Probability Project

Siddiq Syed   
Engineering Sciences (Data Sciences)  
University At BuffaloBuffalo, NY  
UBIT# 50291566

Bhanu Prasanth Yeeli  
Engineering Sciences (Data Sciences)  
University At BuffaloBuffalo, NY  
UBIT#50290104

*Abstract*—To find the Classification accuracies and Error rates for the given data by approximating it as a normal distribution (Multivariate).

# Introduction

In an experiment involving 1000 participants, we recorded two different measurement (𝐹1 and 𝐹2) while participants performed 5 different tasks (𝐶1, 𝐶2, ⋯ 𝐶5). The two measurements are independent and foreach class they can be considered to have a normal distribution. The goal of this project is to construct a classifier such that for any given values of 𝐹1 and 𝐹2, it can predict the performed task (𝐶1, 𝐶2, ⋯ 𝐶5). We assume that the classifier calculates the probability of each class given the measurement data, and output the most probable class as the predicted class.

# Procedure

We first use the given dataset F1’s first 100 rows as training sets and get the mean and standard deviation values. We find the probability for each measurement in test data belonging to all the classes. To calculate this probability we used the principle of Baye;s theorem. We then proceed to predict the class for the test data. For the predictions classification accuracies and error rates are calculated with respected the original class given.

We repeated this above procedure for standardized F1 dataset, F2 dataset and a multivariate with variables F2 and standardized F1.

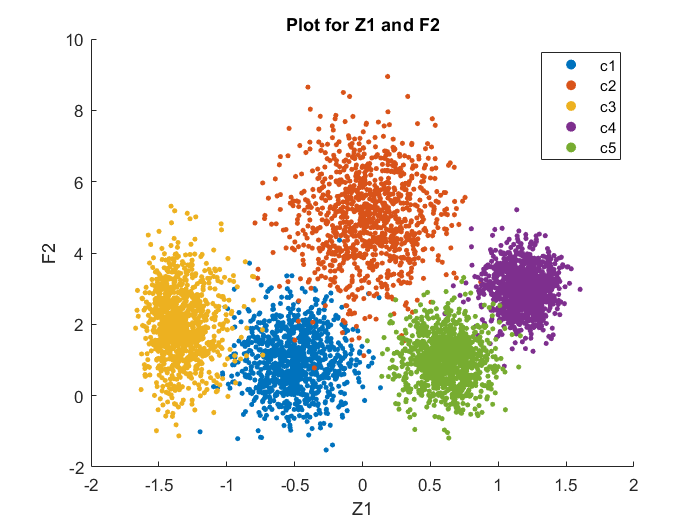
# Implementaion

## Case 1

First we found Z scores, 5 each corresponding to each class of the test data. With the respective means and standard deviation. We used the function normpdf in matlab to calculate the pdf values corresponding to the Z scores. We then found the class that has the maximum conditional probability for the measurement in test data. This is the predicted class. The next step is to predict the accuracy which we did by comparing the predicted with the original class. Accuracy and error rate for the same are below.

Classification Accuracy – 0.5262 Error Rate – 0.4738

Last step is to plot the graph between standardized F1 and F2. Below is the plot for the same.



## Case 2

We repeat the above process with standardized data for F1, for which the accuracies and error rates are as follows.

Classification Accuracy – 0.8838 Error Rate – 0.1162

## Case 3

We repeat the above process with F2, for which the accuracies and error rates are as follows.

Classification Accuracy – 0.5351 Error Rate – 0.4649

## Case 4

In this case we consider two random variables from standardized F1 and F2, as a multivariate normal distribution. Since both the random variables are said to be independent the resultant probability is the product of both the probabilities. So we have multiplied probabilities of case 2 and case 3 to estimate the class. For which the accuracy and error rates are as below.

Classification Accuracy – 9784 Error Rate – 0.0216

# Conclusion

In Case 1 and Case 3 we can see that the error rate is pretty high when compared with case 2 and case 4. Also the multivariate error rate is very low with respect to other cases. Case 2 signifies the importance of standardization of the input form the subject, whereas case 4 emphasizes the importance of having more data.