University of Tennessee

**Two Category Classification Using Bayesian Decision Rule**

Kirolos Shahat

kshahat@vols.utk.edu

ECE 471: Pattern Recognition

Dr. Hairong Qi

02/06/18

**Abstract**

Data is often discovered with little to no prior domain knowledge. For data such as this there needs to be a way to analyze the features to be able to make a decision about new unknown data with a success rate which is superior to randomly guessing. Thus, with some assumptions about the distribution and the use of Bayes Rule, we can create decision rules and analyze how they perform through supervised learning. Three discriminant function cases were used to make a decision on synthetic data, one which only assumed that the data follows a Gaussian distribution and two which also incorporated other assumptions into the calculation. Case one assumed that the features followed a Gaussian distribution, had the same variance, and that they are independent. Case two assumed that the features were Gaussian and shared the same covariance matrix. Case three was more general and only assumed that the features were Gaussian. The result was that case three outperformed case one by about 18% and case two by about 0.3%, based on the assumption that the prior probabilities were equal. This result shows the importance of knowing the relationship between the features of a category and how critical it can be in making a correct decision versus an incorrect one.

**Introduction**

There is so much data currently available and there is a need for it to be analyzed in order to make predictive decisions on new, related, data. It is often difficult to find a mathematical representation of data such that there is a way to be predictive given new data at a rate that is not equivalent to randomly guessing. With the use of probability, namely Bayesian statistics, one can approximate the likelihood of an event occurring as well as the probability of that same event not occuring with the use of field data and a model.

An experiment was performed on synthetic data in which a training and testing set was given. The experiment was one which tests accuracy of models given these different levels of approaches. The first approach, which is labeled as case one, had three basic assumptions: The data follows a Gaussian Distribution, the features are independent of each other, and that the variance among features is equivalent among the features and classes. The second approach, which is labeled as case two, has two basic assumptions: The data follows a Gaussian Distribution and that the covariance matrices among classes is the same. The third approach, which is labeled as case three, has only one assumption: the data follows a Gaussian Distribution. Thus, this experiment is meant to test the accuracy of these approaches as well as compare them to each other to decipher how assumptions lead to different results, which one would logically think would often converge towards inaccuracy. The experiment was very successful and resulted in at least 71% accuracy for the worst performing case and 89% for the best performing case under the assumption that the two classes had equal prior probabilities.

**Technical Approach**

The three cases of discriminant functions which were used all required an estimation of the parameters of a Gaussian Distribution function, thus that was the first step taken. The data was read in and the mean was computed for the x values of class 0, the y values of class 0, the x of class 1, and the y values of class 1. The mean of class 0 was stored as a column vector and the same for the mean of class 1. Using those values of the mean, we were able to obtain covariance matrices for each class in the training set. Because of the assumption that the data follows a Gaussian distribution then we were able to compute the mean by simply averaging the values of the means for each case.

Once those values were obtained, we then could begin our method of comparing the three cases. For all three cases, a class was chosen if the product of its Probability Density Function (PDF) with the prior probability was greater than that of the class not chosen. Thus, for all three cases the discriminant function was the product of the PDF, which was Gaussian Multivariate, and the prior probability.

