|  |
| --- |
| **Maximum Likelihood Estimation and Face Detection** |
| **Josh Gleason and Rod Pickens** |
| **Programming Assignment 2** |
| **Computer Science 679 – Pattern Recognition, UNR, Dr. Bebis** |
| **Due: Monday March 23, 2015** |
| **Submitted: Monday March 23, 2015** |

Contents

[1 Abstract 3](#_Toc414124669)

[2 Technical Discussion 3](#_Toc414124670)

[2.1 Maximum Likelihood Estimation 3](#_Toc414124671)

[2.1.1 Multivariate Normal Distribution 5](#_Toc414124672)

[2.2 Classifiers 7](#_Toc414124673)

[2.2.1 N-Class classifier (JG) 7](#_Toc414124674)

[2.2.2 One class classifier (JG) 8](#_Toc414124675)

[2.2.3 Comparison (JG) 9](#_Toc414124676)

[2.3 Face Detection (RP) 9](#_Toc414124677)

[3 Project 9](#_Toc414124678)

[3.1 Problem 1 & 2 (JG) 9](#_Toc414124679)

[3.1.1 Problem 1 10](#_Toc414124680)

[3.1.2 Problem 2 11](#_Toc414124681)

[3.2 Problem 3 (RP) 12](#_Toc414124682)

[4 Results 12](#_Toc414124683)

[4.1 Maximum Likelihood Estimation (JG) 12](#_Toc414124684)

[4.1.1 Problem 1 (JG) 12](#_Toc414124685)

[4.1.2 Problem 2 (JG) 17](#_Toc414124686)

[4.2 Face Detection (RP) 22](#_Toc414124687)

[4.2.1 Problem 3a (Chromatic Color Space) (RP) 22](#_Toc414124688)

[4.2.2 Problem 3b (YCbCr Color Space) (RP) 23](#_Toc414124689)

[4.2.3 Problem 3c (RGB Color Space) (RP) 23](#_Toc414124690)

[4.2.4 Conclusion 23](#_Toc414124691)

[5 Conclusion 24](#_Toc414124692)

[6 Contributors 24](#_Toc414124693)

**Bayesian Classification of Classes with Features that are Normally Distributed**

# Abstract

This paper describes our research regarding the first class project for the Computer Science (CS) pattern recognition class CS 679 taught by Dr. Bebis Department Chair of the Computer Science Department at the University of Nevada in Reno, Nevada.

The topics included in this main body of this paper are a description of the project, the classifiers used in the project, and the results of the classification. The topic discussed in the appendix are a description of the theory of a Bayesian classifier and the derivation of the two classifiers used in the project, the theoretical errors expected from a minimum risk Bayesian classifier with a zero-one loss function, as well as the generation of normal Gaussian random variables and the evaluation of the Gaussian random numbers using a Goodness of Fitness test.

# Technical Discussion

Bayesian Classifiers generally require knowledge of the both the likelihood and prior probability for all classes. In the previous project, the likelihood and prior probabilities were given. In this project it is assumed that the samples come from distributions with known form and unknown parameters.

## Maximum Likelihood Estimation

Maximum Likelihood (ML) is one method of approximating the likelihood functions given that the parametric form of the distributions is known. Equation (1) describes the parametric form of the likelihood function with representing parameter number for class

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

For simplicity, the parameters will be manipulated in vector form

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

In order to find parameters which most accurately describe the likelihood function, training data is used. One set of training data is used for each class and that training data consists of independently drawn samples. In other words, the training data used to determine the likelihood function for class are samples drawn from a random variable with probability density function . The training data for class is represented as follows

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

The premise of ML is to determine the parameters of the likelihood function that would be most likely to produce the training data. This leads to the following definition for the parameters estimated by maximum likelihood.

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Where is an estimate of the true parameters for the likelihood function.

Because the samples in the training data were drawn independently, the probability density function in the right hand side of Equation (4) may be expressed as

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Because Equation (4) is an optimization, and probability density functions are always positive, we can take the expression in the maximization and apply a log function (or any other monotonically increasing function). By also substituting in Equation (5), Equation (4) becomes

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

The logarithm makes dealing with the maximization problem easier in many cases. For example, when the model for the likelihood function is a Normal distribution as we will see in Section 2.1.1.

In order to solve the Equation (6) we take the gradient of the optimization expression and find where it becomes zero.

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

where

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

Solving for all solutions to Equation (7) will give the position of all local maxima, local minima, and inflection points. If not apparent, it is important to test each solution as well as the bounds of the function to determine where the true global maxima lies.

### Multivariate Normal Distribution

One interesting case to consider for ML is when likelihood takes the form of a Multivariate Normal distribution with an unknown mean and unknown covariance .

The parameter vector then consists of the mean and covariance of the distribution.

To begin the derivation we substitute the equation for the Multivariate Normal distribution into Equation (7)

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

Because is composed of and we express Equation (9) as the following scalar-matrix derivatives

|  |  |  |
| --- | --- | --- |
|  |  | (10) |
|  |  | (11) |

Equation (10) may be solved as follows

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

Similarly, Equation (11) may be solved as follows (substituting for )

|  |  |  |
| --- | --- | --- |
|  |  | (13) |

Together, the results from Equations (12) and (13) express the ML estimates for the parameters of the Multivariate Normal distribution as

|  |  |  |
| --- | --- | --- |
|  |  | (14) |
|  |  | (15) |

which are the estimates that will be tested in this project.

## Classifiers

In this project we again make use of a Bayesian classifier so a brief recap of the theory is warranted.

In pattern recognition a classifier is a machine that makes a decision among candidate choices given the available evidence conditioned on the choice under consideration. Conditioned evidence may be described by an n-dimensional feature vector as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (16) |

Each dimension of the vector is a measurement of one aspect of the object being evaluated prior to that object being classified.

### Multiclass classifier (JG)

The Multiclass Bayesian classifier is the general case of the Bayesian classifier in that it classifies conditioned evidence into one of the states nature with . In general, a state of nature is unpredictable and must therefore be described probabilistically.

The goal of a Bayesian Classifier is to classify conditioned evidence in such a way that overall risk is minimized. Risk is defined by Hart[[1]](#footnote-3) as

|  |  |  |
| --- | --- | --- |
|  |  | (17) |

Where the function is the conditional risk associated a decision rule given a piece of conditioned evidence.

The minimum error classifier is a special version of the minimum risk where

|  |  |  |
| --- | --- | --- |
|  |  | (18) |

And the decision corresponds to the action of classifying as class which can be expressed mathematically as .

This reduces to the following decision rule (the shorthand is used to represent

|  |  |  |
| --- | --- | --- |
|  |  | (19) |

which simply defines the decision rule as choosing the class which yields lowest a posteriori probability.

Because this is a minimization problem, the posteriori probability doesn’t need to be computed directly. Instead we can apply Bayes formula to Equation (19) and remove the scaling factor from the minimization yielding

|  |  |  |
| --- | --- | --- |
|  |  | (20) |

Furthermore, we can apply the logarithm function to the minimization portion of (20) which then becomes

|  |  |  |
| --- | --- | --- |
|  |  | (21) |

The functions which are being minimized is generally known as the set of discriminate functions which serve to simplify our discussion. While function which maintains the minimization may be used, a commonly used discriminate function is the one in Equation (20)

|  |  |  |
| --- | --- | --- |
|  |  | (22) |

When classifying a sample, the classifier assigns the state of nature which yields the lowest value for the corresponding discriminate function.

### One class classifier (JG)

### Comparison (JG)

## Face Detection (RP)

# Project

This project consists of 3 sections, all of which apply Maximum Likelihood estimation for learning models for Bayesian classification.

## Problem 1 & 2 (JG)

Parts 1 and 2 of the project use the random samples generated in the first project and attempt to learn the distributions using Maximum Likelihood estimation. Classification is then performed using the estimated models and the results are compared to results from project 1 where the actual models are known.

The following sequence will be used to implement the experiment.

1. Load two class test data used in Project 1
2. Estimate the probability density functions using Maximum Likelihood Estimation using a subset of the test samples.
3. Classify all test samples using a Bayesian Classifier designed using the estimated likelihood function with equal priors.
4. Count misclassified samples and compare with project 1 results.

