recall = 0.9883
D. Mean F-Measure = 0.9879

3.1.3 Real World Data

3.1.3.1 For Class 1 and Class 2

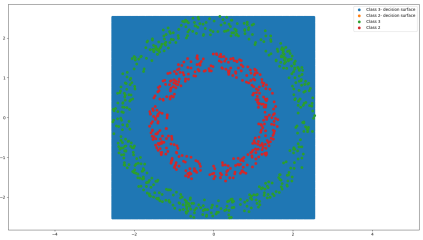


(q) Decision region plot.

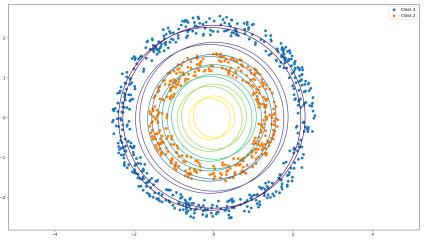


(r) Constant density contour plot.

3.1.3.2 For Class 2 and Class 3

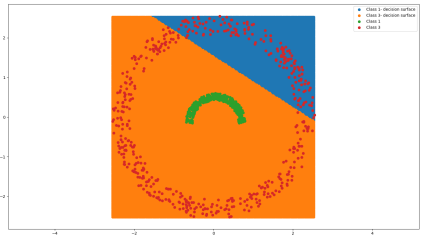


(s) Decision region plot.

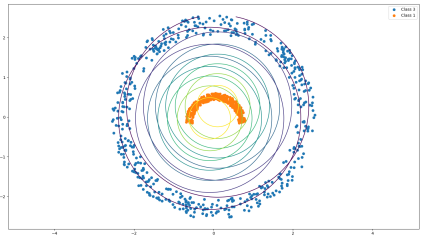


(t) Constant density contour plot.

3.1.3.3 For Class 3 and Class 1

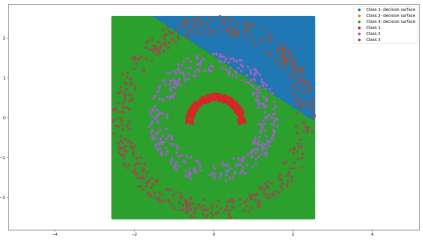


(u) Decision region plot.

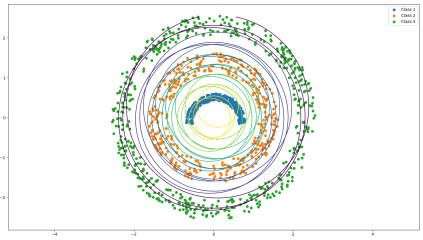


(v) Constant density contour plot.

3.1.3.4 For All Three Classes



(w) Decision region plot.



(x) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	0	0	125
Class 2	17	0	108
Class 3	63	0	112

Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	0.0	0.0	0.64
Recall	NA	NA	NA
F-Measure	NA	NA	NA

A. Classification Accuracy = 26.35%

B. Mean Precision = 0.2133
C. Mean Recall = NA
D. Mean F-Measure = NA

3.1.4 Discussions

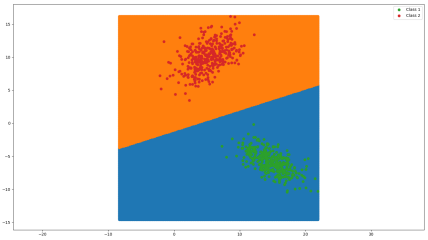
1. Decision boundaries are linear.
2. Contour plots of all the classes is exactly same and circular in shape because the covariance matrix of all the classes are diagonal with equal variances.

3.2 Covariance matrix for all the classes is the same and is Σ.

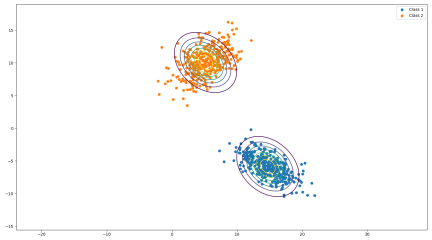
In this case,we obtained the same Covariance matrix for all the classes by taking the average of Covariance matrices of all the classes.

3.2.1 Linearly Separable Data

3.2.1.1 For Class 1 and Class 2

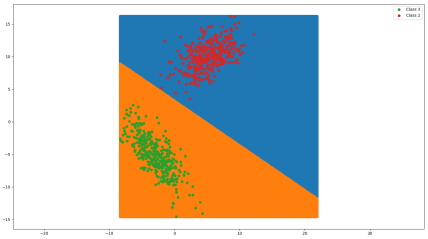


(y) Decision region plot.

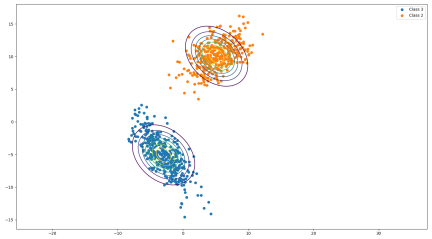


(z) Constant density contour plot.

