Pattern Recognition Assignment 2

Group 12 Athul Vijayan (ED11B004) & KIRAN KUMAR.G.R (AM14D405)

1. **Theory**

   $\mathbf{x} = [x_1, x_2, \cdots, x_d] \in \mathbb{R}^d$ is the *feature vector* in a vector space called *feature space* containing d continuous features in it.

   $\boldsymbol{\omega} = [\omega_1, \omega_2, \cdots, \omega_c]$ be the c finite *state of nature* / catogories.

   We follow from bayes theorem that

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}$$

   In bayesian classification we decide to take action α_i for which Conditional risk $R(\alpha_i|\mathbf{x})$ is minimum. That is, maximum discriminant will correspond to mimimum conditional risk. And for minimum error rate classifier, we can define $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$. Maximum discriminant function corresponds to the maximum posterior probability. Also if $f(.)$ is a monotonously increasing function, then $f(g_i(\mathbf{x}))$ will also give same classifier.

$$\begin{aligned} g_i(x) &= P(\omega_i|\mathbf{x}) \\ &= P(\mathbf{x}|\omega_i)P(\omega_i) \\ &= ln(P(\mathbf{x}|\omega_i)) + ln(P(\omega_i)) \qquad | \quad ln(.) \text{ is monotonously increasing} \end{aligned}$$

   For a dataset with c classes denoted as $\omega_1, \omega_2, \cdots, \omega_c$, we assume the likelihood probability in each class ω_i is distributed as Multivariate Gaussian. i.e. $P(\mathbf{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
   Our General Bayes classifier becomes:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^T - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

   Now depending on the covariance matrix $\boldsymbol{\Sigma}_i$, various cases can be generated.
   One Interesting case is that when $\boldsymbol{\Sigma}_i$ is diagonal, the covariance/ correlation between any two features is zero. That implies that we do not get any information about feature x_i from features x_j if $j \neq i$. In terms of probability, $P(x_i|\omega_k, \{x_j \quad \forall \quad j \neq i\}) = P(x_i|\omega_k)$. This is called **naive bayes** classifier.

   i. **Bayesian Classifier with Covariance same for all classes. $\boldsymbol{\Sigma_i} = \boldsymbol{\Sigma}$.**

      Since $\boldsymbol{\Sigma_i} = \boldsymbol{\Sigma}$, we can drop second and third terms because they are independent of i and will be same for every class. making it

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^T + \ln P(\omega_i) \\ &= \left(\boldsymbol{\Sigma^{-1}\mu}_i\right)^T \mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_i} + \ln P(\omega_i) \end{aligned}$$

      We can see this is a linear classifier. If we assume every class conditional probability is distributed as Multivariate Gaussian with same Covariance $\boldsymbol{\Sigma}$ like this case, the resulting

section 1 continues on the next page...

classifier is called linear discriminant.
Now we need to get estimated $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$. Clearly if we calculate the covariance of measured data in each class, they will not be similar across classes. So we find estimate for parameters by maximizing likelihood.

let sampes in class ω_i be denoted as $\mathcal{D}_i$ such that the total samples can be expressed as $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_c\}$. Let number of samples in $\mathcal{D}_i$ be N_i and number of total samples, i. e number of points in $\mathcal{D}$ is N.

In the generative learning approach, we assume each of $\mathcal{D}_i$ is distributed as gaussian with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$. Note that in LDA, covariance of every $\mathcal{D}_i$ is same.

Log Likelihood function is given by:

$$l(\boldsymbol{\theta}) = \ln P(\mathcal{D}|\boldsymbol{\theta}))$$

and $\boldsymbol{\theta}$ that maximizes $l(\boldsymbol{\theta})$ is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad l(\boldsymbol{\theta})$$

With gaussian aproximation,

$$\begin{aligned}
P(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} P(\mathbf{x}_i|\omega) \\
&= \prod_{k=1}^{c} \prod_{\mathbf{x}_i \in \mathcal{D}_k} P(\mathbf{x}_i|\omega_k) \\
\Rightarrow l(\boldsymbol{\theta}) &= \sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} \ln P(\mathbf{x}_i|\omega_k) \\
&= \sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \frac{1}{2}\ln|\boldsymbol{\Sigma}|
\end{aligned}$$

Now to estimate parameters, we maximize $l(\boldsymbol{\theta})$.

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0}$$

Now to find $\boldsymbol{\mu}_j$ of gaussian pdf of class $P(\mathbf{x}|\omega_j)$,

$$\nabla_{\boldsymbol{\mu}_j} l(\boldsymbol{\theta}) = 0$$

$$\nabla_{\boldsymbol{\mu}_j} \left(\sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \right) = 0$$

For $k \neq j$, we can see all the terms goes to zero. we can denote this estimated mean for class j as $\hat{\boldsymbol{\mu}}_j$.

$$\boxed{\hat{\boldsymbol{\mu}}_j = \frac{1}{N_j} \sum_{\mathbf{x}_i \in \mathcal{D}_j} \mathbf{x}_i} \tag{1}$$

section 1 continues on the next page...

And for $\boldsymbol{\Sigma}$,

$$\begin{aligned}
\nabla_{\boldsymbol{\Sigma}} l(\boldsymbol{\theta}) &= 0 \\
\nabla_{\boldsymbol{\Sigma}} \left(\sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \right) &= 0 \\
\nabla_{\boldsymbol{\Sigma}} \left(\sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2} tr\left((\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T\right) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \right) &= 0 \\
\nabla_{\boldsymbol{\Sigma}} \left(\sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)\right) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \right) &= 0 \\
\nabla_{\boldsymbol{\Sigma}} \left(\sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)\right) \right) - \frac{N}{2}\boldsymbol{\Sigma}^{-1} &= 0
\end{aligned}$$

Above, we have used properties,

- tr(Real number) = Real Number
- $tr(ABC) = tr(BCA)$
- $\nabla_A |A| = A^{-1}$

That gives us,

$$\boxed{\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)} \tag{2}$$

is the best estimate of $\boldsymbol{\Sigma}$ in the maximum likelihood sense. The parameter is biased.

ii. **Bayesian Classifier with Covariance different for all classes** It is straight forward, we can neglect only the constant term in the general discriminant function.

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^T - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

This is a Quadratic Discriminant function. We assume each for class k conditional probabilities belong to a gaussian distribution with $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.
To estimate this parameters using maximum likelihood estimation.

$$\begin{aligned}
P(\mathcal{D}_k|\boldsymbol{\theta}_k) &= \prod_{i=1}^{N_k} P(\mathbf{x}_i|\omega_k) \\
\Rightarrow l(\boldsymbol{\theta}) &= \sum_{\mathbf{x}_i \in \mathcal{D}_k} \ln P(\mathbf{x}_i|\omega_k) \\
&= \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \ln|\boldsymbol{\Sigma}_k|
\end{aligned}$$

Now to estimate parameters, we maximize $l(\boldsymbol{\theta})$.

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0}$$

section 1 continues on the next page...

Now to find $\boldsymbol{\mu}_k$ of gaussian pdf of class $P(\mathbf{x}|\omega_k)$,

$$\nabla_{\boldsymbol{\mu}_k} l(\boldsymbol{\theta}) = 0$$

$$\nabla_{\boldsymbol{\mu}_k} \left(\sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \right) = 0$$

That gives the sample mean of the class as the optimal parameter.

$$\boxed{\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in \mathcal{D}_k} \mathbf{x}_i}$$

Now for Covariance estimation,

$$\nabla_{\boldsymbol{\Sigma}_k} l(\boldsymbol{\theta}) = 0$$

$$\nabla_{\boldsymbol{\Sigma}_k} \left(\sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \right) = 0$$

Using properties used above, we find optimal estimate is the biased covariance of class k.

$$\boxed{\hat{\boldsymbol{\Sigma}}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in \mathcal{D}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)} \tag{3}$$

iii. **Naive Bayes**

With naive bayes, we have $\boldsymbol{\Sigma}_{ij} = 0 \quad \forall \quad i \neq j$ diagonal. Intiuitively, it means that we do not get any information about a feature if we know the value of any other feature vector. We can write our joint class conditional probability as.

$$\begin{aligned} P(\mathbf{x}|\omega_k) &= P(x_1, x_2, \cdots, x_d|\omega_k) \qquad x_i \text{ are individual features} \\ &= P(x_1|\omega_k)P(x_2, x_3, \cdots, x_d|\omega_k, x_1) \\ &= P(x_1|\omega_k)P(x_2|\omega_k, x_1)P(x_3, \cdots, x_d|\omega_k, x_1, x_2) \\ &\vdots \\ &= P(x_1|\omega_k)P(x_2|\omega_k, x_1)\cdots P(x_d|\omega_k, x_1, x_2, \cdots, x_{d-1}) \end{aligned}$$

Now naive payes assumes that, features are independent that is,

$$P(x_i|\omega_k, \{x_j \quad \forall \quad j \neq i\}) = P(x_i|\omega_k)$$

That gives us,

$$\begin{aligned} P(\mathbf{x}|\omega_k) &= P(x_1|\omega_k)P(x_2|\omega_k)\cdots P(x_d|\omega_k) \\ &= \prod_{i=1}^{d} P(x_i|\omega_k) \end{aligned}$$

section 1 continues on the next page...

Now we assumed already that $P(\mathbf{x}|\omega_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Now with independence condition, we can say $P(x_i|\omega_k) \sim \mathcal{N}(\mu_i, \sigma_i^2)$ where $\mu_i = \boldsymbol{\mu}_i$ and $\sigma_i^2 = \boldsymbol{\Sigma}_{ii}$. This means each feature itself is distributed as univariate gaussian with mean and variance corresponding to that feature alone. We will find the optimum estimate of these parameters now:

- **Naive Bayes with $\boldsymbol{C = \sigma^2 \mathbf{I}}$**
  Like in bayesian case, we assume a gaussian distribution with different covariance for every class.
  Let covariance of class ω_k be in the form of $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$. Here again, we find the optimum estimation for scalar σ_k^2 using maximizing likelihood function.
  Let $C = \sigma_k^2$

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{\mathbf{x}_i \in \mathcal{D}_k} \ln P(\mathbf{x}_i|\omega_k) \\ &= \sum_{\mathbf{x}_i \in \mathcal{D}_k} -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)\frac{1}{C}\boldsymbol{I}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \frac{1}{2}\ln C \end{aligned}$$

  Now differentiating and equating to zero, we get estimate of $\boldsymbol{\mu}$ as

$$\boxed{\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k}\sum_{\mathbf{x}_i \in \mathcal{D}_k} \mathbf{x}_i}$$

  and optimal value of σ_k^2 is:

$$\boxed{\hat{\sigma}_k^2 = \frac{1}{N_k}\sum_{\mathbf{x}_i \in \mathcal{D}_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T} \tag{4}$$

  Note that it is a scalar for every class.
- **Naive Bayes with Different covariance for all classes**
  Like in bayesian case, we assume a gaussian distribution with different covariance for every class.
  Now to estimate optimum parameters for the model, we use likelihood maximizing as before.

$$l(\boldsymbol{\theta}_k) = \sum_{\mathbf{x}_i \in \mathcal{D}_k} \sum_{i=1}^{d} \ln P(x_i|\omega_k)$$

  Now, $\frac{\partial l(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\mu}_k} = 0$ gives us,

$$\boxed{\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k}\sum_{\mathbf{x}_i \in \mathcal{D}_k} \mathbf{x}_i}$$

  Now, $\frac{\partial l(\boldsymbol{\theta}_k)}{\partial \sigma_d} = 0$ will give us the optimum variance σ_d^2 of the univariate gaussian distribution where feature x_d belongs.

$$\sigma_d^2 = \frac{1}{N_k}\sum_{\mathbf{x}_i \in \mathcal{D}_k} ((x_d)_i - \mu_d)^2$$

section 1 continues on the next page...

So best $\mathbf{\Sigma}$ is:

$$\boxed{(\hat{\mathbf{\Sigma}}_k)_{lm} = \begin{cases} \frac{1}{N_k}\sum_{\mathbf{x}_i \in \mathcal{D}_k}((x_l)_i - \mu_l)^2 & \text{if } l = m \\ 0 & \text{otherwise} \end{cases}}$$

So the best estimate has variance of each feature along its diagonals and zero everywhere else.

- **Naive Bayes with Same covariance for all classes**
  Like in bayesian case, we maximize the overall likelihood function.

$$\begin{aligned} P(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} P(\mathbf{x}_i|\omega) \\ &= \prod_{k=1}^{c} \prod_{\mathbf{x}_i \in \mathcal{D}_k} \prod_{m=1}^{d} P((x_m)_i|\omega_k) \\ \Rightarrow l(\boldsymbol{\theta}) &= \sum_{k=1}^{c} \sum_{\mathbf{x}_i \in \mathcal{D}_k} \sum_{m=1}^{d} \ln P((x_m)_i|\omega_k) \end{aligned}$$

Maximizing this by differentiating gives us:

$$\boxed{\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in \mathcal{D}_k} \mathbf{x}_i}$$

and optimum $\mathbf{\Sigma}$ is:

$$\boxed{(\hat{\mathbf{\Sigma}})_{lm} = \begin{cases} \frac{1}{N}\sum_{k=1}^{c}\sum_{\mathbf{x}_i \in \mathcal{D}_k}((x_l)_i - \mu_l)^2 & \text{if } l = m \\ 0 & \text{otherwise} \end{cases}}$$

So the best estimate has cumulative variance of each feature along its diagonals and zero everywhere else.

iv. **Performance evaluation methods**

(a) **Confusion matrix and Performance metrics**
Confusion matrix or an error matrix is a specific table layout that allows visualization of the performance of an algorithm. The structure of the matrix is as follows,

- Rows correspond to classes in the test set.
- Columns correspond to classes in the classification result.
- The diagonal elements in the matrix represent the number of correctly classified data points of each class, i.e. the number of data points with a certain class name that actually obtained the same class name during classification.
- The off-diagonal elements represent misclassified data points or the classification errors, i.e. the number of data points that ended up in another class during classification.

section 1 continues on the next page...

- Off-diagonal row elements represent data points of a certain class which were excluded from the respective class during classification. Such errors are known as errors of omission or exclusion. It is the ratio sum of off diagonal row elements to the sum total off the row.
- Off-diagonal column elements represent data points of other classes that were included in a certain classification class. Such errors are known as errors of commission or inclusion. It is the ratio sum of off diagonal column elements to the sum total off the column.

Accuracy is the fraction of correctly classified data points with respect to all data points of the given class. It is ratio of sum of the diagonal elements to the grand total of the matrix elements.

Precision is the proportion of the predicted positive cases that were correct. It is the ratio of the corresponding diagonal element of the class to the sum the off-diagonal column elements.

(b) **Receiver Operating Characteristics (ROC)**
A receiver operating characteristics (ROC) curve is a technique used for visualizing performance of classifiers. ROC graphs are two-dimensional graphs in which True Positive rate is plotted along the Y axis and False Positive rate is plotted on the X axis. An ROC graph depicts relative trade-offs between true positives and false positives. Interpretations from ROC curves are,

- It shows the tradeoff between sensitivity (True positive rate) and class specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the classification.

(c) **Detection error tradeoff (DET)** In the DET curve, plot consists of error rates on both axes, giving uniform treatment to both types of error, and use a scale for both axes which spreads out the plot. It better distinguishes different well performing systems. The DET curves are approximately straight lines, corresponding to normal likelihood distributions. The curves are limited to the lower left quadrant if the performance is good. The closer the curve comes to the x=-y diagonal of the DET plot, more random is the performance.

2. Implementation And results

i. The linearly sepearble data with three classes.

First step is to visualize the data and predict the accuracy of our algorithms intuitively. This will help us identify outliers and bad samples in data and plan ahead accordingly. Figure 1 shows scatter plot.

