Pattern Recognition

Assignment #2

Bayesian Classifier

Professor:
Hema A. Murthy

Nirav Bhavsar (CS17S016)
Sidharth Aggarwal (CS17S012)

1.1 Bayesian Classification

In this assignment we are making a classifier which can classify any point that to which class it belongs to. For this we are designing our model with the training data.In this we will use the most common parametric form i.e, the Normal/Gaussian Probability function for the classification.

$$f(x) = 2\pi^{-\frac{d}{2}} \det \Sigma^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)$$

On all three data sets we have done experiments by plotting the data, their gaussian forms, the contours their eigen vectors according the the shape of the gaussian, the confusion matrix and further the ROC and DET plots. And we have done all these experiments for five different cases on each data set. And the five cases are as follows:-

1. Bayes with Covariance same for all classes
2. Bayes with Covariance different for all classes
3. Naive Bayes with C = $\sigma^2 * I$
4. Naive Bayes with C same for all classes.
5. Naive Bayes with C different for all classes.

1.2 Linearly Separable Data

In this as the name suggests the classes will be separated from each other with the linear decision boundaries. From this we get to know that there is very less overlapping of the point of the different classes.

As we can see from the below plots for the linearly separable data. In these plots we have shown the Gaussian PDF of the data along with the their contours.Another plot is of the confusion matrix which tells the accuracy of the classifier for every classes. The ROC plot tells the accuracy of the classification done by plotting the graph between TPR(True Positive Rate) and FPR(False Positive Rate). The plot with regions is according to the decision boundary which is linear in this case.

Observations:-
From the plot of the contours we observed that if the covariance matrix of the all the classes are same and of the form $\sigma^2 * I$ then the contours will be circular as the variance of the of both axis are same and the axis of the contours are the eigen vectors which are parallel to the standard axis.Further if the covariance matrix is same but not of $\sigma^2 * I$ then contours will be eclipse with will be parallel to coordinate axis.If the covariance matrix is different then the contours may or may not be eclipse with orthogonal principal axis and may or may not be parallel to coordinate axis.

In the ROC plot we know that it is the curve between the False Positive Rate(means the classifier has classified it to be target but actually it is non-target) and the True Positive Rate(means the classifier has accurately classified the point). This curve tells the accuracy level in the system. So as in this case the accuracy is 100% so the curve is along the axis the we know that that threshold is best for the system for which the curve is at the top left corner.

