Probability Project

Sreekar Guggilam
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, USA
email address – sreekarg@buffalo.edu

Rohith Kumar Poshala
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, USA
email address – rposhala@buffalo.edu

***Abstract*— In this project, a classifier is built to predict the class of the measurement taken by the participant. The classifier is built on the concept that for a given point the probability is calculated for each distribution (i.e. for each class) and the most probable class is considered as the output class. That is - *Predicted Class* = argmax [$P(C_i \mid X)$],i=1,2,⋯5).**

I. Introduction

In this project, Measurements (F_1, F_2) for 5 different tasks (C1, C2, C3, C4, C5) were taken by 1000 participants independently. With the help of classifier which is built as desired, we will observe which data of measurement does the classifier fits the most based on the classification accuracy.

II. Methodology

A. Classifier –

A normal distribution has been created using first 100 subjects of each class of a measurement data (say F_1) with the help of the mean and standard deviation of the data. And probability of each class for remaining 900 subjects were calculated using the above defined pdf (probability density function). With the help of Bayes Theorem, we can observe that given a measurement probability of it being in a class is equal to probability of that data point(measurement) given a particular class in relative terms.

$$P(C_i|X) = (P(X|C_i)*P(C_i))/P(X),$$

(X = participant's measurement and which kind of measurement selected i.e. F_1, F_2, Z_1 or [Z_1, F_2])

And $P(C_1) = P(C_2) = P(C_3) = P(C_4) = P(C_5)$ as each class is equally likely.

As $P(C_i))/P(X)$ is same for every class.

$$P(C_i|X) = \text{constant} * (P(X|C_i) \text{ for all } i = 1,2,3,4,5$$

So, a data point of a subject is classified as a class by comparing the probability of the data point of that subject across distributions of all classes and the class with maximum probability is predicted as the class of that data point.

$$\mathbf{argmax}[P(C_i\,|X)\,]\mathbf{=argmax}[P(\mathrm{X}|\mathrm{C_i})]$$

for all i = 1,2,3,4,5

and X = F_1, F_2, Z_1 or [Z_1 , F_2] in this case.

Later, every data point of a measurement which is classified as its own class to said to be predicted accurately and is used to get the 'Classification accuracy' along with error rate of the measurement data.

Classification accuracy = correct predictions / total predictions (which is 4500 in this case)

Error rate = incorrect predictions / total predictions

B. Cases I ($X = F_1$):

For Measurement F_1, Classification accuracy is being calculated with the same flow of above-mentioned classifier.

C. Cases II ($X = Z_1$):

For Measurement Z_1, Z_1 measurement is obtained by normalizing each measurement data point of F1 across data points of each subject to remove the effect of individual differences. And the same flow of methods as above classifier is performed to obtain Classification accuracy.

D. Cases III ($X = F_2$):

For Measurement F_2, Classification accuracy is being calculated with the same flow of above-mentioned classifier.

E. Cases IV ($X = [Z_1 , F_2]$):

For the Multivariant normal distributed Measurement data Z_1 and F_2, Classification accuracy is calculated with a similar flow of classifier except in case of pdf we have to consider pdf obtained by combining pdf of both the measurements which can be obtained by computing mean and standard deviation of first 100 subjects of each class for both measurements (Z_1 and F_2) individually. As the distribution of Z_1 and F_2 are considered to be independent, the combined pdf is obtained by multiplying the individual pdf of Z_1 and F_2. Then similar set of steps are performed as mentioned in the classifier to obtain Classification accuracy of this multivariant data.

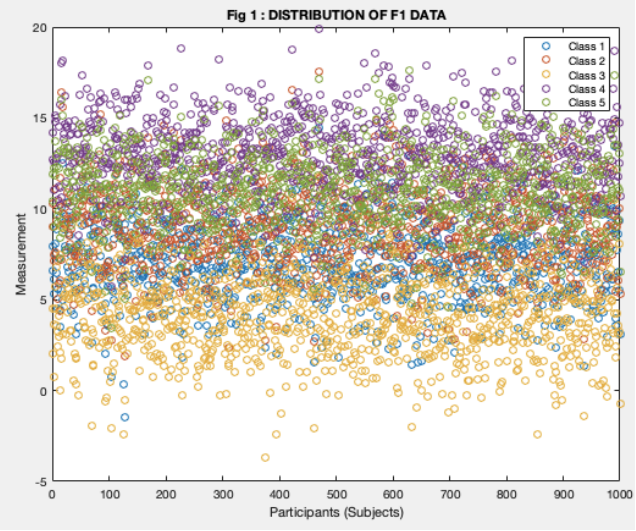
III. Results

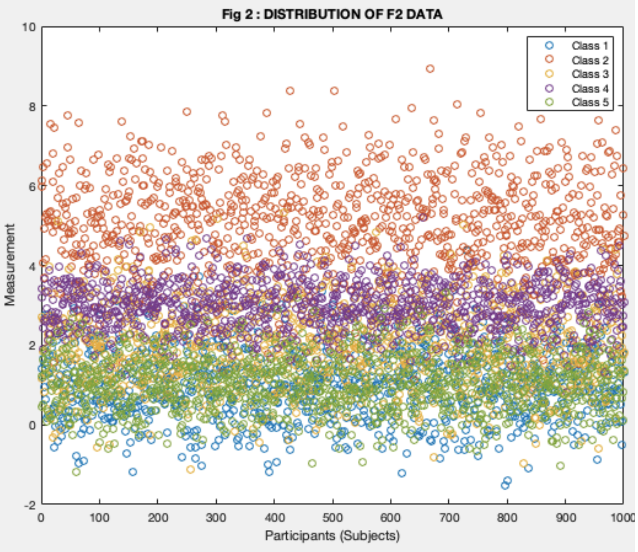
The accuracies for the classifier for different measurement cases are shown below-