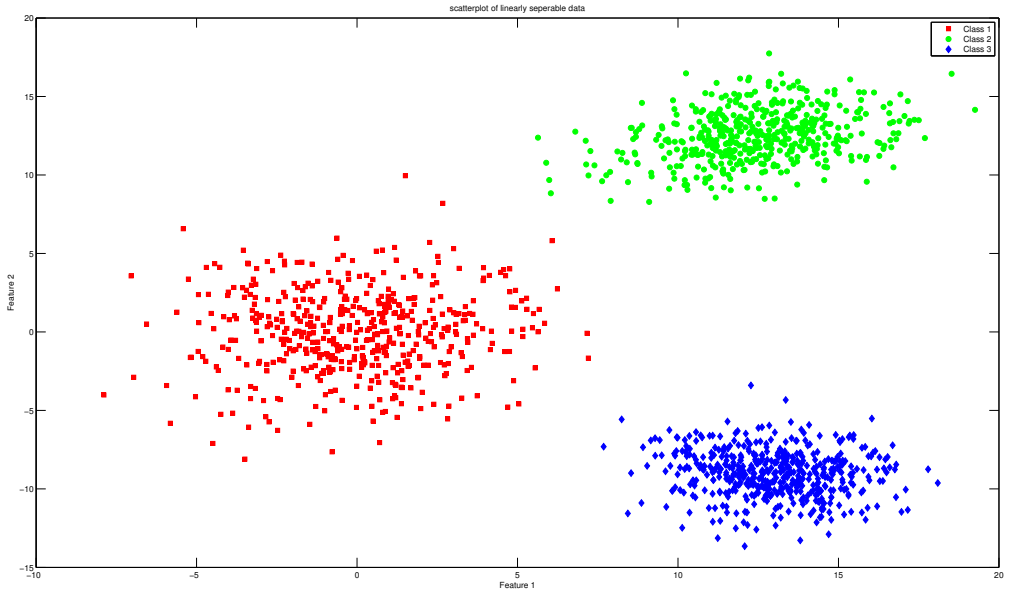


Figure 1: Scatter plot of Linearly sepearble data

As it is evident from visual inspection, the data looks clean in the sense that it does not have much outliers or missing feature etc.. So we will apply our algorithms to this data with different choices of covariance. As the each classes are well seperated, we can expect a good accuracy for our model.

(a) **Case 1: Bayes with Covariance same for all classes**
We have used the estimates for optimum mean and covariance as we derived earlier. We constructed the gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 2 shows the plot.
We can visualize the decision boundary better by examining the contour plot and the Discriminant boundary. The test data is plotted, the contours of each class conditional probabilities are also plotted. The plot is Figure 3.
As we can see, The shape of all the gaussians are similar. This is caused by the same covariance of each classes. As the gaussians are same in shape, only their mean difffer. So it gives us a linear discriminant function of which boundaries are shown.
The decision boundary passes through the midpoint of line segment joining the means of any pair of classes. This is because of equal prior probability.

(b) **Case 2: Bayes with Covariance different for all classes**
Gaussian pdf of class conditional probability is plotted and the three classes and

section 2 continues on the next page...

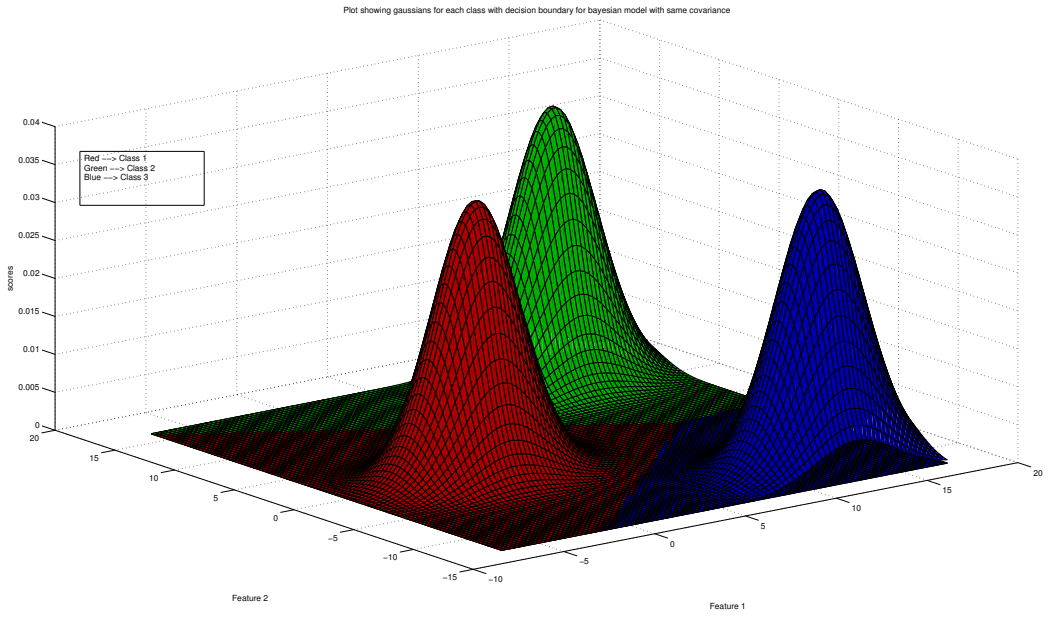


Figure 2: Gaussian pdf of posterior probability showing decision boundary of linearly sepearble data with bayesian model with same covariance

decision boundary is seperated by different color. Figure 4 shows the plot.
Now the decision boundary can be seen better in a contour plot.The plot is Figure 5.

As we can see, The shape of all the gaussians are dissimilar as the covariance is different for each class. So it gives us a non linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

(c) **Case 3: Naive Bayes with $\Sigma_k = \sigma_k^2 \boldsymbol{I}$**
We expect a symmetric gaussian. Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 6 shows the plot.
We expect concentric circles in the contour plot. The plot is Figure 7. As we can see, The shape of all the gaussians are symmetrical / circular contours.The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

(d) **Case 4: Naive Bayes with same covariance for all classes**
As the covariance is a diagonal matrix with arbitrary elements, we expect a gaussian whose contours are elliptical. Again, since covariance of each class is same, we expect a linear boundary. Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 8 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric ellipses in the contour plot. The plot is Figure 9. The covariance is same for each

section 2 continues on the next page...

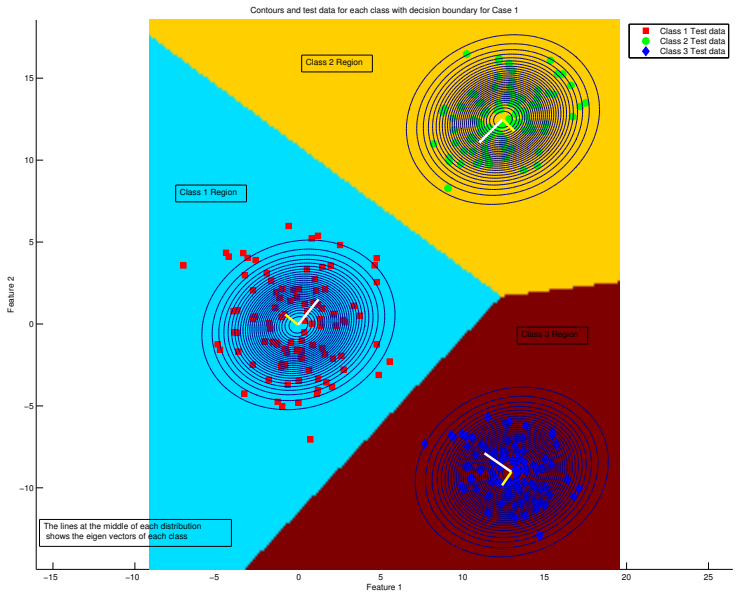


Figure 3: Contour plots and test data points showing decision boundary for linear sepearble data with bayesian model with same covariance

class, So it gives us a linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

(e) **Case 5: Naive Bayes with different covariance for all classes**
Since covariance of each class is different, we expect a non linear boundary. Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 10 shows the plot.
Now the decision boundary can be seen better in a contour plot. The plot is Figure 11. The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

(f) **Performance evaluation** The confusion matrix and all the performance metrics are same for every algorithm we used. It is given as:

		Prediction				
		Class 1	Class 2	Class 3	Total	Incl. Error
Truth	Class 1	150	0	0	150	0
	Class 2	0	150	0	150	0
	Class 3	0	0	150	150	0
	Total	150	150	150	450	0
	Excl. Error	0	0	0	0	

Table 1: Confusion matrix for Linearly seperable data, All Algorithm

	Precision	Accuracy
Class 1	1.00	100 %
Class 2	1.00	
Class 3	1.00	

Table 2: Performance metric

And ROC curve is not necessary in this case since all algorithms give same full accuracy.

section 2 continues on the next page...

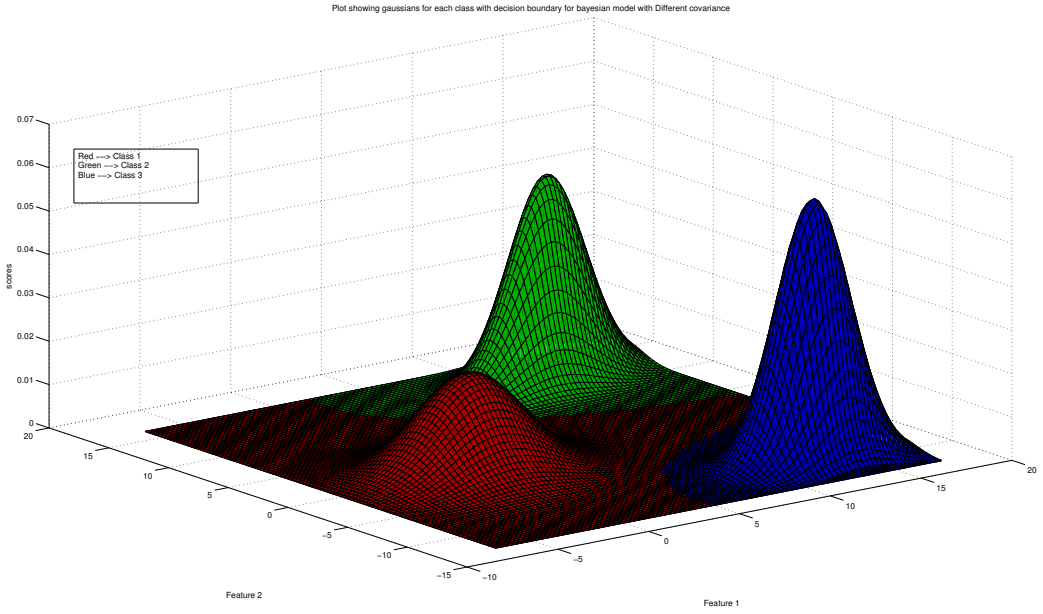


Figure 4: Gaussian pdf of posterior probability showing decision boundary of linearly seperable data for bayesian model with different covariance

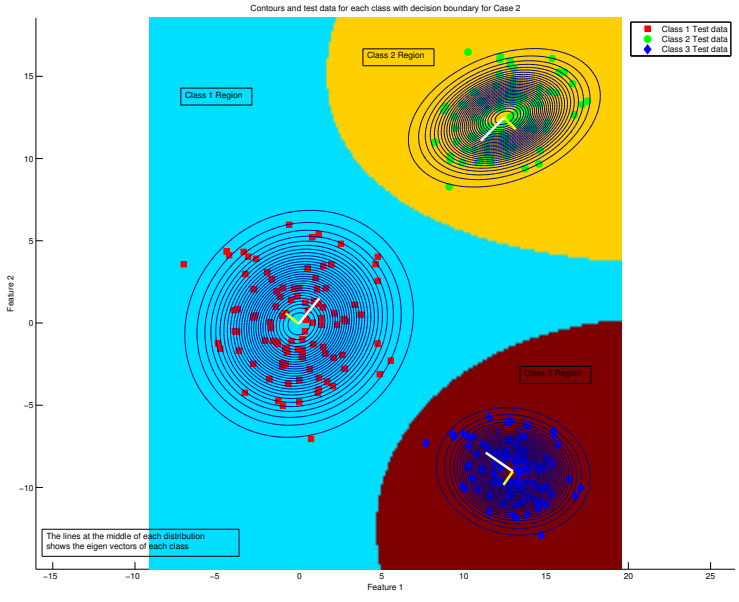


Figure 5: Contour plots and test data points showing decision boundary for linear sepearble data for bayesian model with different covariance

section 2 continues on the next page...

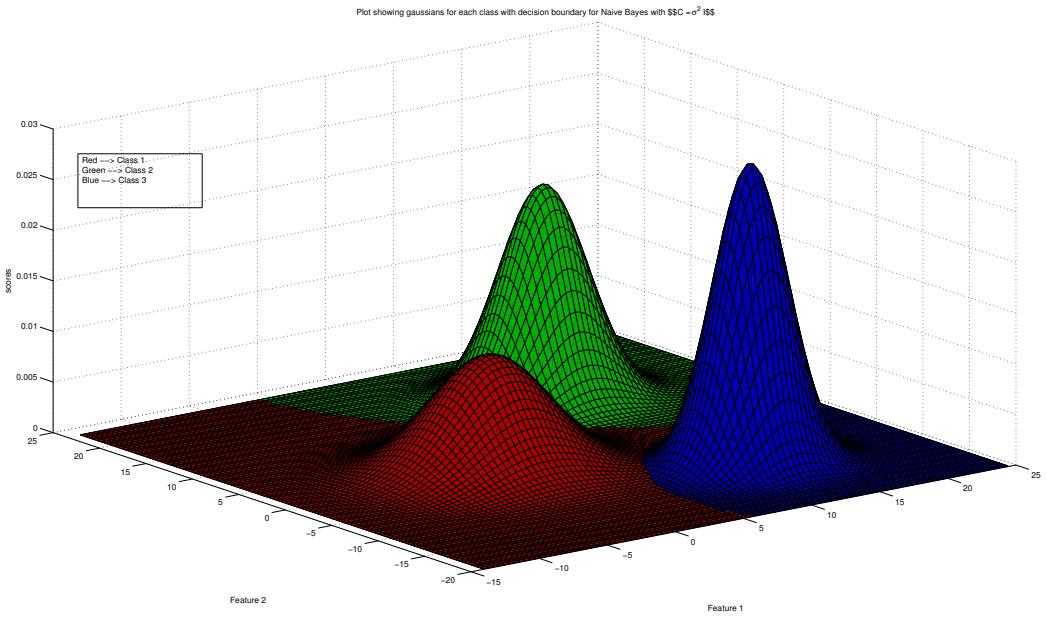


Figure 6: Gaussian pdf of posterior probability showing decision boundary of linearly seperable data for Naive Bayes with $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

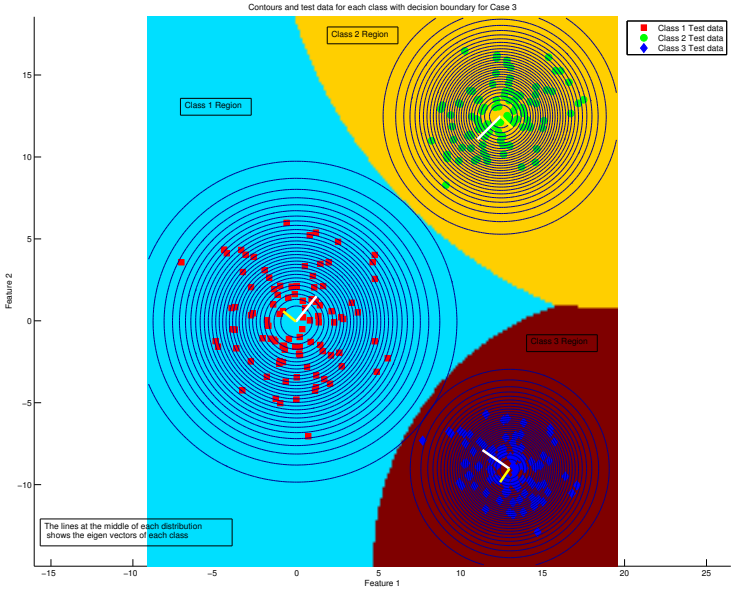


Figure 7: Contour plots and test data points showing decision boundary for linear seperble data for Naive Bayes with $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

section 2 continues on the next page...

