CS 669 Assignment 1

Rohit Patiyal
Devang Bacharwar

September 14, 2015

1 Objective

To build Bayes and Naive-Bayes classifiers for different types of data sets :

1.1 2-D artificial Data of 3 or 4 classes

1. Linearly separable data set
2. Nonlinearly separable data sets (3 Data sets)
3. Overlapping data set

1.2 Real World data set

2 Procedure

1. Data for each class is partitioned into 75 % for training and 25 % for testing
2. Mean and Covariances are calculated for each class using the training .
3. For points in a grid, likelihood is calculated for each class and is labeled as of the class with the maximum likelihood probability.

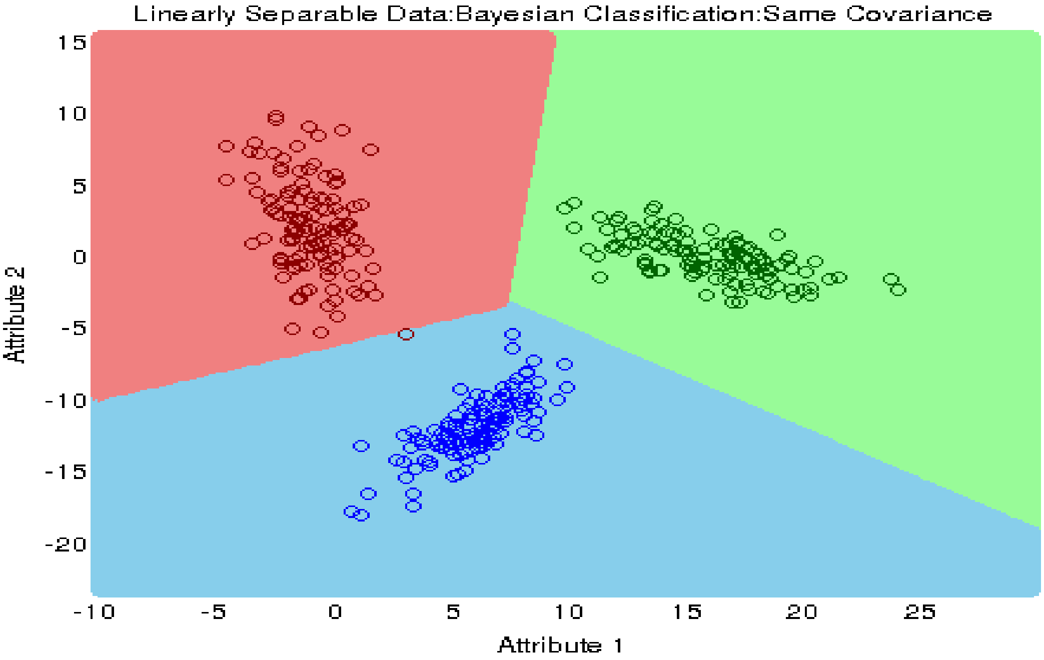
   For bayes classifier, the likelihood is assumed to be a multivariate gaussian distribution
4. These labelled points are plotted with different colors to see the different regions separated by the decision boundaries.
5. The testing data is also plotted over the regions, and observations a re made.

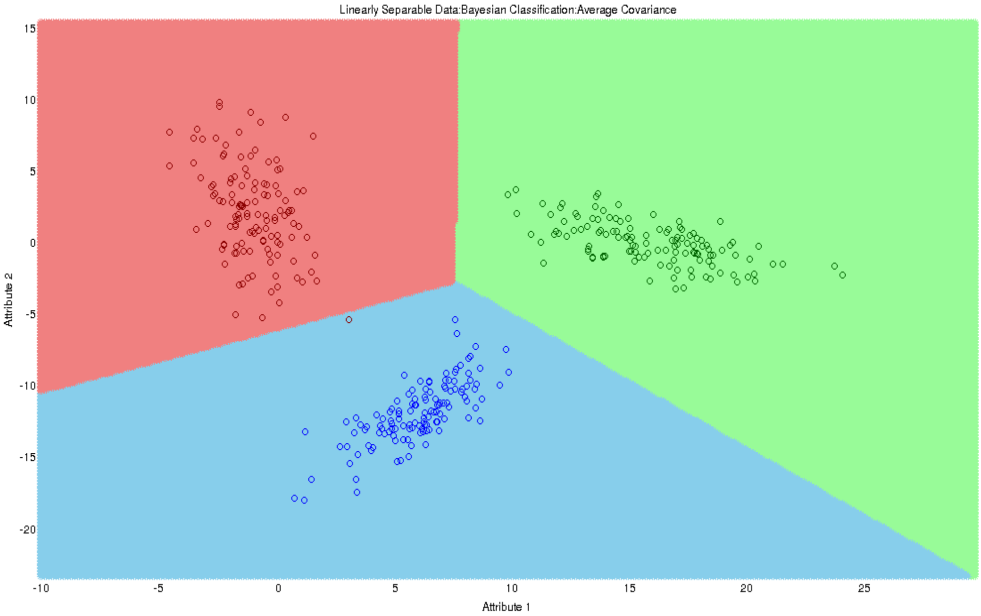
3 Observations

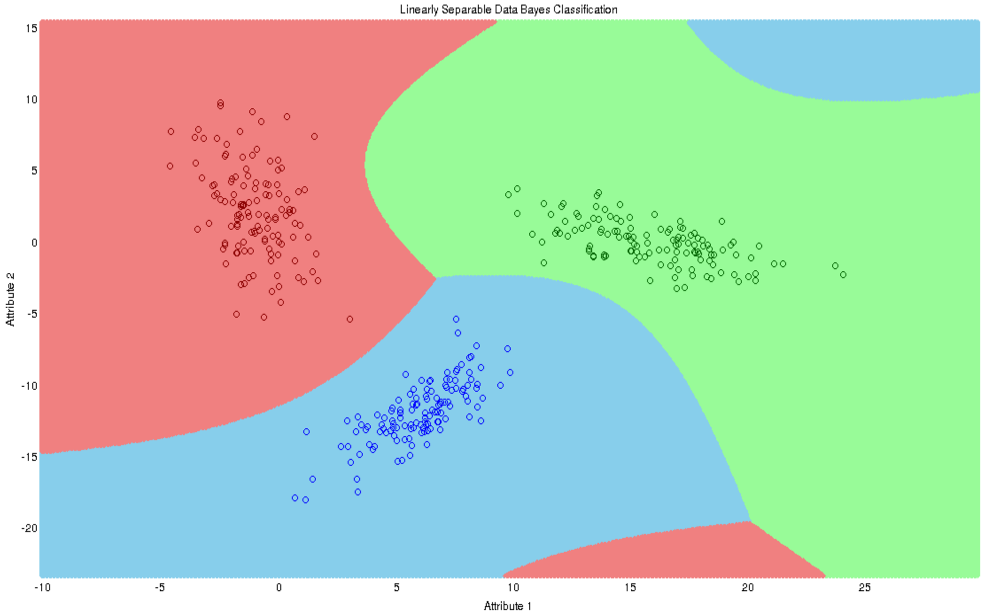
3.1 Bayes Classifier

3.1.1 Linearly separable data set

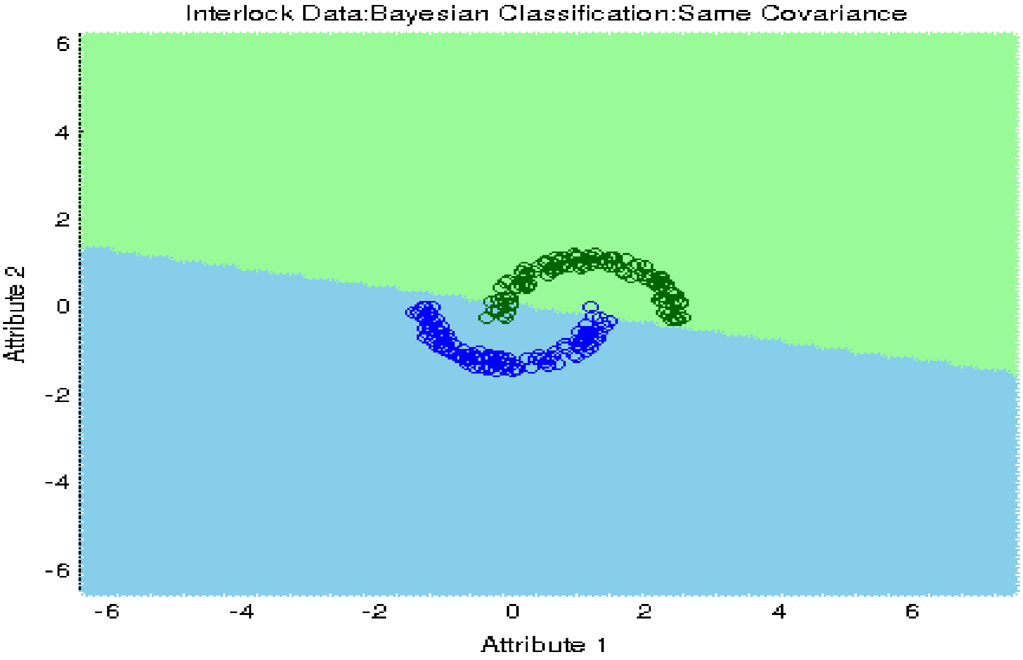
The decision boundary clearly separates the testing data as per classes as the data forms widely separated clusters.