As in project 1, the Bayesian classifier used will be a dichotomizer which examines the difference between discriminant functions for class and as follows

|  |  |  |
| --- | --- | --- |
|  |  | (23) |

Where the general discriminant function for class is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (24) |

Where

|  |  |  |
| --- | --- | --- |
|  |  | (25) |

and

|  |  |  |
| --- | --- | --- |
|  |  | (26) |

and

|  |  |  |
| --- | --- | --- |
|  |  | (27) |

where the parameters and are the Maximum Likelihood estimation parameters for class .For Multivariate Normal Distributions, the Maximum Likelihood estimation turns out to be the sample mean and sample covariance. The derivation of the Maximum Likelihood solution is given in Section 2.1.1.

Also, in this experiment we are no longer using the linear version of the discriminant function in problem 1 since the Maximum Likelihood estimate is not guaranteed to result in parameters with the same properties as the true distributions.

The required experiments for the project are to use 10,000 and 1,000 samples for the Maximum Likelihood estimation. However, we will perform experiments using both 100 and 10 training samples as well. The primary goal behind the extra experiments to gain a better understanding of the effect different sized training data sets have on a simple two dimensional, two-class classifier.

As an extra experiment, all the experiments for Problems 1 and 2 will be run 1000 times with different sets of random data. The classification error rate for each experiment will be recorded as the result of each experiment. The sample mean and variance of the classification error rates will then be visualized and examined in order to gain insight into how reliably a classifier may be built given a particular number of training samples.

### Problem 1

Problem 1 uses samples drawn from multivariate distributions with independent features and unity variance.

The true parameters for the normal distribution for class are:

|  |  |  |
| --- | --- | --- |
|  |  | (28) |

and the true parameters for the normal distribution for the class are:

|  |  |  |
| --- | --- | --- |
|  |  | (29) |

As a comparison, Figure 1 shows the classification results from Project 1 where the classifier was constructed with the true likelihood function known with equal priors. If the estimation is working correctly, the classification error for this project should be close to the 1.72% error rate achieved in project 1.

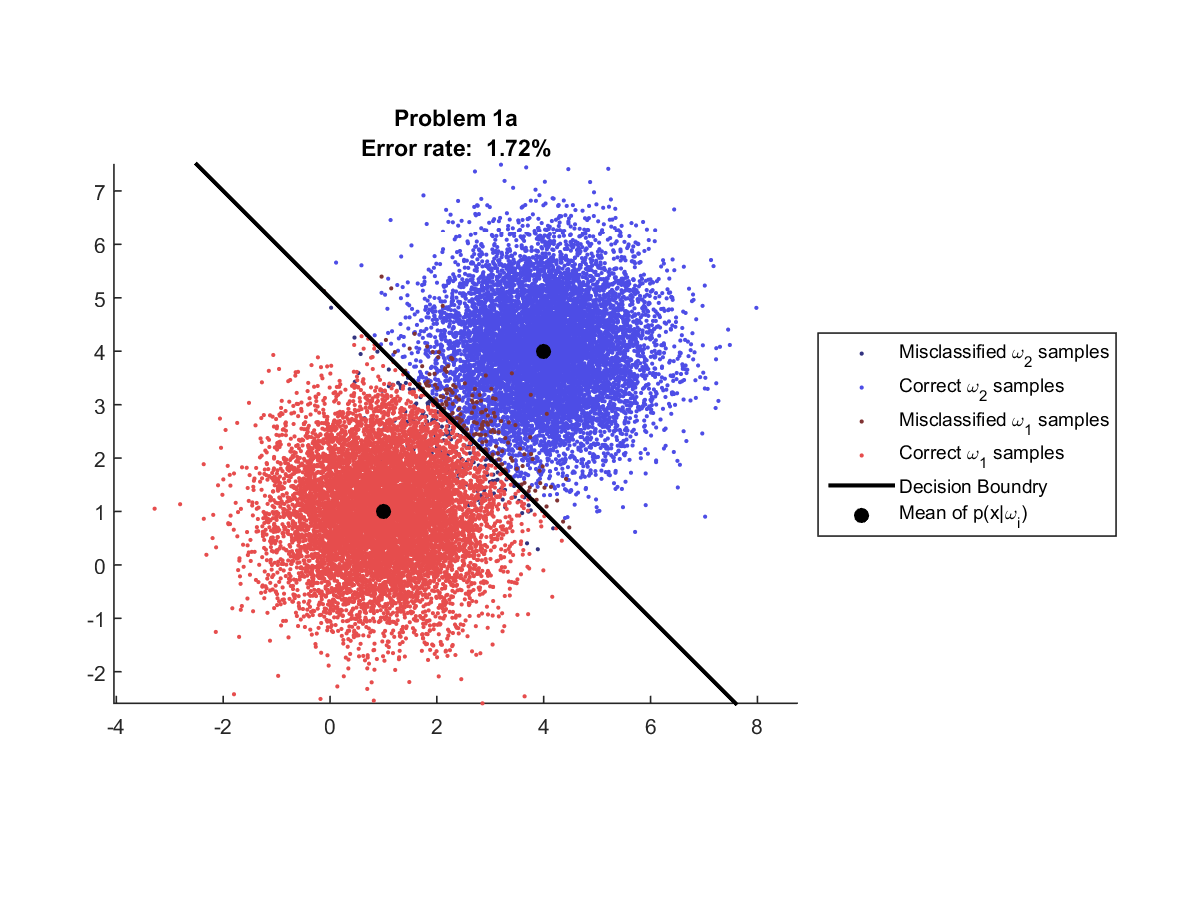


Figure : Project 1, Problem 1 classification results with equal priors

### Problem 2

Problem 2 uses samples drawn from multivariate normal distributions with independent features.

The true parameters for the normal distribution for class are:

|  |  |  |
| --- | --- | --- |
|  |  | (30) |

and the true parameters for the normal distribution for the class are:

|  |  |  |
| --- | --- | --- |
|  |  | (31) |

As a comparison, Figure 1 shows the classification results from Project 1 where the classifier was constructed with the true likelihood function known with equal priors. If the estimation is working correctly, the classification error for this project should be close to the 7.03% error rate achieved in project 1.