3.2.1.2 For Class 2 and Class 3

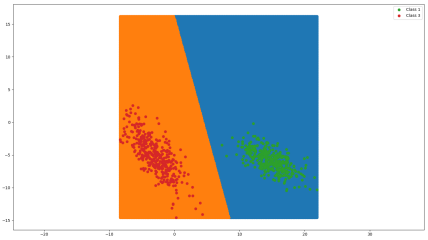


(aa) Decision region plot.

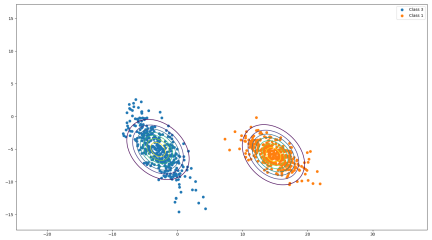


(ab) Constant density contour plot.

3.2.1.3 For Class 3 and Class 1

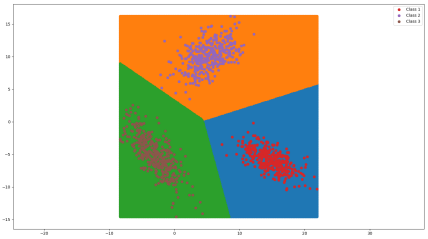


(ac) Decision region plot.

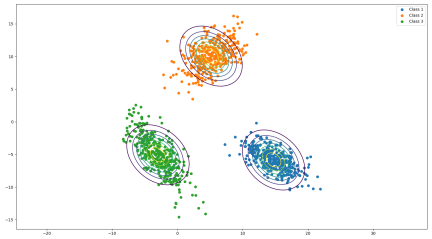


(ad) Constant density contour plot.

3.2.1.4 For All Three Classes



(ae) Decision region plot.



(af) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	125	0	0
Class 2	0	125	0
Class 3	0	0	125

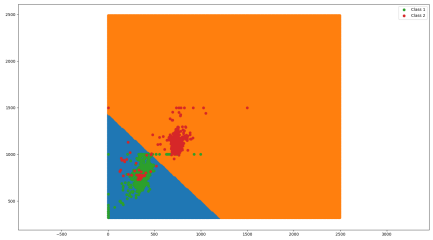
Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F-Measure	1.00	1.00	1.00

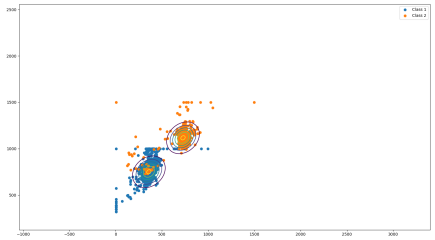
A. Classification Accuracy = 100%
B. Mean Precision = 1.00
C. Mean Recall = 1.00
D. Mean F-Measure = 1.00

3.2.2 Non-Linearly Separable Data

3.2.2.1 For Class 1 and Class 2

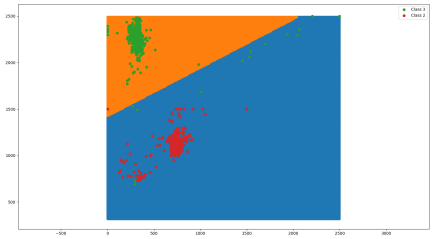


(ag) Decision region plot.

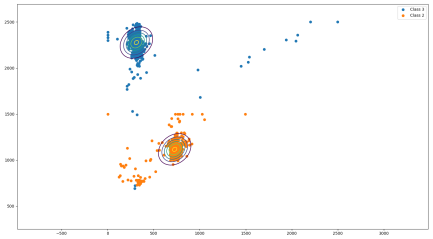


(ah) Constant density contour plot.

3.2.2.2 For Class 2 and Class 3

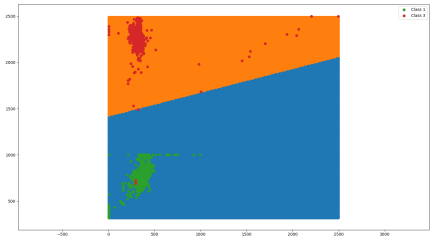


(ai) Decision region plot.

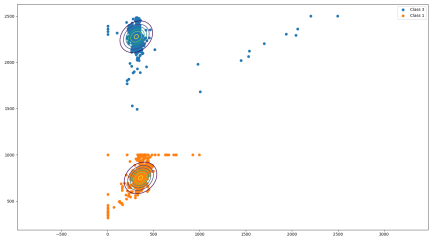


(aj) Constant density contour plot.