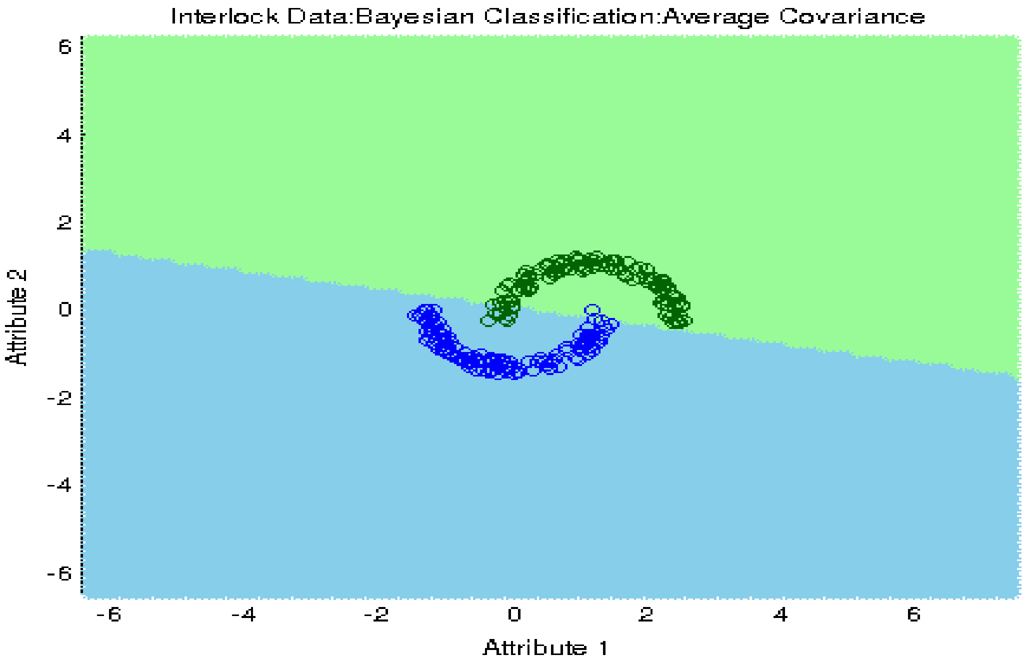


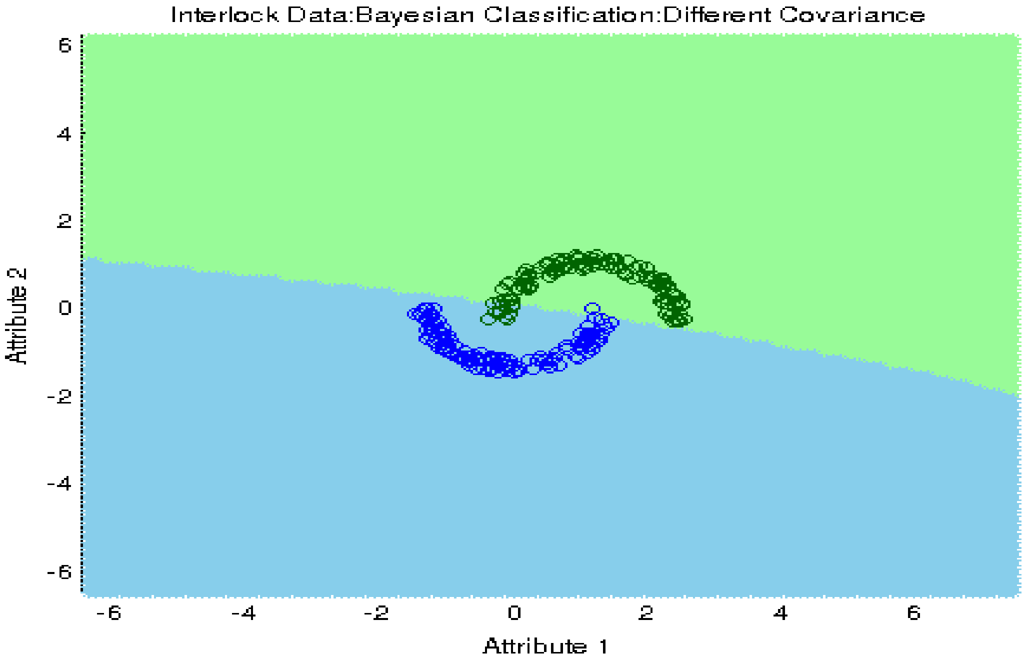




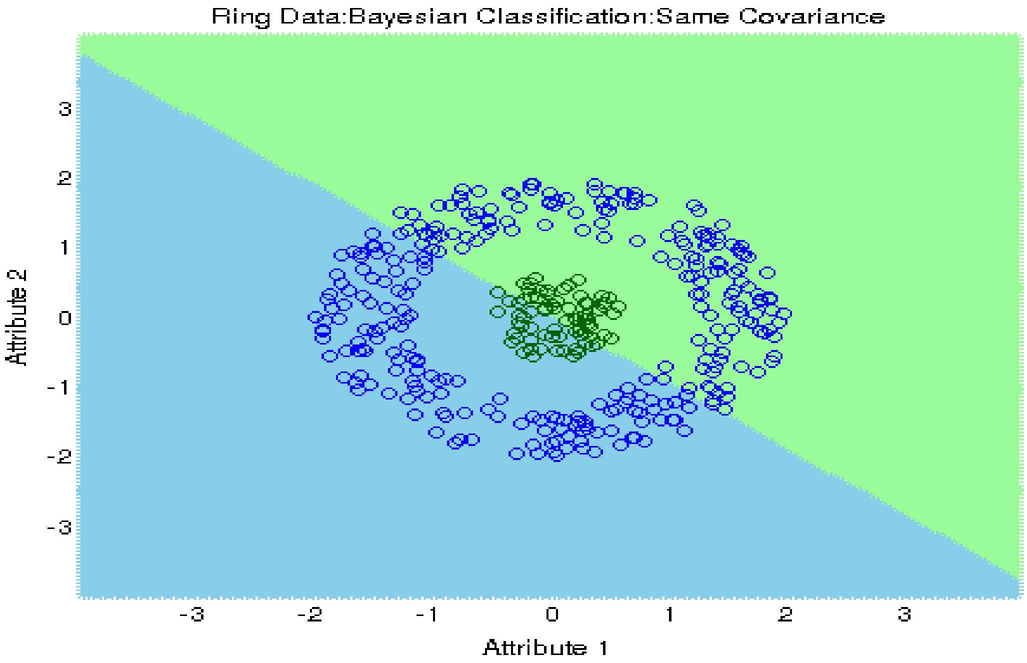
3.1.2 Non-Linearly separable data set

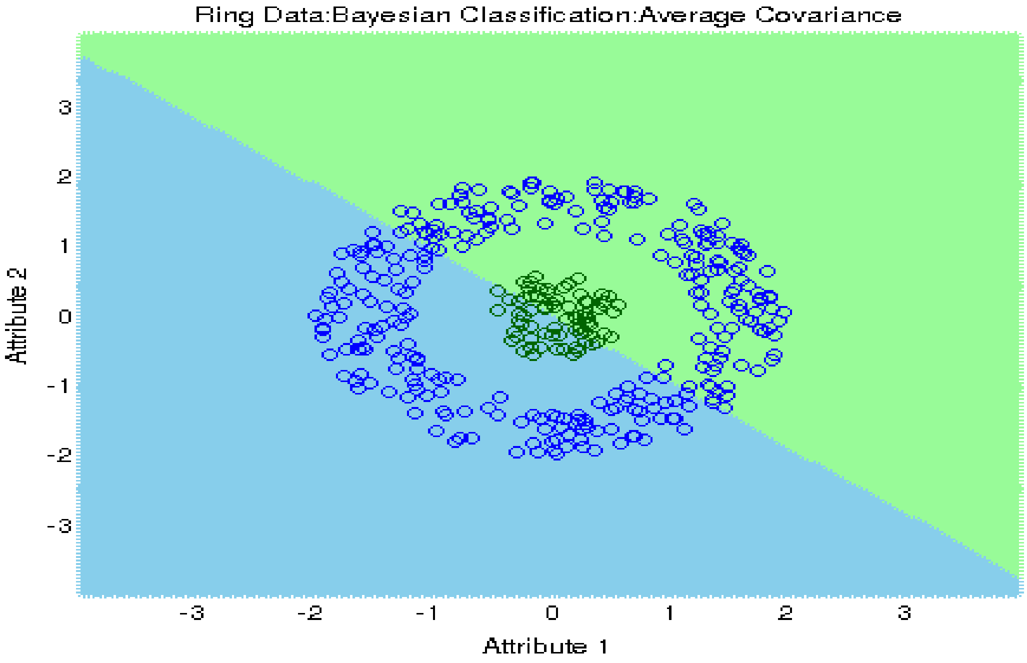


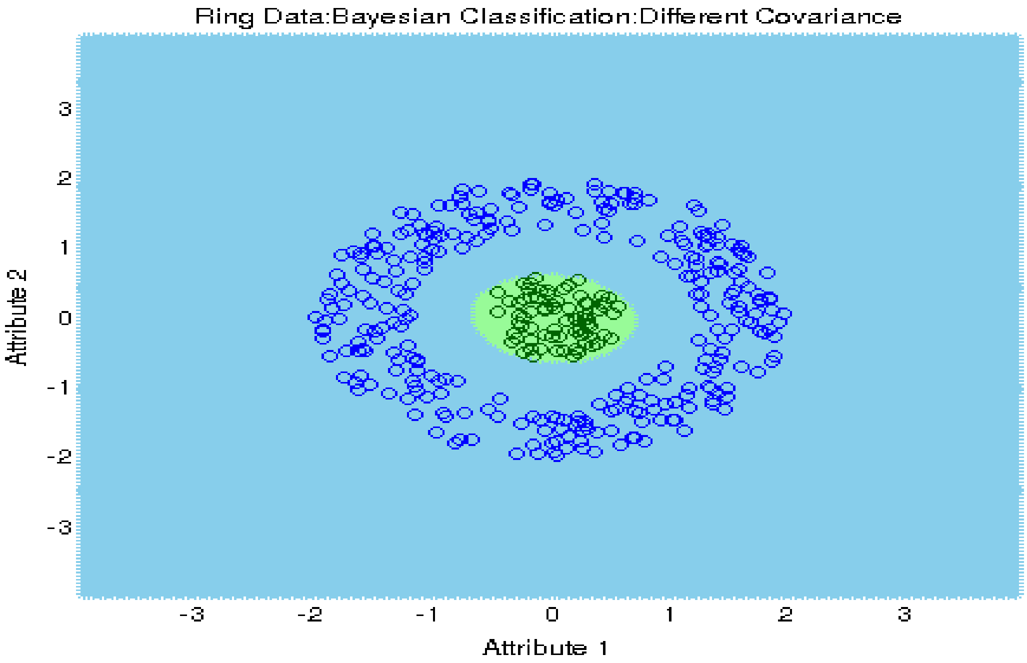




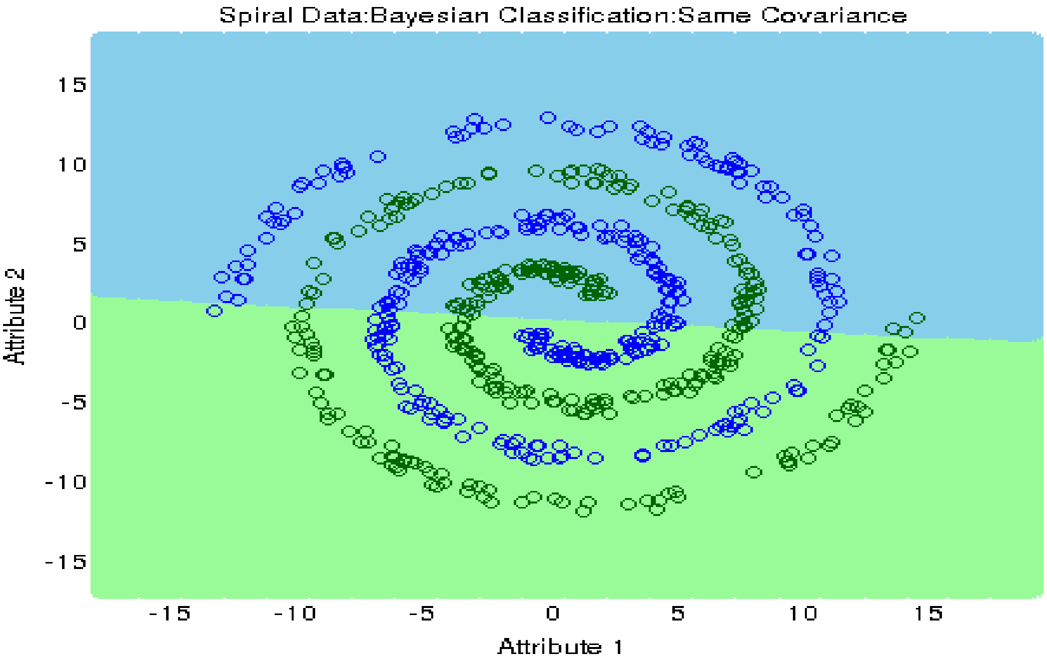
3.1.2.1 Data of Interlocking Classes

