# EAS-595 Introduction to Probability Theory Project report

Shreyas SN
Department of Data Science
University Of Buffalo
*Buffalo,New York*
sshanubh@buffalo.edu

***Abstract*—This document throws light on the behavior and variation in predictions due to few modifications of two independent measurements following Gaussian distribution.It mainly consists of 4 cases and the prediction accuracy of the 4 cases.**

## I. Introduction

In the world of Data Science , accuracy with minimum computational cost is the thing which is most sought after. Here we try to analyze the accuracy of a classification model.

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 for each 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). The data contains measurements F1 and F2 that are both matrices with the size of 1000x5. Each column contains the information of one of the subjects and each row corresponds to one of the tasks (1st row: 1st task, 2nd row: 2nd task, etc.)

We considered 4 different cases , started with classifying based on F1 , followed by Z1(normalizing F1) and then F2 and ended with the multivariate distribution consisting of Z1 and F2

## II. Implementation

We followed the following steps

*Step 1:*

We used the data from the first 100 subjects to determine the mean and standard deviation, which would assist us in calculating the probability distribution function of the classes

*Step 2:*

In this step we calculated the probability distribution function of the classes knowing their mean and standard deviation of the first 100 measurements we applied the Bayes' theorem to provide us with the inference. We considered 900 data measurements (from 101 - 1000). In this step we considered the F1 measurements and accuracy was calculated by the number of correct predictions by the total number of predictions . We assigned a data point to a particular class based on the most probable class given by the predicted class

*Step 3:*

In this step we considered F1 distribution , with a slight change , we decided to normalize the data set

*Step 4:*

Here we repeated the step 2 for different cases where X = F2, and joint distribution of Z1 and F2

## III. Results

Upon implementing Gaussian Bayes' classifier on 4 different cases , we found out that the lease accuracy was obtained upon considering F1 and F2 (53% and 55% respectively).

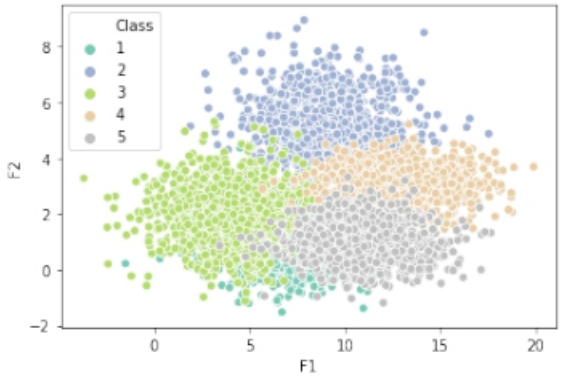
The second best accuracy was obtained when applying

accuracy was obtained while we considered the multivariate Gaussian distribution of Z1 and F2(98%)

## IV. INFERENCE

*Case 1 and 3:*

In this case we considered the two different measurements without any changes and the prediction was accuracy was not great



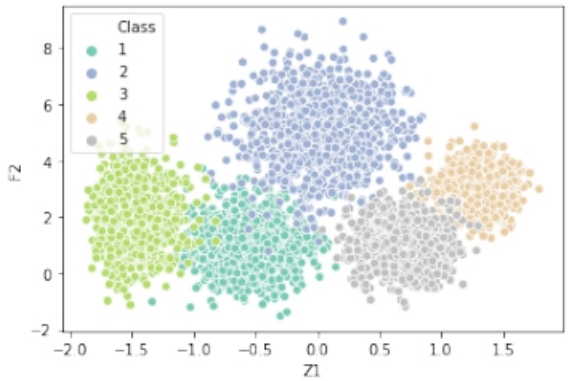

And also the three different classes were not clearly separated , which makes prediction harder

*Case 2:*

In this case accuracy was far greater , this has to do with the fact that in the F1 measurements the data was widely spread out, so a data point with higher values can influence the prediction widely , to reduce this inherent Bias normalization was applied causing to reduce the bias and hence provided a better accuracy

Case 3:

Since the two distributions were not correlated , considering their joint distribution yields greater accuracy because their is much more information with reduced Bias and variance , the prediction is much better.




And also it can be seen when the distribution was normalized there is a clear separation of the classes and it makes prediction easier

### C. *Equations*

To classify the data point into a class following equation was made use

$Predicted\ Class = \text{argmax}[P(Ci\ |X\ )]\ ,\ i = 1,2,\ldots\ 5$

The following bayes theorem was used for bayes classifier

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

For the probability density function , since the distribution was normal the following equation was used

f X (x) = 1 σ √ 2π e ^(− 1 /2 ( x−μ σ ))

## IV. CONCLUSION

To conclude we can say that joint distribution of independent gaussian distribution provides a better accuracy and also if the data points are scattered , it would be a good idea to normalize the data points so as to remove the bias , and to provide a better accuracy

## V REFERENCES

Nothing such in particular , random web search