Bayesian Classifier

Abstract -- In this project we are performing an experiment with 1000 participants and two independent measurements (F_1 and F_2) are recorded while the participants performed 5 different tasks ($C1, C2, \cdots C5$). *Here we are finding the best Classifier of this data by using Bayes theorem by following the steps: training, testing and calculating the accuracy of the classifier, calculating standard normal of F_1 and F_2, considering 4 cases($X=F_1$, $X= Z_1$, $X= F_2$ and $X= [Z_1; F_2]$) and comparing the classification rates of all the 4 cases.*

1. Data Description

The dataset has 1000 participants where each of them are performing 5 different tasks. There were two different measurements taken namely F1 and F2 for each task. The two measurements are independent and were considered to have a normal distribution. Using Bayes' theorem, the probabilities of a data point is computed for each class and the class with the maximum probability is the one predicted for that data point.

2. Procedure

Step-1 Training the Data:

Here we have built a model which takes first 100 observations of F_1 data and calculates the mean and variance and then we repeat the same for F_2

Step-2.1 Testing:

Here, we will assign F_1 data to a variable X and calculate the probability of each class for data of the remaining subjects (columns 101-1000 of F_1) and consequently predict the class of 4500 data points.

Step-2.2 Calculating the accuracy of the classifier:

Here we need to check the percentage of the data whose class are correctly predicted.

Classification accuracy = correct predictions / total predictions (which is 4500 in this case) Error rate = incorrect predictions / total predictions == 1- accuracy

Step-3 Standard Normal (Z-Score)

Since F1 can be a subjective measure, it is highly probable that the measurements reported by different participants can have different mean and variance of values. Hence, we normalized the measurements for each class, reported by every individual. The normalized measurements were named Z1 (Z-Score) which were (F1i − μi)/σi , i = 1, ..., 1000.we then plotted before and after normalization which can be observed in fig 1.1 and fig 1.2. We can see that after normalization, the 5 different classes became very well separated and there were well defined boundaries between them. This is because the values reported by different individuals, for each class, were having different mean and standard deviations. This led to collide of variance of values which can be seen in Fig 1.1. The classification before and after normalization, were compared.

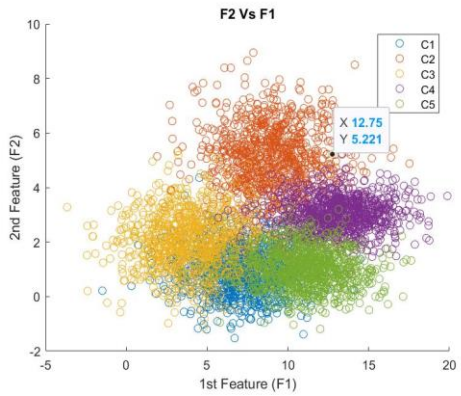

Fig 1.1: Before-Normalization

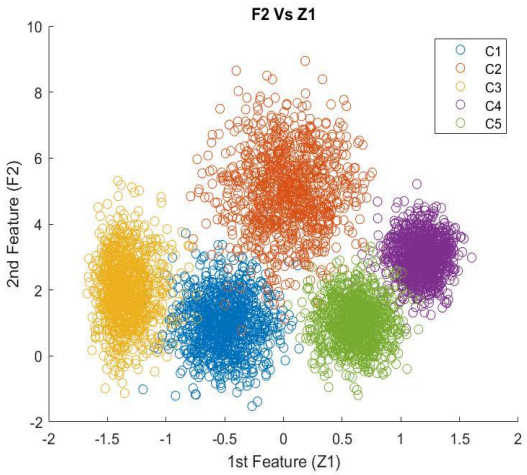

Fig 1.2: After-Normalization

Step-4:

Here we have 4 cases:

Case 1: we have implemented this case in step-1.

Case 2: Implemented step-1 and Step 2 for normalized data Z_1 and calculated the prediction and accuracy

Case 3: Implemented step-1 and Step 2 for normalized data F_2 and calculated the prediction and accuracy

Case 4: Here we combined Z_1 and F_2 to form a multivariate normal distribution and then implemented step 1 and 2 to calculate the prediction and accuracy.

3. Results

Step-1: mean and variance values of F_1 for first 100 observations of 5 subjects are:

Mean: 7.0933 9.1445 4.2877 13.3375 11.2419

Variance: 2.0700 2.3060 2.2669 1.9490 2.0157

A. Case: X = F_1

- Accuracy = 53%
- Error Rate = 47%

B. Case: X = Z_1

- Accuracy = 88.31%
- Error Rate = 11.69%

C. Case: X = F_2

- Accuracy = 55.09%
- Error Rate = 44.91%

D. Case: X = [Z_1 F_2]

- Accuracy = 97.98%
- Error Rate = 2.02%

4. Comparison

When we used F_1 and F_2, we did not achieve high accuracy, because they were not consistent across all the individuals and hence our model was not trained properly in these cases. However, after normalizing, Z_1 alone was able to achieve a good accuracy of prediction as we were able to get clearing separated values for each class. The accuracy further improved when we used both Z_1 and F_1 for modeling and prediction, since F_2 was also a significant predictor for our classification problem.

5. Conclusion

The project successfully shows us how we can use Bayes' theorem to classify the data points and how we can improve the accuracy of our Bayes' classifier through Z-transformation of the reported data. In particular, we saw that normalized data had a well-defined score boundaries which improved classification accuracy. Additionally, taking two variables together further improved it since both predictors had significant contributions.

6. References

Introduction to Probability (2nd Ed.) by D.P. Bertsekas and J. N. Tsitsiklis (Athena Scientific) Devore (Main Textbook)