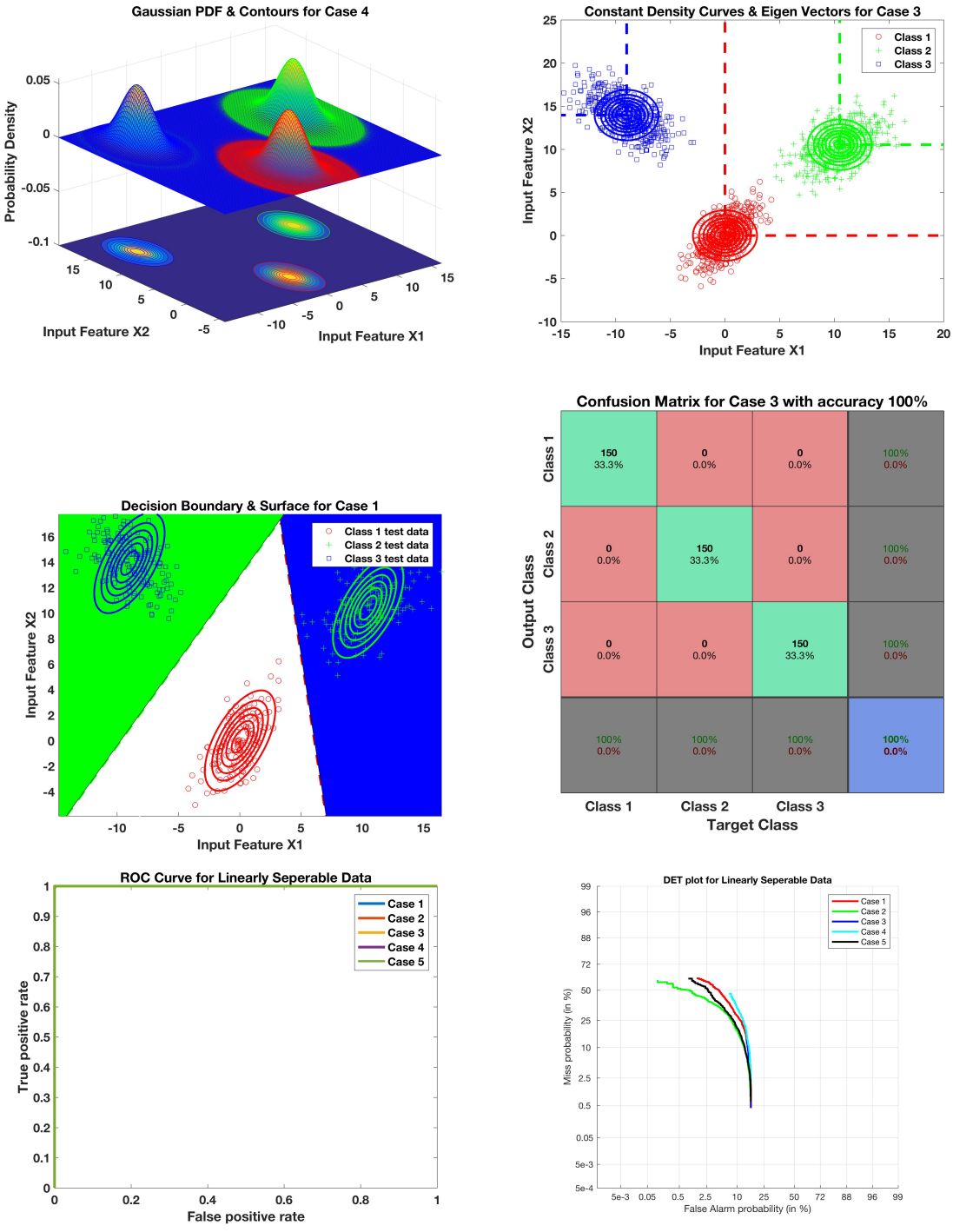

Figure 1.1: Plots for Linearly Separable Data

1.3 Non-Linearly Separable Data

In this as the name suggests the classes will be separated from each other with the non-linear decision boundaries. It means that the points of different classes will overlap with each other due to which it will will be difficult to predict with linear boundary that the point belongs to the class or not.

To find the decision boundary we have used the following discriminant functions:-

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^t\bigg(-\frac{1}{2}\Sigma_i^{-1}\bigg)\boldsymbol{x} + \bigg(\Sigma_i^{-1}\boldsymbol{\mu}_i\bigg)^t\boldsymbol{x} + \bigg(-\frac{1}{2}\boldsymbol{\mu}_i^t\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln(|\Sigma_i|)\bigg)$$

As we can see from the below plots for the linearly separable data. In these plots we have shown the Gaussian PDF of the data along with the their contours.Another plot is of the confusion matrix which tells the accuracy of the classifier for every classes. The ROC plot tells the accuracy of the classification done by plotting the graph between TPR(True Positive Rate) and FPR(False Positive Rate). The plot with regions is according to the decision boundary which is linear in this case.

Observations:-
From the plot of the contours we observed that if the covariance matrix of the all the classes are same and of the form $\sigma^2 * I$ then the contours will be circular as the variance of the of both axis are same and the axis of the contours are the eigen vectors which are parallel to the standard axis.Further if the covariance matrix is same but not of form $\sigma^2 * I$ then contours will be eclipse with orthogonal axis. As observed when the covariance matrix is different then the contours eigen vectors may or may not be parallel to coordinate axis but they would be orthogonal with each other.

In the ROC plot we know that it is the curve between the False Positive Rate(means the classifier has classified it to be target but actually it is non-target) and the True Positive Rate(means the classifier has accurately classified the point). In the ROC plot for the non-linear data and corresponding to the five cases the curves are coming. As we know that to get the non-linear decision boundaries the covariance matrix should be different and for that the accuracy of classification is good. So here in the ROC of non-linear data we observed that for the cases in which the cocariance matrix is different is having much more accuracy and among naive bayes and bayesian, and further bayesian case is having more accuracy.

