EAS 595 Intro to Probability Theory Project

Jing Chen
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, NY, USA
jchen445@buffalo.edu

Kunal Mehta
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, NY, USA
kunalkam@buffalo.edu

***Abstract*—The dataset involved in this project is the measurements of Features 1 and 2, recorded in matrices F1 and F2 respectively, for 1000 individual participants who perform 5 different tasks, namely C1, C2, C3, C4 and C5. Based on subsets of the data, the project predicts the classes of all of the input points. Compared in this experiment are the results of the predictions for different transformations of the data in order to determine which of them are the most suitable, which is done by comparing the accuracy of the predictions. To calculate each prediction, the Bayes' theorem is used. As a special case, a multivariate normal distribution is also used in which it is assumed that features are independent of each other.**

***Index Terms*—Bayes' Theorem, Classification, Normal Distribution**

I. Introduction

In an experiment that involved 1000 participants, measurements F1 and F2 were recorded for each task and stored in a .mat file. These are 1000x5 matrices, and are considered to have a normal distribution, abd the Bayes' theorem was used to calculate the probabilities of each data point in each class. The class with the highest probability score was then taken as the prediction for that data point. The accuracy was calculated by simply dividing the number of correct predictions by the total number of data points.

The problem statement states that each column represents the information of a subject, and each row corresponds to a task. However, the matrices being 1000x5, and not 5x1000, instead of transposing both of the matrices, we simply logically associated each column with a task and each row with a participant.

II. Training

The first step was to find the mean and variance for each class using the recorded measurements of F1 and F2 for the first 100 participants. This is an estimation, not the true mean and variance for each of the classes in F1 and F2, and uses only one tenth of the data for estimation. This step was performed by adding all the measurements for each class and dividing by the total by the number of participants, which is 100, to get the mean. The variance was calculated using the built-in MATLAB function var().

III. Testing, Prediction and Accuracy calculation

A. Testing

Using the means and variances calculated in the previous step, the classes for each of the remaining data points of F1 was to be predicted. Since we used the top 100 participants to estimate the mean and variance, the predictions were only to be made for the remaining 900 participants, which was a total of 900x5 = 4500 predictions. This task was performed in the following manner. For each individual data point, the probability for each class was calculated using the probability density function of a normal distribution, as given below:

$$\frac{\exp(-0.5((x-\mu)/\sigma)^2)}{(\sigma\sqrt{2\pi})} \tag{1}$$

After the probabilities are calculated, the class with the highest probability is chosen as the prediction for that data point. In a similar fashion, all the 4500 data points have classes predicted for them.

B. Calculation of Classifier Accuracy

The predictions are compared with the true class of the data, which is mentioned as being the row number of the data. However, since we worked with the original matrices without transposing them, the true class corresponds to the column number. The classification accuracy is calculated simply by dividing the number of correct predictions by the total predictions. The error rate is simply the accuracy rate subtracted from 1. Accuracy rate: 0.5300 Error rate: 0.4700

IV. Standard Normal (Z-Score)

The F1 measurements can very well be subjective measures. Therefore, the mean and range of data reported by one participant could very well be inconsistent with other subjects. Therefore, to remove the bias, to wit the effect of individual differences, the next step was to normalize the data of each subject using the standard normal formulation. This means subtracting the mean and dividing by the standard deviation for each point. This standard normal formulation was computed for F1 and stored in matrix Z1 which had the same size as F1, 1000x5. To compare how this affected the measurements, a scatter diagram of Z1 vs F2 was plotted. The two diagrams are displayed below.