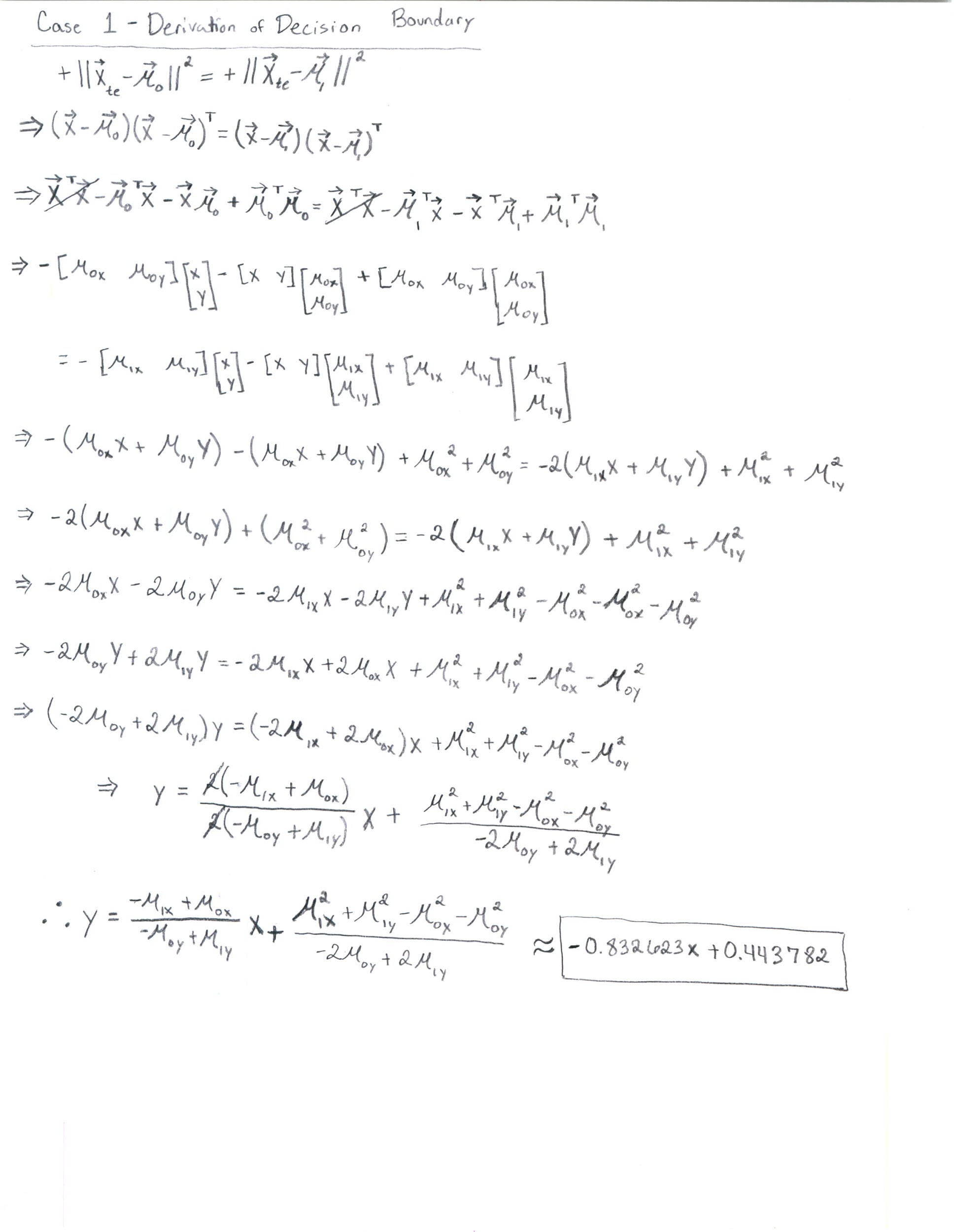
For case one, the assumptions were that the features had no correlation among each other and that the covariance matrix i **= 2**I, where 2 is the variance of one of the classes. Thus, this allows for simplifications in the PDF in which it becomes the Euclidean Distance. The variance we chose was that of x for class zero. For case two, the assumption was that i on top of the original assumption of the data being Gaussian. Thus, this allows for the simplification of the discriminant function to become Mahalanobis distance. The covariance matrix which was chosen for this was that of class 0. Case three was the last discriminant function which was considered. This case only had the assumption that the data followed a Gaussian Distribution and thus could not viably be simplified.