3.2.2.3 For Class 3 and Class 1

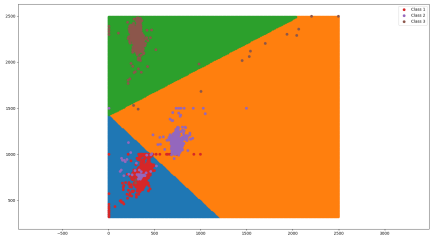


(ak) Decision region plot.

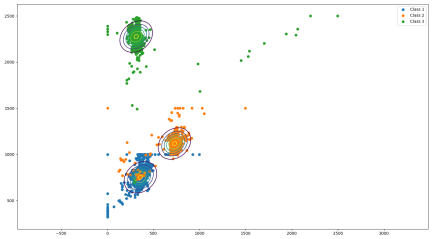


(al) Constant density contour plot.

3.2.2.4 For All Three Classes



(am) Decision region plot.



(an) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	617	5	0
Class 2	14	527	0
Class 3	1	4	592

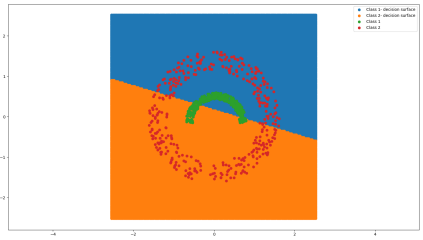
Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	0.991	0.974	0.991
Recall	0.976	0.983	0.978
F-Measure	0.984	0.978	0.995

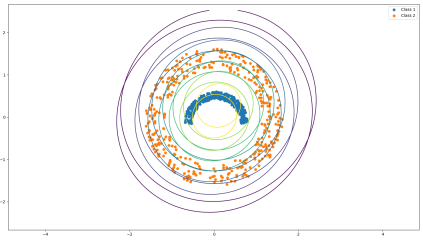
A. Classification Accuracy = 98.63%
B. Mean Precision = 0.9859
C. Mean Recall = 0.9864
D. Mean F-Measure = 0.9861

3.2.3 Real World Data

3.2.3.1 For Class 1 and Class 2

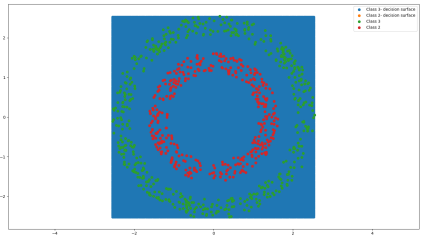


(ao) Decision region plot.

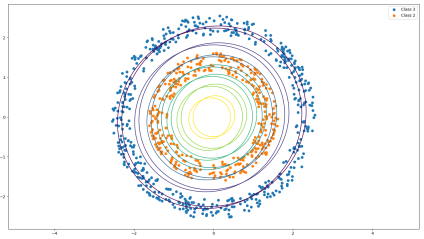


(ap) Constant density contour plot.

3.2.3.2 For Class 2 and Class 3

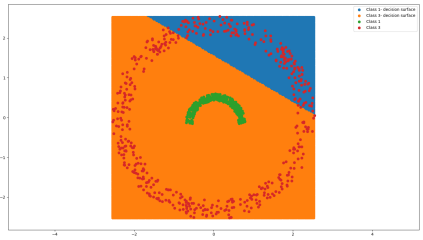


(aq) Decision region plot.

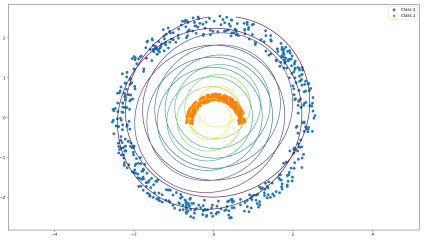


(ar) Constant density contour plot.

3.2.3.3 For Class 3 and Class 1

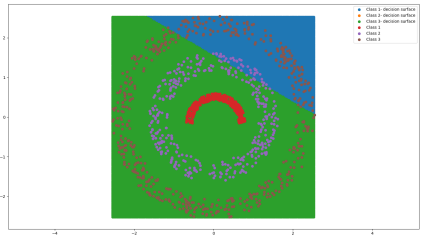


(as) Decision region plot.

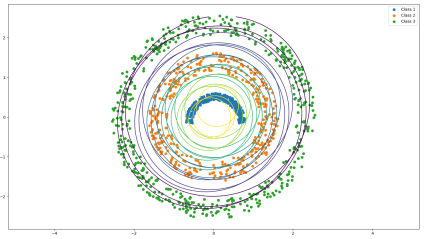


(at) Constant density contour plot.

3.2.3.4 For All Three Classes



(au) Decision region plot.