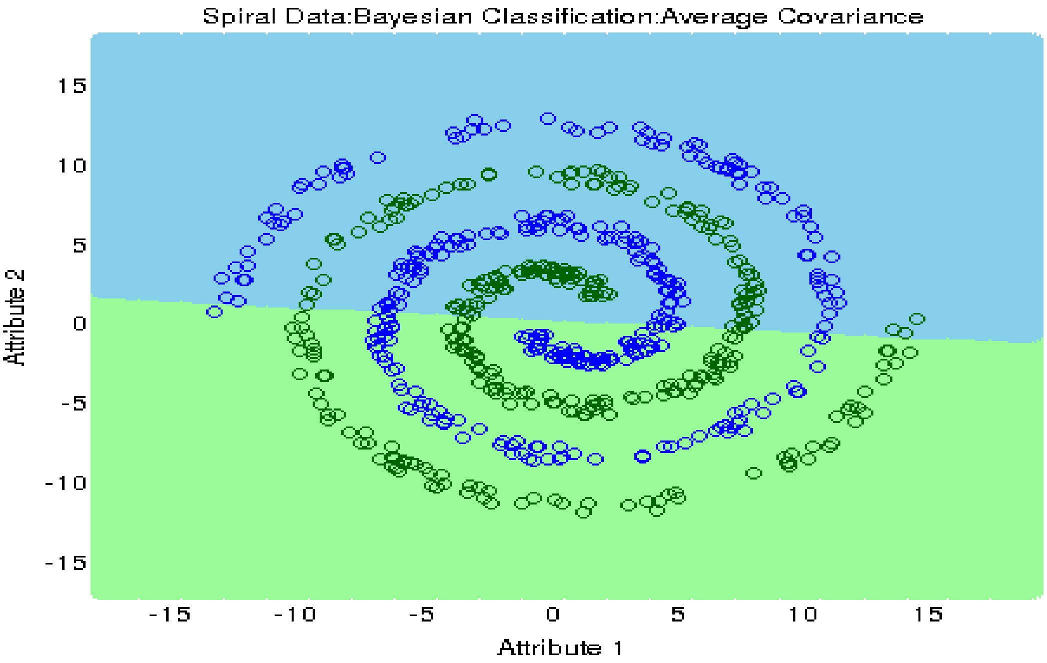

Ring Data:Bayesian Classification:Same Covariance
3
2
1
0
-1
-2
-3
Attribute 2
-3
-2
-1
0
1
2
3
Attribute 1

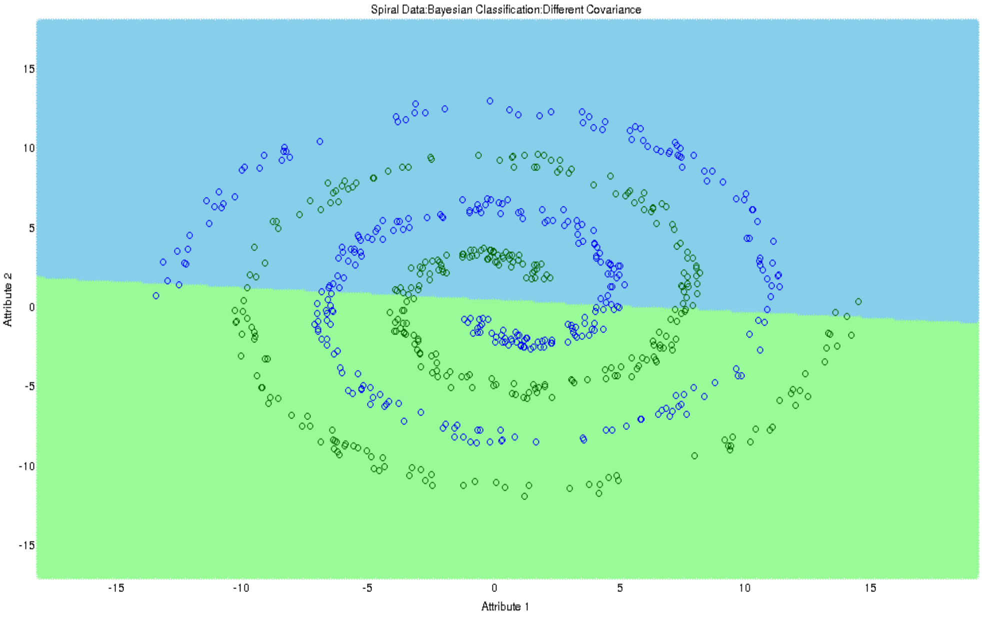

Ring Data:Bayesian Classification:Average Covariance
3
2
1
0
-1
-2
-3
Attribute 2
-3
-2
-1
0
1
2
3
Attribute 1