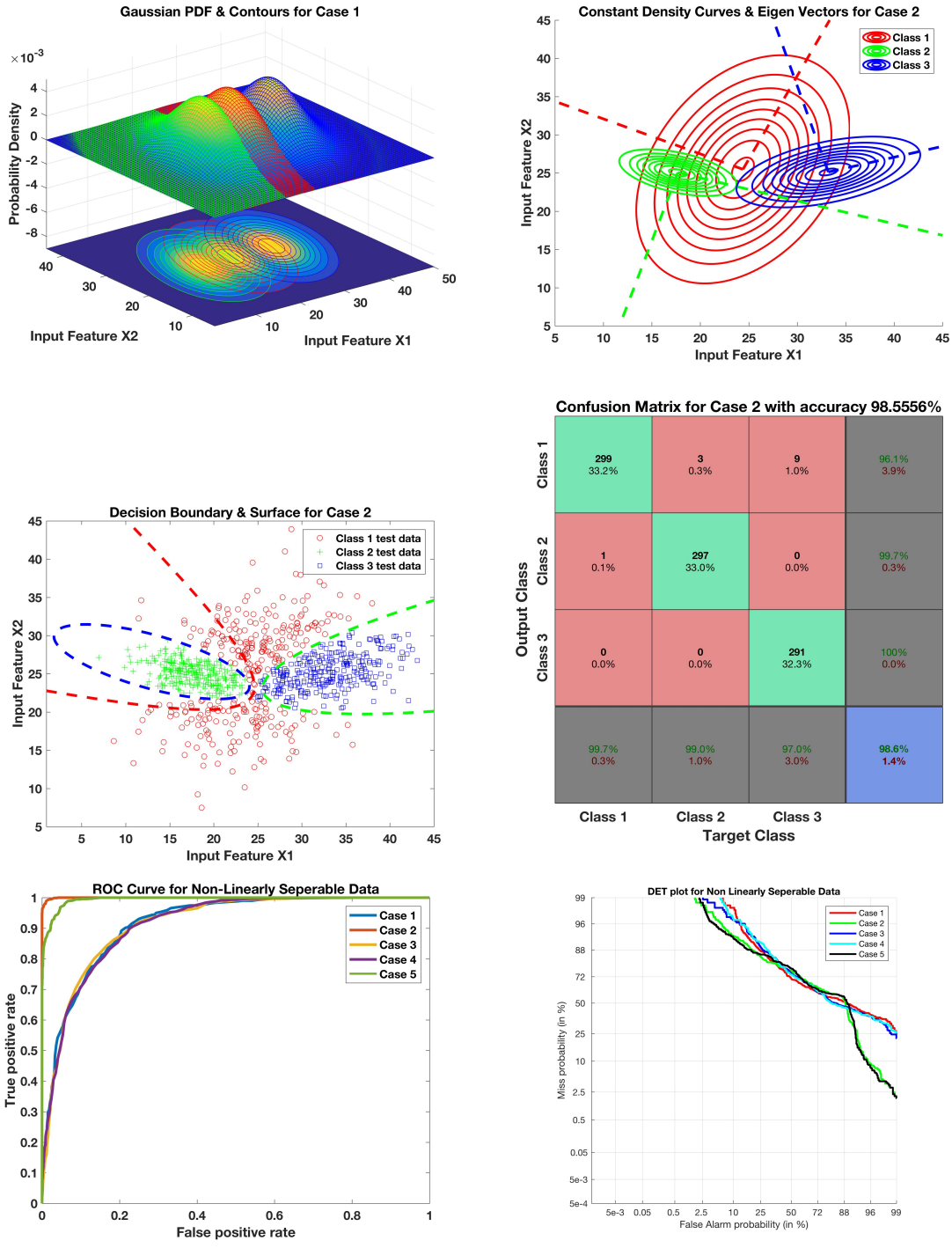

Figure 1.2: Plots for Non Linearly Separable Data

1.4 Real Data

In this the data is real data so the output could be any, either the linear boundaries or non-linear boundaries.

As we can see from the below plots for the linearly separable data. In these plots we have shown the Gaussian PDF of the data along with the their contours.Another plot is of the confusion matrix which tells the accuracy of the classifier for every classes. The ROC plot tells the accuracy of the classification done by plotting the graph between TPR(True Positive Rate) and FPR(False Positive Rate). The plot with regions is according to the decision boundary which is linear in this case.

Observations:-
From the plot of the contours we observed that if the covariance matrix of the all the classes are same and of the form $\sigma^2 * I$ then the contours will be circular as the variance of the of both axis are same and the axis of the contours are the eigen vectors which are parallel to the standard axis.Further if the covariance matrix is same but not of form $\sigma^2 * I$ then contours will be eclipse with orthogonal axis.If the covariance matrix is different then the contours may or may not be eclipse with orthogonal principal axis and may or may not be parallel to coordinate axis.