(av) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	0	0	125
Class 2	11	0	114
Class 3	59	0	116

Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	0.O	0.0	0.662
Recall	NA	NA	NA
F-Measure	NA	NA	NA

A. Classification Accuracy = 27.29%
B. Mean Precision = 0.2209
C. Mean Recall = NA
D. Mean F-Measure = NA

3.2.4 Discussions

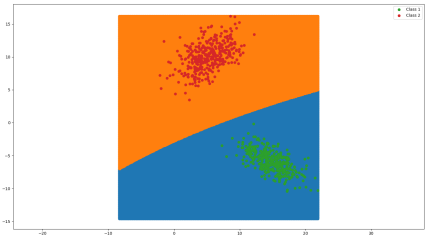
1. Decision boundaries are linear.
2. Contour plots of all the classes are exactly same but elliptical and tilted in shape.

3.3 Covariance matrices are diagonal and are different for each class.

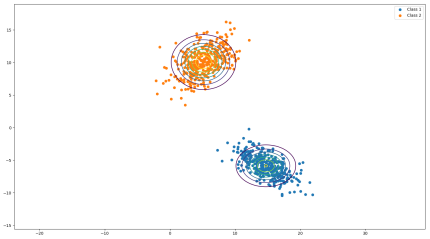
In this case, the non-diagonal terms of the matrices are taken to be zero.

3.3.1 Linearly Separable Data

3.3.1.1 For Class 1 and Class 2

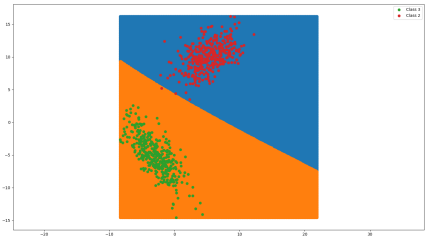


(aw) Decision region plot.

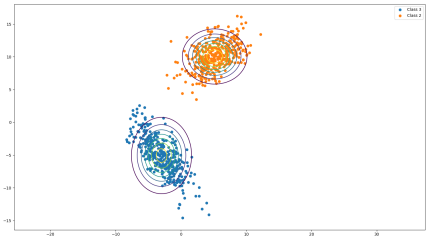


(ax) Constant density contour plot.

3.3.1.2 For Class 2 and Class 3

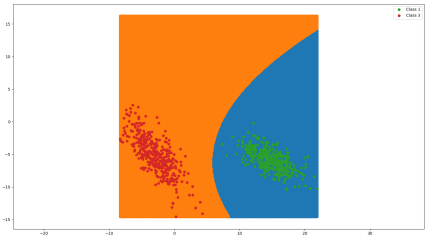


(ay) Decision region plot.

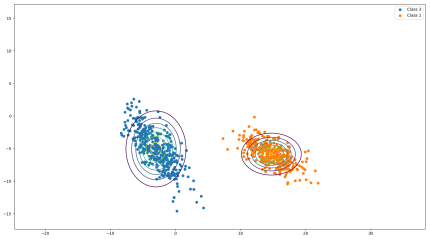


(az) Constant density contour plot.

3.3.1.3 For Class 3 and Class 1

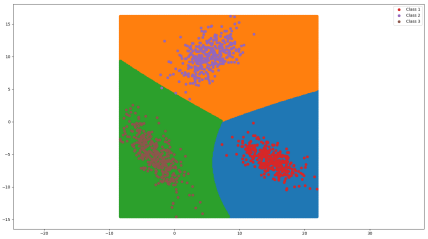


(ba) Decision region plot.

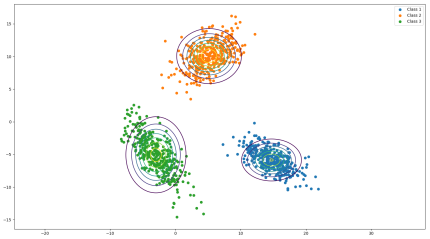


(bb) Constant density contour plot.

3.3.1.4 For All Three Classes



(bc) Decision region plot.



(bd) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	125	0	0
Class 2	0	125	0
Class 3	0	0	125

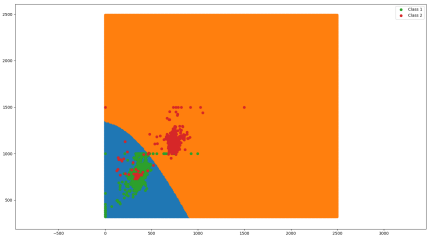
Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F-Measure	1.00	1.00	1.00

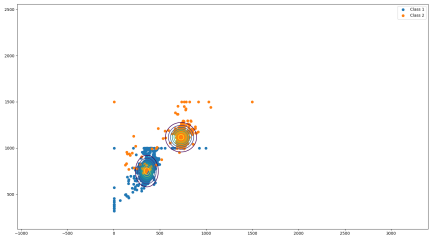
A. Classification Accuracy = 100%
B. Mean Precision = 1.00
C. Mean Recall = 1.00
D. Mean F-Measure = 1.00

3.3.2 Non-Linearly Separable Data

3.3.2.1 For Class 1 and Class 2

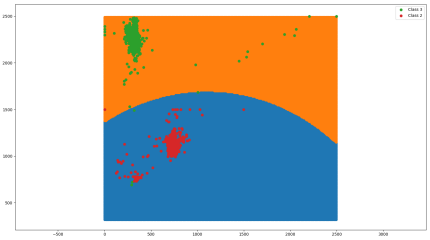


(be) Decision region plot.