3.1.2.2 A ring with a central mass

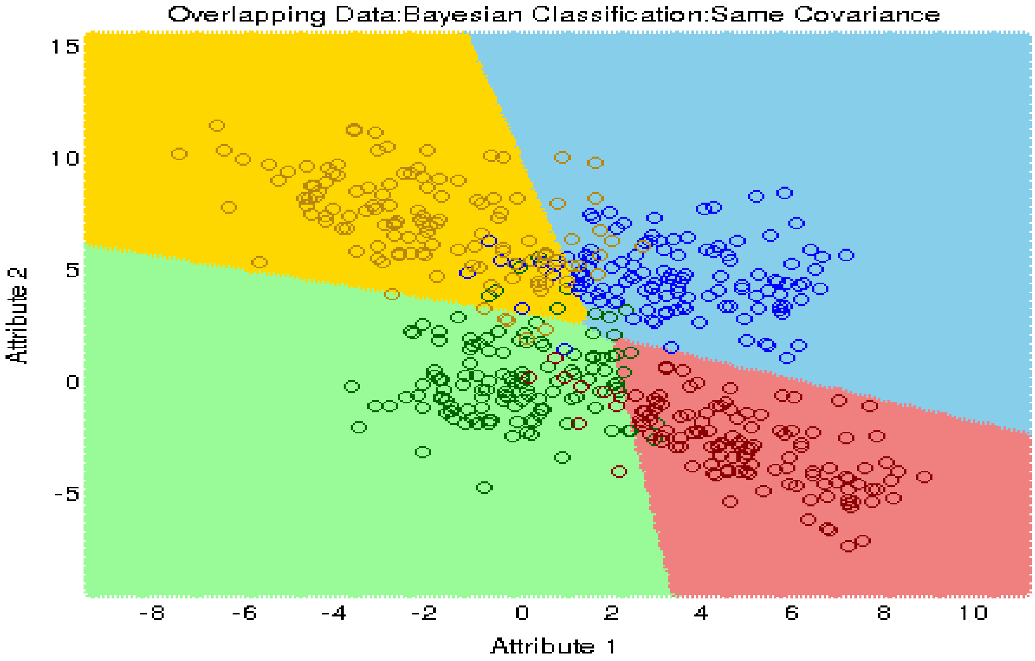


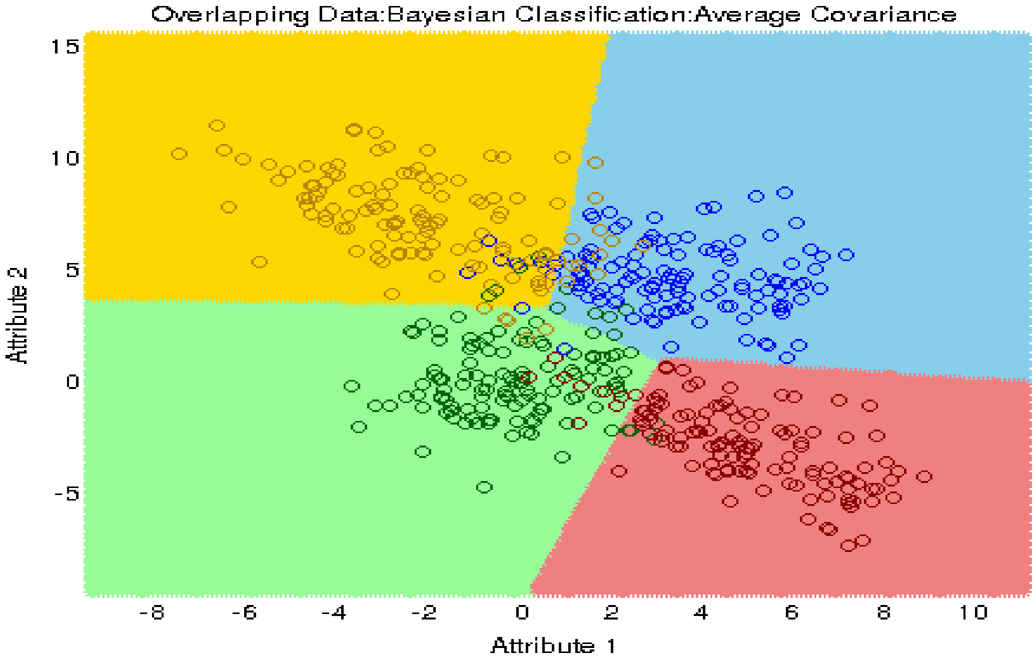


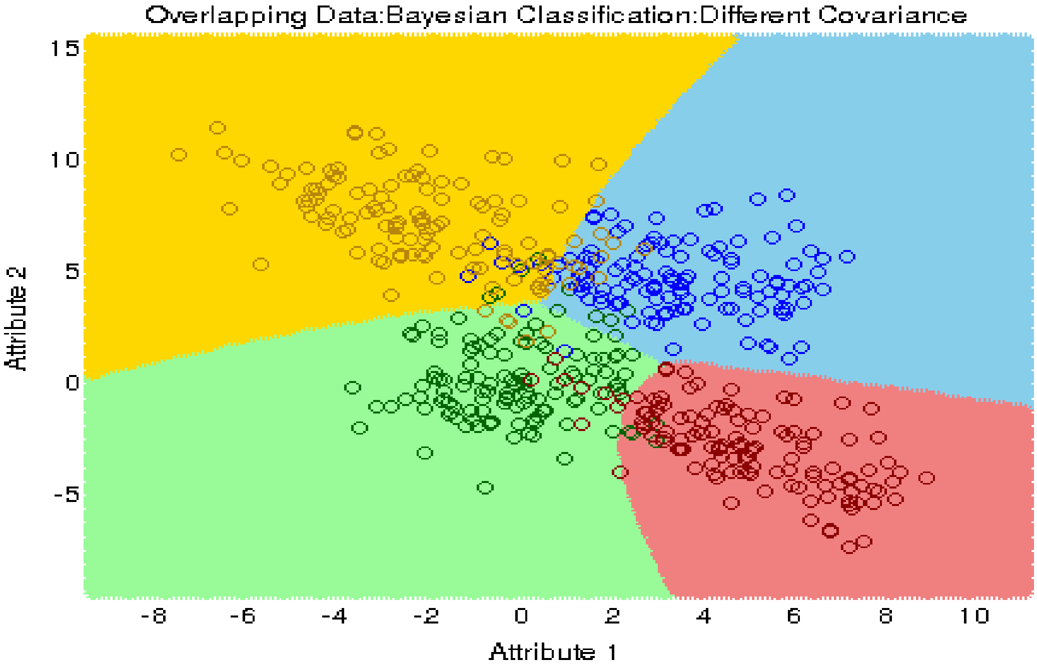


3.1.2.3 Spiral Dataset

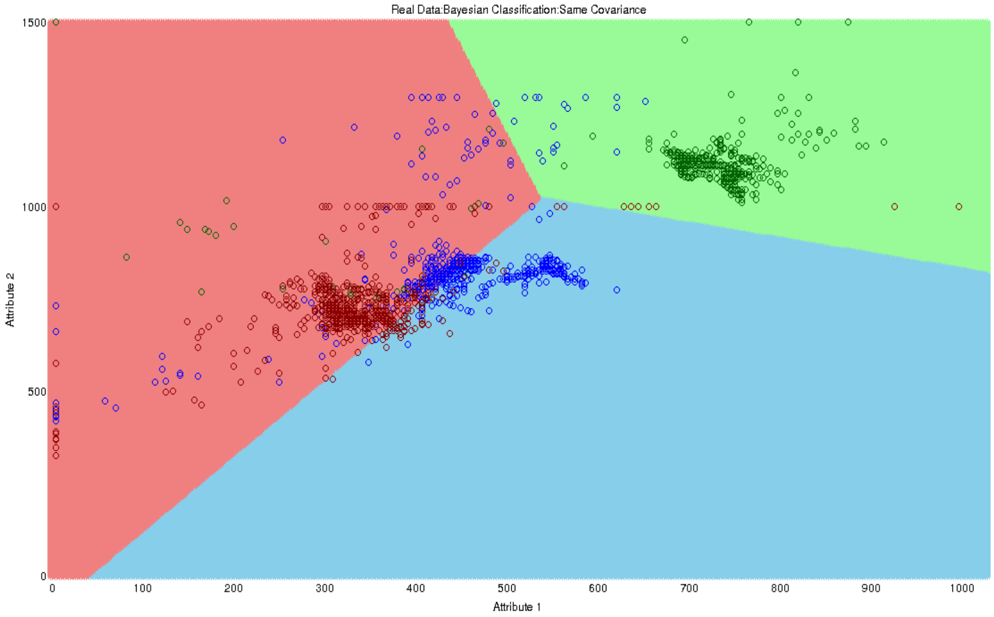
3.1.3 Overlapping data set

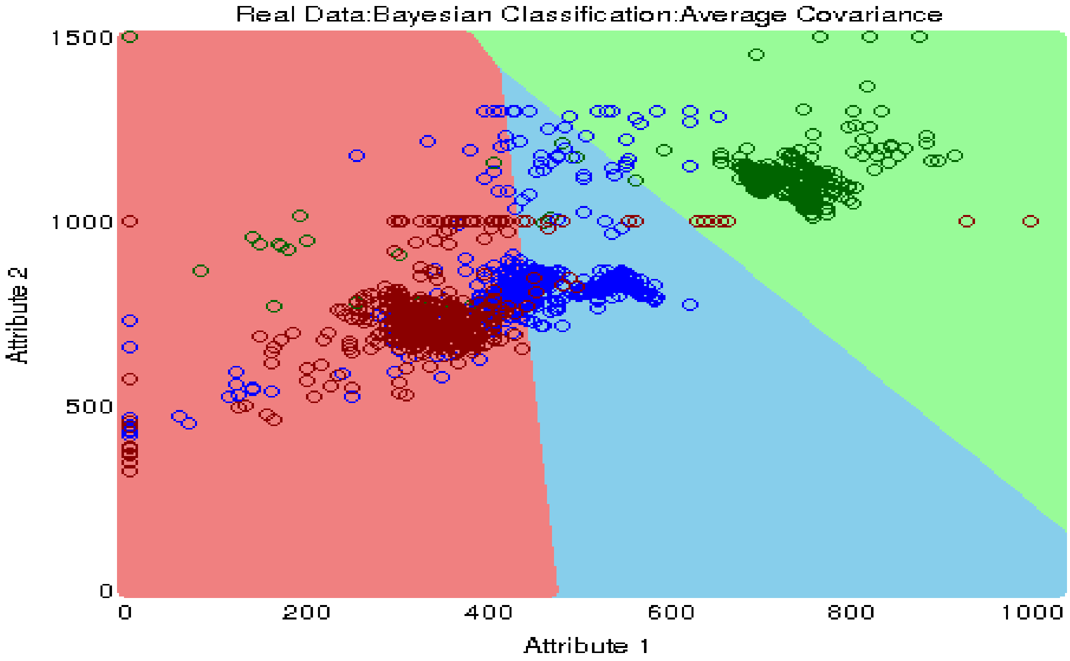


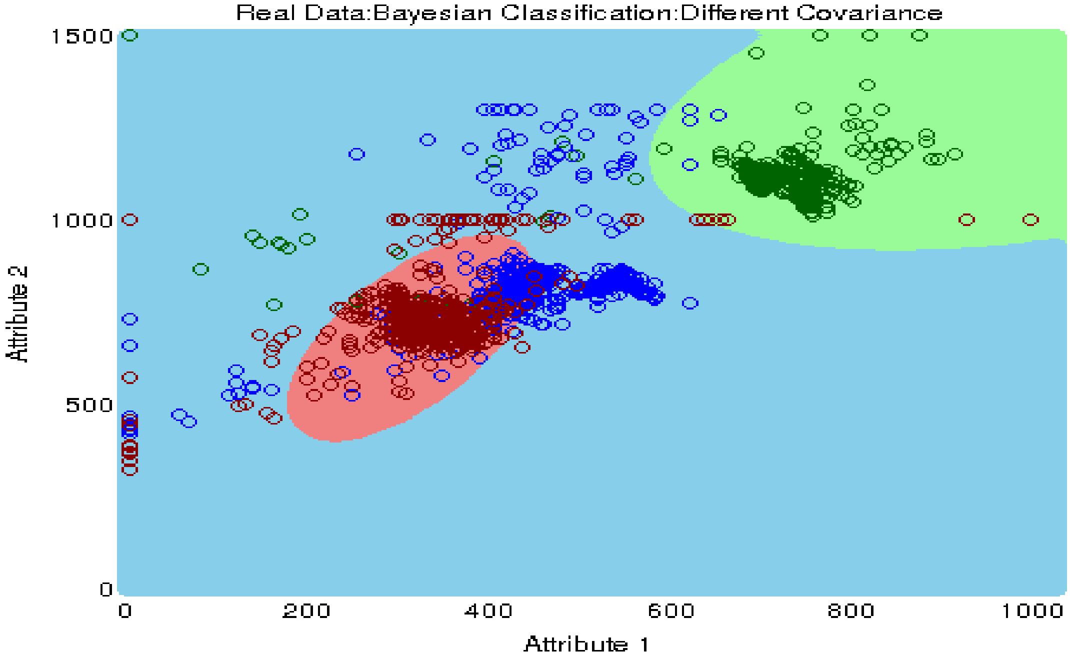




3.1.4 Real world data set