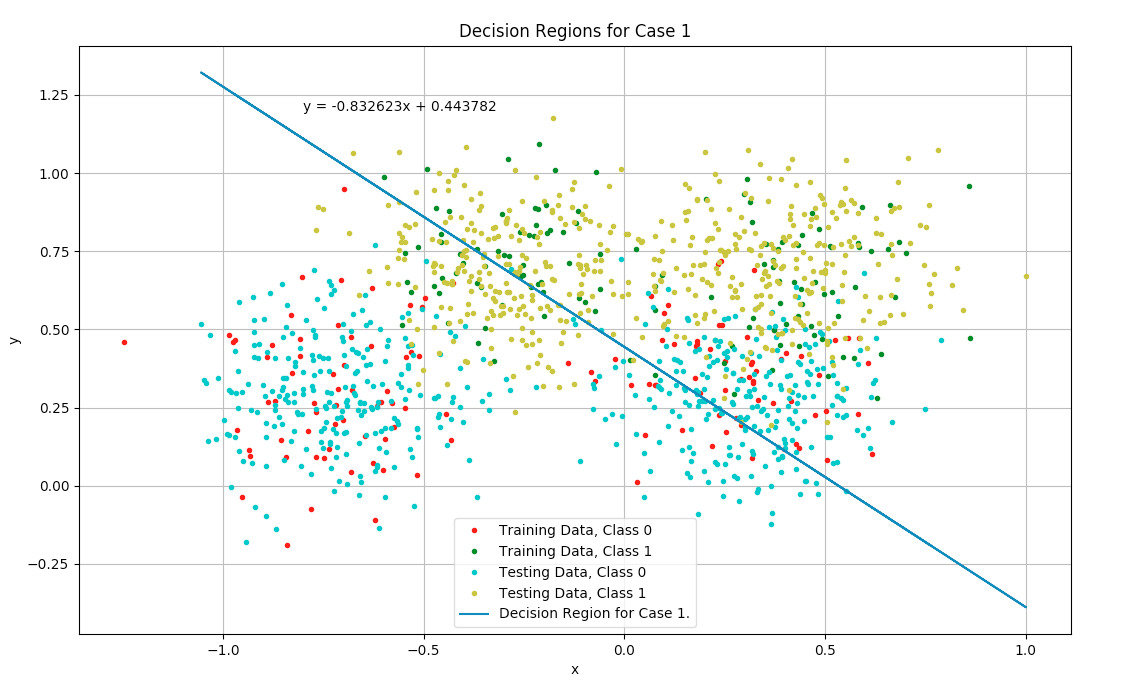
**Results**

After obtaining the derivation of the mean and covariance matrices, we were able to estimate the parameters of the Multivariate Gaussian Distribution. The values obtained were as follows:

|  |  |
| --- | --- |
| **0** | 0, x= -0.22147  0, y = 0.325755 |
| **1** | 1, x= 0.0759543  1, y = 0.682969 |
| **0** | 0 (x,x) = 0.27681 0 (x,y) = 0.0112287  0 (y,x) = 0.0112287 0 (y,y) = 0.0361191 |
| **1** | 1 (x,x) = 0.159748 1 (x,y) = -0.015575  1 (y,x) = -0.015575 1 (y,y) = 0.0299584 |

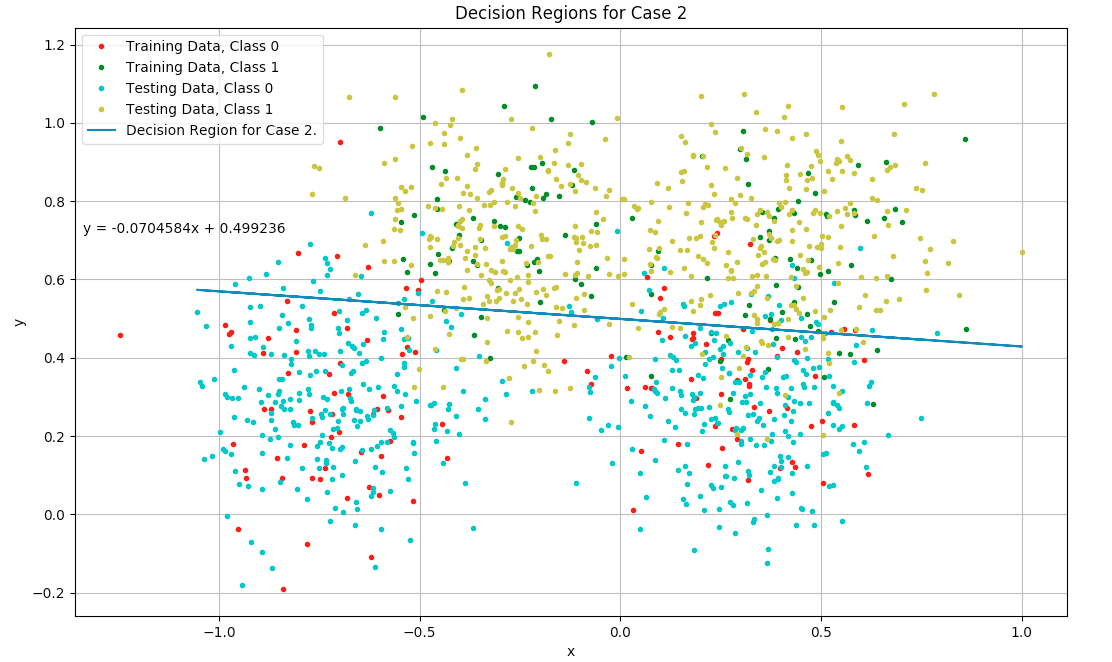
This result shows that the two features are correlated in some sense because of non-zero off-diagonal elements in the two covariance matrices. Using this result and the knowledge that the case one discriminant function being the Euclidean Distance formula, we were able to derive an equation for the decision boundary which is as follows:



Thus, noticing that the equation is one of linearity we can sketch the result and obtain the following: 

