Bayes’ Classifier for Gaussian distributions

Manoj Madabhushi   
*School of Engineering and Applied Sciences*  
*University at Buffalo*New York, United States  
lmadabhu@buffalo.edu

Karun Parashar  
*School of Engineering and Applied Sciences*  
*University at Buffalo*New York, United States  
karunpar@buffalo.edu

***Abstract*— The aim of this project is to build a classifier which uses Bayes’ theorem for predictions. In the real-world, we often encounter data that suffers from scaling issues. In this paper, we study ways to tackle this with normalization and its impact on classification accuracy.**

# Introduction

In an experiment involving 1000 participants, two different measurements (F1 and F2) were recorded, each of which involved 5 different tasks (C1-C5) performed by all the participants. The task data for each participant is independent of each other and from the other measurement. The data is normally distributed across each task for both measurements as follows:

P(F1|Ci)=N(m1iσ1i2), P(F2|Ci)=N(m2i,σ2i2) for i=1,2,3,4,5 (1)

where, m1i, σ1i2 are the mean and variance of F1 for the ith class and m2i, σ2i2 are the mean and variance of F2 for the ith class.

Our objective is to build a Bayesian classifier to predict the class of each participant given the measurement.

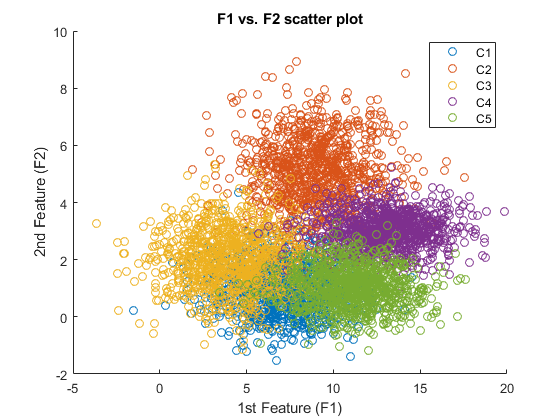
# METHODS

## Data Preparation

The F1 measurement is assumed to be a subjective measure i.e., the values and the range of data reported by one participant might not be in the same scale as others. These differences could impact the classification accuracy and hence, normalizing the data of each subject by using the standard normal formulation will alleviate this problem. To normalize the data, we remove the mean of all tasks of a participant data from that participant’s reported task values and divide it by the standard deviation of all tasks. This process will make sure that all the participant data is centered at zero with a standard deviation of one. The new transformed (Z-score) measure is named Z1.

* 1. *Visualization*

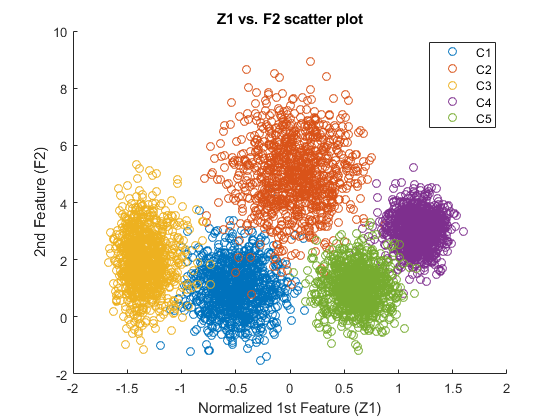
Prior to normalization of F1, we visualize the separation of classes in the following scatter plots with respect to F2 (see Fig. 1).



1. F1 vs. F2 distribution plot by tasks

We see that the classes are not well separated in the F1 vs. F2 distribution plot.

After normalization of F1 to Z1, we visualize the separation of classes in the following scatter plot with respect to F2 (see Fig. 2).



1. Z1 vs. F2 distribution plot by tasks

We see that classes are well separated in the Z1 vs. F2 distribution plot. This suggests that normalization of F1 has increased the separation between classes which could lead to better prediction accuracy.

## Training

The first 100 participant data was used to train the model on F1, F2, and Z1 measurements. For each measurement, the model was trained by computing means and variances using this data.

## Testing

To test out the accuracy of the model, we use Bayes’ theorem. The formulation is as follows:

P(Ci|X) = P(Ci)\*P(X|Ci)/P(X) (2)

where Ci-Class (i=1,2,3,4,5), X- Measurement

P(Ci) is called *a priori* probability. This probability is same for all classes. The denominator P(X) is common for all P(Ci|X). Hence, we can simplify the formulation (2) to the following proportionality:

P(Ci|X) *α* P(X|Ci) (3)

For a given measurement, we simply calculate the normal probabilities of each data point with means and standard deviations of all the five tasks and assign the task with highest probability as the prediction for that data point. This is performed on three univariate measurements F1, F2 and Z1, and on one multivariate measurement [Z1 F2]. For the multivariate case, we assume independence between the measurements. This means we calculate the probability for each task of a participant in both measurements (Z1 and F2) and multiply the two probabilities. The task with the highest probability thus achieved is assigned as the prediction of that data point.

## Classification Performance Metrics

As a metric to study the performance of each measurement in classifying the tasks of participants, we look at accuracy and error rate which are defined as follows:

Accuracy=Correct predictions/Total predictions (4)

Error rate=Incorrect predictions/Total predictions (5)

1. RESULTS

After testing out the four different cases listed above, we calculate the accuracy and error rates of each case using formulae (4) and (5). Table 1 summarizes the results of different measurements. The results reveal that models F1 and F2 performed poorly on task classification as they suffer from scaling issues. After normalization of F1, we see that Z1 alone has produced a high classification accuracy rate of 88.31%. This further corroborates the assumption that the original data was inconsistent across participants. Finally, we observe that the multivariate case has performed the best among all suggesting that F2 is also a significant predictor that provides additional information in successful classification of tasks.

1. CLASSIFICATION ACCURACY AND ERROR RATES OF DIFFERENT MEASUREMENTS

|  |  |  |
| --- | --- | --- |
| **Case** | **Classification Accuracy (%)** | **Classification Error (%)** |
| X = F1 | 53% | 47% |
| X = F2 | 55.09% | 44.91% |
| X = Z1 | 88.31% | 11.69% |
| X = [Z1 F2] | 97.98% | 2.02% |

##### CONCLUSION

In this study, we call attention to the significance of Bayes inference in statistics. The accuracy of the Bayes classifier is highly dependent on consistency of data and we see the significant impact of normalization in separation of tasks. This in turn led to high accuracy levels in classification of tasks. Finally, we conclude that both measurements play a significant role in predicting the tasks of participants.

##### References

Dimitri P. Bertsekas and John N. Tsitsiklis, Massachusetts Institute of Technology