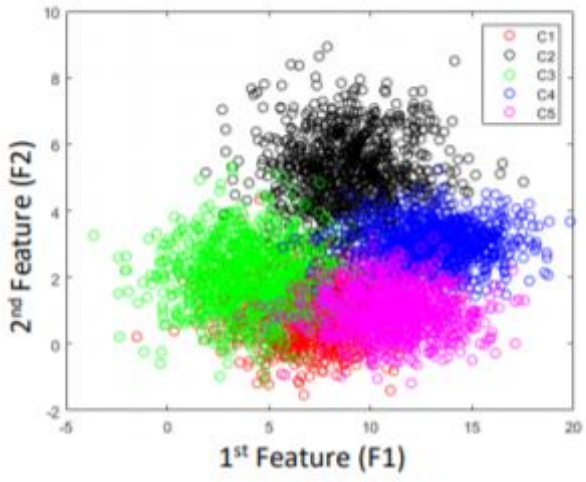

Fig. 1. Fig 1: Scatter Plot 1: F1 vs F2

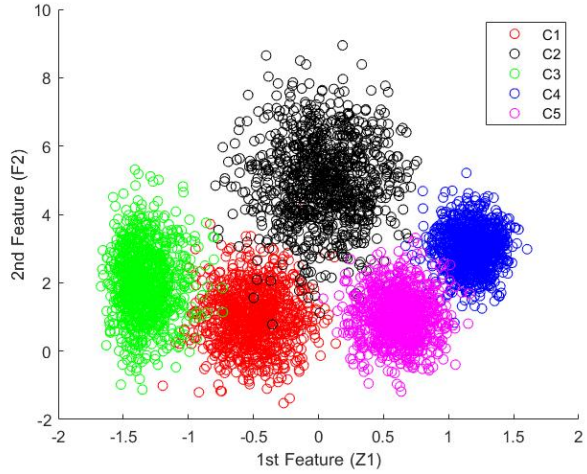

Fig. 2. Fig 2: Scatter Plot 2: Z1 vs F2

As can be easily observed, the Standard Normal formulation caused the 5 classes to become much more distinct and separate from the other classes. This was most likely a result of the removal of the bias.

V. Testing, Prediction and Accuracy Calculation for other Matrices

The testing, prediction and accuracy calculation steps were performed for the following cases:

- X = Z1
- X = F2
- $X = [\frac{Z1}{F2}]$

The last case is a multivariate distribution, for which we have assumed independence of features. Therefore the probability for each class is simply the probability of that for Z1 multiplied by the probability of that for F2. The rest of the cases were handled in an identical manner as described in the Testing, Prediction and Classification Accuracy calculation section above. For the first case, X = Z1, the Training step is performed as well. The accuracy for each of the 3 cases, as well as the original is displayed below:

- Accuracy X = F1: 53%
- Accuracy X = Z1: 88.3111%
- Accuracy X = F2: 55.0889%
- Accuracy $X = [\frac{Z1}{F2}]$: 97.9778%

As can be observed, the Accuracy rates for Z1 is much higher than that of un-normalized data in F1 and F2, but the multivariate distribution has nearly a 98% accuracy, making it by far the best classifier. The reason for the low accuracy of X = F1 and X = F2 is the inconsistency of measurements across all the individuals. This caused an overlap of values, clearly evident in Figure 1 (Scatter Plot 1: F1 vs F2). In contrast, X = Z1 has a significantly higher accuracy due to removal of bias. Using both Z1 and F2 was the best predictor out of the chosen cases, because both F1 and F2 were the driving factors of class outcome. Z1 is derived from F1 itself, so having F2 contribute to the probability calculation as well is intuitively logical, and did improve accuracy empirically to nearly 98%. Since we were intrigued by the whopping increase in accuracy due to normalizing the data, despite it not being part of the problem statement, we calculated Z2 and conducted the accuracy test for $X = [\frac{Z1}{Z2}]$. As we surmised, the accuracy was even higher than that of $X = [\frac{Z1}{F2}]$ due to even F2 being normalized in this case, and we received an accuracy of 98.36%.

VI. Conclusion

This project involved building classifiers to predict the class of data points using Bayes' theorem. It also brought into perspective the significance of the Standard Normal formulation (Z-score) of the given data and how important the removal of bias is to classification. Another important component was the use of multivariate distribution to better predict the class of a data point.