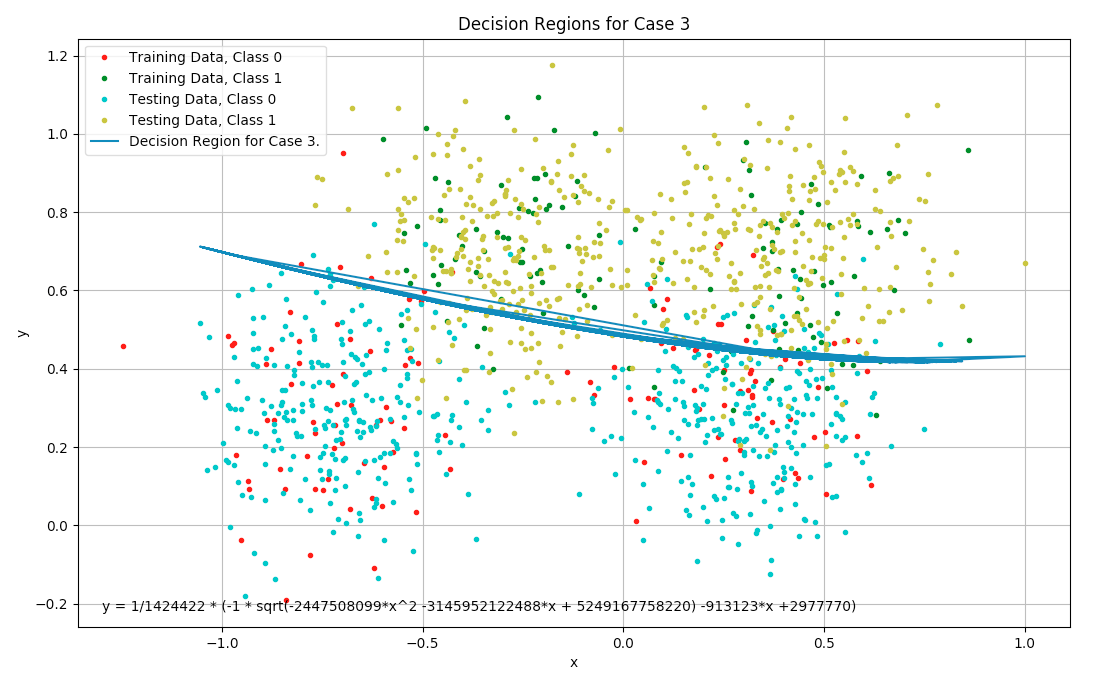
This result already seems to be fairly accurate. The accuracy of this result was 71.2% on the testing data which accomplishes the goal of attaining a success rate of greater than randomly guessing where the variance used was that of x in class zero.

Considering case two, which assumed that the covariance matrices were all the same. The matrix we chose was that which produced the most accurate results which is the one based on class zero. An equation for the boundary was also derived for that one using Mahalanobis Distance and the equation of the line which was derived was:

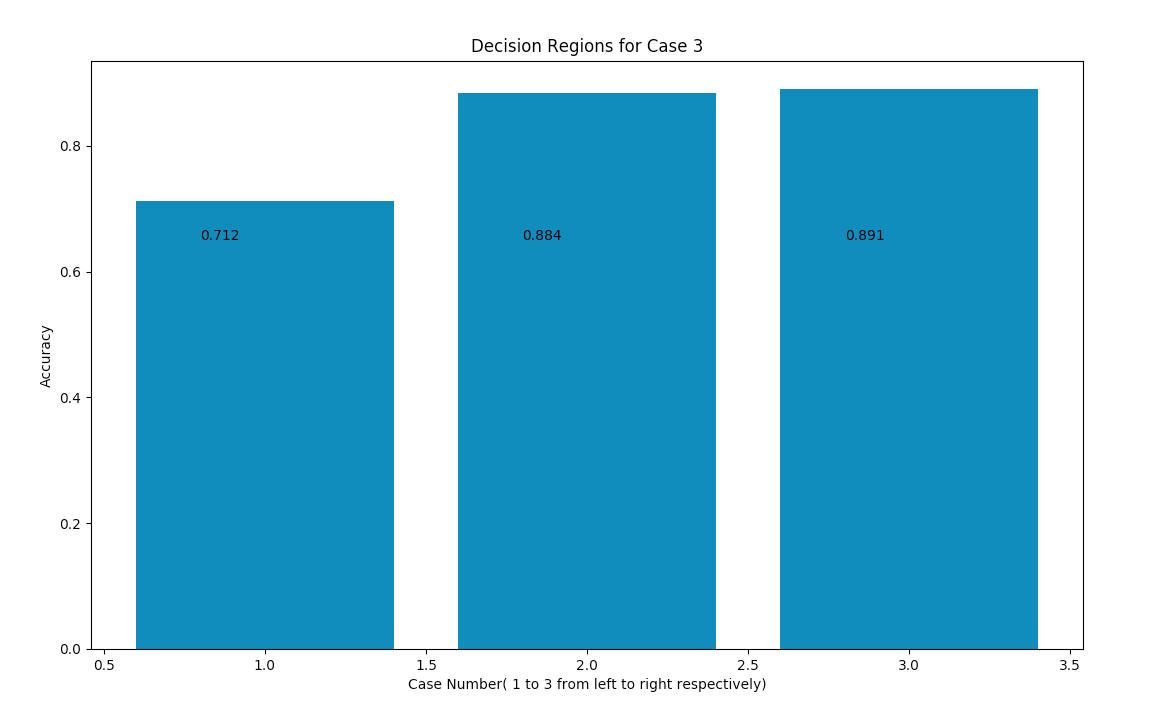
When the boundary was plotted the following was the result: 