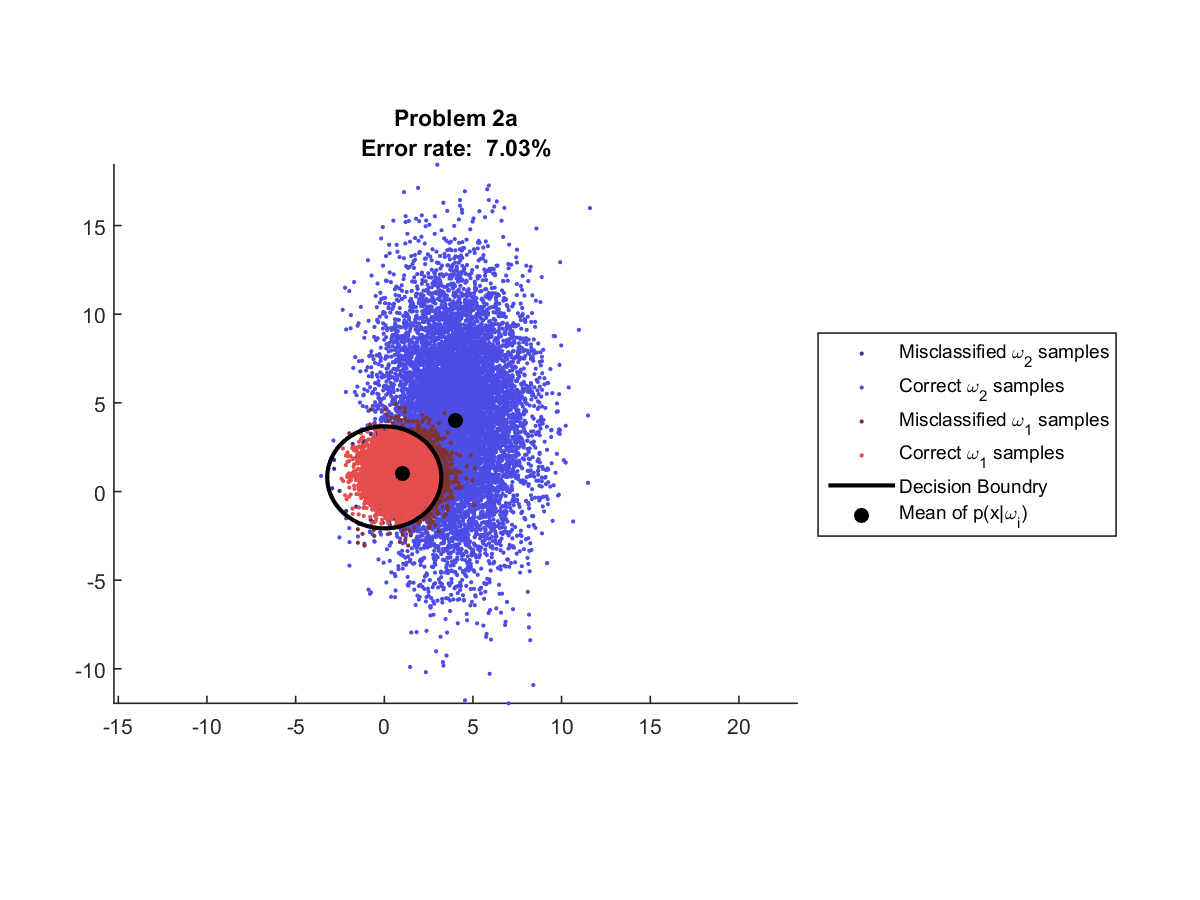


Figure : Project 1, Problem 2 classification results with equal priors

## Problem 3 (RP)

# Results

## Maximum Likelihood Estimation (JG)

### Problem 1 (JG)

For Problem 1, we use the same samples used for each of the two classes from Project 1 problem 1. The true distributions which the samples were taken from are defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (32) |
|  |  | (33) |

The a-priori probabilities are as follows

|  |  |  |
| --- | --- | --- |
|  |  | (34) |

In Figure 3, the two classes are colored differently: class 1 (red dots) is and class 2 (blue dots) is The decision boundary between the two classes is very close to the an affine function (a linear function with an additive intercept) which matches the expectation from Project 1 show in Figure 1.

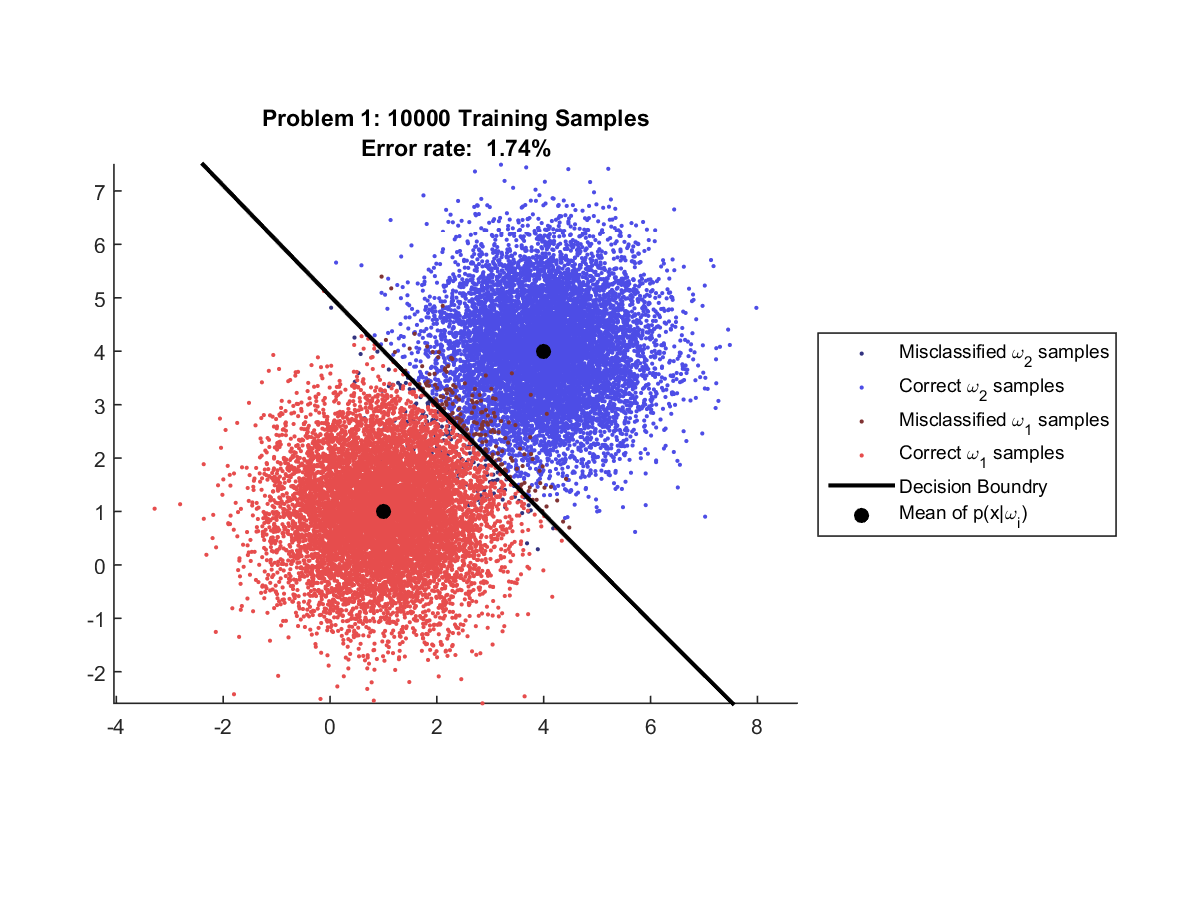


Figure : Classification with 10,000 training samples

The error rate is 1.74% which is very close to the 1.72% reported using the project 1 classifier. This indicates that the Maximum Likelihood estimate using 10,000 is fairly accurate.

The estimated Maximum Likelihood parameters for 10,000 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (35) |
|  |  | (36) |

As expected from the low error rate, these estimates are very close to the true parameters.

Figure 4 shows the classification results using 1,000 training samples. This case appears nearly identical to the 10,000 sample experiment with an even better error rate of 1.73%. The fact that the error rate is very slightly better, even though less training samples were used is due to the random nature of the sample data. This result appears to indicate that 1,000 samples is sufficient to estimate the likelihood distributions and using 10,000 training samples is only marginally better; if at all.

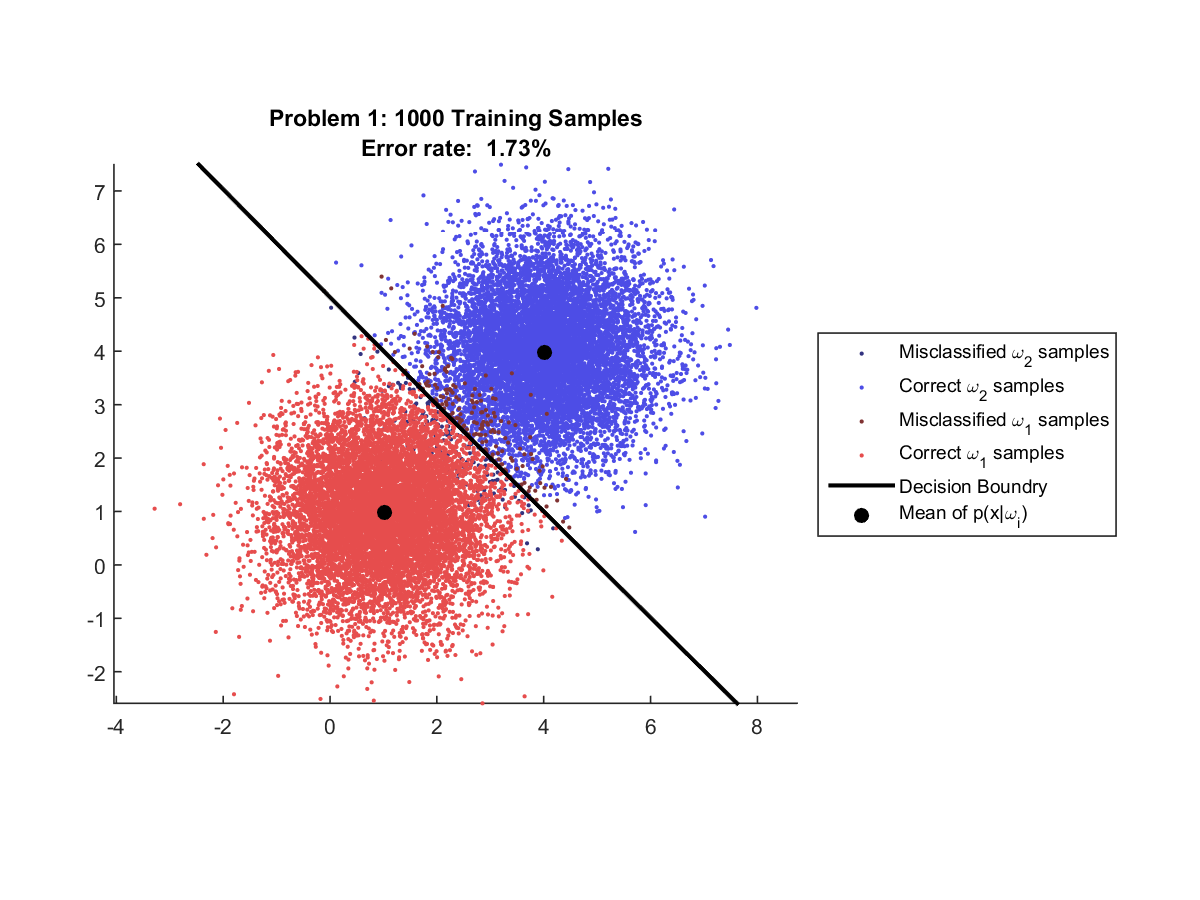


Figure : Classification with 1,000 training samples

The estimated Maximum Likelihood parameters for 1,000 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (37) |
|  |  | (38) |