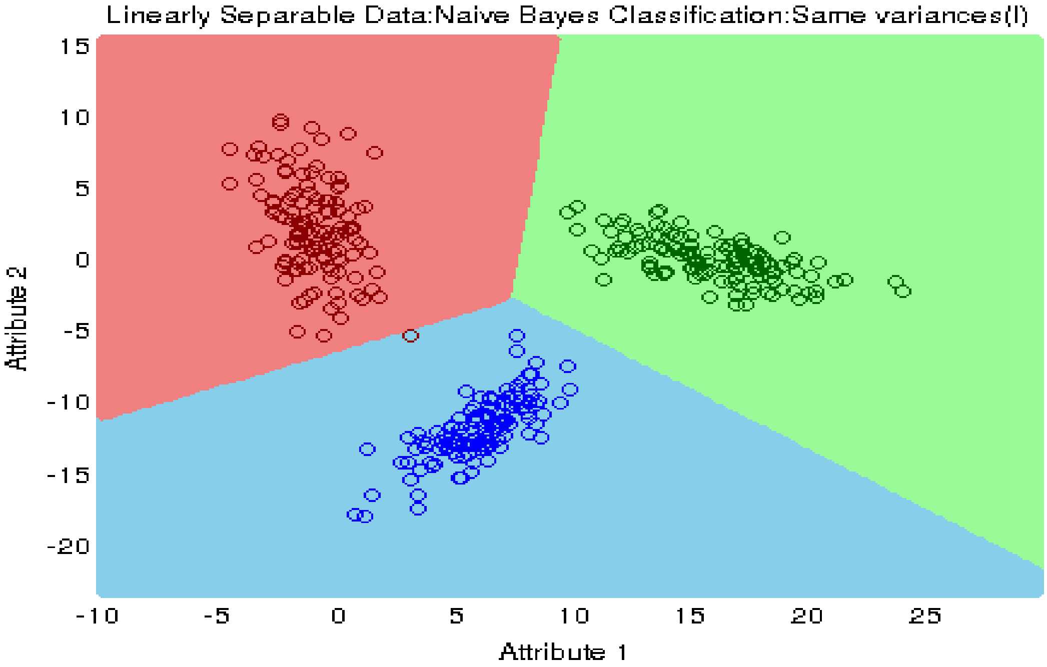


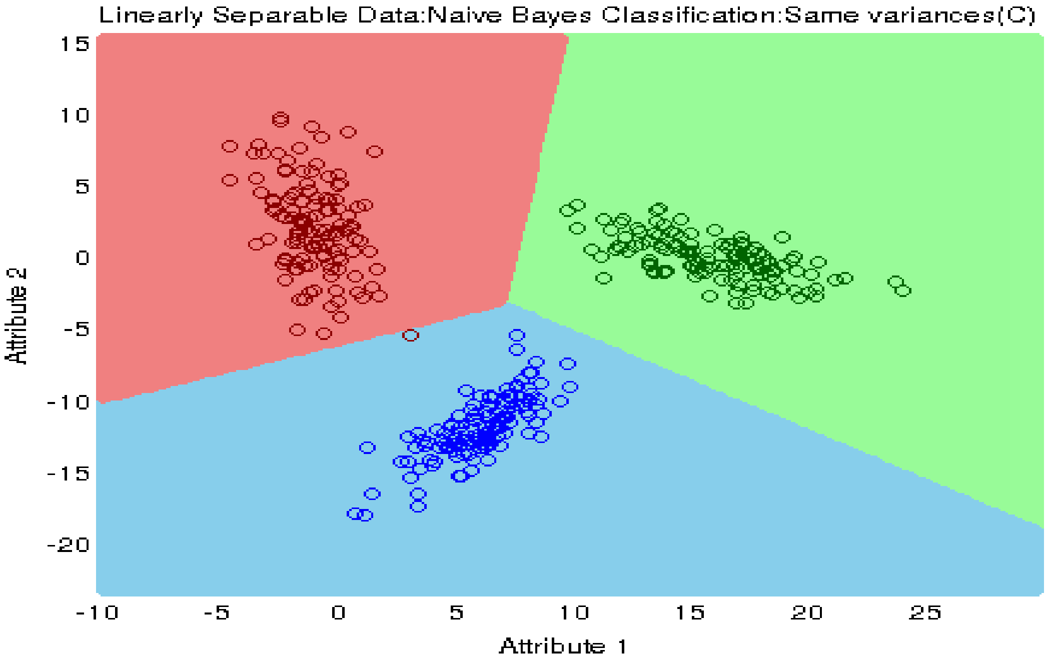


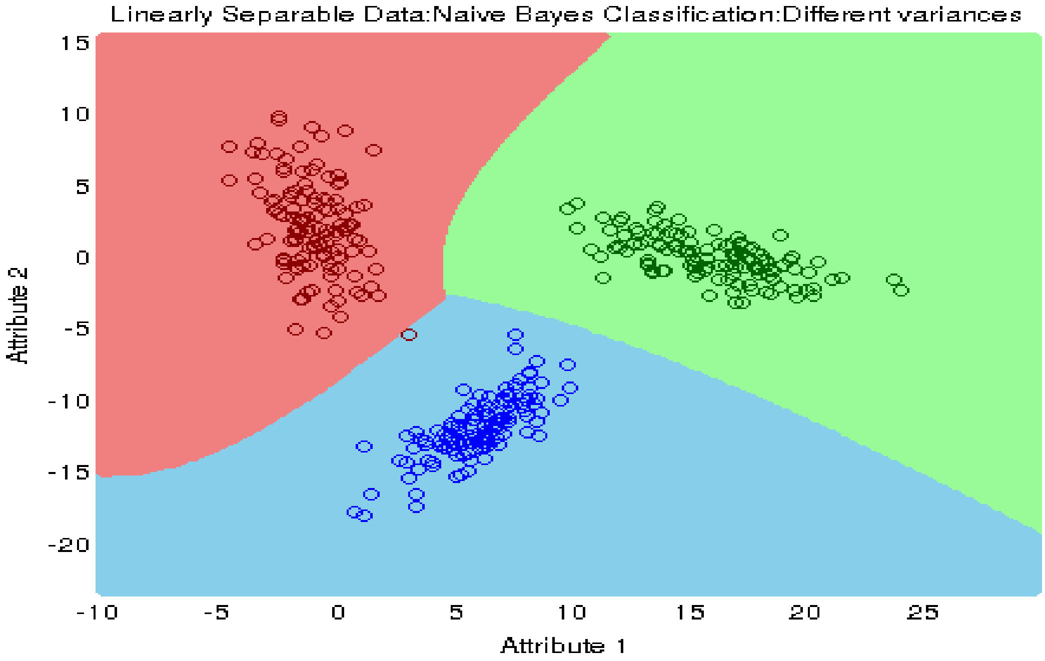
3.2 Naive-Bayes classifier

3.2.1 Linearly separable data set

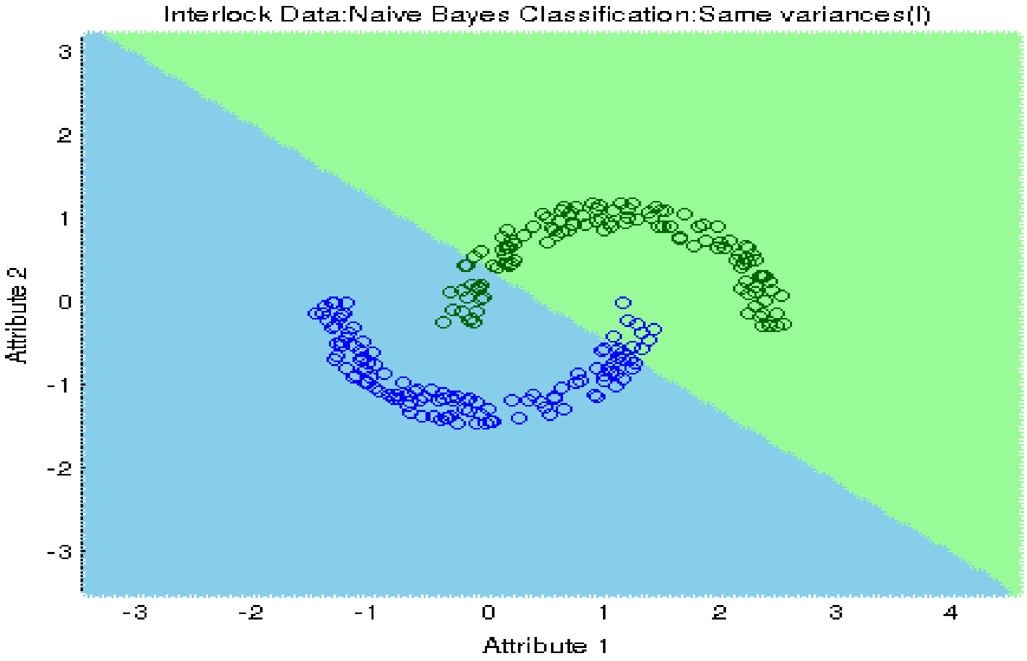
The decision boundary clearly separates the testing data as per classes as the data forms widely separated clusters.