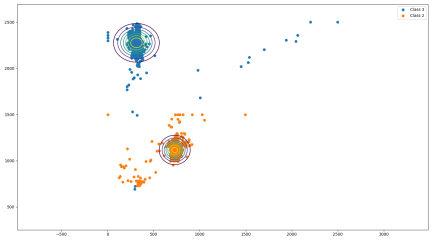


(bf) Constant density contour plot.

3.3.2.2 For Class 2 and Class 3

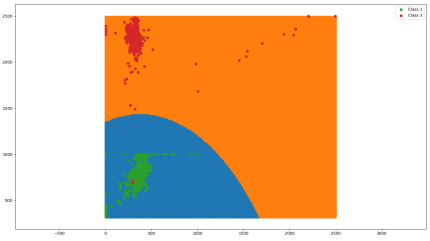


(bg) Decision region plot.

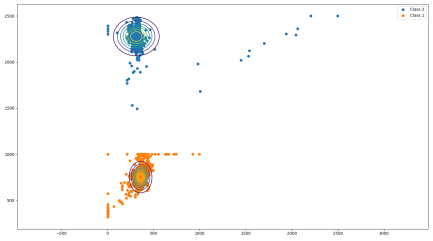


(bh) Constant density contour plot.

3.3.2.3 For Class 3 and Class 1

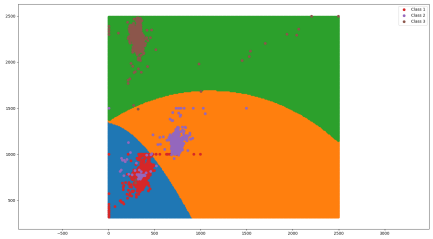


(bi) Decision region plot.

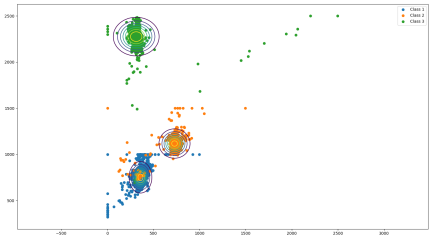


(bj) Constant density contour plot.

3.3.2.4 For All Three Classes



(bk) Decision region plot.



(bl) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	617	5	0
Class 2	14	527	0
Class 3	1	0	596

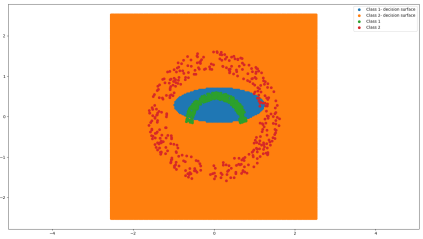
Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	0.991	0.974	0.998
Recall	0.976	0.990	1.00
F-Measure	0.984	0.982	0.999

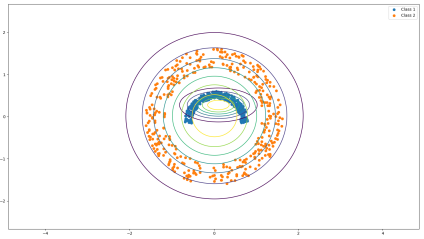
A. Classification Accuracy = 98.86%
B. Mean Precision = 0.9881
C. Mean Recall = 0.9889
D. Mean F-Measure = 0.9885

3.3.3 Real World Data

3.3.3.1 For Class 1 and Class 2

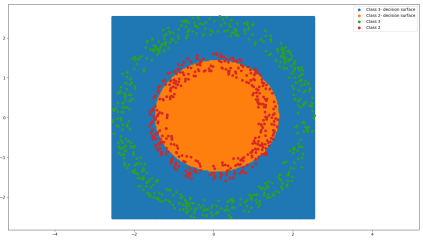


(bm) Decision region plot.

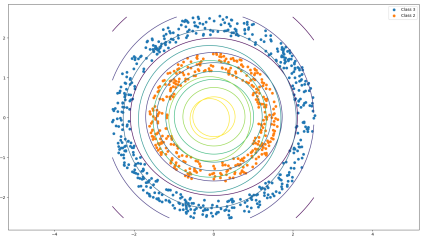


(bn) Constant density contour plot.

3.3.3.2 For Class 2 and Class 3

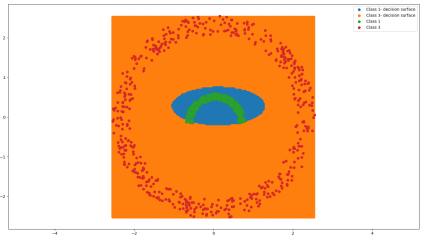


(bo) Decision region plot.