Surprisingly, even though the parameters estimated with 1,000 samples are further from the truth on average than the 10,000 sample experiment, the error rate is better. This again is probably due to the random nature of the data.

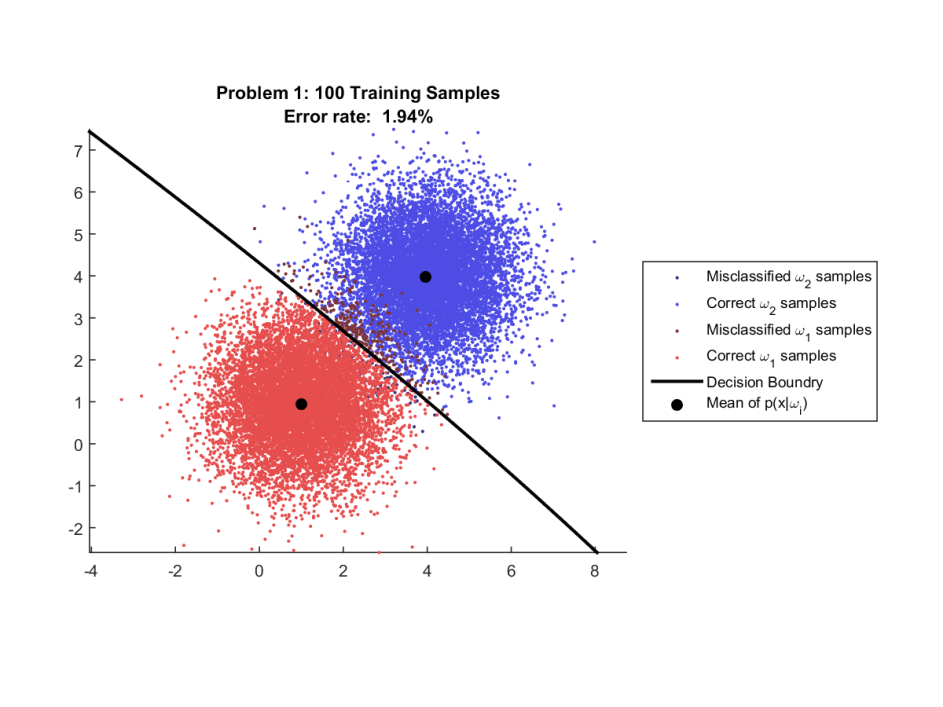
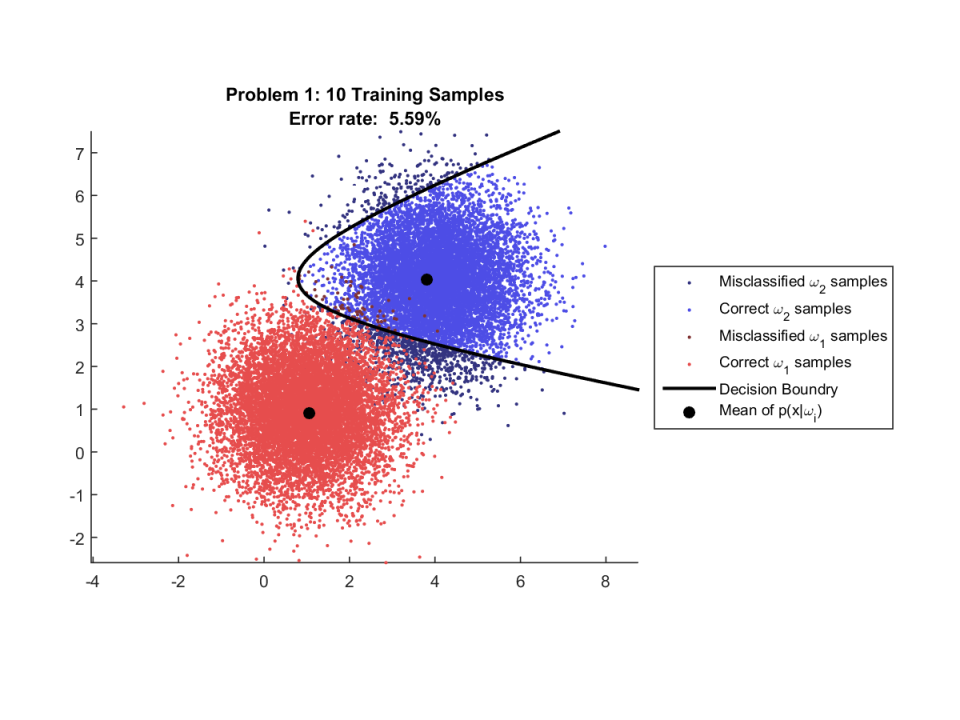
 

Figure : Classification with 100 (left) and 10 (right) training samples

In Figure 5 the classification results using 100 and 10 training samples can be seen. These experiments show the error rate clearly getting higher as the number of samples used for training decreases.

The estimated Maximum Likelihood parameters for 100 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (39) |
|  |  | (40) |

In this experiment, the mean is fairly well estimated, but most of the error appears in the standard deviation estimate. This indicates that an accurate estimate of standard deviation requires more samples than the mean. This makes sense intuitively because the sample variance is calculated using the sample mean. Because of the dependency, errors in the mean will be compounded with the errors in the covariance estimate making the estimate less precise. Also, the covariance is composed of units squared which will make a direct comparison with errors in the less meaningful.

While 100 samples do not lead to a near-truth classifier, the results in this experiment are far from horrible and it is this authors opinion that 100 samples would be sufficient for certain real-world applications.

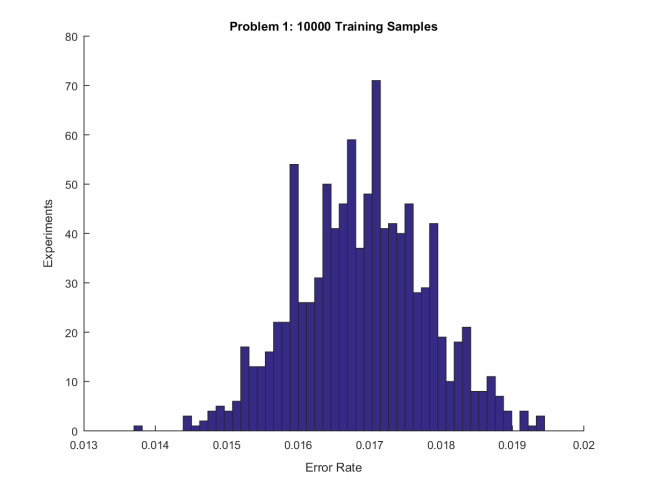
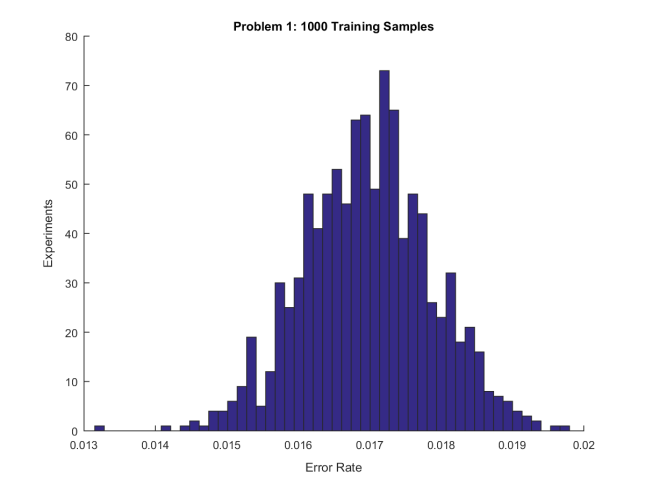
The estimated Maximum Likelihood parameters for 10 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (41) |
|  |  | (42) |

The experiment using 10 training samples gives a very poor estimate of the distribution especially for the covariance estimate. This can be seen by the highly non-linear decision boundary constructed by the classifier. It appears that 10 samples is far from ideal and that generally more samples would be required to obtain an adequate estimate of the likelihood distribution.

