EAS 595

Introduction to Probability, Spring 2020

Analysis of classifier performance by normalization of data points

Kashyap Balasubramanian

UB ID: 50325440

kashyapb@buffalo.edu

University of Buffalo, New York

Ashish Sanghvi

UB ID: 50318479

ashishsa@buffalo.edu

University of Buffalo, New York

We use of Bayes' Rule for predicting the class of the input data. We also experiment with transformation of the data in order to improve the accuracy of the classifier. The improvement in the accuracy is observed and visualized to understand that collected data often need to be normalized in order to make it more meaningful. Finally, a multivariate classifier is also constructed and we compare its results with univariate counterpart for various cases.

The data was collected from an experiment involving 1000 participants who performed 5 different tasks (also called class). There were two different measurements taken namely F1 and F2 for each class. 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.

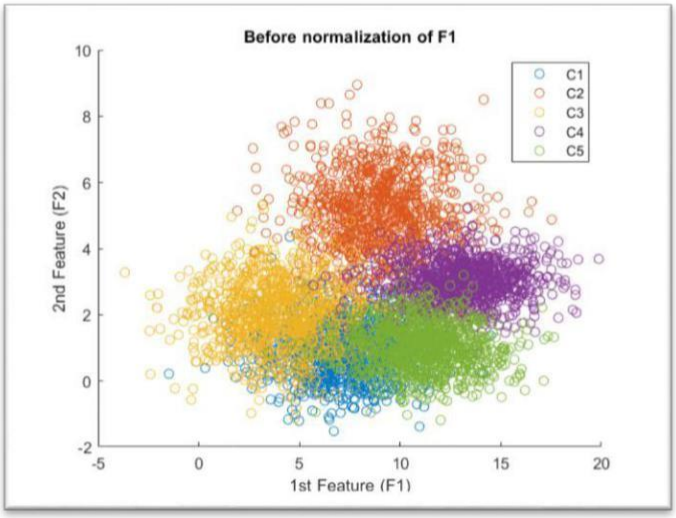
The first 100 observations were used to train the model. The training consisted of calculating the mean and variance of each column (class). Using Bayes' theorem, we know that: -

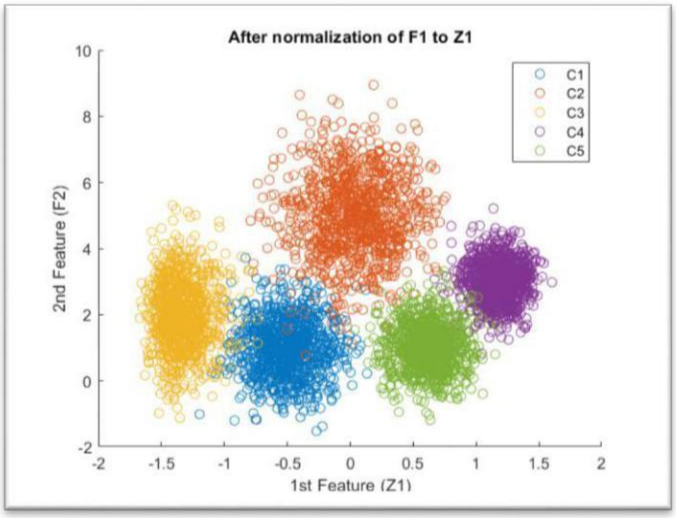
$$P\,(F_1 \mid C_i) = N\,(m_{1i} \mid \sigma_i^2\,) \qquad (1)$$

$$P\,(F_2 \mid C_i) = N\,(m_{2i} \mid \sigma_i^2\,) \qquad (2)$$

Since all other factors are common for the classes. Therefore, we can use the probability density functions, for each of the class, to compare the probabilities and hence classify the data points.

Since F1 can be a subjective measure, it is highly probable that the measurements reported by different participants can have different means and range of values. Hence, we normalized the measurements for each class, reported by every individual. The normalized measurements were named Z1 (ZScore) which were (F1i − μi)/σi ,i = 1, ..., 1000.





The pre-normalization and post-normalization can be visualized in Fig. 1 and 2 respectively. We can see that after normalization, the 5 different classes became very well separated and there were well defined boundaries between them. This was because the values reported by different individuals, for each class, were having different means and ranges. This led to overlap of range of values as evident in Fig. 1. This enabled a much better classification as seen later. The classification with and without normalization, were compared.

The univariate and bivariate classifiers were tested on the remaining 4500 (900 × 5) data points, for different cases and the accuracy, in each case, was calculated as (the number of correct predictions / total number of predictions). Similarly, the error rate was calculated as (the number of incorrect predictions / total number of predictions). The following cases were considered, and their performance obtained:

A. Case: X = F1

• Accuracy = 53%

• Error Rate = 47%

B. Case: X = Z1

• Accuracy = 88.31%

• Error Rate =11.69%

C. Case: X = F2

• Accuracy = 55.09%

• Error Rate =44.91%

D. Case: X = [Z1 F2]

• Accuracy = 97.98%

• Error Rate = 2.02%

When we used F1 and F2, 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, Z1 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 Z1 and F1 for modeling and prediction, since F2 was also a significant predictor for our classification problem.

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. We saw that normalized data had a well-defined score boundaries which greatly improved classification accuracy. Additionally, taking two variables together further improved it since both predictors had significant contributions.

Citations and References

[1] Introduction to Probability, Second Edition - Dimitri P. Bertsekas and John N. Tsitsiklis, Massachusetts Institute of Technology