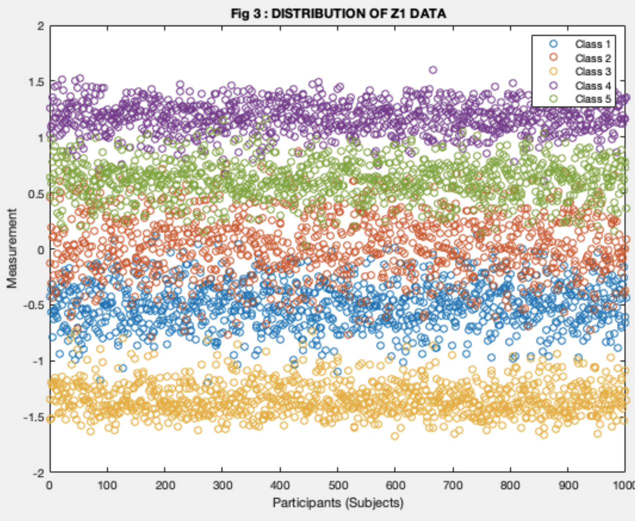
Measurement Data (X)	Classification accuracy
F_1	53.0000
Z_1	88.3111
F_2	55.0889
[Z_1 , F_2]	97.9778

IV. Discussions

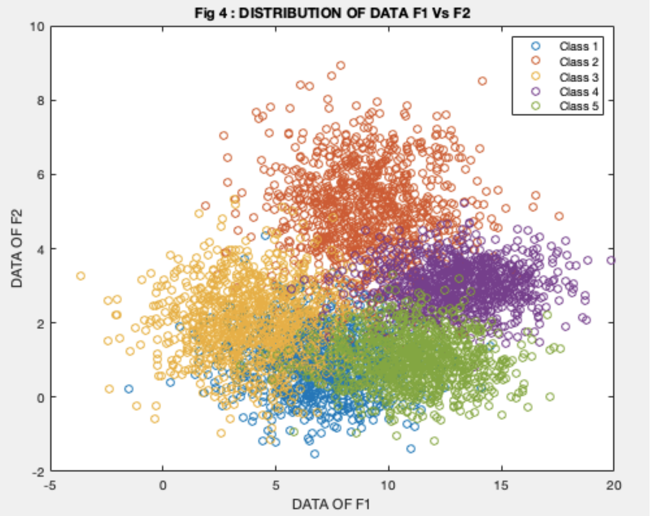
From the distribution of F_1data (Fig:1) it can be seen that the distribution of each class is overlapping significantly when considered across all subjects, similar pattern of distribution can be observed for the distribution of F_1data (Fig: 2). But when we observe the distribution of Z_1 data (normalized data of F_1) (Fig:1), the distributions for each class is distinct (classified) from each other as Z_1 data does not consider the effects of individual differences. Hence the classifier built for the F_1or F_2data has lower accuracy when compared with the classifier built for the Z_1 data.

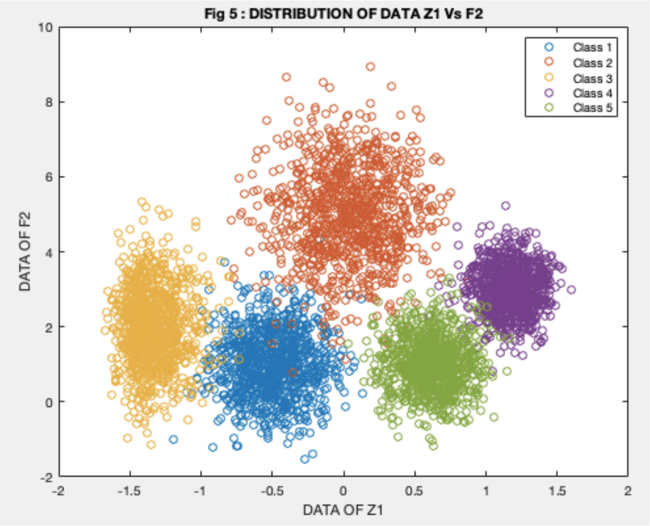






From the Scatter plot for distributions of F_1vs F_2data (Fig : 4), it can be observed that the distributions of each class are distinguishable from each other but not that significantly compared to the scatterplot for distributions of Z_1vs F_2data (see fig:5). From this distribution, we can see that all the classes are clearly distinguishable. This is the why the classifier built for this multivariant normal distribution [Z_1 , F_2] has higher accuracy than the other classifiers. Because this classifier considers the distribution of both Z_1 and F_2 data, the accuracy obtained by considering Z_1 data is increased as the data point can be more specifically classified with the help F_2 data along with Z_1 data.





V. CONCLUSION

To conclude that, the classifier built for multivariant normal distribution [Z_1, F_2] is observed to have more accuracy among remaining cases. This is because by combining F_2 data with Z_1 data (Z_1 data has normalized properties of F_1), we now have a distribution which comprises almost all properties (information) which can help classify a data point to its corresponding class even more accurately.