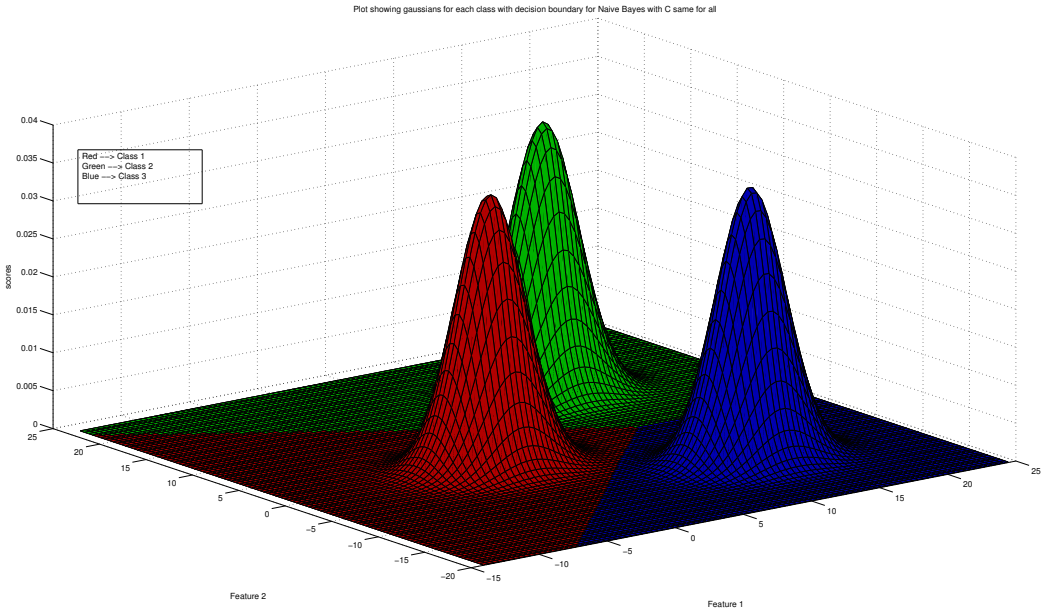


Figure 8: Gaussian pdf of posterior probability showing decision boundary of linearly seperable data for Naive Bayes with same covariance

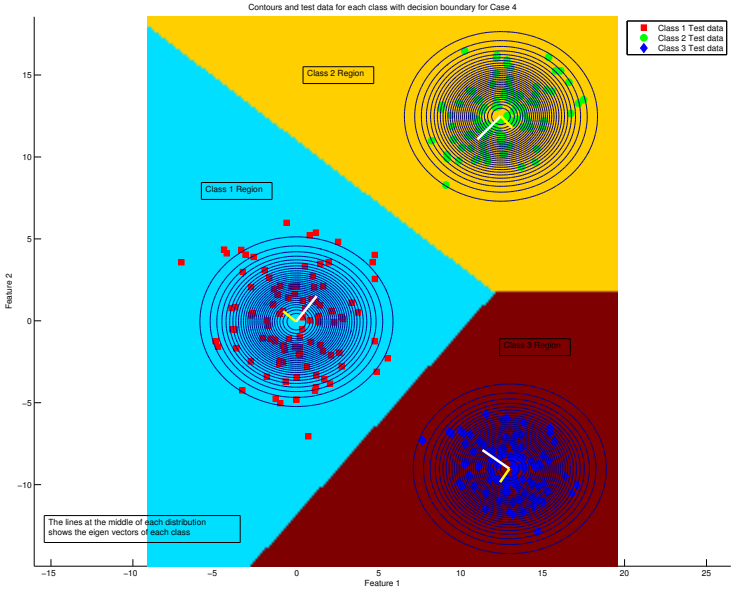


Figure 9: Contour plots and test data points showing decision boundary for linear seperble data for Naive Bayes with different covariance

section 2 continues on the next page...

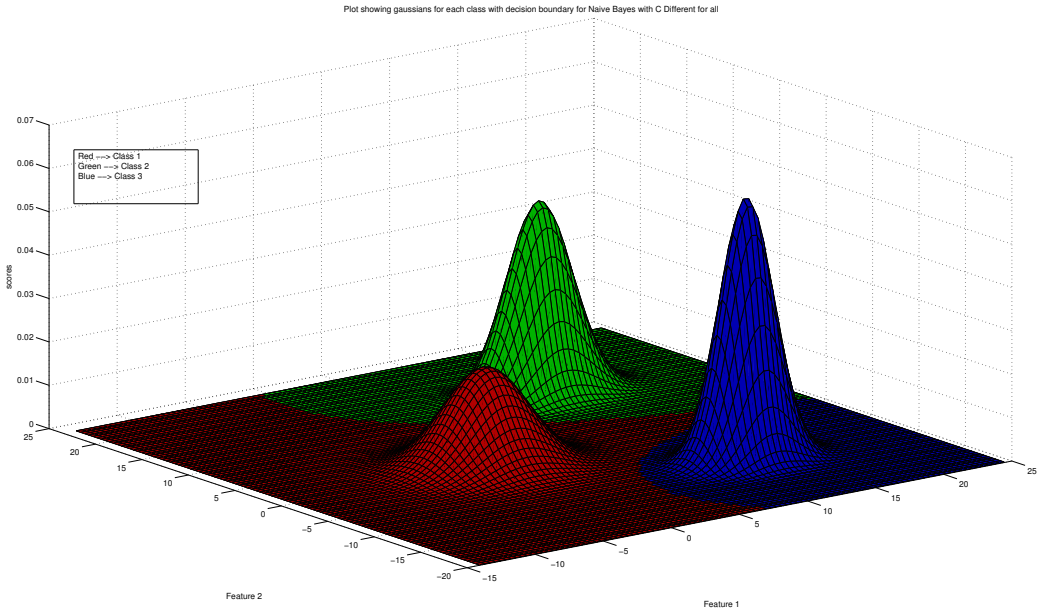


Figure 10: Gaussian pdf of posterior probability showing decision boundary of linearly seperable data for Naive Bayes with different covariance

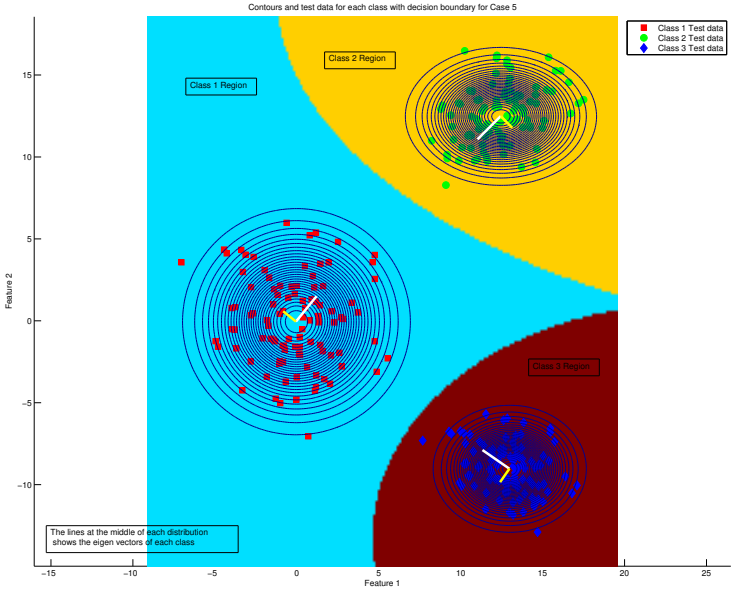


Figure 11: Contour plots and test data points showing decision boundary for linear seperble data for Naive Bayes with different covariance

section 2 continues on the next page...

ii. **Non linearly seperable data with three classes**.
Like before, first we visualize data and draw some conclusions. This will help us identify outliers and bad samples in data and plan ahead accordingly. Figure 12 shows scatter plot.

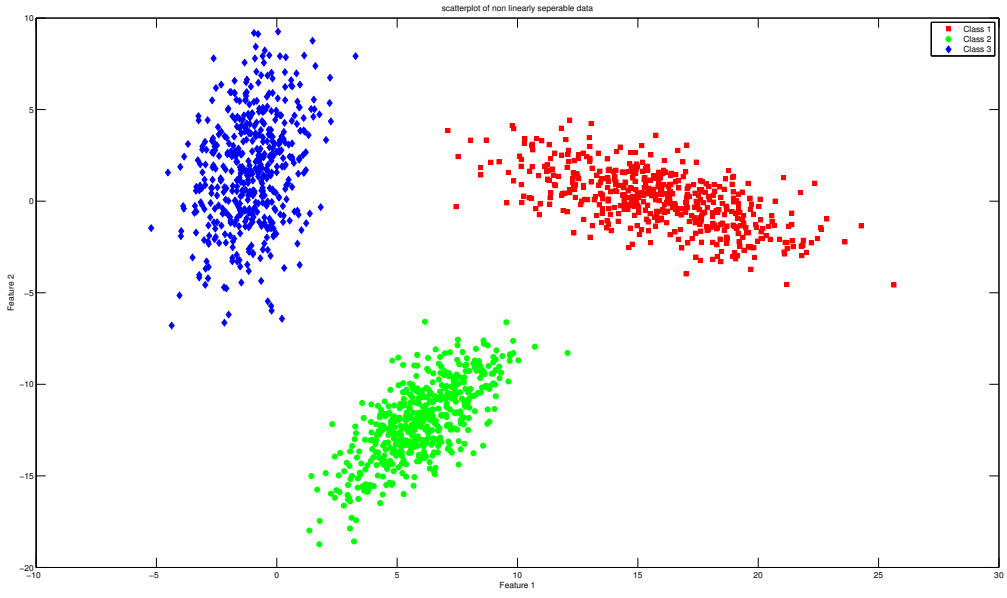


Figure 12: Scatter plot of Linearly sepearble data

Since there are less overlapping data, we can tell that accuracy will be good generally.

(a) **Case 1: Bayes with Covariance same for all classes**
We have used the estimates for optimum mean and covariance as we derived earlier. We constructed the gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 13 shows the plot.

We can visualize the decision boundary better by examining the contour plot and the linear Discriminant boundary. The test data is plotted, the contours of each class conditional probabilities are also plotted. The plot is Figure 14.

It gives us a linear discriminant function of which boundaries are shown.

(b) **Case 2: Bayes with Covariance different for all classes**
gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 15 shows the plot.

Now the decision boundary can be seen better in a contour plot.The plot is Figure 16.

It gives us a non linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

section 2 continues on the next page...

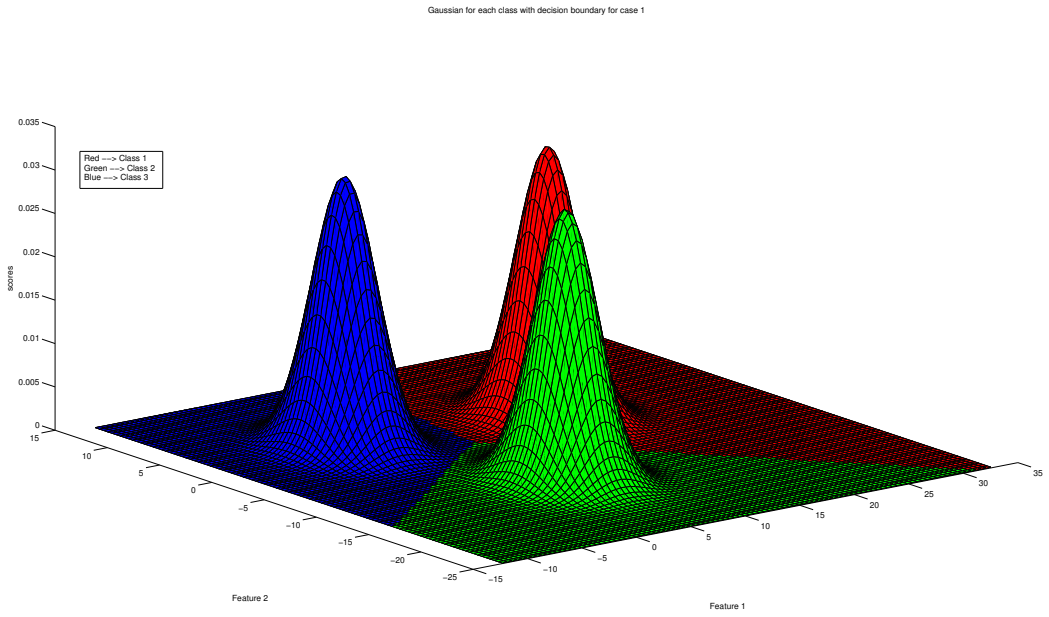


Figure 13: Gaussian pdf of posterior probability showing decision boundary of non linearly sepearble data with bayesian model with same covariance

(c) **Case 3: Naive Bayes with $\mathbf{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 17 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric circles in the contour plot. The plot is Figure 18.
As we can see, The shape of all the gaussians are symmetrical / circular contours.The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

(d) **Case 4: Naive Bayes with same covariance for all classes**
Again, since covariance of each class is same, we expect a linear boundary. Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 19 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric ellipses in the contour plot. The plot is Figure 20.
The covariance is same for each class, So it gives us a linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

(e) **Case 5: Naive Bayes with different covariance for all classes**
we expect a non linear boundary. Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 21 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric ellipses in the contour plot. The plot is Figure 22.

section 2 continues on the next page...

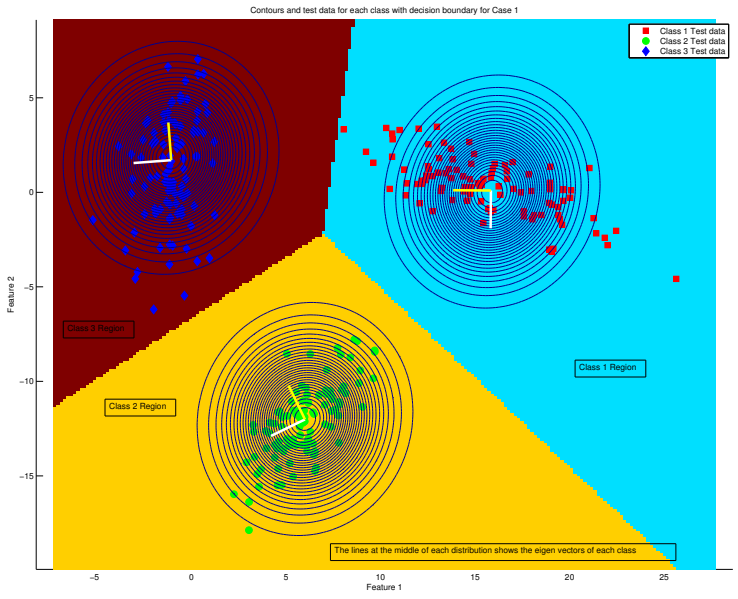


Figure 14: Contour plots and test data points showing decision boundary for non linear sepearble data with bayesian model with same covariance

The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown. As the data set is well seperated, the predictor gives a good accuracy.

(f) **Performance evaluation** The confusion matrix and all the performance metrics are same for every algorithm we used. It is given in Table 3:

		Prediction				
		Class 1	Class 2	Class 3	Total	Incl. Error
Truth	Class 1	150	0	0	150	0
	Class 2	0	150	0	150	0
	Class 3	0	0	150	150	0
	Total	150	150	150	450	0
	Excl. Error	0	0	0	0	

Table 3: Confusion matrix for Non Linearly seperable data, All Algorithm

	Precision	Accuracy
Class 1	1.00	100 %
Class 2	1.00	
Class 3	1.00	

Table 4: Performance metric

And ROC curve is not necessary in this case since all algorithms give same full accuracy.

section 2 continues on the next page...

