EAS595 - Probability Project Report

Anchal Sojatiya – 50321238 – (anchalso@buffalo.edu)
Samratsinh Dhumal – 50321053 – (samratsi@buffalo.edu)
School of Engineering and Applied Sciences, SUNY Buffalo
Buffalo, NY

***Abstract*— 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 ($C1$, $C2$, C3, C4, $C5$). We used the powerful Bayes Theorem for classification.**

***Keywords*—Bayes Rule, Naïve Bayes', z-score, normal distribution, multivariate normal**

I. Introduction

In an experiment involving 1000 participants, we recorded two different measurement (F_1 and F_2) while participants performed 5 different tasks (C_1, C_2, ..., C_5). The two measurements are independent and for each class they can be considered to have a normal distribution as follow:

$$P\ (F_1 \mid C_i) = N\ (m_{1i}, \sigma^2{}_{1i}) \text{ and } P\ (F_2 \mid C_i) = N\ (m_{2i}, \sigma^2{}_{2i})$$
$$\text{for } i = 1, 2, \dots ,5$$

where m_{1i}, $\sigma^2{}_{1i}$ are the mean and variance of F_1 for the i^{th} class and m_{2i}, $\sigma^2{}_{2i}$ are the mean and variance of F_2 for the i^{th} class.

Using Bayes Theorem to build a Naive Bayes classifier to calculate the probability of each class given the measurement data, and output the most probable class as the predicted class.

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

Bayes Rule

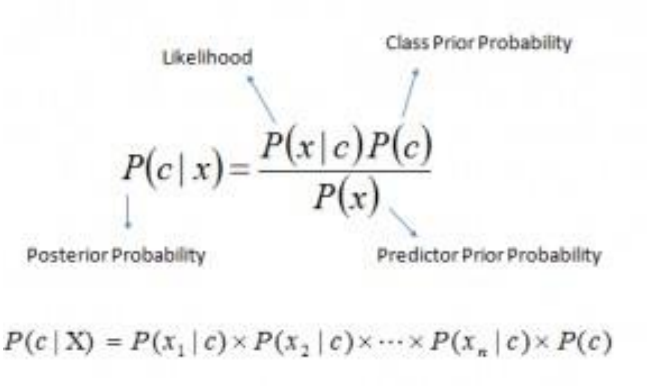

Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$

Fig. 1. Naive Bayes Classifier

II. Dataset Description

The file 'data.m' 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.)

III. Implementation

1) In order to make the results comparable we used the first 100 observations to find their mean and standard deviation to normalize the observation.
2) As per the problem we used the rest 900 observations as a test dataset to classify the observations into different tasks.
3) We calculated the accuracy and error rates comparing the predicted class with the true class to understand the model.

   The accuracy and rate were calculated as follows:

   Classification Accuracy = true predictions/ total predictions

   Error rate = 1- Classification Accuracy
4) To remove the effect of individual differences, we had to normalize the data of each subject using the standard normal formulation (removing the mean and dividing by standard deviation) to find Z_1 from F_1
5) We did all the above for F1, Z1, F2 and $\begin{pmatrix} Z1 \\ F2 \end{pmatrix}$. We plotted a graph for the same to compare results (Fig 4).

IV. Case Methods

Case 1:

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

Case 2:

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

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

where, μ is the Mean and σ is the Standard deviation.

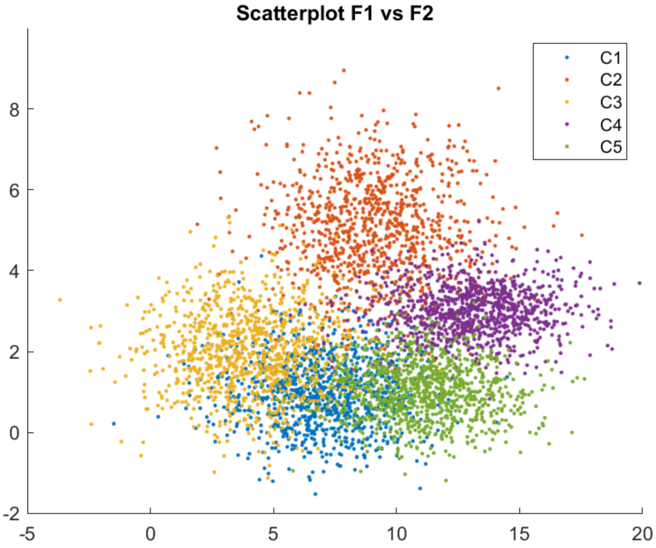
Case 3:

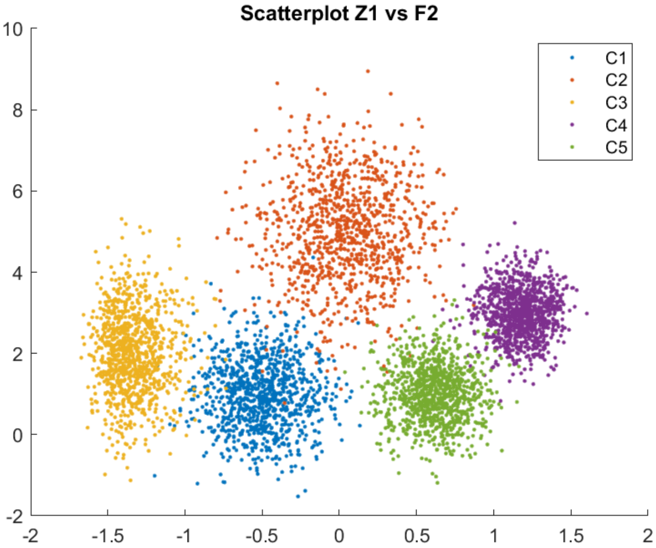
For case 3 we used Bayes' classifier to calculate the probability of each class test set for F2 and consequently computed accuracy and error rate.

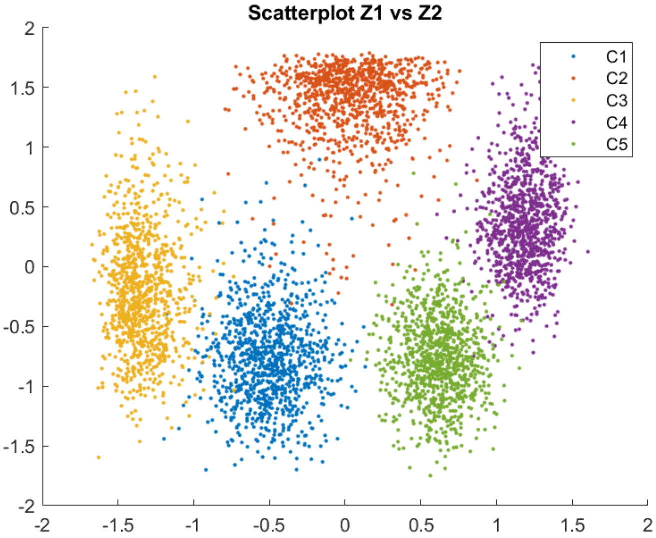
Case 4:

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

V. Results

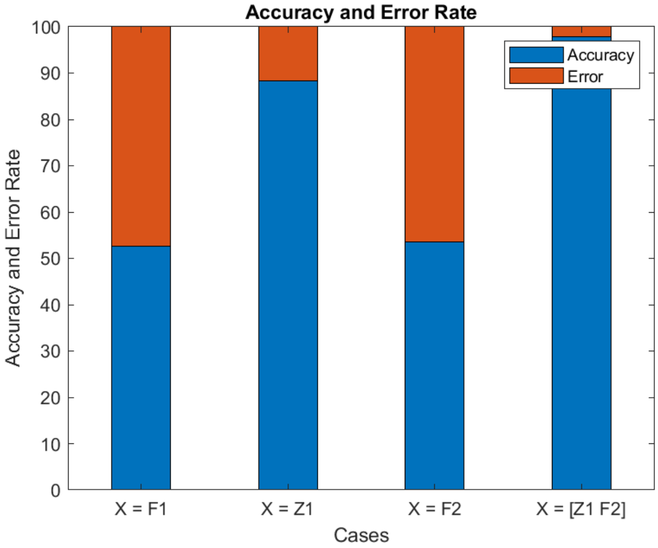






VI. Conclusions

Measurements	Classification Accuracy	Error Rate
F1	52.622	47.378
F2	53.511	46.489
Z1	88.378	11.622
$\begin{pmatrix}Z1\\F2\end{pmatrix}$	97.844	2.156



F1, F2 give around 52-53% accuracy which can be increased using the normalization of F1 into Z1 thereby making the values in different classes comparable the accuracy of the prediction class can be increased.

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.

VII. References

[1] https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54

[2] https://en.wikipedia.org/wiki/Multivariate_normal_distribution

[3] http://www.statsoft.com/textbook/naive-bayes-classifier

[4] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained