EAS595-Probability Project

Tanya Jain
tanyajai@buffalo.edu

Pranjal Jain
pranjalj@buffalo.edu

***Abstract*—This project describes the implementation of Bayes' theorem for classification problem. The goal is to construct a classifier with two measures F1 and F2 so as to predict the performed tasks (C1, C2, C3, C4, C5)**

Keywords—Bayes Rule, Naïve Bayes', Normalization, Z-Score

I. Introduction

In an experiment involving 1000 participants, we recorded two different measurements ($F1$ and $F2$) while participants performed 5 different tasks ($C1$, $C2$, C3, C4, $C5$). The two measurements are independent and for each class they can be considered to have a normal distribution as follows:

$$P(F1\ |Ci\ ) = N(m1i\ , \sigma1i\ 2\ )$$

$$P(F2\ |Ci\ ) = N(m2i\ , \sigma2i\ 2\ ) \qquad \text{for } i = 1,2, \cdots 5$$

where, $m1i$, $\sigma1i$ 2 are the mean and variance of $F1$ for the ith class. Similarly, $m2i$, $\sigma2i$ 2 are the mean and variance of $F2$ for the ith class.

Bayes' Rule:

Given a hypothesis H and evidence E, Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence $P(H)$ and the probability of the hypothesis after getting the evidence $P(H|E)$ is

$$P(\mathrm{H|E}) = \frac{P(E|H)P(H)}{P(E)}$$

Naïve Bayes':

Naïve Bayes' is a conditional probability model: given a problem instance to be classified, represented by a vector

$x= (x_1,x_2,\ldots,\ x_n)$ representing some n features (independent variables), it assigns to this instance probabilities

$$p(C_k \mid x_1,x_2,\ldots,x_n)$$

for each of K possible outcomes or *classes* $C_{k.}$

II. Data Analysis

A. Dataset Description

Dataset contains 1000 records of each participant performing 5 different tasks. This data is classified into 2 measurement groups. There are total 5000 data points which are partitioned into train and test data with 500 and 4500 data points. Data is normalized for F1 measure with different mean and standard deviation for different classes describing those 5 tasks.

B. Data Normalization

We used Z-score for data normalization:

$$Z= X- \mu/\ \sigma$$

where, μ is the Mean and σ is the Standard deviation. For standard normal distribution, $\mu = 0$ and $\sigma = 1$.

After analysing the dataset, we found that the values of records range from -3.6691 to 19.87. Normalizing these values gives us -1.67 as lower bound and 1.6 as upper bound.

III. Method

1. We used the first 100 records and calculated their mean and standard deviation.
2. For the rest 900 records, test data matrix was created to classify the data into different tasks.
3. Every data point was normalized for each class so as to determine which class it belongs to.
4. After predicting class for each data point, accuracy and error rates were calculated.
5. We performed above steps for F1, Z1, F2 and $\binom{Z1}{F2}$.
6. For comparison of classification rates of these 5 classes, we plotted a bar-graph.

IV. Results

The accuracy and rate were calculated as follows:

Classification Accuracy = correct predictions/total predictions

Error rate = 1- Classification Accuracy

A. Case 1:

In this we used Bayes' theorem to calculate the probability of each class test set for F1and consequently predicted the class for each data point.

B. Case 2:

In this the F1 data was normalized by calculating the Z-score. Accuracy and error rate were determined.

C. Case 3:

In this we used Bayes' classifier to calculate the probability of each class test set for F2 and consequently computed accuracy and error rate.

D. Case 4:

In this we used Naïve Bayes' classifier on Z1 and F2 data to depict its use on multivariate data. Accuracy and error rate were determined.