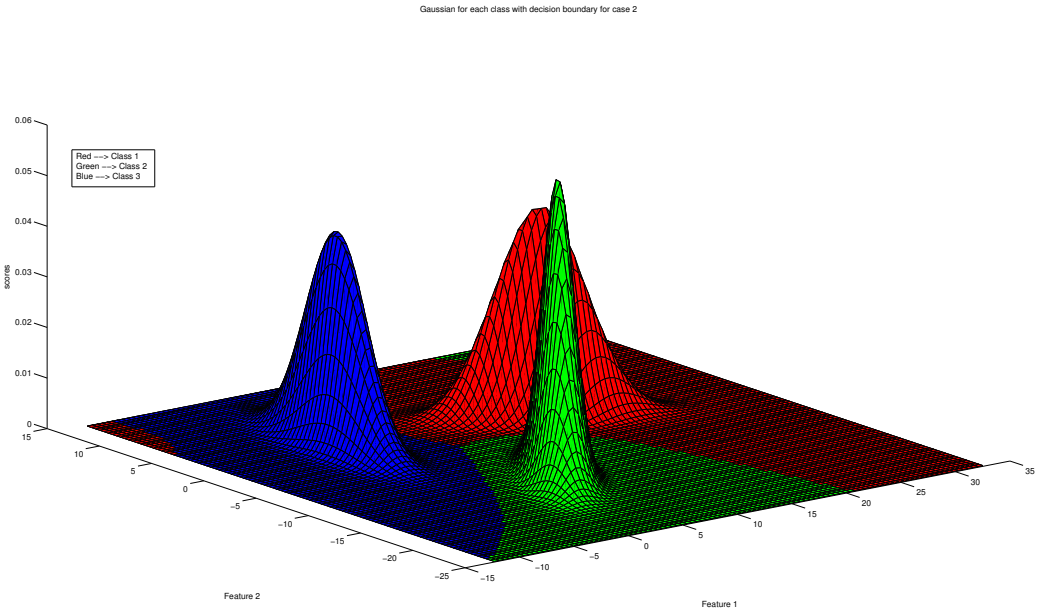


Figure 15: Gaussian pdf of posterior probability showing decision boundary of non linearly seperable data for bayesian model with different covariance

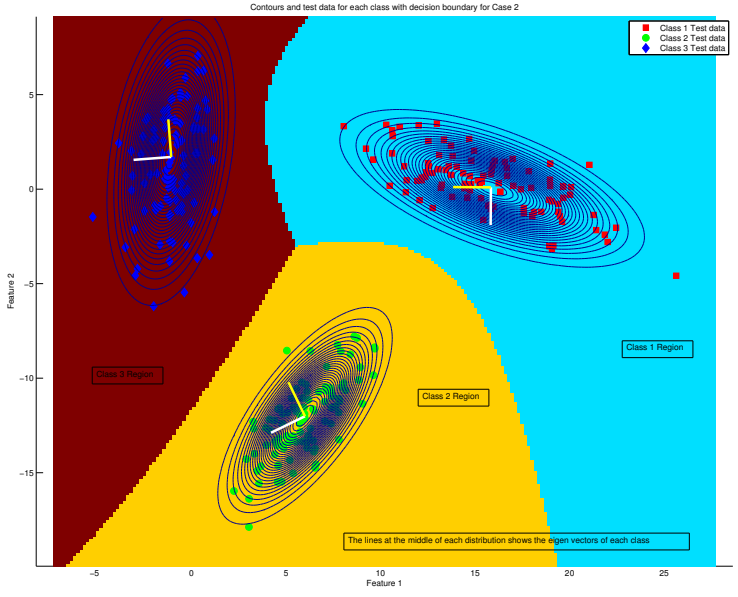


Figure 16: Contour plots and test data points showing decision boundary for non linear sepearble data for bayesian model with different covariance

section 2 continues on the next page...

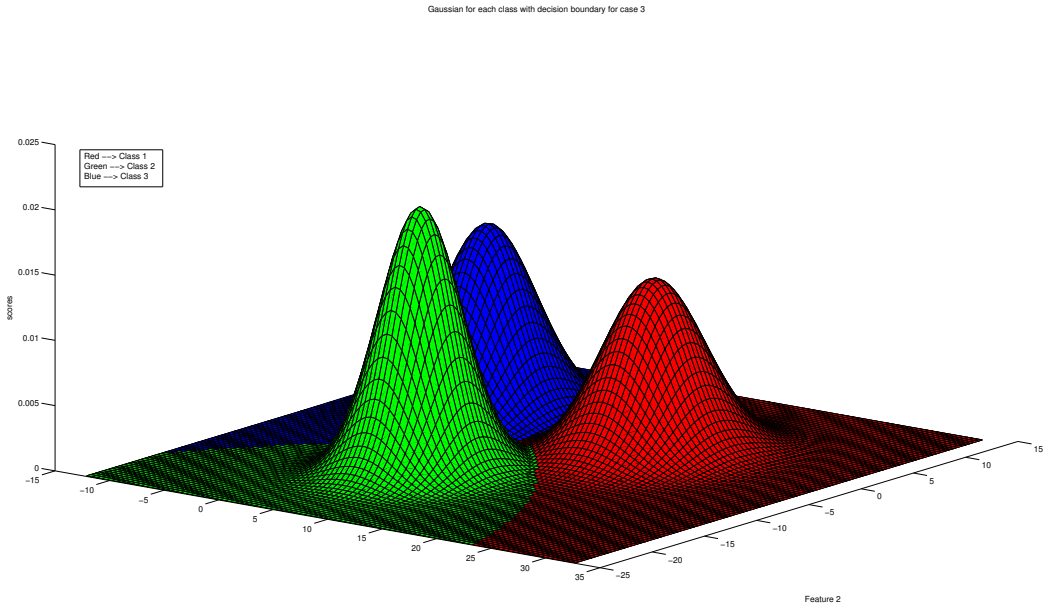


Figure 17: Gaussian pdf of posterior probability showing decision boundary of non linearly seperable data for Naive Bayes with $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

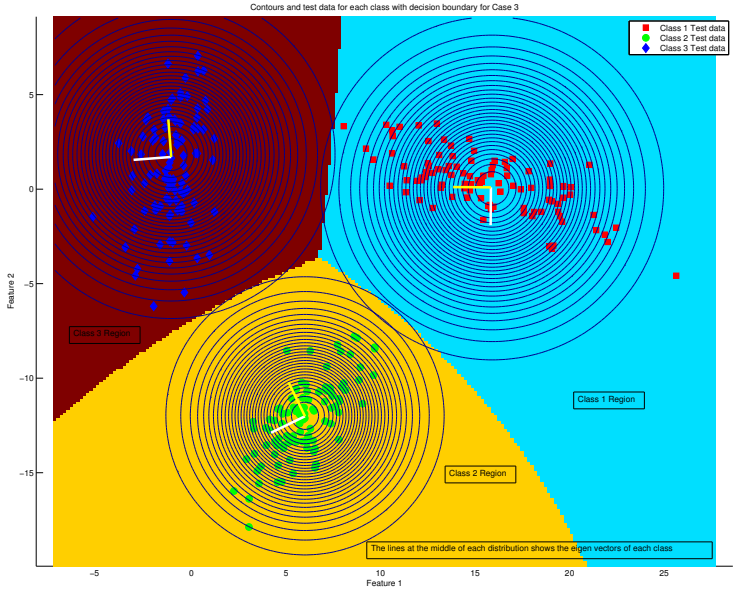


Figure 18: Contour plots and test data points showing decision boundary for non linear seperble data for Naive Bayes with $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

section 2 continues on the next page...

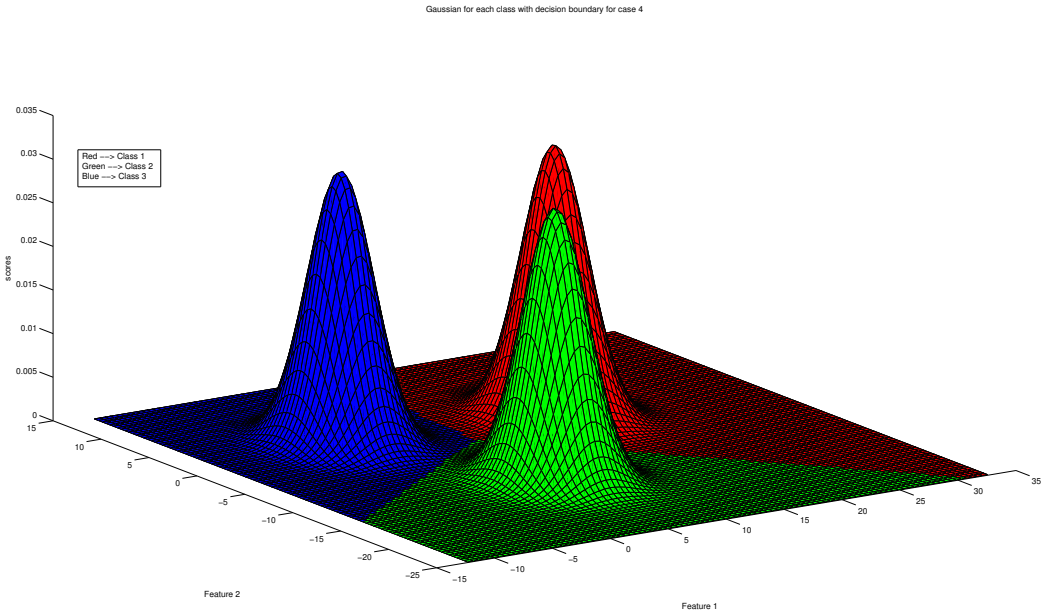


Figure 19: Gaussian pdf of posterior probability showing decision boundary of non linearly seperable data for Naive Bayes with same covariance

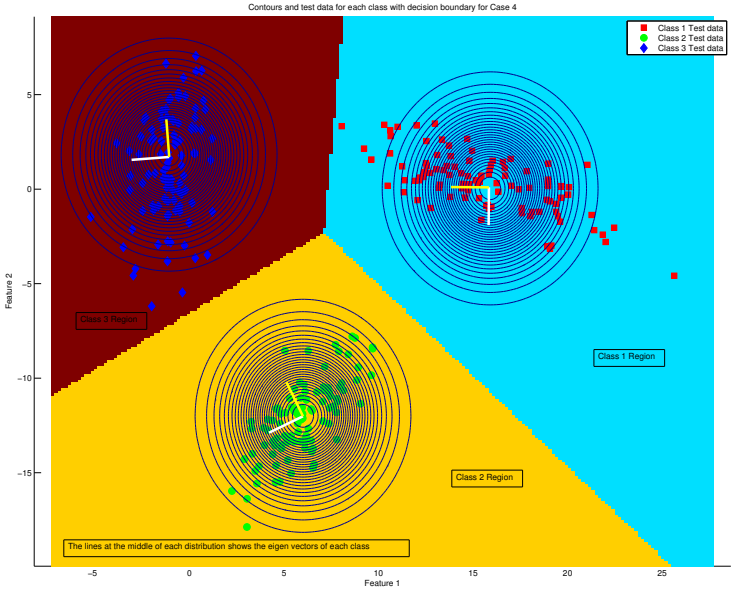


Figure 20: Contour plots and test data points showing decision boundary for non linear seperble data for Naive Bayes with different covariance

section 2 continues on the next page...

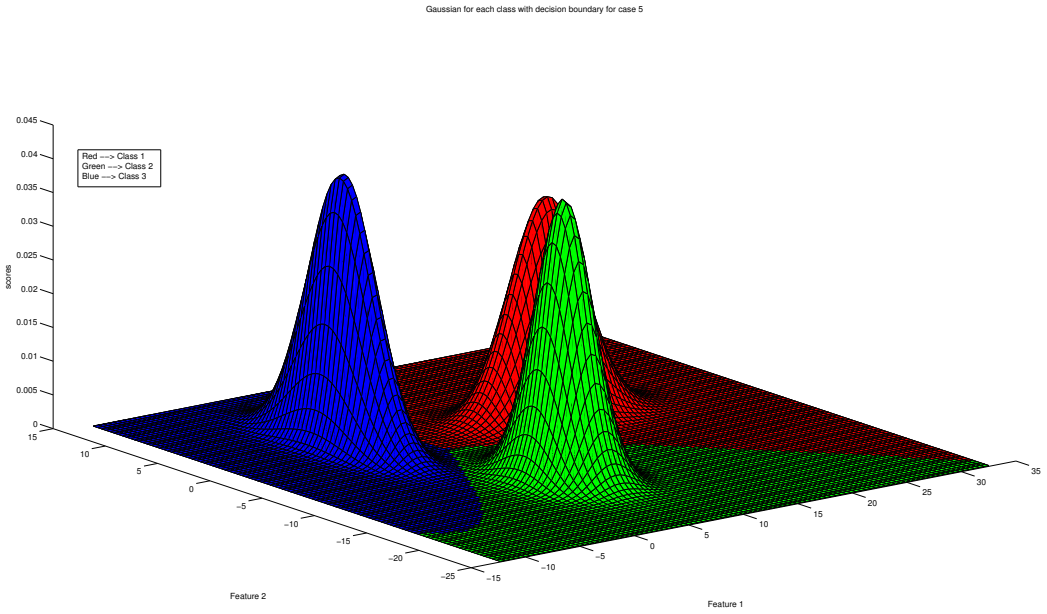


Figure 21: Gaussian pdf of posterior probability showing decision boundary of non linearly seperable data for Naive Bayes with different covariance

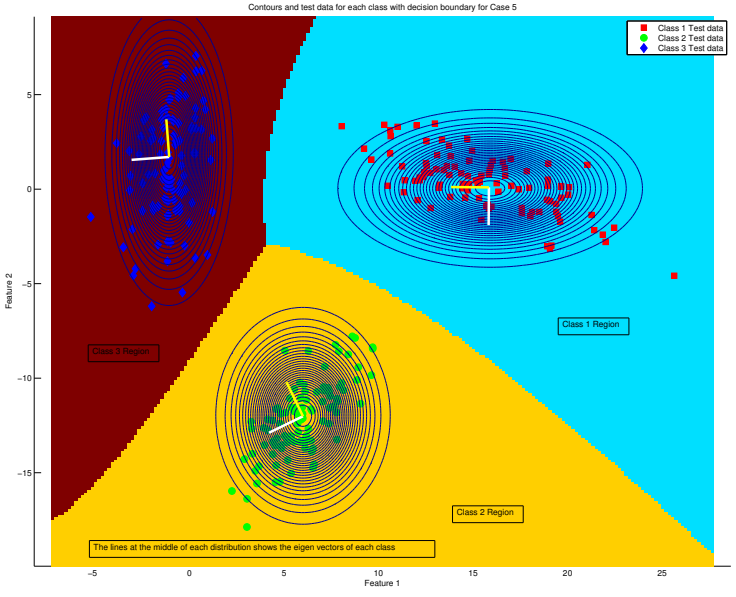


Figure 22: Contour plots and test data points showing decision boundary for non linear seperble data for Naive Bayes with different covariance

section 2 continues on the next page...

iii. **Overlapping Data with three classes**

As always, first step is to visualize the data.Lets scatterplot the data with different markers for each class. The scatterplot is given in Figure 23

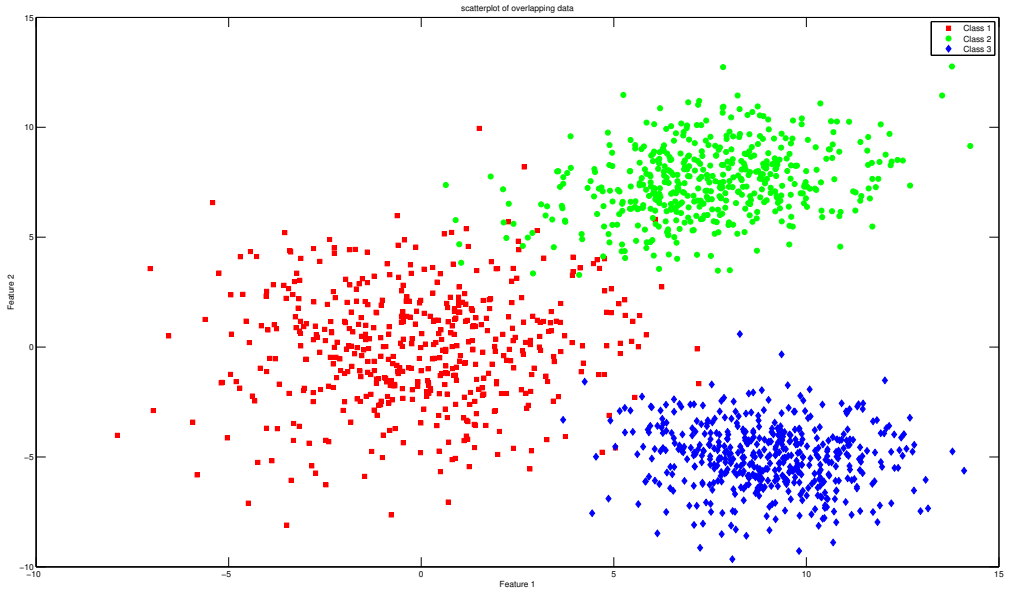


Figure 23: Scatter plot of Overlapping data

As we can see there is no clear boundary seperating each class, the accuracy will be less than than previous cases. But it is evident that majority of data belonging to each class is at the centre. The number of data points which are in the boundary are less. Even though data is overlapping, there aren't much outliers/ missing features. So we skip the preprocessing of data.