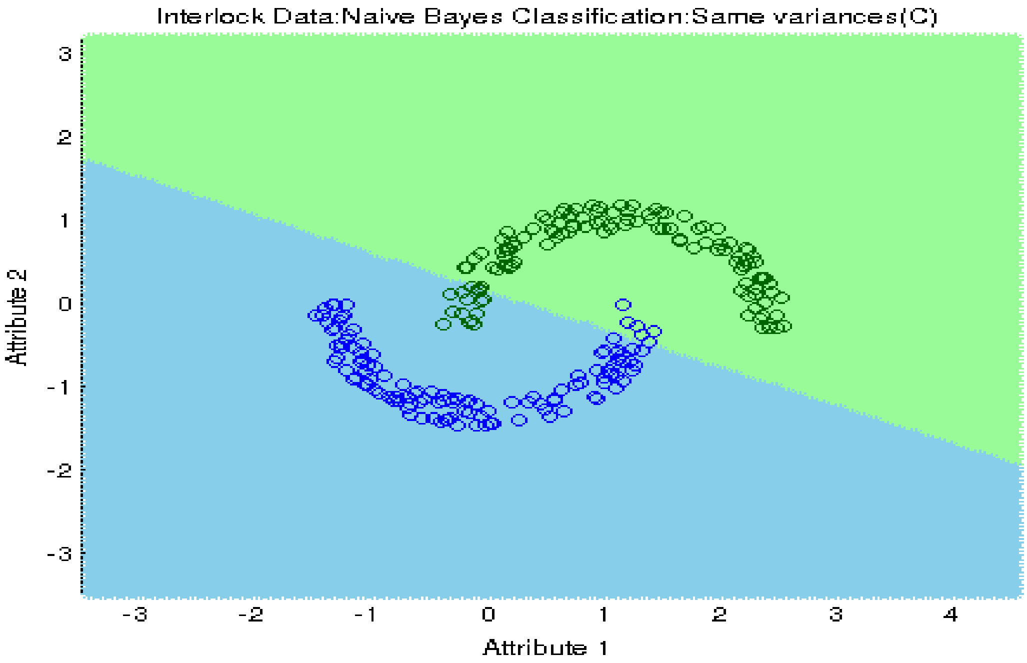

Linearly Separable Data:Naive Bayes Classification:Same variances(I)
15
10
5
0
-5
-10
-15
-20
Attribute 2
-10
-5
0
5
10
15
20
25
Attribute 1


Linearly Separable Data:Naive Bayes Classification:Same variances(C)
15
10
5
0
-5
-10
-15
-20
Attribute 2
-10
-5
0
5
10
15
20
25
Attribute 1