#### Problem 1 Extra Experiments

Based on the experiments in Section 4.1.1 we can make certain assumptions about the necessary amount of training data for the simple two class classifier. However, it is not impossible to imagine that the results for the 10 sample experiment were unusually bad or unusually good. If we were to run the experiment again, how well would we expect the classifier to perform? The purpose of this experiment is to answer that question. In this section we run the entire training/classification problem 1000 times using different, randomly generated data sets consisting of 10,000 samples from each class.

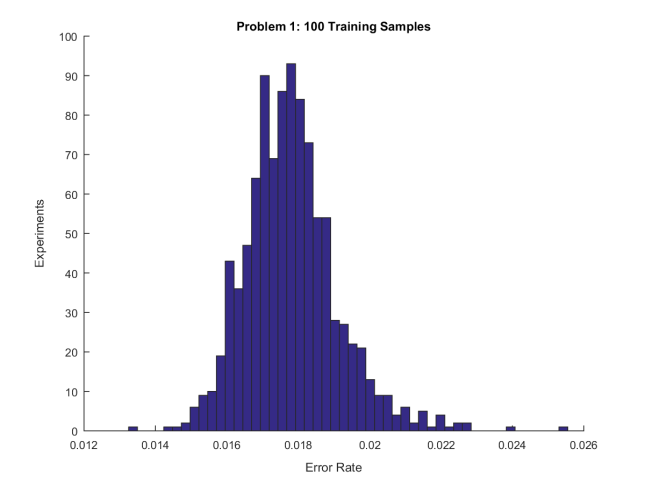
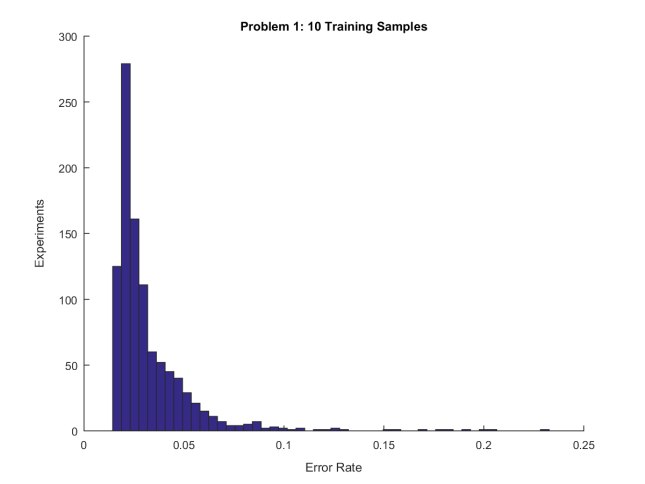
 

Figure : Distribution of Error Rates after 1000 experiments using 10,000 training samples (upper left), 1,000 training samples (upper right), 100 training samples (lower left), and 10 training samples (lower right).

From examining the histograms in Figure 6 we can see that the variance in the classification error when using only 10 training samples is very high, however the variance and expected error in the 100 training sample case is actually quite good. In fact the 100, 1,000, and 10,000 training sample cases appear to be nearly the same. The expected error in the 100 sample case is about 0.08% less than the 1,000 and 0.09% worse than the 10,000 training sample case but this seems to be marginal even with multiple orders of magnitude less in training data. Because this author was still curious about what a good amount of training data would be, we ran one more experiment.

In the final experiment we performed the classification experiment 1000 times but this time used only 3 training samples and worked our way up to 10,000 training samples, plotting the mean and standard deviation of the error rate for various amounts of training data.

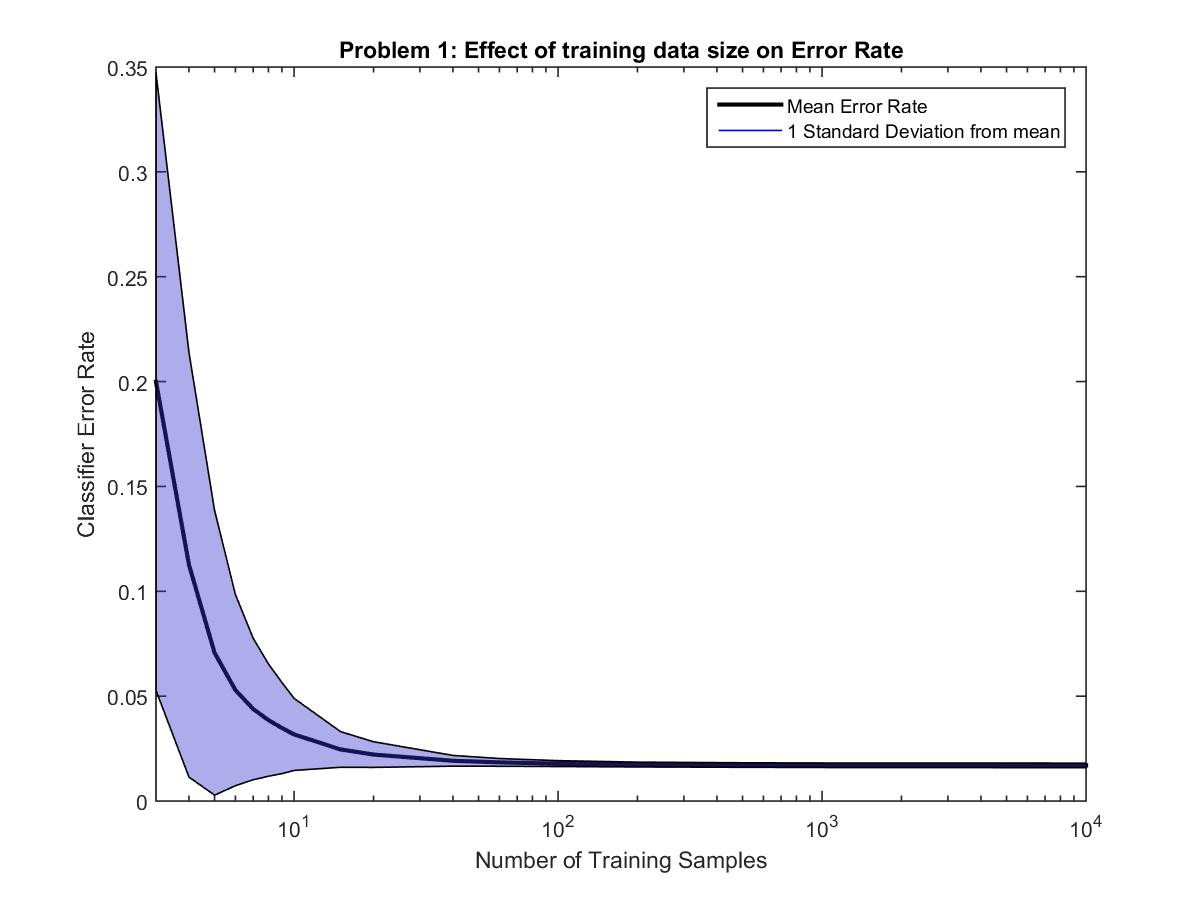


Figure : Effect of amount of training data on Error Rate

Figure 7 shows the results of the experiment. The black line indicates the mean error rate given some number of samples. The blue region is filled one standard deviation above and below the black line.

From the figure it is clear that 100, 1,000, and 10,000 training samples result in a very similar classification accuracy. However, from inspection we see that less than 10 training samples results in very poor performance, with 3 training samples having an average error rate of 20%. The accuracy of the classifier also seems to converge fairly quickly, with 20 or more samples giving surprisingly good results.

### Problem 2 (JG)

For Problem 2, we use the same samples used for each of the two classes from Project 1 Problem 2. The true distributions which the samples were taken from are defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (43) |
|  |  | (44) |

The a-priori probabilities are as follows

|  |  |  |
| --- | --- | --- |
|  |  | (45) |

Figure 8 shows the classification results using 10,000 training samples. The decision boundary in this case appears to be nearly identical to the decision boundary from Project 1 show in Figure 2.