Now lets make classification models for the data and analyse the results.

(a) **Case 1: Bayes with Covariance same for all classes**

Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 24 shows the plot.

Note the gaussians are not well seperated like previous case, which is expected of overlapping data.We can visualize the decision boundary better by examining the contour plot and the linear Discriminant boundary. The test data is plotted, the contours of each class conditional probabilities are also plotted. The plot is Figure 25.

As the gaussians are same in shape, only their mean difffer. So it gives us a linear discriminant function of which boundaries are shown. The contour lines are crossing each other because of overlapping data.

The performance matrix is given in Table 5:

(b) **Case 2: Bayes with Covariance different for all classes**

Gaussian pdf of class conditional probability is plotted and the three classes and

section 2 continues on the next page...

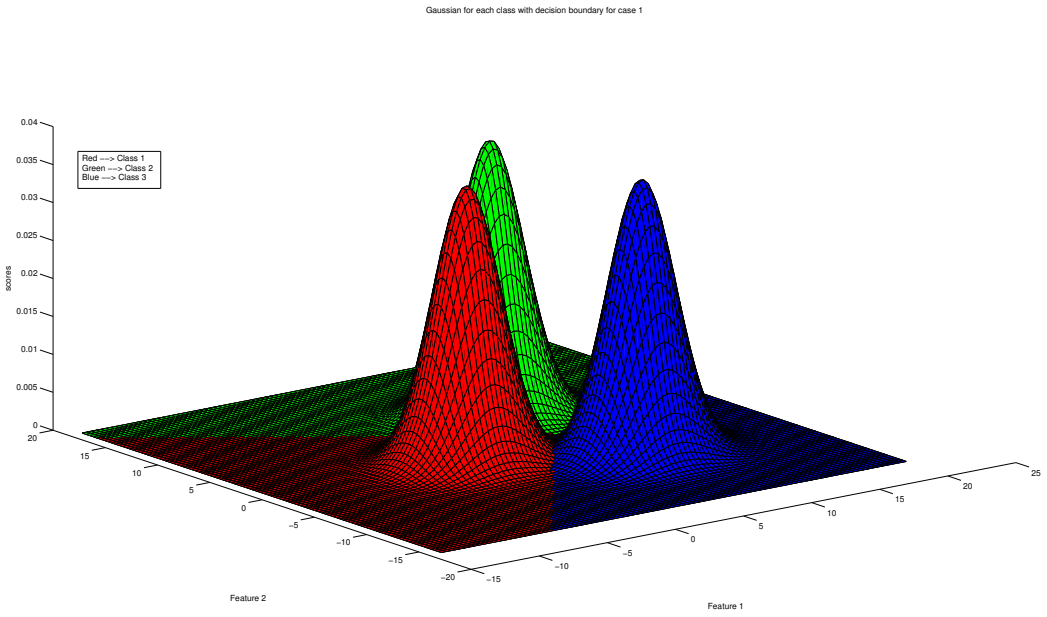


Figure 24: Gaussian pdf of posterior probability showing decision boundary of overlapping data with bayesian model with same covariance

		Prediction				
		Class 1	Class 2	Class 3	Total	Incl. Error
Truth	Class 1	144	3	3	150	0.04
	Class 2	1	149	0	150	0.0066
	Class 3	1	0	149	150	0.006
	Total	146	152	152		
	Excl. Error	0.013	0.019	0.019		

Table 5: Confusion matrix for Overlapping data, Case 1 Algorithm

	Precision	Accuracy
Class 1	0.96	98.22 %
Class 2	0.993	
Class 3	0.993	

Table 6: Performance metric

decision boundary is seperated by different color. Figure 26 shows the plot.

Now the decision boundary can be seen better in a contour plot.The plot is Figure 27.

As the data set is overlapping, the classification accuracy is expected to be less than previous cases. The performance matrix is given in Table 7:

(c) **Case 3: Naive Bayes with $\Sigma_k = \sigma_k^2 \boldsymbol{I}$**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 28 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric circles in the contour plot. The plot is Figure 29.
As we can see, The shape of all the gaussians are symmetrical / circular contours.The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown.

section 2 continues on the next page...

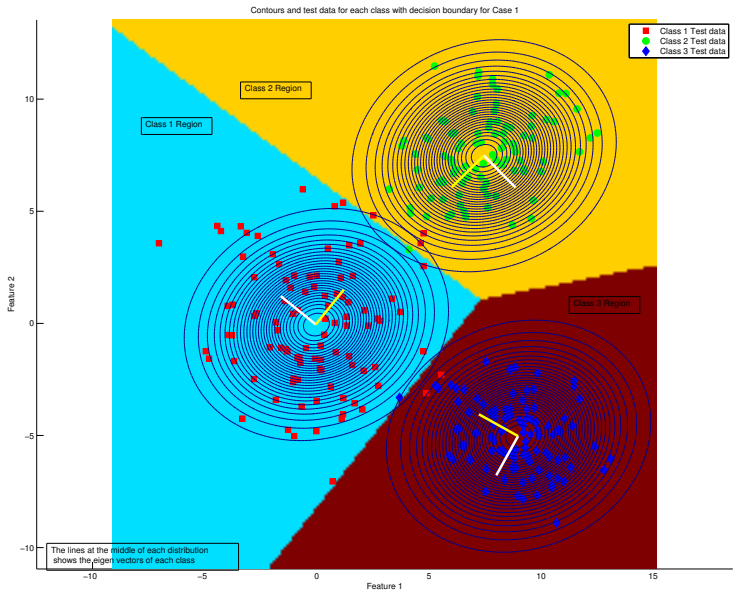


Figure 25: Contour plots and test data points showing decision boundary for non Overlapping data with bayesian model with same covariance

Truth \ Prediction	Class 1	Class 2	Class 3	Total	Incl. Error
Class 1	147	1	2	150	0.02
Class 2	2	148	0	150	0.0133
Class 3	1	0	149	150	0.006
Total	150	149	151		
Excl. Error	0.02	0.0067	0.013		

Table 7: Confusion matrix for Overlapping data, Case 2 Algorithm

	Precision	Accuracy
Class 1	0.98	98.66 %
Class 2	0.986	
Class 3	0.993	

Table 8: Performance metric

The performance matrix is given in Table 9:

(d) **Case 4: Naive Bayes with same covariance for all classes**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 30 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric ellipses in the contour plot. The plot is Figure 31.
The covariance is same for each class, So it gives us a linear discriminant function of which boundaries are shown. The performance matrix is given in Table 11:

(e) **Case 5: Naive Bayes with different covariance for all classes**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 32 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric ellipses in the contour plot. The plot is Figure 33.

section 2 continues on the next page...

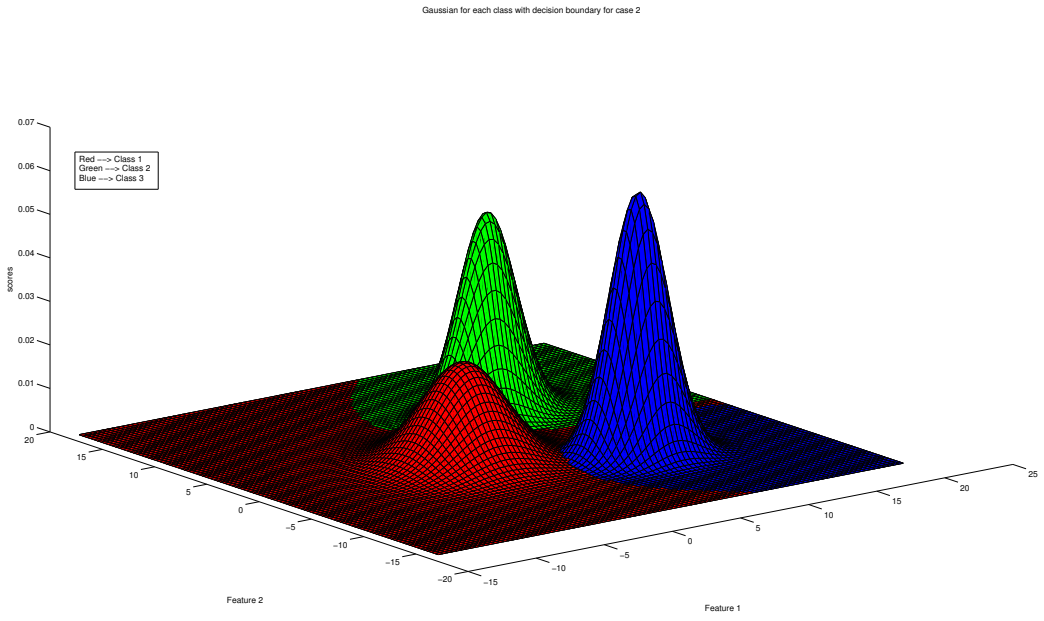


Figure 26: Gaussian pdf of posterior probability showing decision boundary of overlapping data for bayesian model with different covariance

		Prediction				
		Class 1	Class 2	Class 3	Total	Incl. Error
Truth	Class 1	146	2	2	150	0.02
	Class 2	2	148	0	150	0.0133
	Class 3	1	0	149	150	0.006
	Total	149	150	151		
	Excl. Error	0.02	0.013	0.013		

Table 9: Confusion matrix for Overlapping data, Case 3 Algorithm

	Precision	Accuracy
Class 1	0.973	98.44 %
Class 2	0.986	
Class 3	0.993	

Table 10: Performance metric

The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown. The confusion matrix and all the performance metrics are same for every algorithm we used. It is given in Table 13:

(f) **Performance evaluation** Now with ROC and DET curves, we can evaluate the performance. There are two possibilities, either I can see the performance of each algorithm in a particular class or i can take an algorithm and see which class achieved best performance for that model.
We plot both cases; Table 15 shows ROC and DET curve for each algorithm in each class.
Table 16 shows difference in performance of binary classification of each class for a particular algorithm.

section 2 continues on the next page...

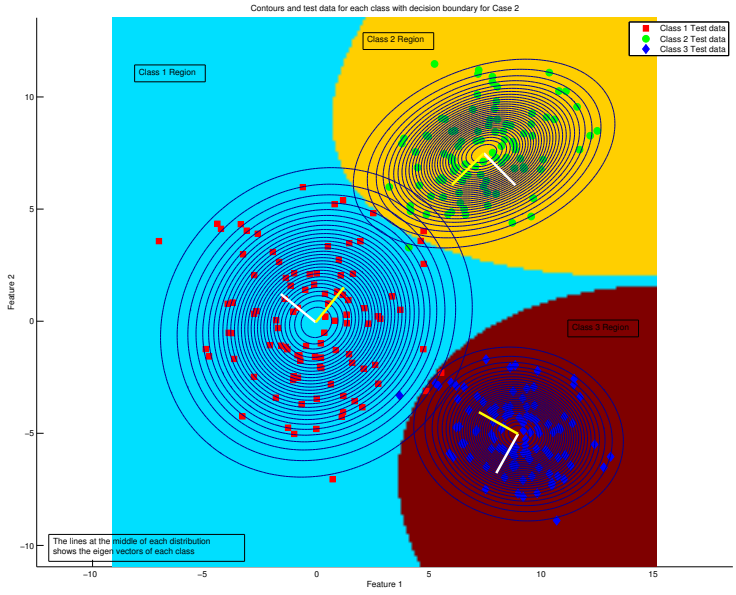


Figure 27: Contour plots and test data points showing decision boundary for overlapping data for bayesian model with different covariance

Truth \ Prediction	Class 1	Class 2	Class 3	Total	Incl. Error
Class 1	145	3	2	150	0.033
Class 2	1	149	0	150	0.066
Class 3	1	0	149	150	0.006
Total	147	152	151		
Excl. Error	0.013	0.019	0.013		

Table 11: Confusion matrix for Overlapping data, Case 4 Algorithm

	Precision	Accuracy
Class 1	0.966	98.44 %
Class 2	0.993	
Class 3	0.993	

Table 12: Performance metric

Truth \ Prediction	Class 1	Class 2	Class 3	Total	Incl. Error
Class 1	148	1	1	150	0.013
Class 2	2	148	0	150	0.013
Class 3	1	0	149	150	0.066
Total	151	149	150		0
Excl. Error	0.0198	0.0061	0.0067	0	

Table 13: Confusion matrix for overlapping data, Algorithm 5

	Precision	Accuracy
Class 1	0.9866	98.89 %
Class 2	0.9866	
Class 3	0.993	

Table 14: Performance metric

section 2 continues on the next page...

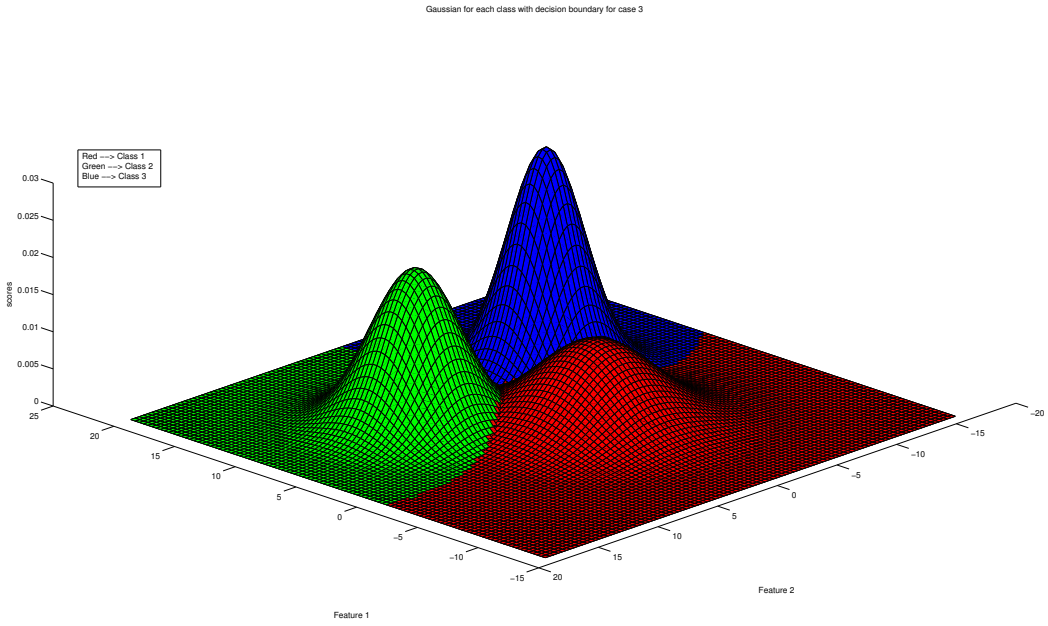


Figure 28: Gaussian pdf of posterior probability showing decision boundary of overlapping data for Naive Bayes with $\mathbf{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

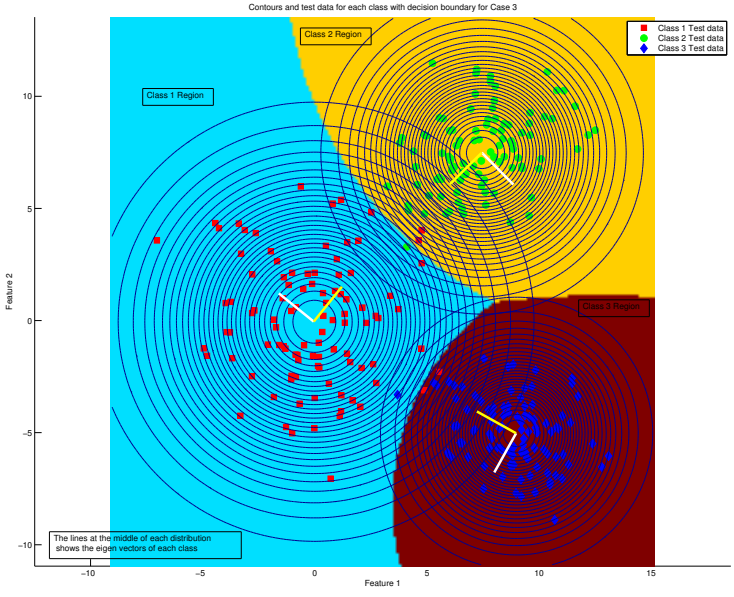


Figure 29: Contour plots and test data points showing decision boundary for overlapping data for Naive Bayes with $\mathbf{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

section 2 continues on the next page...