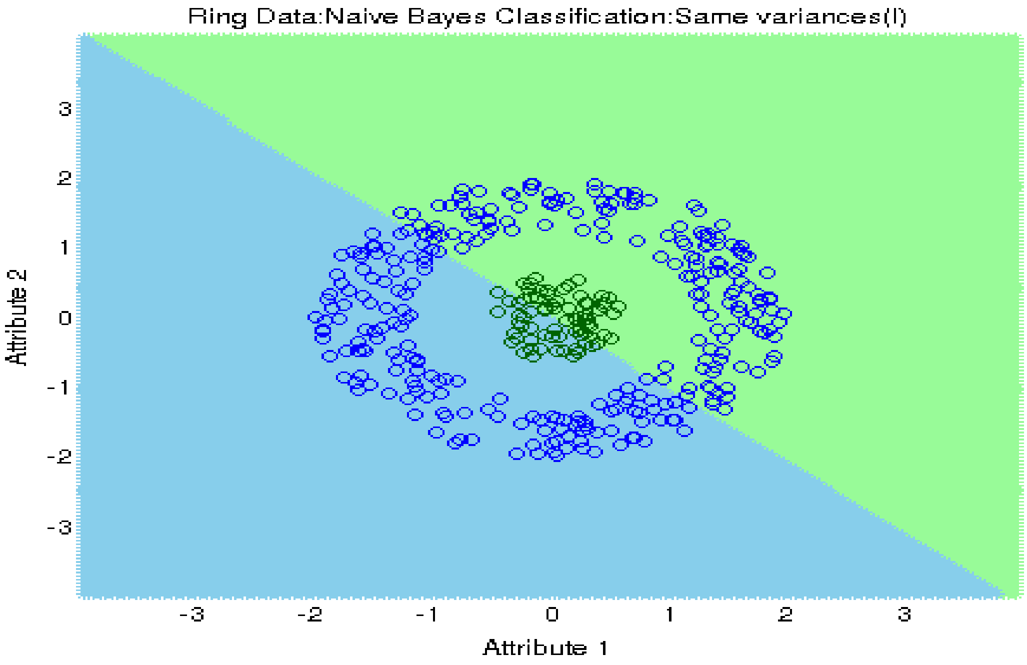
3.2.2 Non-Linearly separable data set

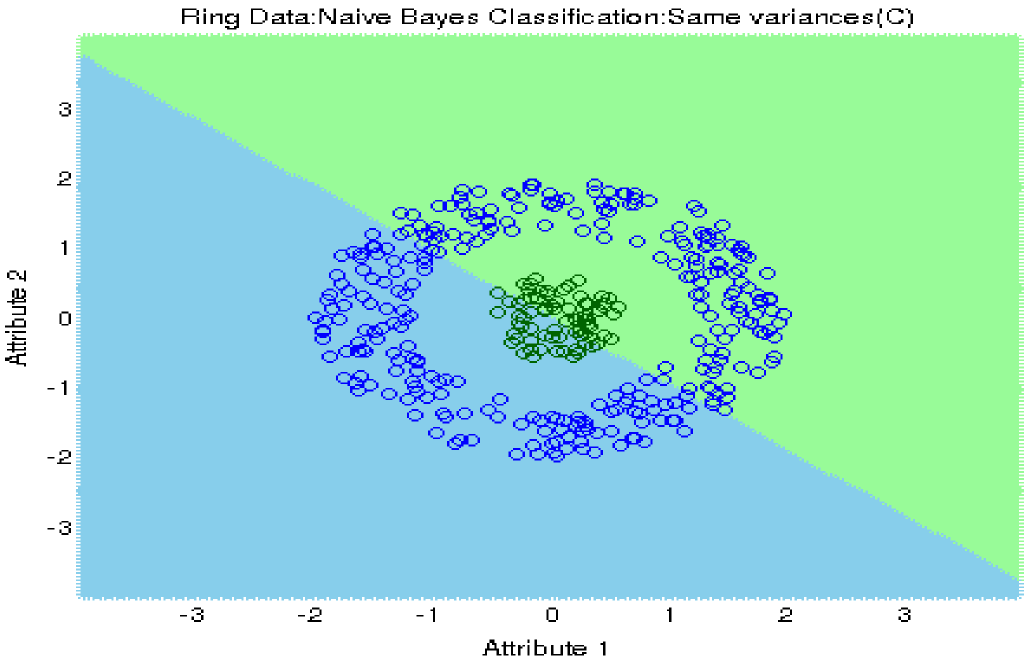


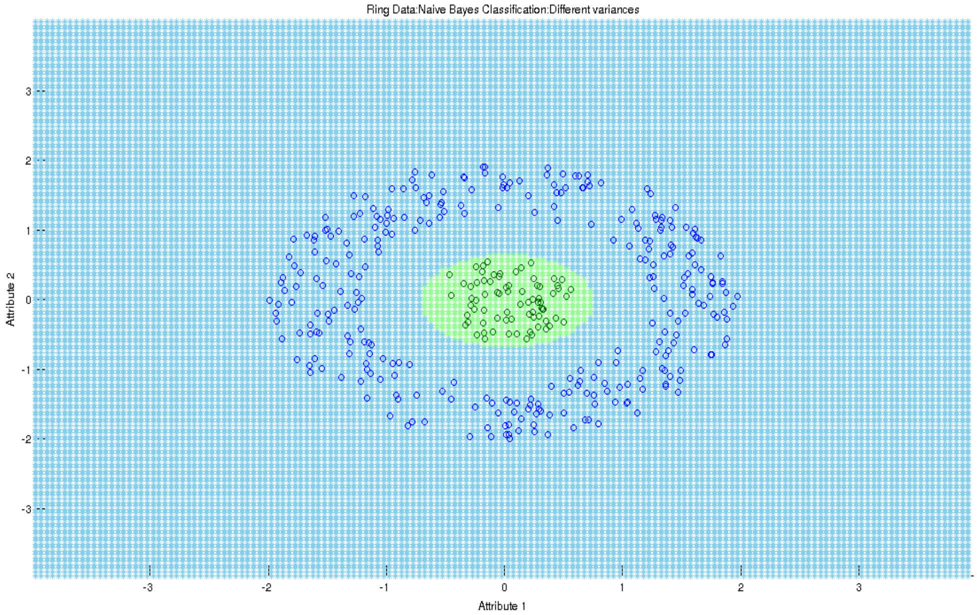


plots/naivebayes/nls/interlock/diff_var.png

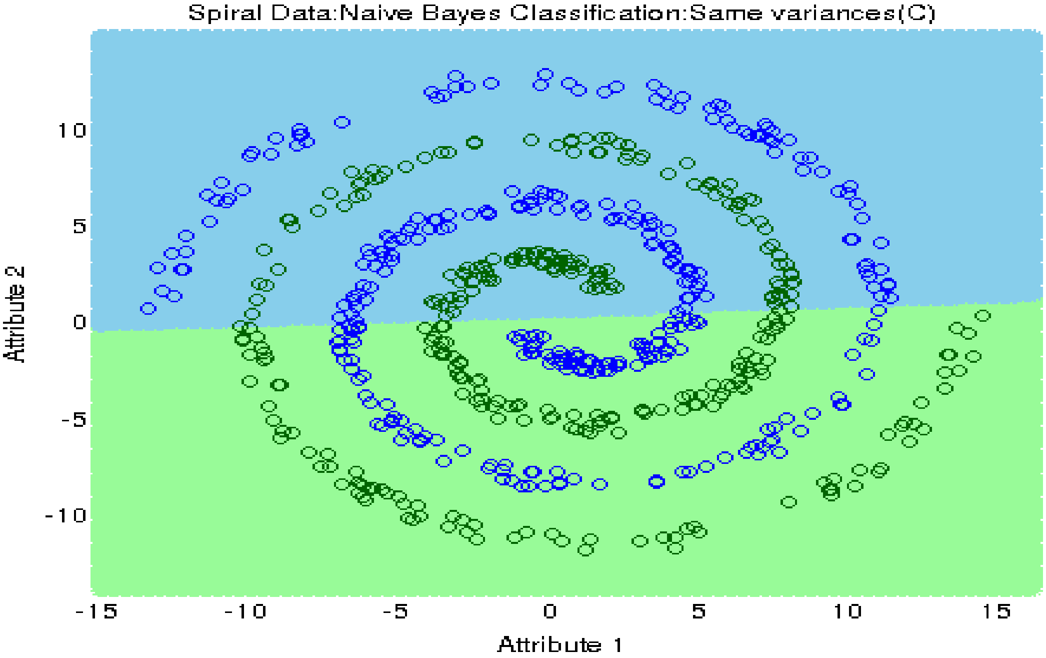
3.2.2.1 Data of Interlocking Classes

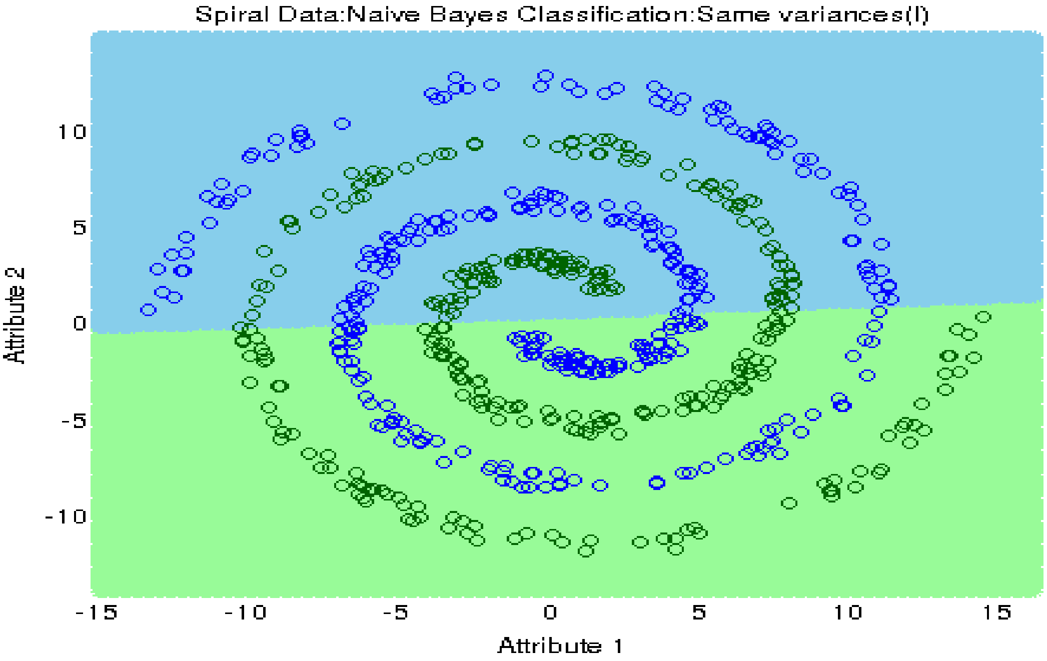

Ring Data:Naive Bayes Classification:Same variances(I)
Attribute 2
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Attribute 1

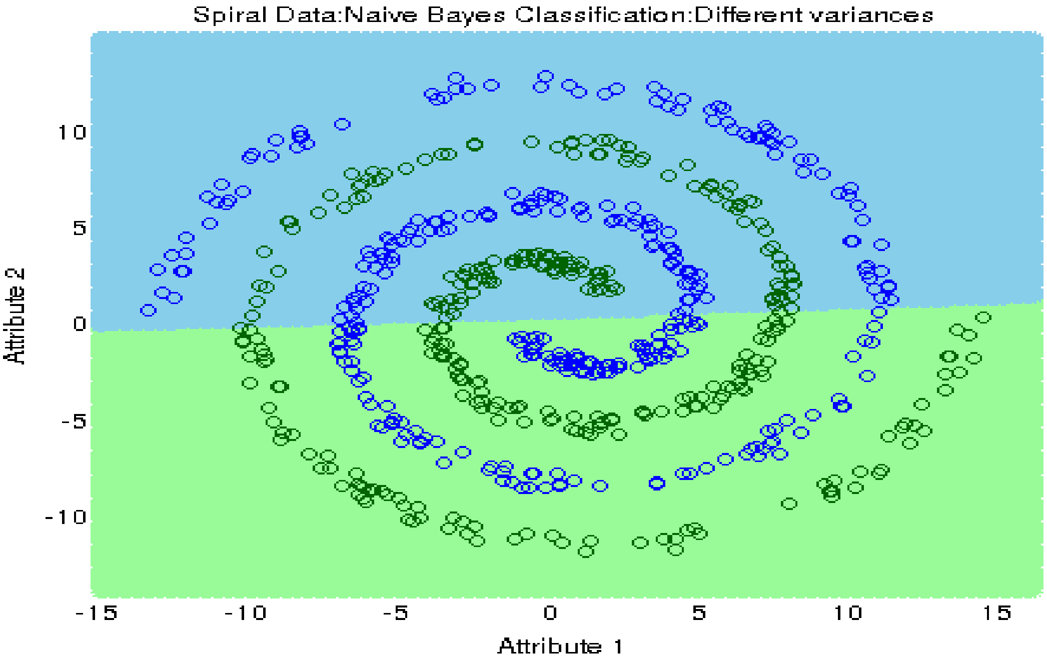

Ring Data:Naive Bayes Classification:Same variances(C)
Attribute 2
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Attribute 1