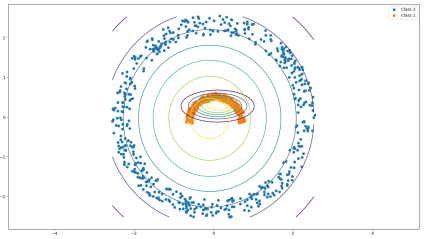


(bp) Constant density contour plot.

3.3.3.3 For Class 3 and Class 1

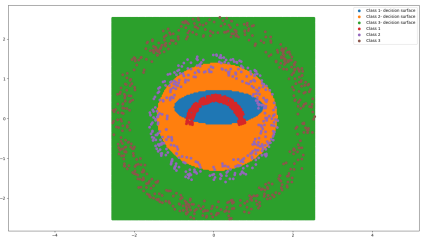


(bq) Decision region plot.

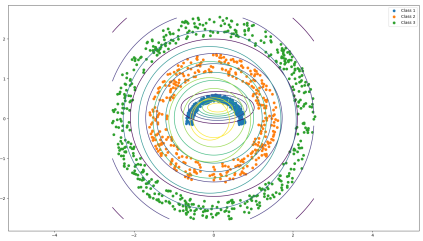


(br) Constant density contour plot.

3.3.3.4 For All Three Classes



(bs) Decision region plot.



(bt) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	116	9	0
Class 2	2	76	47
Class 3	0	0	175

Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	0.928	0.608	1.0
Recall	0.983	0.894	0.788
F-Measure	0.954	0.723	0.881

A. Classification Accuracy = 86.35%
B. Mean Precision = 0.8453
C. Mean Recall = 0.8884
D. Mean F-Measure = 0.8533

3.3.4 Discussions

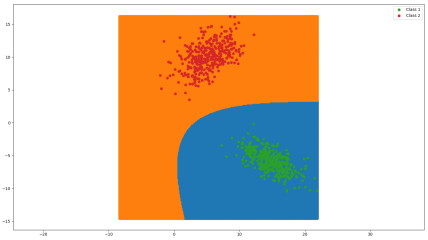
1. Decision boundaries are non-linear.
2. Contour plots of all the classes are different(because of diff. covariance matrices), elliptical and aligned with the feature axes(because non diagonal elements of cov. matrices are zero).

3.4 Covariance matrices are different for each class.

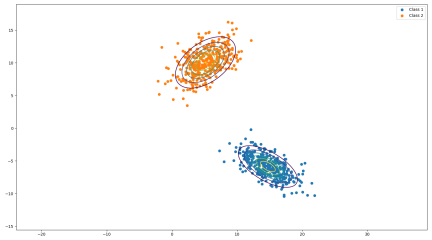
In this case, the covariance matrices are taken as such.

3.4.1 Linearly Separable Data

3.4.1.1 For Class 1 and Class 2

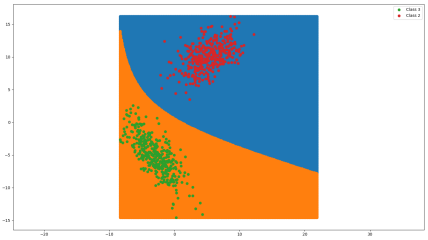


(bu) Decision region plot.

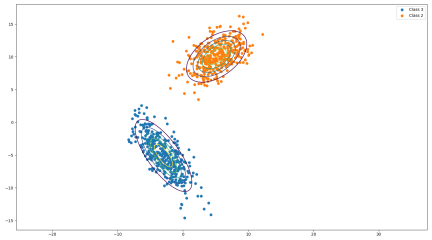


(bv) Constant density contour plot.

3.4.1.2 For Class 2 and Class 3

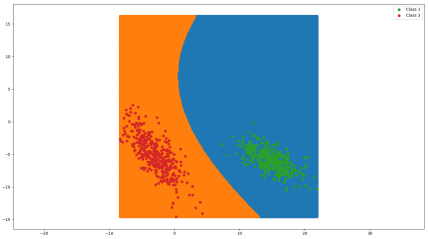


(bw) Decision region plot.

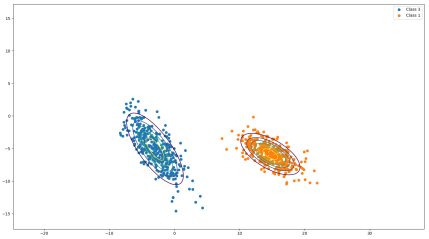


(bx) Constant density contour plot.

3.4.1.3 For Class 3 and Class 1

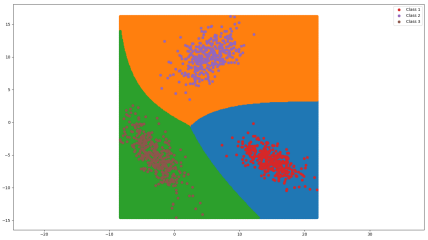


(by) Decision region plot.