In the ROC plot we know that it is the curve between the False Positive Rate(means the classifier has classified it to be target but actually it is non-target) and the True Positive Rate(means the classifier has accurately classified the point). In the ROC plot for the non-linear data and corresponding to the five cases the curves are coming. As we know that to get the non-linear decision boundaries the covariance matrix should be different and for that the accuracy of classification is good. So here in the ROC of non-linear data we observed that for the cases in which the covariance matrix is different is having much more accuracy and among naive bayes and bayesian, bayesian case is having more accuracy.

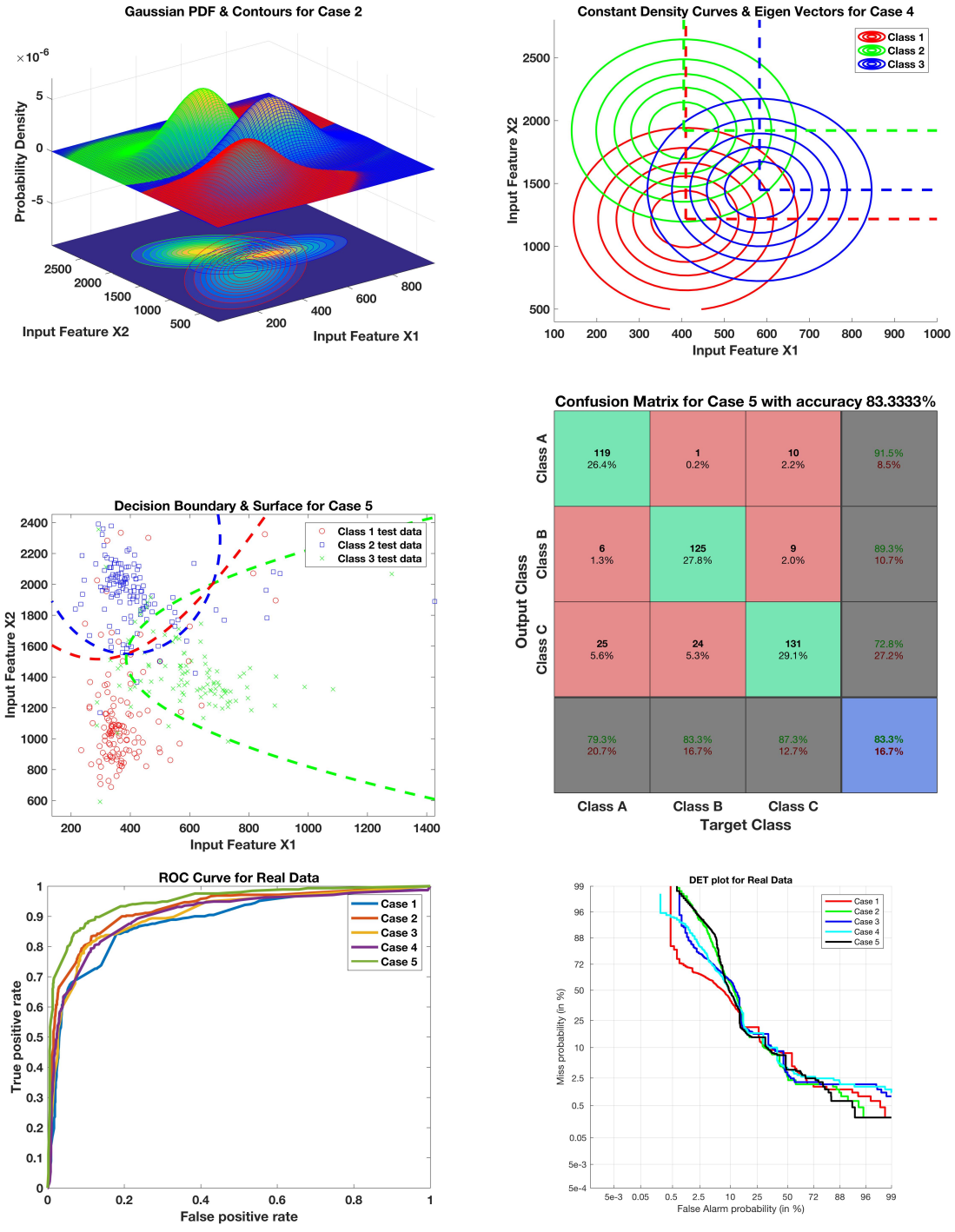

Figure 1.3: Plots for Real Data