3.2.2.2 A ring with a central mass

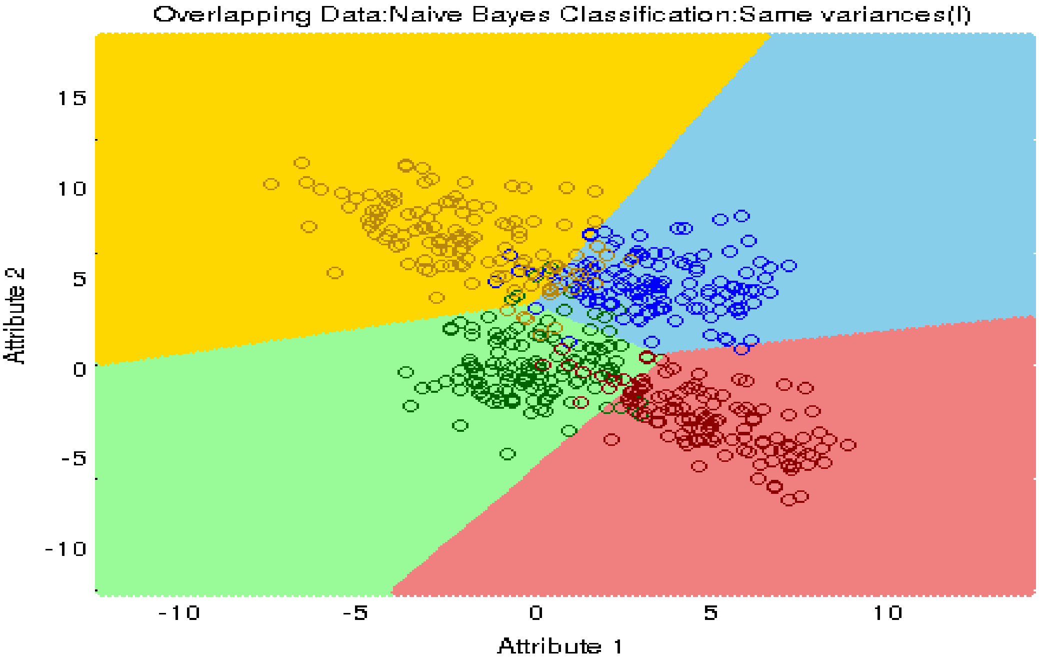


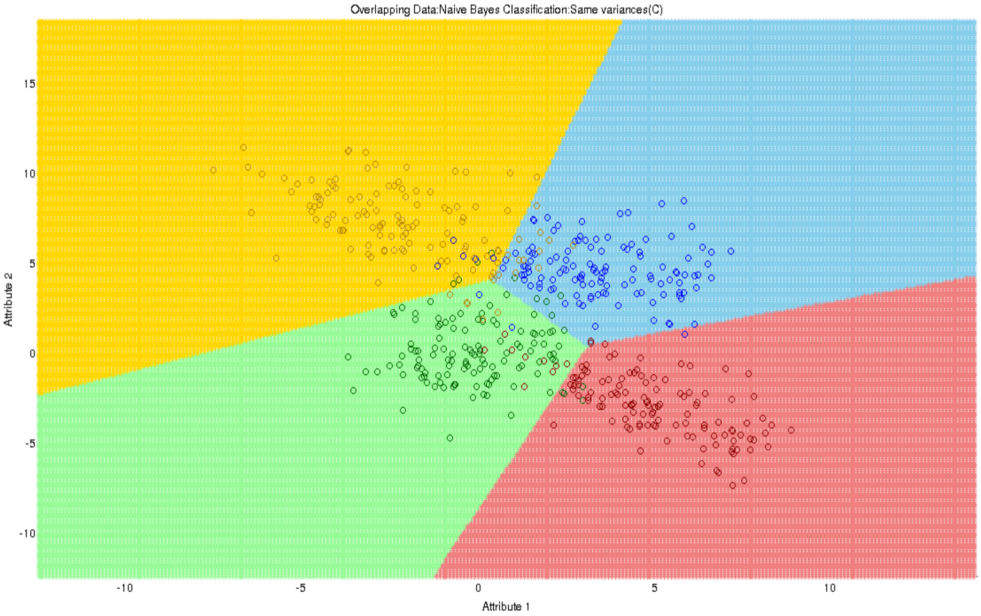


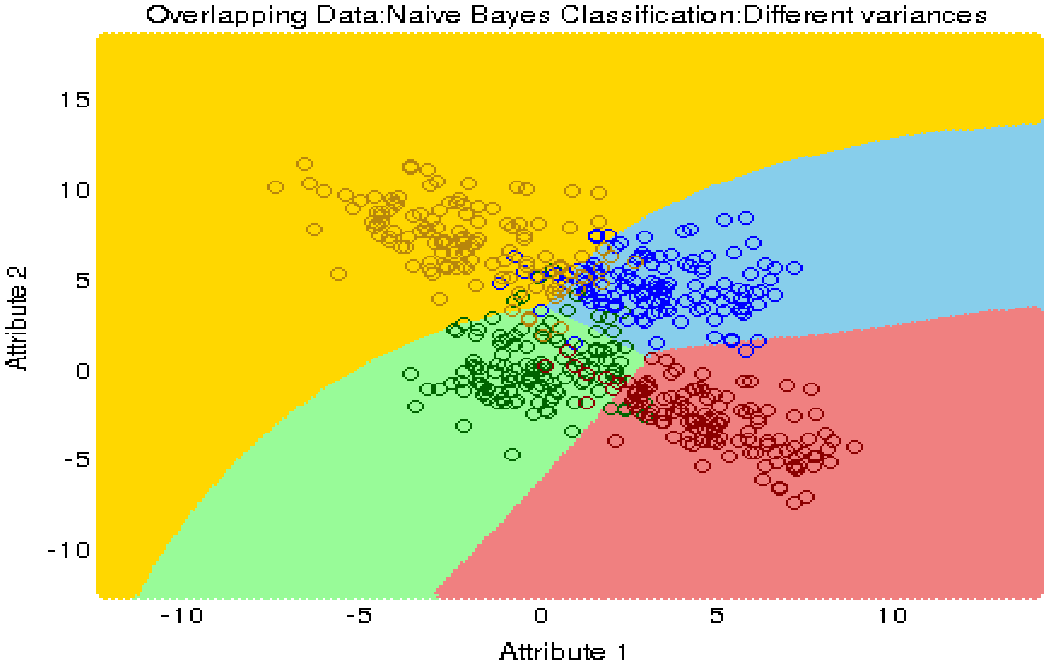


3.2.2.3 Spiral Dataset

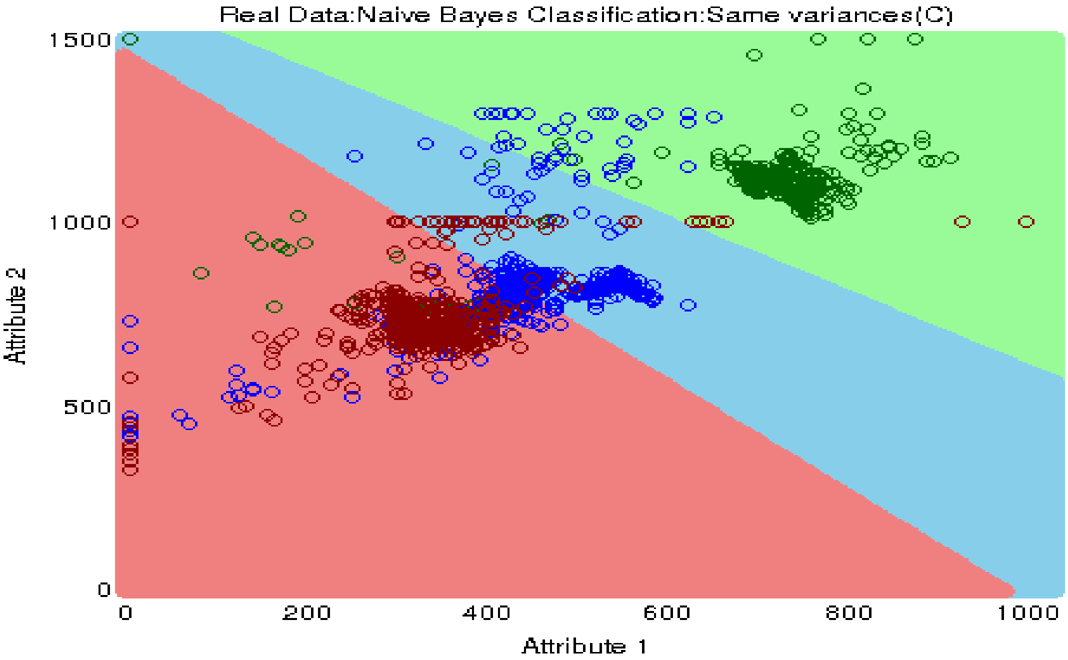
3.2.3 Overlapping data set

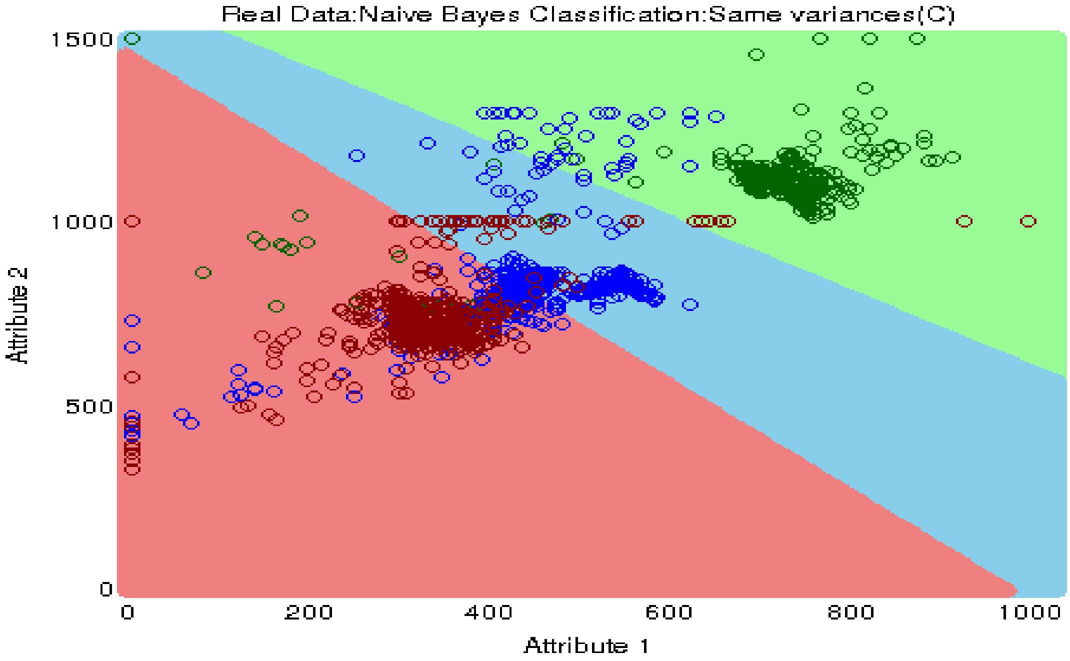


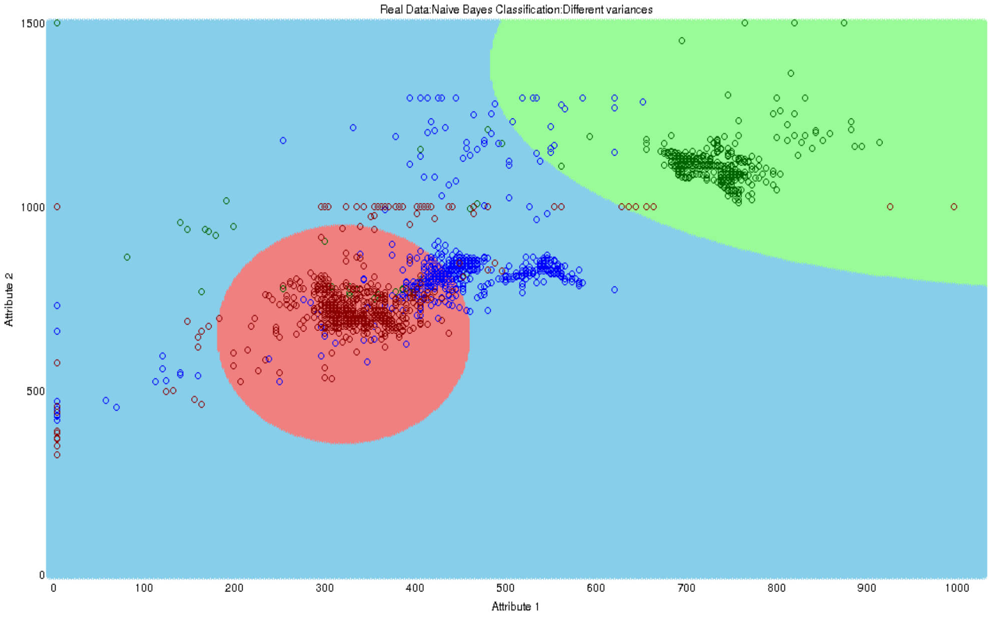




3.2.4 Real world data set







4 Conclusion

As per the observations, we can make the following conclusions :

1. The Decision Boundaries are more accurate in the case of different covariance for different classes as compared to the other cases.
2. The curvature of the decision boundaries is due to the covariance term in the likelihood probabilty which makes the surface quadratic.
3. The Decision Boundaries are better in cases where data is not overlapping and is separable either linearly or non linearly.
4. In case of real data, the data is more overlapping and non linear, resulting in lesser accuracy of the testing data.

```
> data=read.table("hw2_chol.txt")
> hist(data$V1,xlab='Cholesterol (mg/dL)',main='Histogram of Total Cholesterol')
> boxplot(data$V1,main='Total Cholesterol',ylab='Cholesterol (mg/dL)')
```