EAS 595 Project

Akhil Koppera
UB Data Science Fall 2019
Person#: 50318499
UBIT: akhilkop

Akshaykumar Rao Racherla
UB Data Science Fall 2019
Person #: 50320372
UBIT: aracherl

***Abstract*—This report summarizes and compares the results of classifiers built to classify the measurements of an experiment based on four cases. The constructed classifiers calculate the probability of each class and outputs the most probable class as the predicted class.**

I. Introduction

In an experiment, two measurements (F1 and F2) were recorded for 1000 participants while they perform 5 different tasks (C1, C2,...,C5). These two measurements are independent and are considered to have a normal distribution for each class. The goal is to construct a classifier such that for any given values of F1 and F2, it can predict the most probable class of the performed task(C1,C2,..,C5) based on the formula:

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

Out of 1000x5 observations, the first 100 subjects were used for training and the remaining 101-1000 observations were used for testing/predictions. This report summarizes and compares the classification rate of the below four cases:

- Case 1: X=F1
- Case 2: X=Z1
- Case 3: X=F2
- Case 4: *X*=[Z1 F2]

II. Data Description

The given data set contains two measurements F1 and F2, each of size 1000x5 containing information of one of the subjects in the column and each row correponds to one of the tasks. Measurements for each class in F1 and F2 follow a normal distribution.

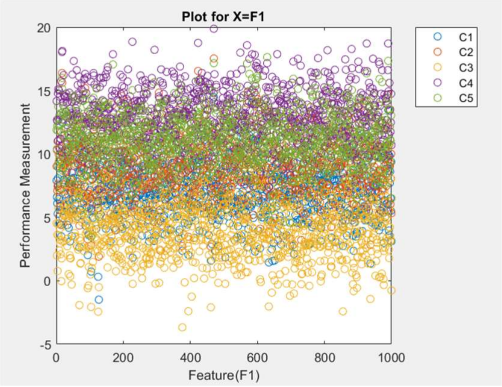

Figure 1. Plot for X=F1

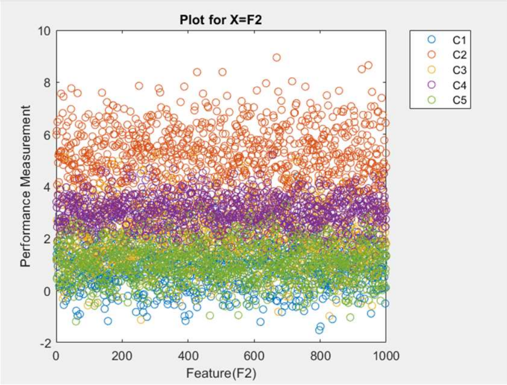

Figure 2. Plot for X=F2

III. Methodology

1. Mean and standard deviation was calculated for each class in F1 and F2 for the first 100 subjects and then normal PDF was created.
2. Using the PDF of each class, probability of each measurement belonging to a class was calculated using Bayes' theorem for the remaining 900 rows.

   Bayes' theorem P(Ci|X) = (P(X|Ci)*P(Ci))/P(X)

   P(Ci) is equal to 1/5 for all *i* =1, 2, … 5

   P(X) = 1/900 as total test points are 900
3. Each measurement is classified based on the max probability (P(X|Ci)) of belonging to classes C1, C2...C5 as P(Ci))/P(X) is same for the measurements.
4. Classification accuracy and error rate for classifier was calculated based on the below formula:

   Classification accuracy = correct predictions / total predictions (5*900)

   Error rate = incorrect predictions / total predictions
5. The above steps were performed for all the four cases. Additionally, for cases 2 and 4, measurements for all the classes by each person is normalized before calculating the PDF.

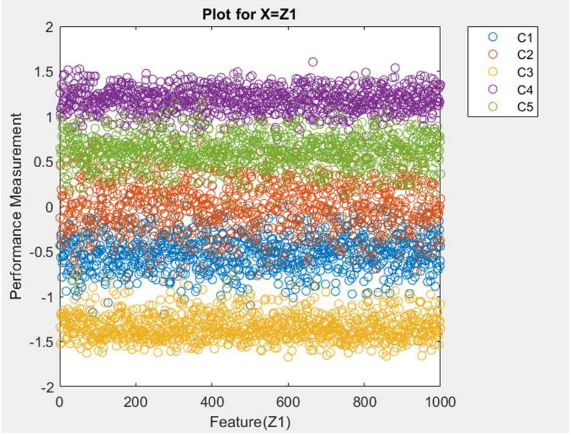

Figure 3. Distribution of data after normalizing

(Case: X=Z1)

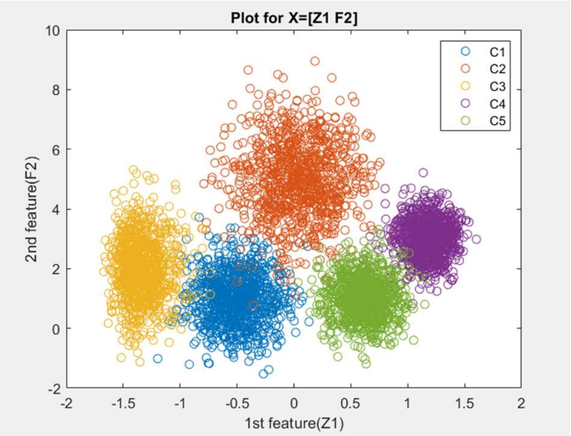

Figure 4. Distribution of data after normalizing (Case: X= [Z1 F2])

IV. Results and Summary

Based on the methodology described above, the results are obtained as below

Cases	Accuracy (%)	Error Rate (%)
X = F1	53.00	47.00
X = Z1	88.31	11.69
X = F2	55.09	44.91
X = [Z1 F2]	97.98	2.02

Table 1. Summary of classifiers for each case.

The accuracy for cases 2 and 4 is high as compared to other cases as the data was normalized. It was observed that the distributions were overlapping in these two cases, indicating that the scale of measurements was different for each person. This difference in the scales was removed by normalizing the measurements of each person in F1. It can be seen from Figure 4 that the accuracy was improved by removing the overlap between the measurements.

V. Conclusion

It can be seen from Table 1, the highest accuracy was obtained for Case 4. X = [Z1 F2]. Hence, it can be concluded that irregularities like differences in the scale/error in measurements can be removed by normalization.