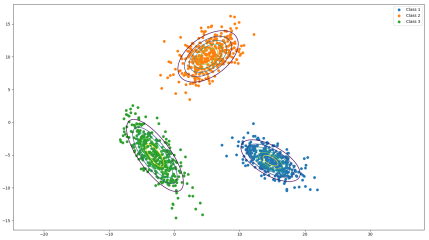


(bz) Constant density contour plot.

3.4.1.4 For All Three Classes



(ca) Decision region plot.



(cb) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	125	0	0
Class 2	0	125	0
Class 3	0	0	125

Statistical Analysis

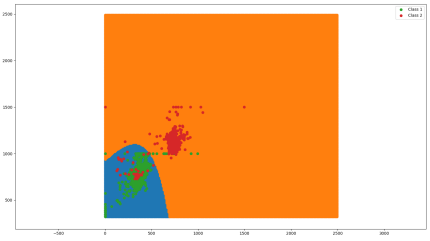
	Class 1	**Class 2**	**Class 3**
Precision	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F-Measure	1.00	1.00	1.00

A. Classification Accuracy = 100%

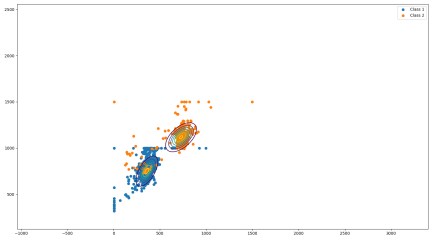
B. Mean Precision = 1.00
C. Mean Recall = 1.00
D. Mean F-Measure = 1.00

3.4.2 Non-Linearly Separable Data

3.4.2.1 For Class 1 and Class 2

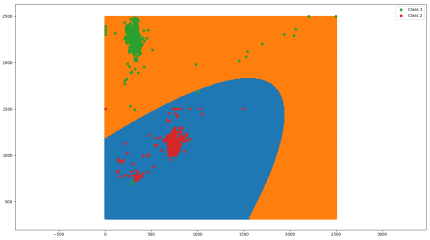


(cc) Decision region plot.

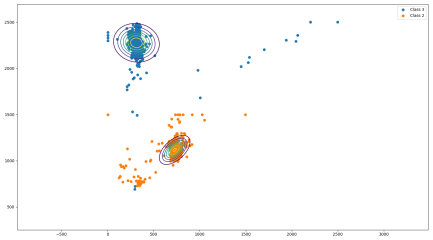


(cd) Constant density contour plot.

3.4.2.2 For Class 2 and Class 3

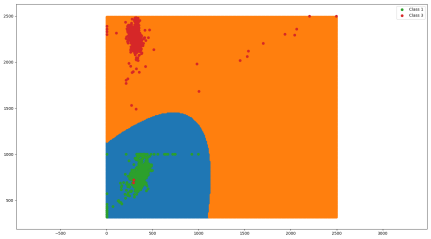


(ce) Decision region plot.

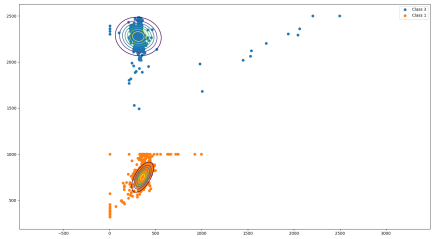


(cf) Constant density contour plot.

3.4.2.3 For Class 3 and Class 1

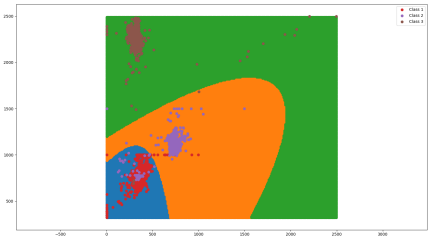


(cg) Decision region plot.

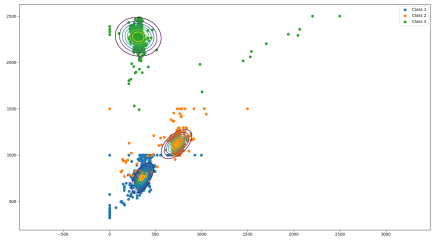


(ch) Constant density contour plot.

3.4.2.4 For All Three Classes



(ci) Decision region plot.