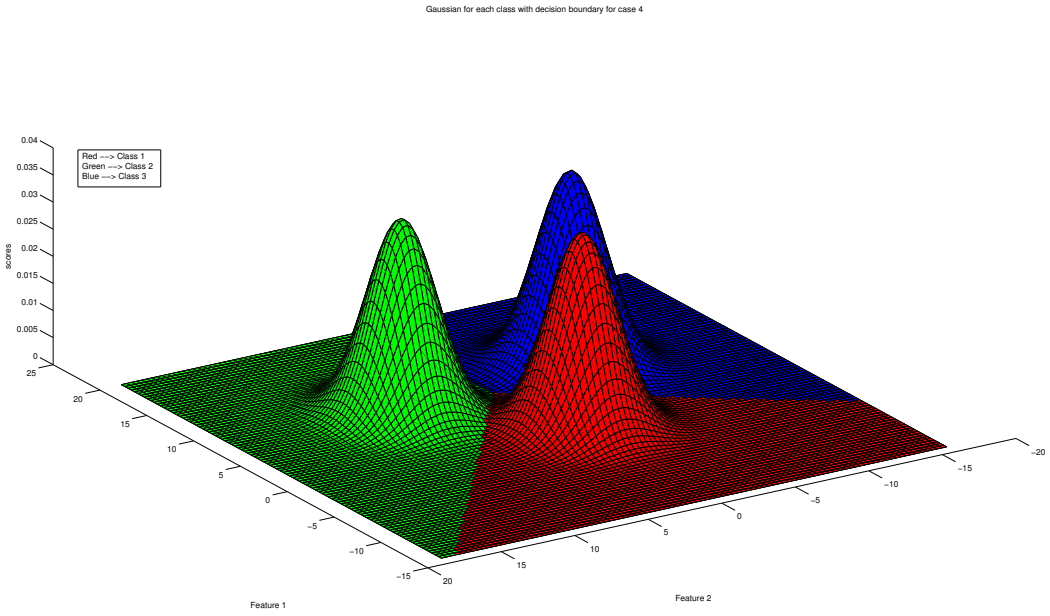


Figure 30: Gaussian pdf of posterior probability showing decision boundary of overlapping data for Naive Bayes with same covariance

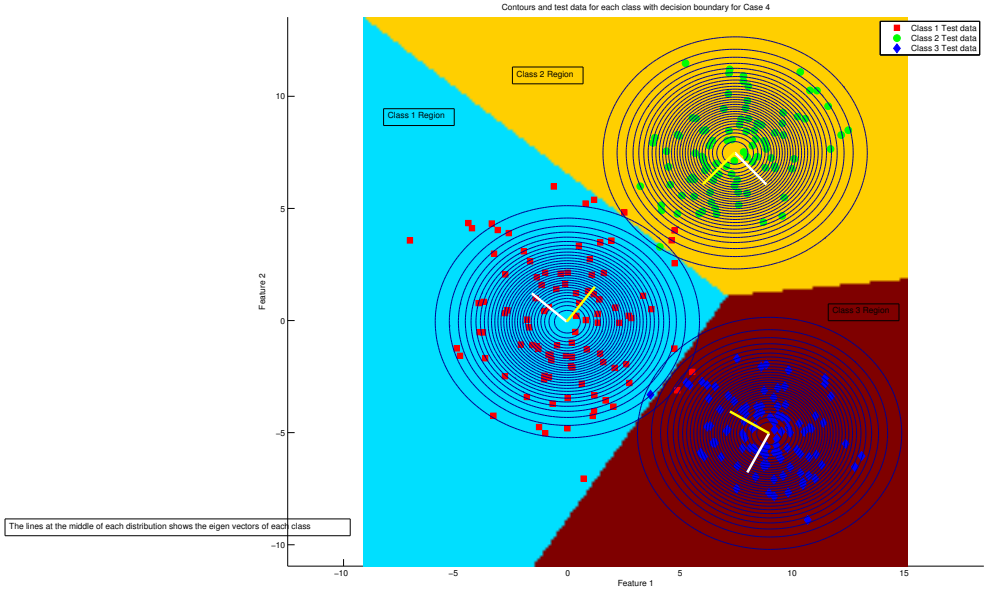


Figure 31: Contour plots and test data points showing decision boundary for overlapping data for Naive Bayes with different covariance

section 2 continues on the next page...

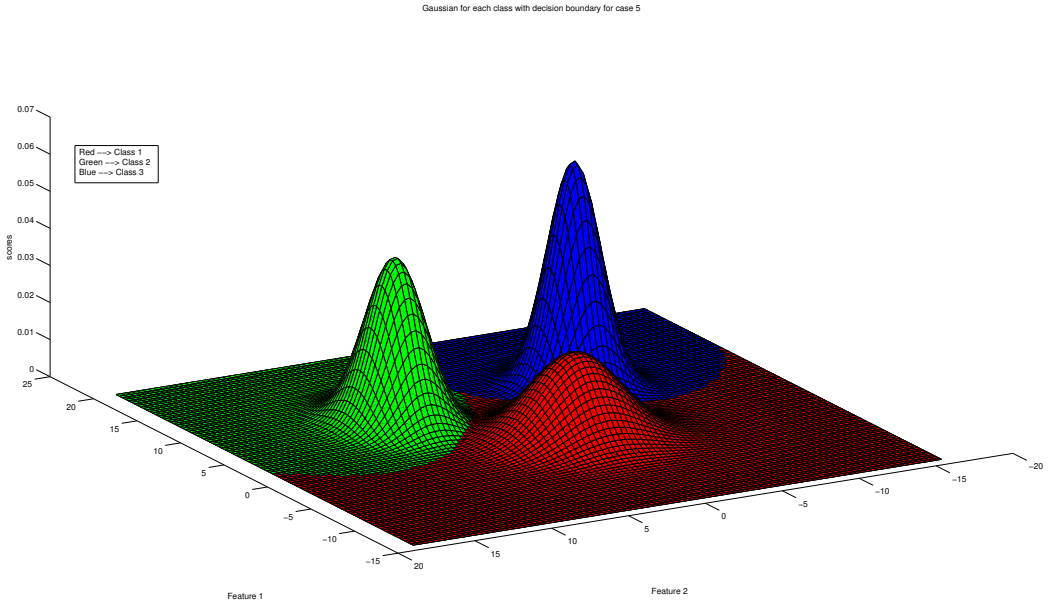


Figure 32: Gaussian pdf of posterior probability showing decision boundary of overlapping data for Naive Bayes with different covariance

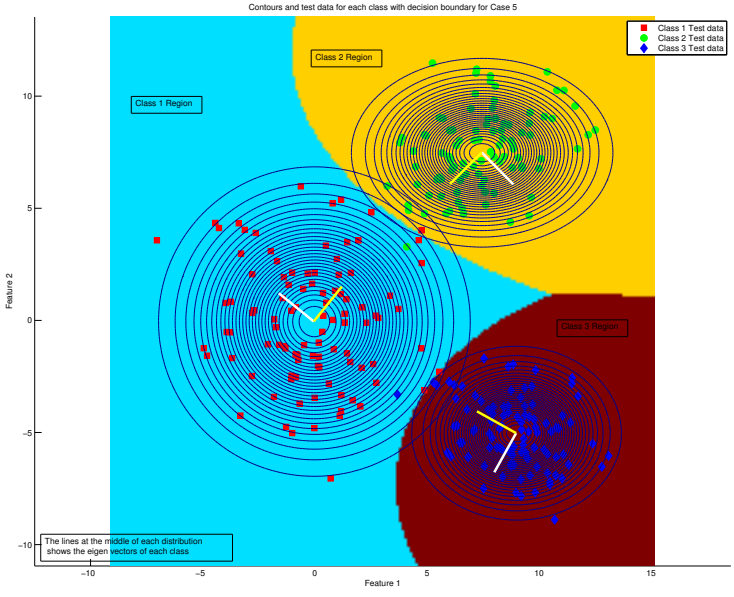
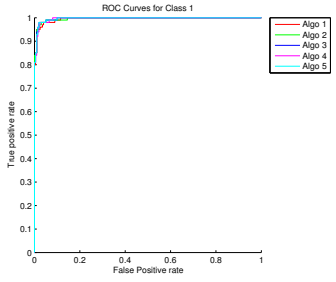
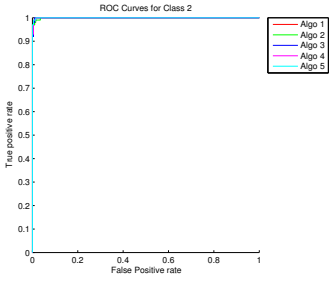


Figure 33: Contour plots and test data points showing decision boundary for non linear seperble data for Naive Bayes with different covariance

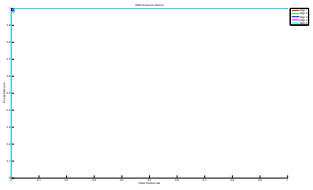
section 2 continues on the next page...



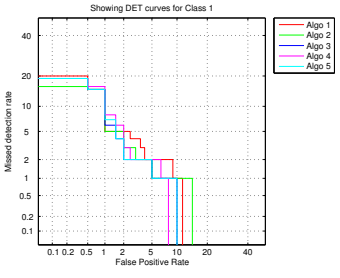
(a) Class 1 ROC



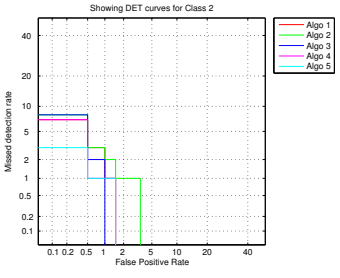
(b) Class 2 ROC



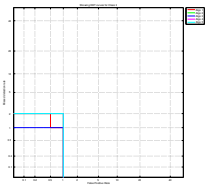
(c) Class 3 ROC



(d) Class 1 DET



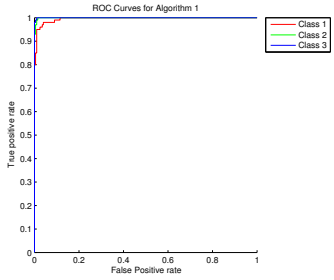
(e) Class 2 DET



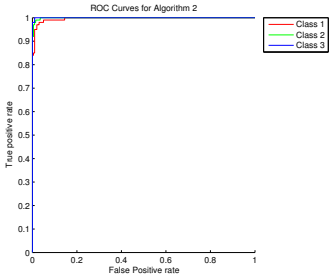
(f) Class 3 DET

Table 15: For overlapping data, DET and ROC curves for each class

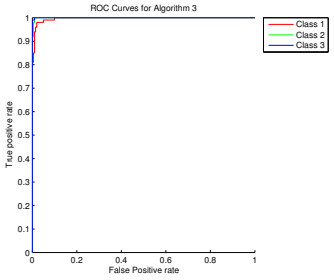
section 2 continues on the next page. . .



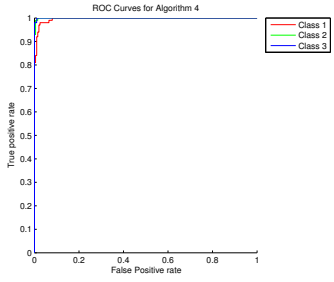
(a) Algo 1 ROC



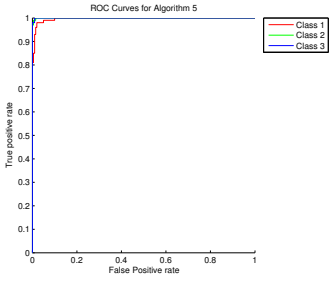
(b) Algo 2 ROC



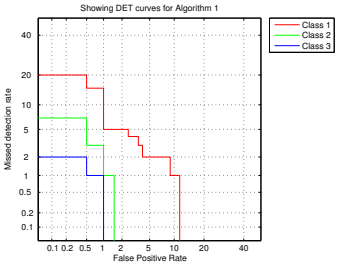
(c) Algo 3 ROC



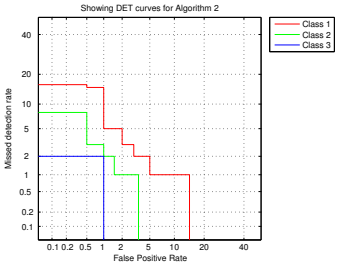
(d) Algo 4 ROC



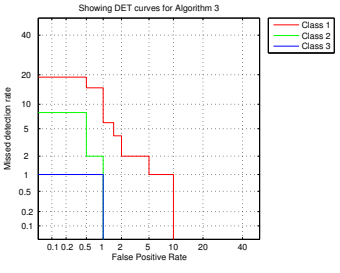
(e) Algo 5 ROC



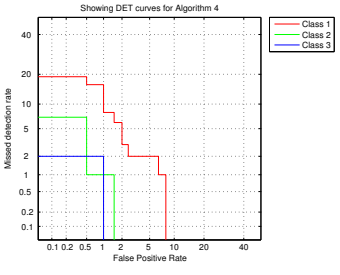
(f) Algo 1 DET



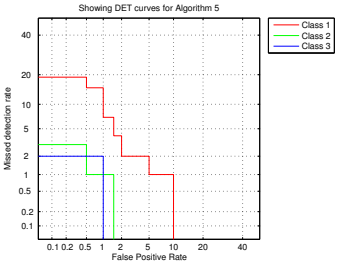
(g) Algo 2 DET



(h) Algo 3 DET



(i) Algo 4 DET



(j) Algo 5 DET

Table 16: For overlapping data, DET and ROC curves for each class

section 2 continues on the next page...

iv. **Real world data**

As always, first lets do a scatterplot of data and draw some inferences. The scatterplot is Figure 34

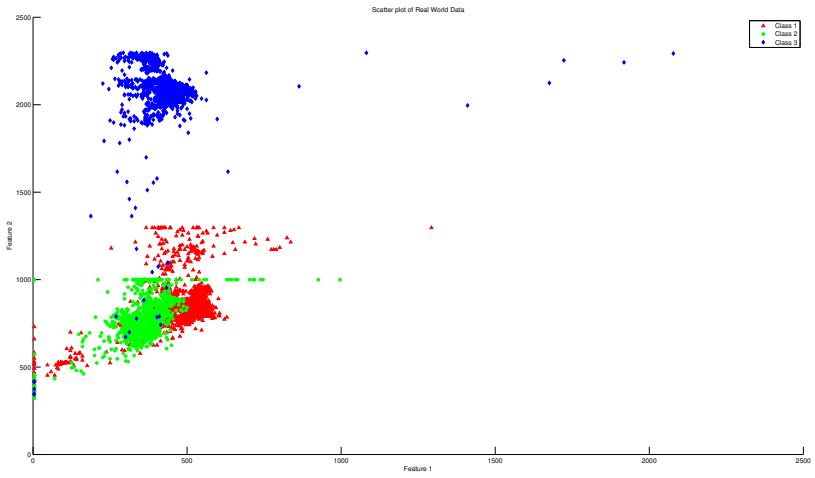


Figure 34: The scatterplot of real world data showing different classes

We can observe from data that:

- The data contains outliers. If we proceed without removing outliers, our mean calculation might get currupt since outliers can change mean significantly.
- The data has missing features, we can see it from the points lying on the left axis.
- The class 1 and class 2 are overlapping. It will be hard to discriminate between these two classes.
- For class 1 and class 2, there seems to be an upper bound for Feature 2 .

Outliers

It is important to remove outliers from the training data so that the outliers do not affect the parameters of the model much, in our case mean and covariance. Lets draw the Boxplot for each class in each feature. Figure 35 shows box plots.

It is evident from the data that there are outliers. Unlike univariate, it is hard to classify outliers in Multivariate. In Multivariate case the distance from the mean to a point is measured using *Mahalanobis Distance*. Distribution of Mahalanobis distance is given by chi squared distribution of d Degrees of freedoms where d is the feature dimension.