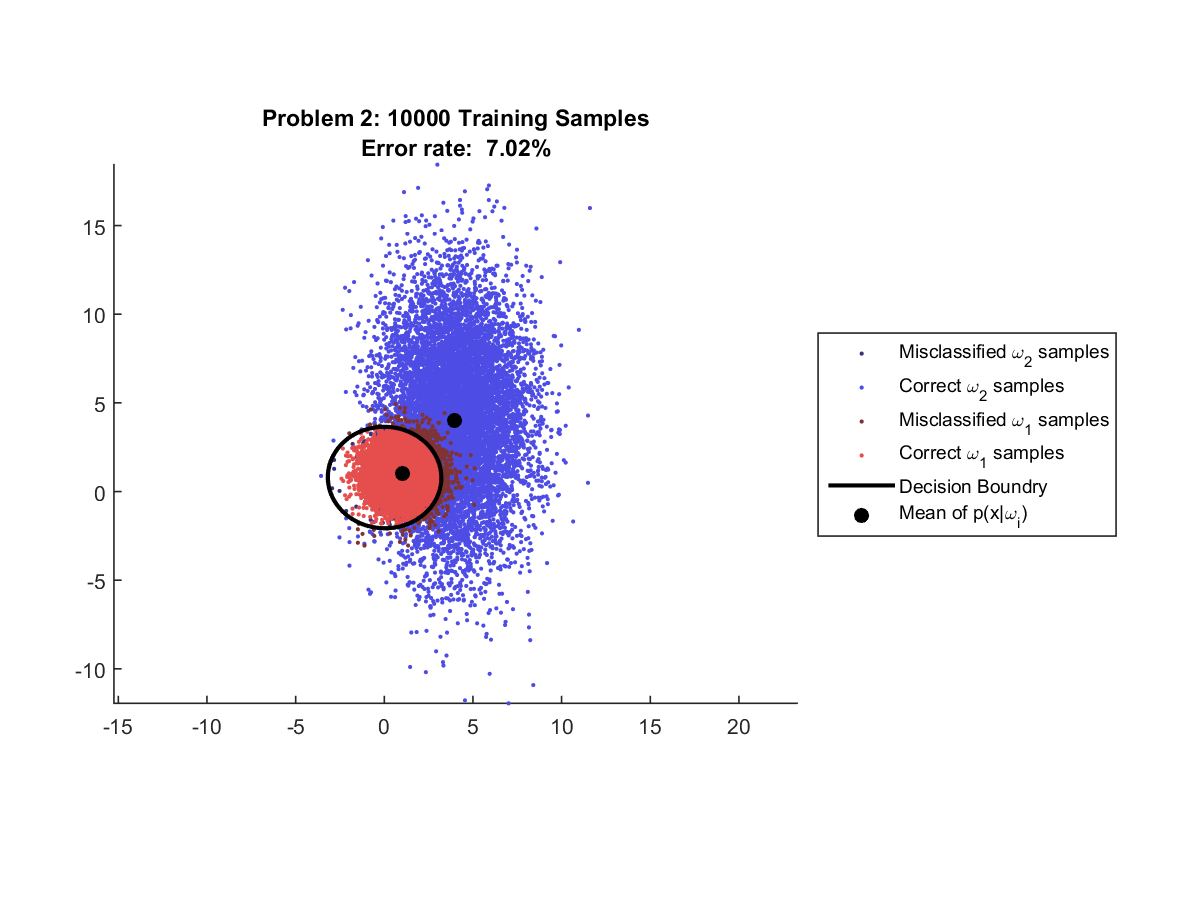


Figure : Classification with 10,000 training samples

The error rate is 7.02% which is actually better than the 7.03% reported using the project 1 classifier. This indicates that the Maximum Likelihood estimate using 10,000 is very accurate.

The estimated Maximum Likelihood parameters for 10,000 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (46) |
|  |  | (47) |

As expected from the low error rate, these estimates are very close to the true parameters. It is worth noting that the mean for class two has more error than the mean for class 1. This is extra error is almost certainly due to the higher variance in class two.

Figure 9 shows the classification results using 1,000 training samples. Unlike the first experiment, the error rate here is somewhat larger than in the 10,000 sample experiment with an even better error rate of 7.22%. This behavior is discussed in more detail in Section 4.1.2.1.

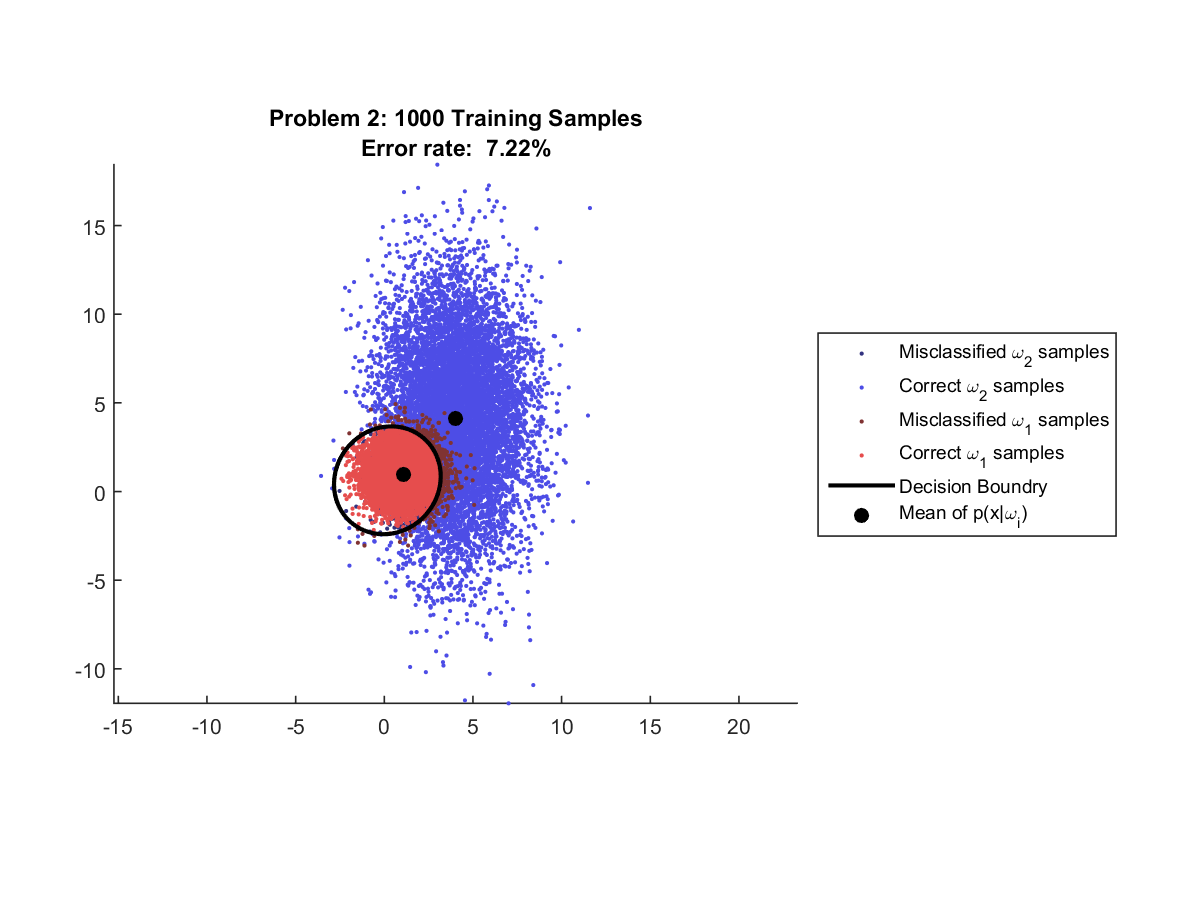


Figure : Classification with 1,000 training samples

The estimated Maximum Likelihood parameters for 1,000 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (48) |
|  |  | (49) |

Similar to problem 1, the parameters estimated with 1,000 samples are further from the truth on average than the 10,000 sample experiment, however, in this case the error rate is noticeably worse.