(cj) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	616	6	0
Class 2	13	528	0
Class 3	1	0	596

Statistical Analysis

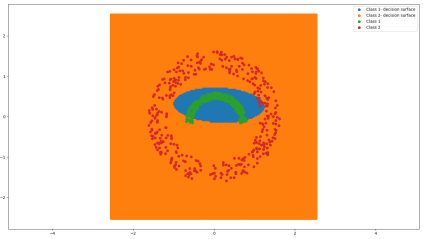
	Class 1	**Class 2**	**Class 3**
Precision	0.990	0.975	0.998
Recall	0.977	0.988	1.00
F-Measure	0.984	0.982	0.999

A. Classification Accuracy = 98.86%

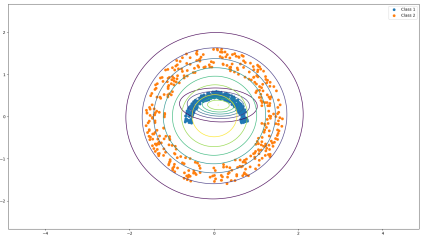
B. Mean Precision = 0.9882
C. Mean Recall = 0.9888
D. Mean F-Measure = 0.9885

3.4.3 Real World Data

3.4.3.1 For Class 1 and Class 2

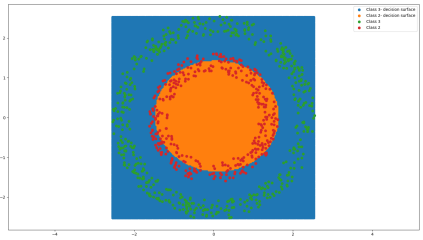


(ck) Decision region plot.

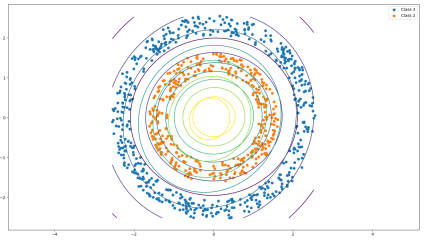


(cl) Constant density contour plot.

3.4.3.2 For Class 2 and Class 3

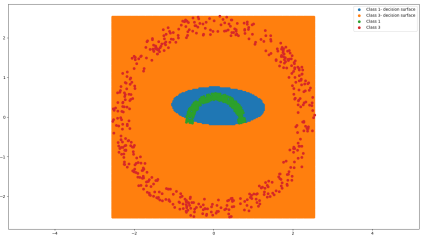


(cm) Decision region plot.

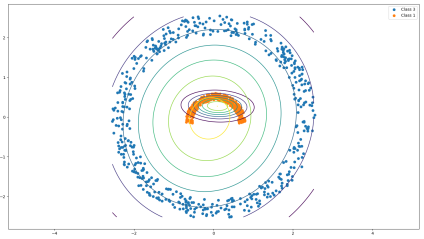


(cn) Constant density contour plot.

3.4.3.3 For Class 3 and Class 1

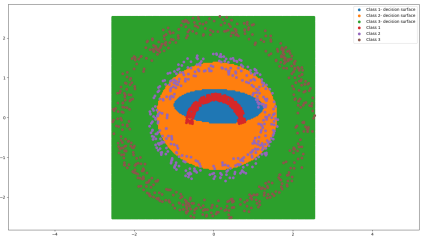


(co) Decision region plot.

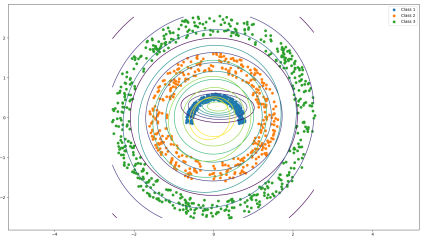


(cp) Constant density contour plot.

3.4.3.4 For All Three Classes



(cq) Decision region plot.



(cr) Constant density contour plot.

Confusion Matrix

	Class 1	**Class 2**	**Class 3**
Class 1	115	10	0
Class 2	0	79	46
Class 3	0	0	175

Statistical Analysis

	Class 1	**Class 2**	**Class 3**
Precision	0.92	0.632	1.0
Recall	1.0	0.887	0.791
F-Measure	0.958	0.738	0.883

A. Classification Accuracy = 86.82%

B. Mean Precision = 0.8506
C. Mean Recall = 0.8931
C. Mean F-Measure = 0.8601

3.4.4 Discussions

1. Decision boundaries are non-linear.
2. Contour plots of all the classes are different, elliptical and titled because all the covariance matrices are different and arbitrary

4 Conclusion

1. Shape of discriminant boundary and contour plots depends on the nature of covariance matrix.

2. In case of linearly separable data, all the classifiers showed 100% accuracy on the test set.

3. In case of non-linearly separable data, the accuracy dropped to around 98% for all cases.

4. In case of real-world data, only classifier 3 and 4 had an acceptable accuracy of 86.35% and 86.82%, others failed miserably with approx. 27% accuracy.

5. Hence, we conclude that Bayes classifier is highly accurate in classification of linearly separable data but has lower accuracy when it comes to non-linear and real world data, as the decision boundary can be atmost quadratic in nature which may not be sufficient in handling different types of class distributions.

References

[1] Covariance Matrix
https://en.wikipedia.org/wiki/Covariance_matrix