Thus it appears to visually be more accurate and the accuracy score defends that notion. Case two had an accuracy 88.4% when covariance matrix of class zero was used which is already a tremendous leap in success.

Case 3 only has the assumption of the data obeying a Gaussian distribution, thus there were no simplifications to be made in the Multivariate Gaussian equation, thus attempting to find the decision boundary of case three was found by using the original Multivariate Gaussian Equation. The result of this was the following solution:

When plotted, this equation leads to the following curve:

Which seems yet more accurate than case 2 and the accuracy score defends that notion. Case three had an accuracy score of 89.1% which is the highest of the three. The three cases are compared below to compare performance based on equal priors:



The last assumption that was made was that all cases assumed equal prior probability and thus a test was performed where different prior probabilities were tested with case three in steps of 0.1 and the result was as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 = 0.1  1 = 0.9  Performance = 76.9% | 0 = 0.2  1 = 0.8  Performance = 83.4% | 0 = 0.3  1 = 0.7  Performance = 87.1% | 0 = 0.4  1 = 0.6  Performance = 88.3% | 0 = 0.5  1 = 0.5  Performance = 89.1% |
| 0 = 0.6  1 = 0.4  Performance = 88.4% | 0 = 0.7  1 = 0.3  Performance = 86.7% | 0 = 0.8  1 = 0.2  Performance = 82.6% | 0 = 0.9  1 = 0.1  Performance = 75.2% | 0 = 1  1 = 0  Performance = 50% |

Thus, coincidently, our assumption of equal prior probabilities gave us the best performance.

**Discussion**

Comparing the results of the three cases one notices that case one was significantly outperformed by cases two and three and one can wonder why that is. Assuming that there is no relationship between features is a very big assumption to make because it disconnects two features completely from one another and thus cannot infer data from one that links to another. Knowing how two features connect together or revolve around each other allows for a significant increase in accuracy and thus classifier performance. Case two and three had very little change in accuracy difference and thus implies that adding more covariance matrices do not add much to the classifiers. This could also be a result of low dimensionality or other unknown reasons but the percentage change fairly insignificant but usually still worth the extra effort.

All of the cases assumed equal prior probabilities originally and thus allowed us to test our cases on an even playing field. But afterwards it was important to test other prior probabilities so that one can be sure the best results are being achieved. It turned out that equal priors was the optimal assumption but that seemed very coincidental and thus not something that should regularly be assumed. Perhaps, when domain knowledge is lacking, one can use equal priors as an initial guess and expand to finding an optimal solution. One should try to avoid assumptions in general and thus a future goal would be to find a way to not rely on assumptions of distribution type and potentially limit our performance but to attempt and find an optimal distribution which supports the type of the data.

**Conclusion**

To conclude, a discriminant function was tested on synthetic data. The training process included estimating the parameters for the multivariate Gaussian distribution. Once those were calculated we were able to compare the performance of each case and discover that with less assumptions made, the better the classifier performs. There were many assumptions put into case one and it had the lowest performance of the three cases, there were less assumptions put into case two and it significantly boosted performance with respect to case one, and case three had very little assumptions and that led to slight increase with respect to case two. Luckily, the assumption that the prior probabilities were equal as well as the assumption that the data obeyed a Gaussian Distribution were fine assumptions for this data, in general it is better to be able to use domain knowledge to get an advantage in classification. With these results, one can hope to be able to significantly improve implementation and performance by knowing the features that affect performance and attempting to not make assumptions on those.