The (squared) Mahalanobis Distance is given by $\text{MD}_i = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$

A point is said to be a outlier if the (squared) Mahalanobis distance is greater than $inv\left(\sqrt{\mathcal{X}_d^2(0.975)}\right)$ where $\mathcal{X}_d^2$ is chi squared distribution of DOF d.

Once they are detected, we use the attribute mean to fill in the feature which is detected as outlier. The figures 38 shows the scatterplot of training data after and before removing outliers.

Now with the cleaned training data, we make 5 algorithms.

section 2 continues on the next page...

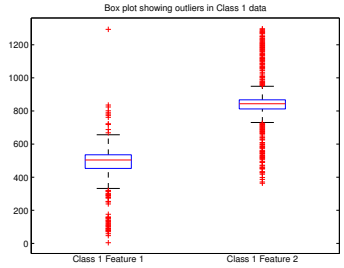


Figure 35: Boxplot for Class 1 data

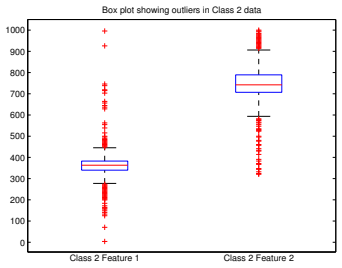


Figure 36: Boxplot for Class 2 data

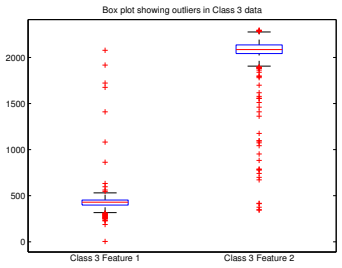


Figure 37: Boxplot for Class 3 data

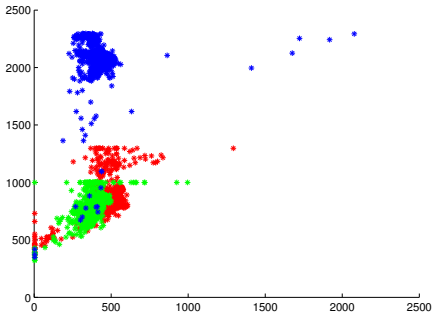


Figure 38: Scatterplot of training data before outlier removal

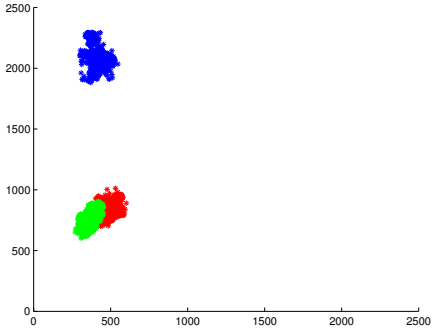


Figure 39: Scatterplot of training data after outlier removal

(a) **Case 1: Bayes with Covariance same for all classes**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 40 shows the plot.

The test data is plotted, the contours of each class conditional probabilities are also plotted. The plot is Figure 41.

As the gaussians are same in shape, only their mean difffer. So it gives us a linear discriminant function of which boundaries are shown. The contour lines are crossing each other because of overlapping data.
The performance matrix is given in Table 17:

(b) **Case 2: Bayes with Covariance different for all classes**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 42 shows the plot.

Now the decision boundary can be seen better in a contour plot.The plot is Figure 43.

The performance matrix is given in Table 19:

section 2 continues on the next page...

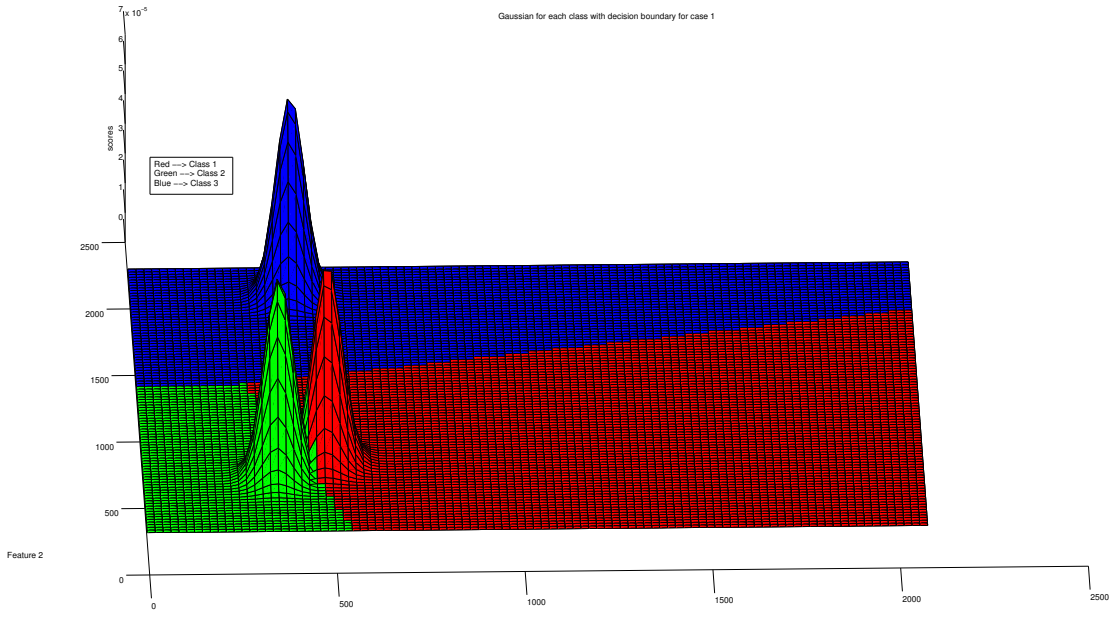


Figure 40: Gaussian pdf of posterior probability showing decision boundary of Real data with bayesian model with same covariance

Truth \ Prediction	Class 1	Class 2	Class 3	Total	Incl. Error
Class 1	1614	225	0	1839	0.122
Class 2	175	1691	0	1866	0.093
Class 3	8	15	2268	2291	0.01
Total	1797	1931	2268		
Excl. Error	0.101	0.124	0		

Table 17: Confusion matrix for Real, Case 1 Algorithm

	Precision	Accuracy
Class 1	0.877	92.95 %
Class 2	0.906	
Class 3	0.989	

Table 18: Performance metric

(c) **Case 3: Naive Bayes with $\Sigma_k = \sigma_k^2 \boldsymbol{I}$**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 44 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric circles in the contour plot. The plot is Figure 45.
As we can see, The shape of all the gaussians are symmetrical / circular contours.The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown.
The performance matrix is given in Table 21:

(d) **Case 4: Naive Bayes with same covariance for all classes**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 46 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric ellipses in the contour plot. The plot is Figure 47.

section 2 continues on the next page...

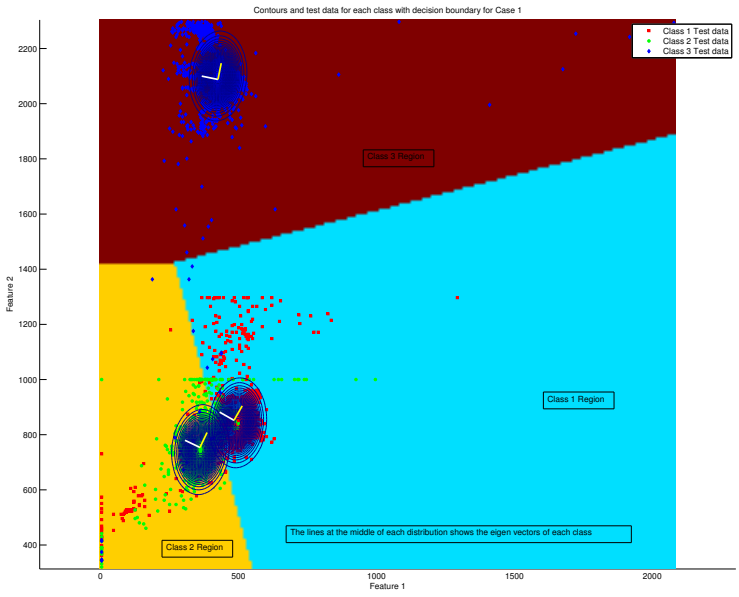


Figure 41: Contour plots and test data points showing decision boundary for real data with bayesian model with same covariance

		Prediction				
		Class 1	Class 2	Class 3	Total	Incl. Error
Truth	Class 1	1618	203	18	1839	0.12
	Class 2	219	1647	0	1866	0.117
	Class 3	10	14	2267	2291	0.010
	Total	1847	1864	2285		
	Excl. Error	0.123	0.116	0.007		

Table 19: Confusion matrix for Real data, Case 2 Algorithm

	Precision	Accuracy
Class 1	0.87	92.26 %
Class 2	0.886	
Class 3	0.989	

Table 20: Performance metric

The covariance is same for each class, So it gives us a linear discriminant function of which boundaries are shown. The performance matrix is given in Table 23:

(e) **Case 5: Naive Bayes with different covariance for all classes**
Gaussian pdf of class conditional probability is plotted and the three classes and decision boundary is seperated by different color. Figure 48 shows the plot.
Now the decision boundary can be seen better in a contour plot. We expect concentric ellipses in the contour plot. The plot is Figure 49.
The covariance is different for each class, So it gives us a non linear discriminant function of which boundaries are shown. The confusion matrix and all the performance metrics are same for every algorithm we used. It is given in Table 25:

(f) **Performance evaluation** Now with ROC and DET curves, we can evaluate the performance. There are two possibilities, either I can see the performance of each algorithm in a particular class or i can take an algorithm and see which class achieved

section 2 continues on the next page...

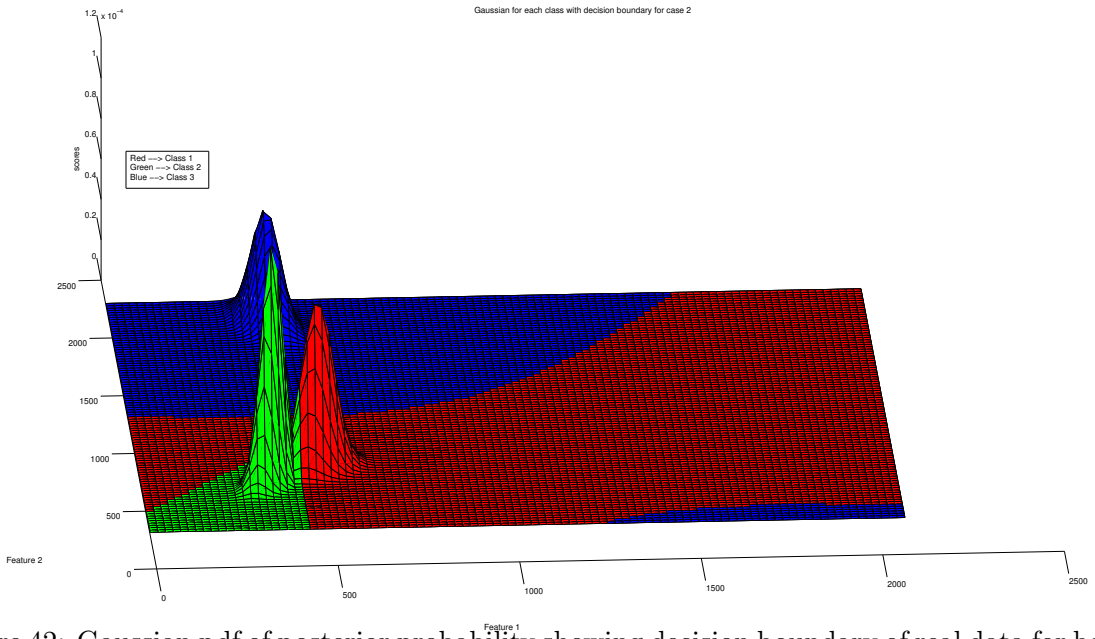


Figure 42: Gaussian pdf of posterior probability showing decision boundary of real data for bayesian model with different covariance

Truth \ Prediction	Class 1	Class 2	Class 3	Total	Incl. Error
Class 1	1628	211	0	1839	0.114
Class 2	291	1575	0	1866	0.1553
Class 3	6	14	2271	2451	0.008
Total	1925	1900	2271		
Excl. Error	0.154	0.125	0		

Table 21: Confusion matrix for Real data, Case 3 Algorithm

	Precision	Accuracy
Class 1	0.88	91.24 %
Class 2	0.84	
Class 3	0.99	

Table 22: Performance metric

best performance for that model.
We plot both cases; Table 27 shows ROC and DET curve for each algorithm in each class.

Table 28 shows difference in performance of binary classification of each class for a particular algorithm.

section 2 continues on the next page...

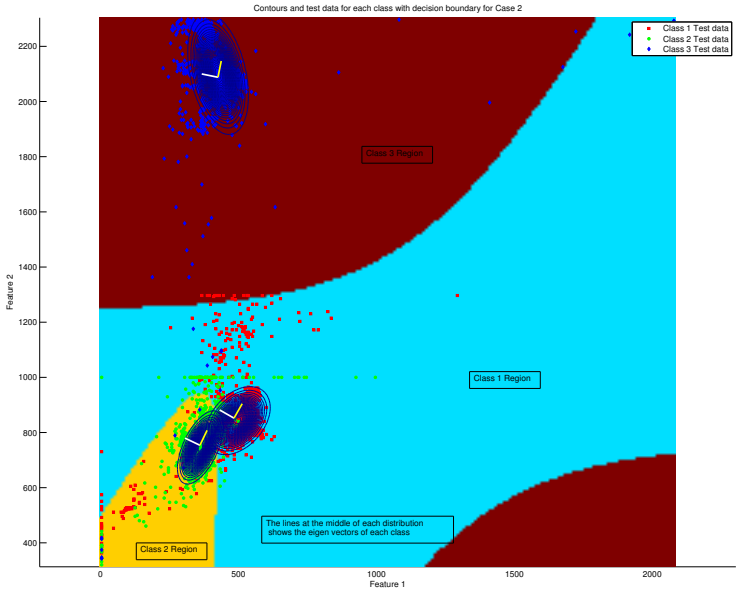


Figure 43: Contour plots and test data points showing decision boundary for Real data for bayesian model with different covariance

Truth \ Prediction	Class 1	Class 2	Class 3	Total	Incl. Error
Class 1	1615	224	0	1839	0.121
Class 2	190	1676	0	1866	0.101
Class 3	8	15	2268	2291	0.01
Total	1813	1915	2268		
Excl. Error	0.109	0.124	0		

Table 23: Confusion matrix for real data, Case 4 Algorithm

	Precision	Accuracy
Class 1	0.87	92.71 %
Class 2	0.89	
Class 3	0.98	

Table 24: Performance metric

Truth \ Prediction	Class 1	Class 2	Class 3	Total	Incl. Error
Class 1	1641	193	5	150	0.107
Class 2	261	1605	0	150	0.139
Class 3	6	14	2271	150	0.008
Total	151	149	150		0
Excl. Error	0.139	0.114	0.0022	0	

Table 25: Confusion matrix for Real data, Algorithm 5

	Precision	Accuracy
Class 1	0.89	92.01 %
Class 2	0.86	
Class 3	0.991	

Table 26: Performance metric

section 2 continues on the next page...