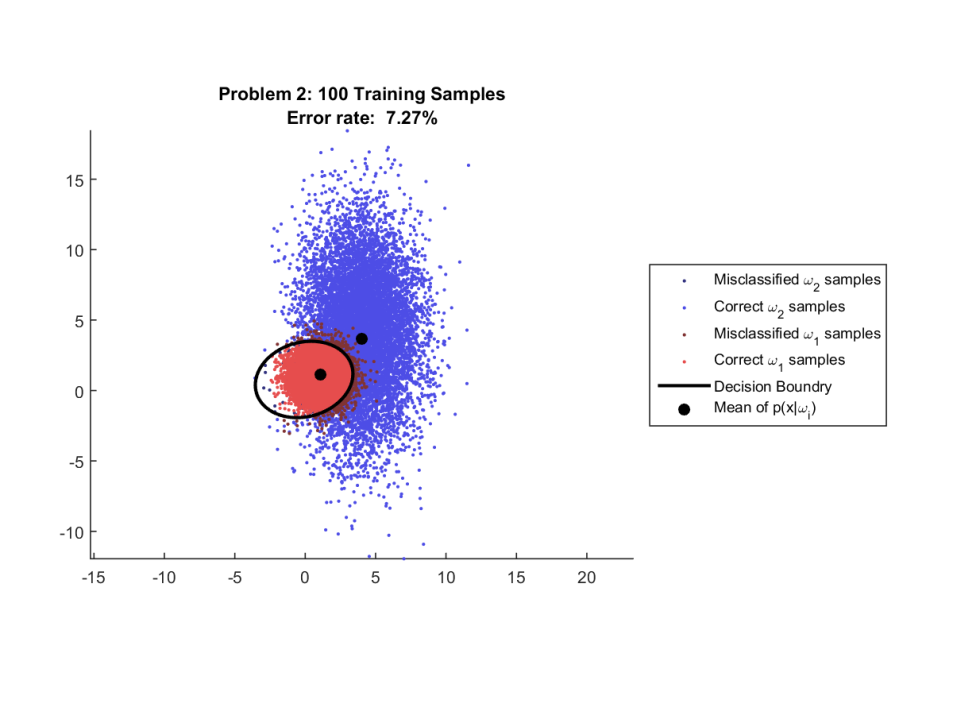
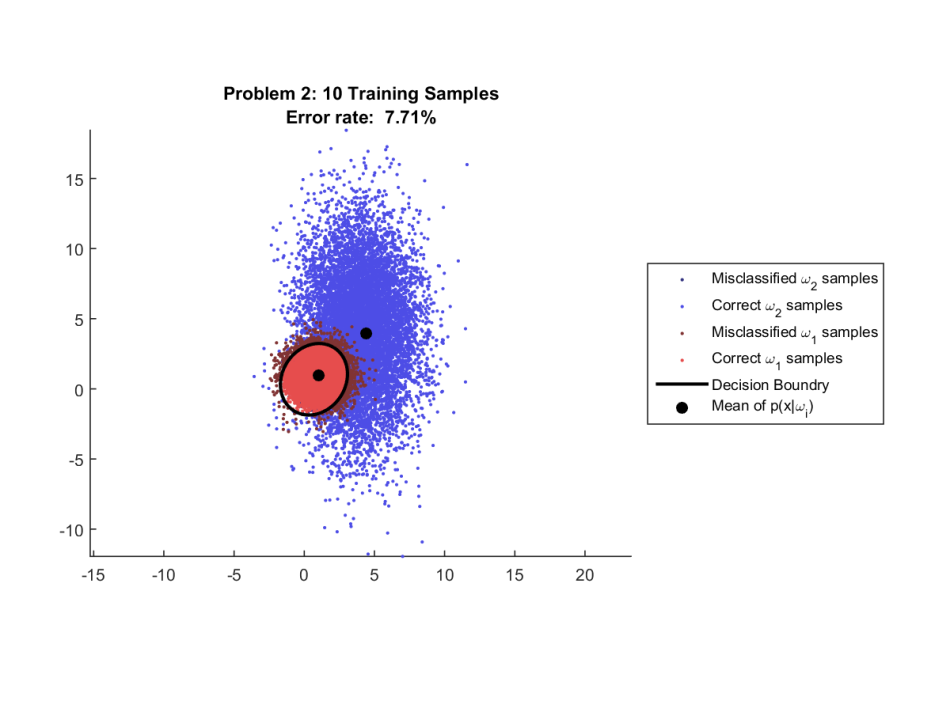
 

Figure : Classification with 100 (left) and 10 (right) training samples

In Figure 10 the classification results using 100 and 10 training samples can be seen. These experiments show the error rate clearly getting higher as the number of samples used for training decreases, however, the magnitude of the error rate increase is actually quite small compared to the drastic 3% increase in problem 1. This result seems to indicate that the number of training samples does not have as drastic of an affect on the classification results as in problem 1.

The estimated Maximum Likelihood parameters for 100 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (50) |
|  |  | (51) |

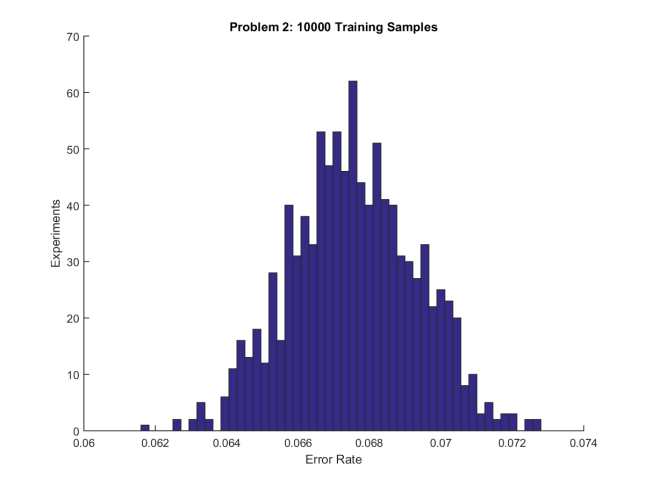
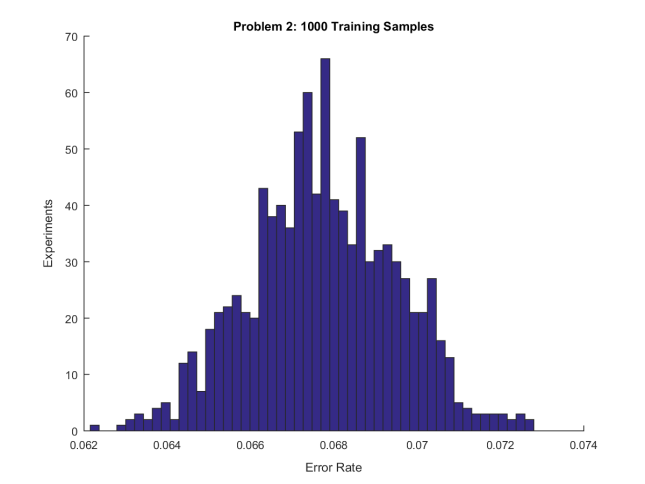
The estimated Maximum Likelihood parameters for 10 training samples are:

|  |  |  |
| --- | --- | --- |
|  |  | (52) |
|  |  | (53) |

The experiment using 10 training samples gives a very poor estimate of the distribution especially for the covariance estimate. However, unlike in problem 1, the decision boundary looks relatively sane. It appears that while 10 samples leads to a less accurate classifier, the difference is much less severe than the problem 1 case.

#### Problem 2 Extra Experiments

The same extra experiments performed in problem 1 (see 4.1.1.1) were also performed for problem 2. By running each of the training plus classification experiments 1000 times we can examine the stability of the classifier as a function of the amount of training data.