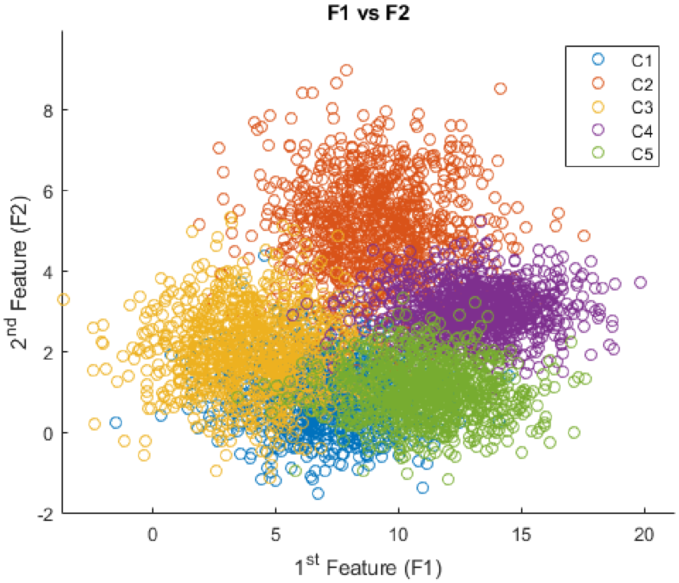

Fig.1: Distribution of Data using F1 and F2

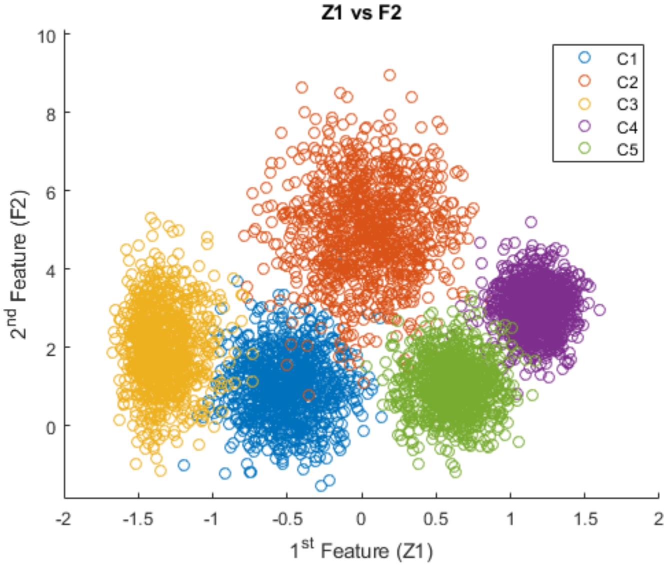

Fig.2: Distribution of Data using Z1and F2

In figure 1, we plotted F1 against F2 measure for five different classes, each class distinguished by different colors. As the data is not normalized, a lot of overlapping is observed, affecting the performance of the classifier.

In Figure 2 scatter plot can be seen of normalized data (Z1) with less overlapping i.e. less variance and more distinguished classes that are tightly clustered, showing the significance of data normalization.

V. Summary

Table below summarizes the results for all the four cases:

Measurements	Classification Accuracy	Error Rate
F1	52.62%	47.38%
Z1	88.38%	11.62%
F2	53.51%	46.49%
Z1,F2	97.84%	2.16%

From the graph below, it was observed that data normalization increases the accuracy of classification.

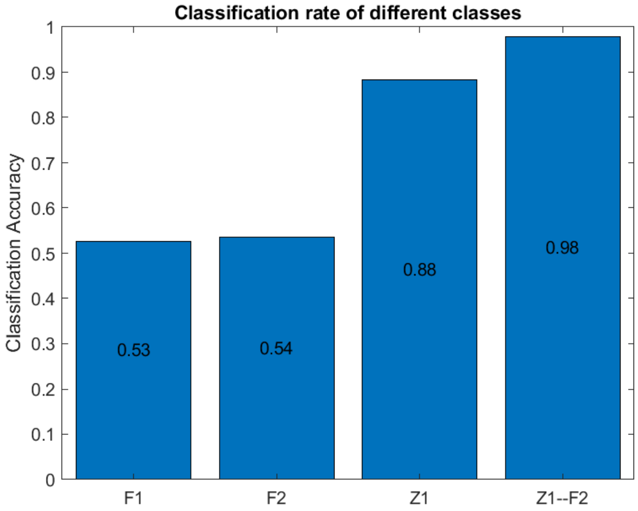

Fig.3: Comparison of Classification Accuracy and Error Rate

VI. Conclusion

In this project, we implemented Bayes' classifier for univariate and multivariate cases. It was observed that normalization improved the performance of classifier.

From the results, it was observed that Multivariate performs as the best classifier due to highest accuracy of 98% and minimum error rate. It is because multivariate normal considers relationship amongst different features in multiple datasets and this property makes it predict better than univariate normal.

References

[1] https://en.m.wikipedia.org/wiki/Multivariate_normal_distribution

[2] https://www.mathworks.com

[3] https://www.towardsdatascience.com