EAS 595: Probability Project

1^{st} Vineel Kurma
MS Data Science
University at Buffalo
Buffalo, US
vineelku@buffalo.edu

2^{nd} Eswar Chand Jonnakuti
MS Data Science
University at Buffalo
Buffalo, US
ejonnaku@buffalo.edu

***Abstract*—The purpose of this project is to determine the class of a participant. The maximum probability was taken as the parameter to determine the class. This process was repeated for both the data in F1 and F2. And then finally compared with the normalized part of F1. It was very interesting to see how normalization has impacted the classification of classes.**

I. Introduction

We initially estimated the mean and variance for each class in F1 & F2. Only 100 records were used for the above training. The probabilities of each of the 4500 records were estimated using the above training parameters. The maximum probability was taken to the class decider. This process was repeated for F1, F2, z1, F1&F2 and F2&Z1.

II. Process Flow

Mean and Variance were calculated for each of the first 100 records in the training process. These mean and variance were used to get the probabilities of each record from 101 to 1000. The maximum probability was used to get the final class of the record.

$$Predicted\ Class = argmax[P(Ci/X)] \tag{1}$$

After the predicted class is determined for each record in F1, we replicated this process for F2. Both of them were compared using their classification accuracy. The formula goes as below:

$$Classification\ accuracy = CP/TP \tag{2}$$

$$CP = Correct\ Predictions \tag{3}$$

$$TP = Total\ Predictions \tag{4}$$

The Classification accuracy came out as:
F1 = 53%
F2 = 55%

After predicting for both F1 & F2, we normalized all the records in F1 by using the formula:

$$z_1 = X_{Normalized} = (X - \mu)/\sigma \tag{5}$$

Now we will have a distribution of N(0,1) for z1. This is done in order to remove the effect of individual differences of each participant. Now we repeat the classification process for the combination of z1&F2. This time we put the mean and standard deviation values in a matrix and proceed with our accuracy estimation. The accuracy for this case came out to be:
z1&F2 = 96%

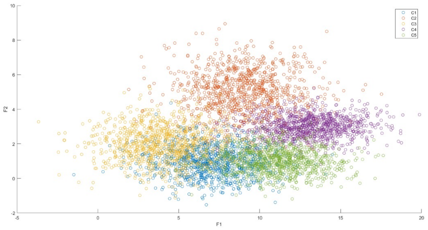

Fig. 1. F1 vs. F2 (color represents class)

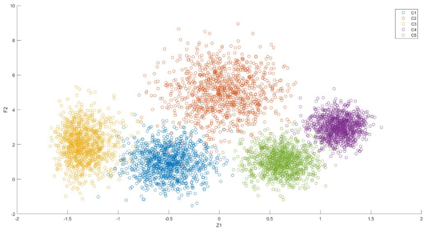

Fig. 2. z1 vs F2 (color represents class)

III. Understanding & Conclusion

We can see from Fig.2 that after taking the individual differences into account, there is a clear split between the classes. All the classes are now clearly distinguished from each other. We also got an accuracy of about 88.34% for z1 alone. This shows that before normalization the differences in variable ranges could potentially affect negatively to the performance of our algorithms.

Standardizing the features so that they are centered around 0 with a standard deviation of 1 is not only important if we are comparing measurements that have different units, but it is also a general requirement for many machine learning algorithms. Intuitively, we can think of gradient descent as a prominent example (an optimization algorithm often used in logistic regression, SVMs, perceptrons, neural networks etc. with features being on different scales, certain weights may update faster than others since the feature values ’X’ play a role in the weight updates.

A. *Comparison between all the classifications done*

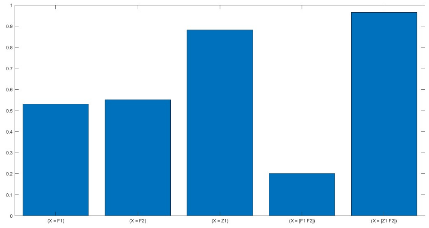

Fig. 3. Models vs Accuracy

Fig 3. is an illustration of the accuracy of all the cases we’ve experimented with. As stated above z1&F2 had the highest accuracy and F1&F2 had the lowest accuracy.

REFERENCES

[1] https://sebastianraschka.com/Articles/2014_about_feature_scaling.htmlabout-standardization
[2] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[3] https://www.mathworks.com/help/stats/normal-distribution.html