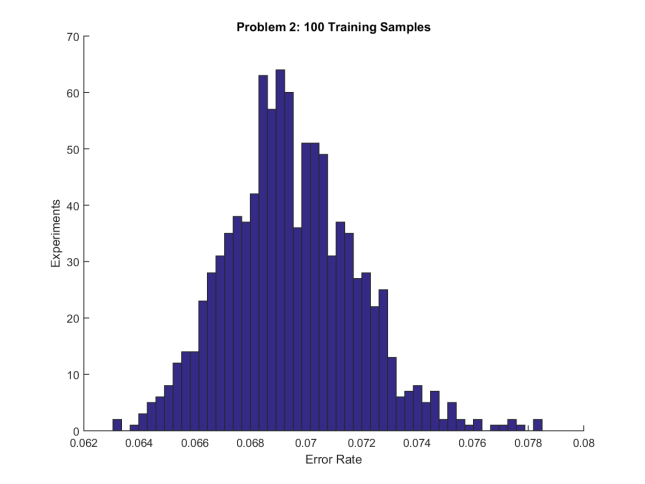
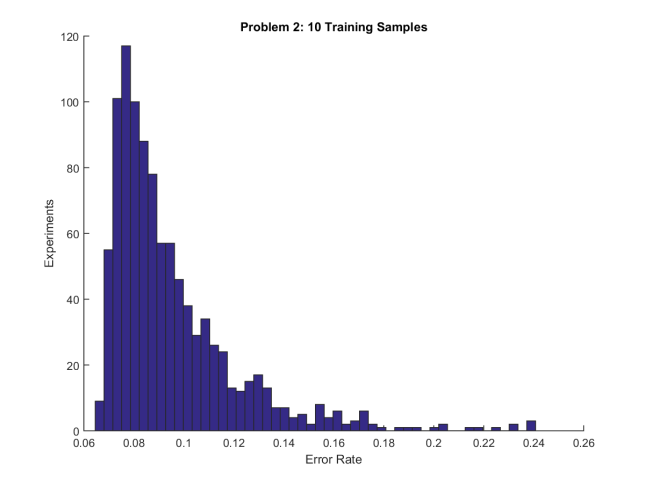
 

Figure : Distribution of Error Rates after 1000 experiments using 10,000 training samples (upper left), 1,000 training samples (upper right), 100 training samples (lower left), and 10 training samples (lower right).

From examining the histograms in Figure 11 we can see that the variance in the classification error when using only 10 training samples is very high, however the variance and expected error in the 100 training sample case is actually quite good. As in problem 1, the 100, 1,000, and 10,000 training sample cases appear to be nearly identical. The variance of the classifier error rate is double what it was in problem 1, even with 10,000 samples. This leads to the behavior of the 1,000 training sample experiment was noticeably worse than the 10,000 training sample experiment. From these histograms it would not be unexpected for the reverse to be true and the 1,000 training sample experiment could have noticeably outperformed the 10,000 training sample experiment. The expected error in the 100 sample case is about 0.18% less than the 1,000 and 0.19% worse than the 10,000 training sample case but this seems to be marginal even with multiple orders of magnitude less in training data.

In the final experiment we performed the classification experiment 1000 times but this time used only 3 training samples and worked our way up to 10,000 training samples, plotting the mean and standard deviation of the error rate for various amounts of training data.

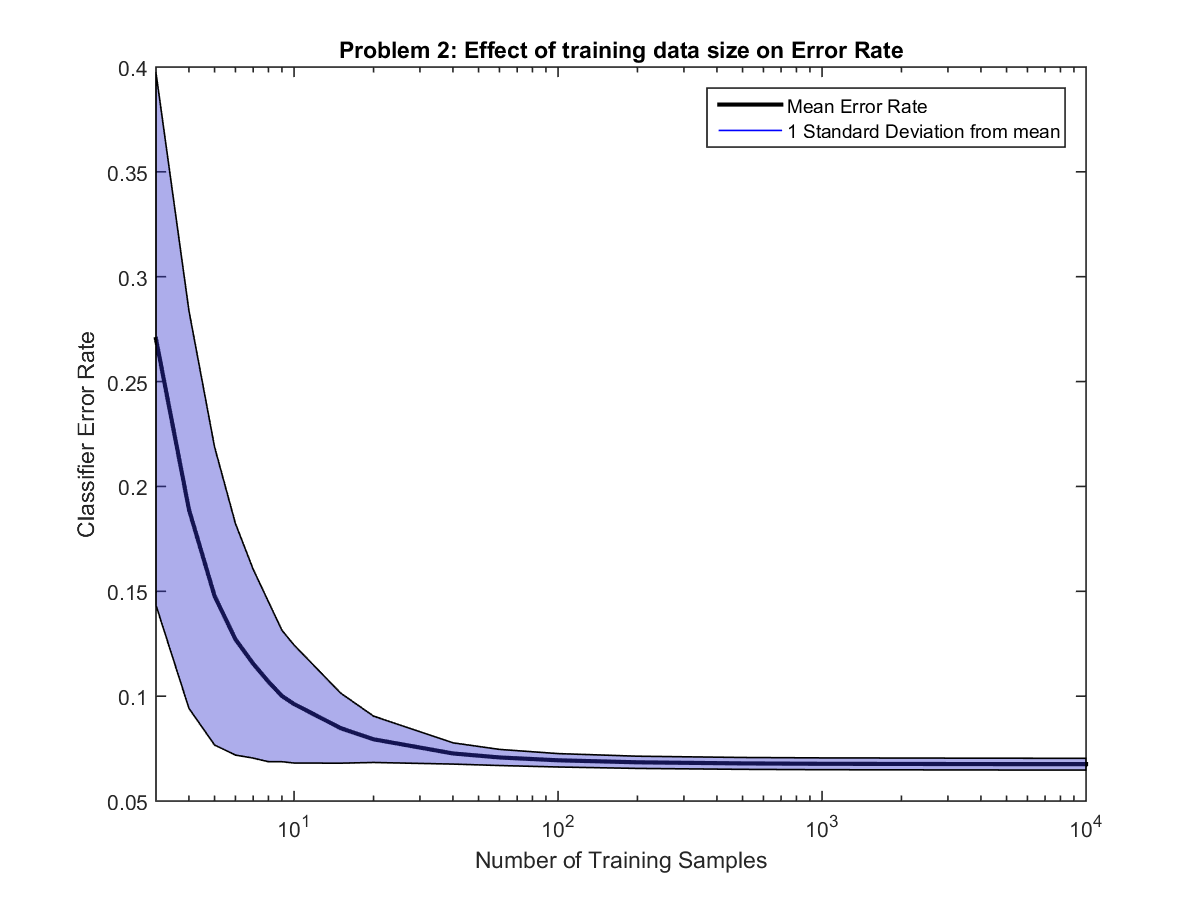


Figure : Effect of amount of training data on Error Rate

Figure 12 shows the results of the experiment. The black line indicates the mean error rate given some number of samples. The blue region is filled one standard deviation above and below the black line.

From the figure it is clear that 100, 1,000, and 10,000 training samples result in very similar classification accuracy. However, from inspection we see that less than 10 training samples results in poor performance, with 3 training samples having an average error rate of 27%. The accuracy of the classifier also seems to converge slightly more slowly than the classifier in problem 1.

## Face Detection (RP)

### Problem 3a (Chromatic Color Space) (RP)

One class

Two class

### Problem 3b (YCbCr Color Space) (RP)

One class

Two class

### Problem 3c (RGB Color Space) (RP)

One class

Two class

### Conclusion

# Conclusion

The Bayesian classifier for all test cases in this project produces results in which the error is minimized over all classifiers. This minimization occurs because the classifier was designed as a minimum risk classifier using a zero-one loss function.[[2]](#footnote-4) The classifier is known as a Bayesian classifier because the decision rule is based on estimating a-posteriori class probabilities using the Bayes rule that

|  |  |  |
| --- | --- | --- |
|  |  | (54) |

The resulting classifier is both a Bayes minimum risk classifier and minimum probability of error classifier.[[3]](#footnote-5)

# Contributors

Josh Gleason and Rod Pickens each wrote their own MATLAB software to perform the classification, the maximum likelihood estimation, the error estimation, the calculation of the Bayes error, the Chernoff error bound, and the Bhattacharyya error bound.

Josh generated the maximum likelihood performance charts, and Rod generated the face detection performance charts.

Josh wrote the theory section on maximum likelihood along with the results section, and as an added feature, Josh wrote the section on one class classification using as reference Fukunaga[[4]](#footnote-6), as suggested by Rod. Rod Pickens wrote the theory section on face detection along with the results section, and Rod also read and discussed with Josh the theory of one class classification.

1. Duda, Richard O., Hart, Peter O. and Stork, David G, “Pattern Classification,” Wiley Interscience, Second Edition, page 25, equation 12. [↑](#footnote-ref-3)
2. Ibid., page 26, Equation 19. [↑](#footnote-ref-4)
3. Ibid., page 27, Equation 21. [↑](#footnote-ref-5)
4. Keinosuke Fukunaga, Raymond Hayes, and Leslie Novak, “The Acquisition Probability for a Minimum Distance One-Class Classifier,” IEEE Transactions on Aerospace and Electronic Systems, Vol. AES-23, No. 4, July 1987, pages 493-499. [↑](#footnote-ref-6)