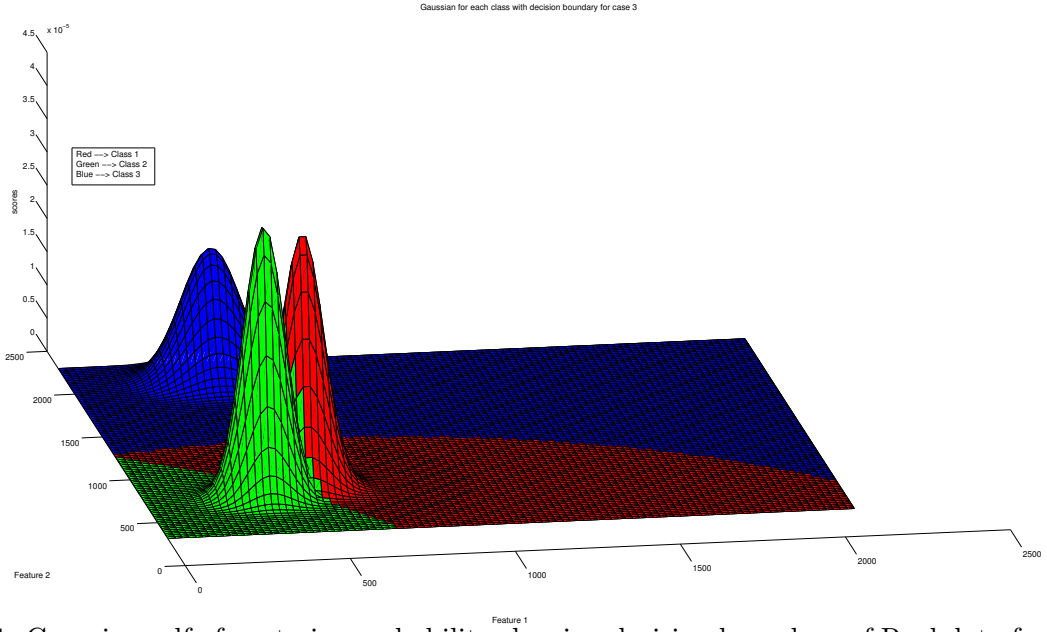


Figure 44: Gaussian pdf of posterior probability showing decision boundary of Real data for Naive Bayes with $\mathbf{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

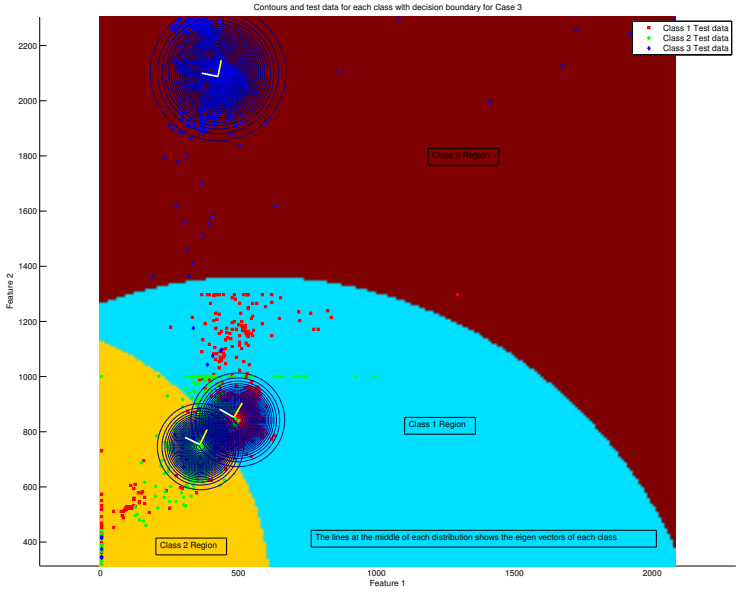


Figure 45: Contour plots and test data points showing decision boundary for Real data for Naive Bayes with $\mathbf{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

section 2 continues on the next page...

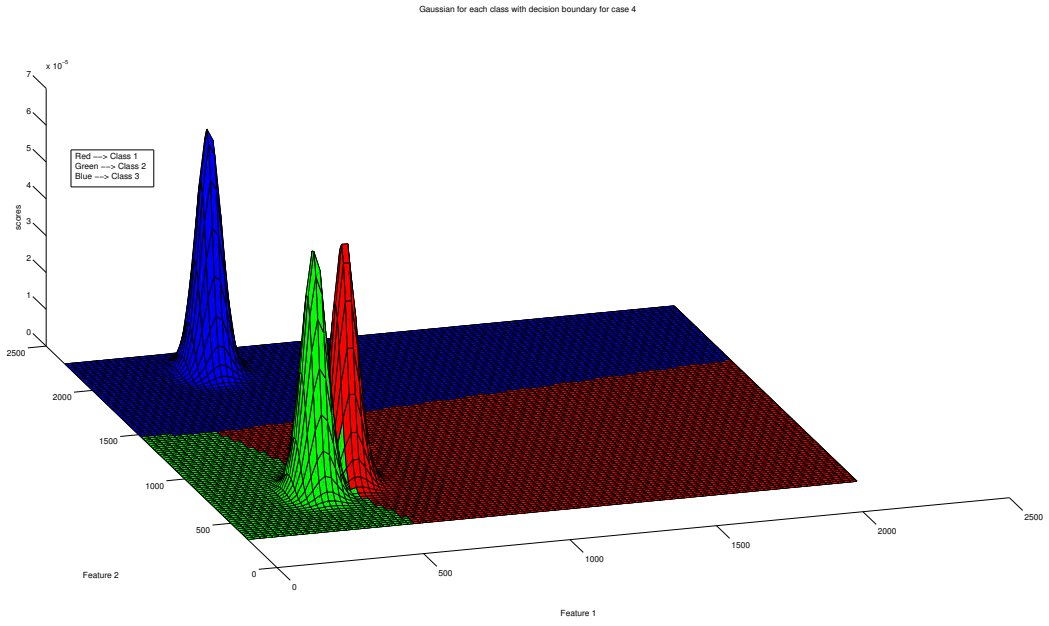


Figure 46: Gaussian pdf of posterior probability showing decision boundary of real data for Naive Bayes with same covariance

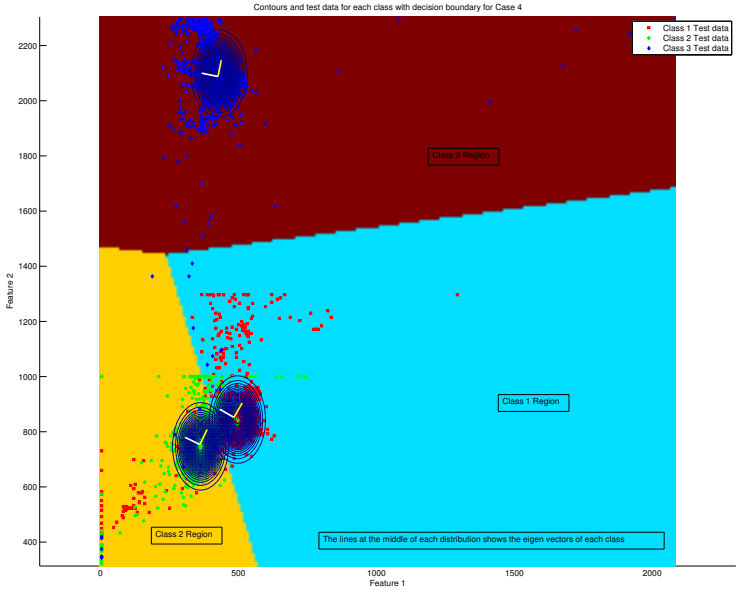


Figure 47: Contour plots and test data points showing decision boundary for real data for Naive Bayes with different covariance

section 2 continues on the next page...

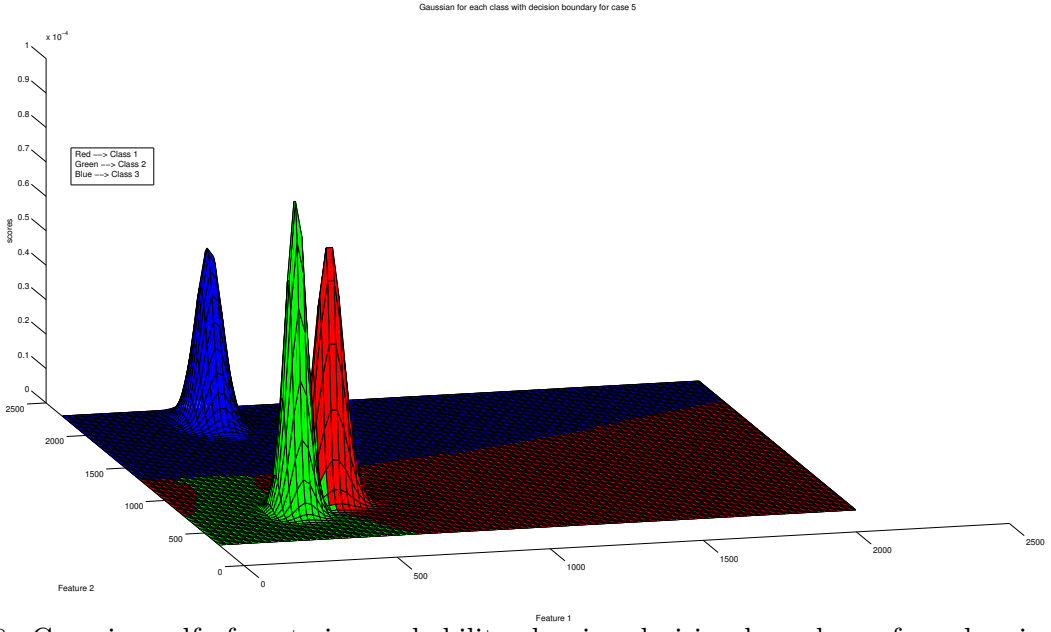


Figure 48: Gaussian pdf of posterior probability showing decision boundary of overlapping data for Naive Bayes with different covariance

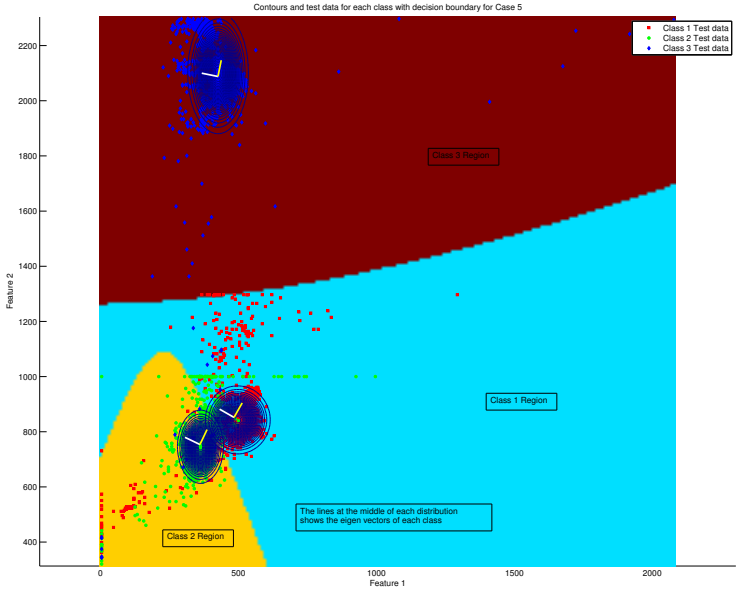
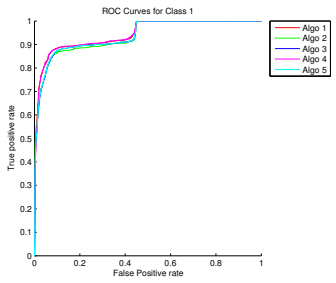
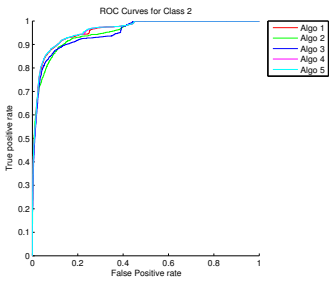


Figure 49: Contour plots and test data points showing decision boundary for real data for Naive Bayes with different covariance

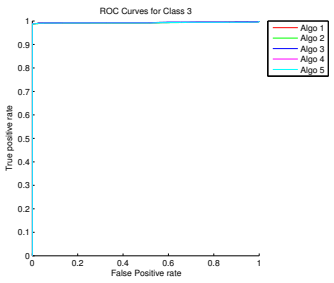
section 2 continues on the next page...



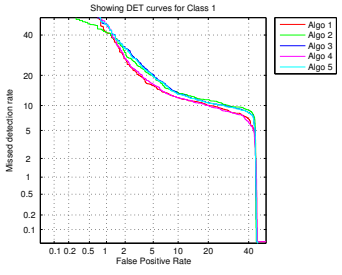
(a) Class 1 ROC



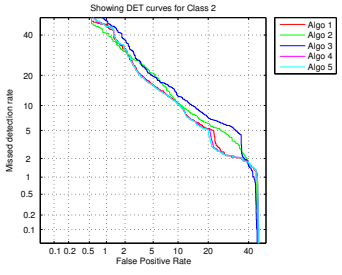
(b) Class 2 ROC



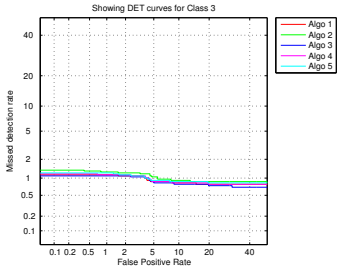
(c) Class 3 ROC



(d) Class 1 DET



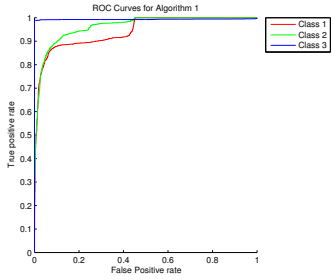
(e) Class 2 DET



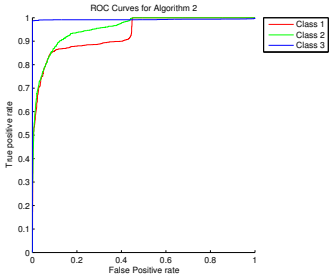
(f) Class 3 DET

Table 27: For Real data, DET and ROC curves for each class

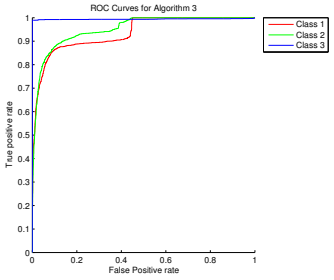
section 2 continues on the next page...



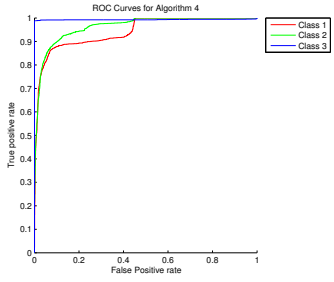
(a) Algo 1 ROC



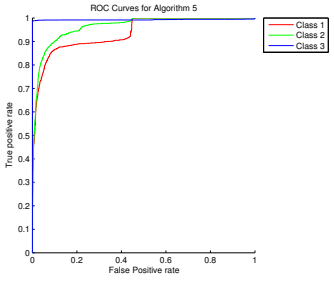
(b) Algo 2 ROC



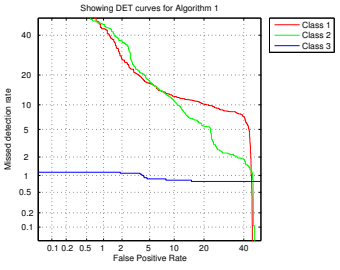
(c) Algo 3 ROC



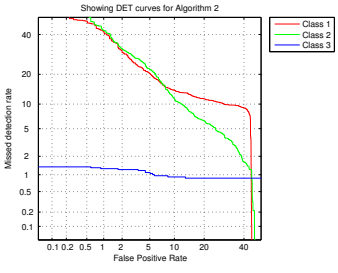
(d) Algo 4 ROC



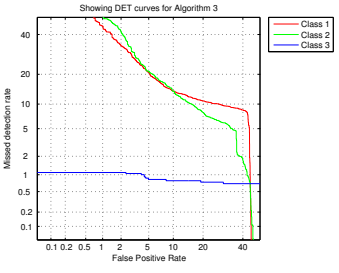
(e) Algo 5 ROC



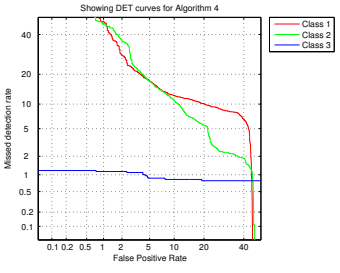
(f) Algo 1 DET



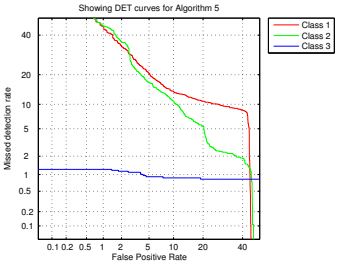
(g) Algo 2 DET



(h) Algo 3 DET



(i) Algo 4 DET



(j) Algo 5 DET

Table 28: For Real data, DET and ROC curves for each class