EAS 595 Final Project

Exploration of the Normal Distribution

Matthew Sah
Engineering Science (Focus on Data Science)
School of Engineering and Applied Sciecce
Buffalo, NY
msah@buffalo.edu

Ye Shi
Engineering Science (Focus on Data Science)
School of Engineering and Applied Sciecce
Buffalo, NY
yshi23@buffalo.edu

Abstract

Through some simple process of training, predicting and visualizing data, we show some characteristics of the Normal Distribution

I. Introduction

With a intention of observing the characteristics of the normal distribution. We try observing the characteristics of two dfferent features(F1, F2) as a raw data, then we process the data to normalized form then replot to see how much dfference it made. With some processing we also categorize difference values to it's resective class.

II. Data Description

A. *data.mat*

The data using in this project is a .mat file named data. There are five contents in this file, some meta data regarding the production and the version of the data, and the data regarding the two features(F1, F2) that we will be doing our processes on.

B. *Data loading and processing*

Under the python platform, data loading was prompt and was as simple as using the more traditional Matlab language. Data Processing wasn't required in our particular test case.

III. Methods and Tools

A. *Tools*

Python for ease of use and the abundance of library and community support.

Jupyter Notebook/Lab to immediate recognize errors in code and results in processes.

B. *Normal Distribution*

The two measurements are independent and for each class they can be considered to have a normal distribution as follow:

$P(F_1|C_i) = N(m_{1i}, \sigma_{1i}^2)$ and $P(F_2|C_i) = N(m_{2i}, \sigma_{2i}^2)$ for $i = 1,2,\cdots 5$

C. *Predict the class*

the classifier calculates the probability of each class given the measurement data, and output the most probable class as the predicted class.

$$Predicted\ Class = \text{argmax}[P(C_i|X)]\ , i = 1,2,\cdots 5$$

IV. Process and Results

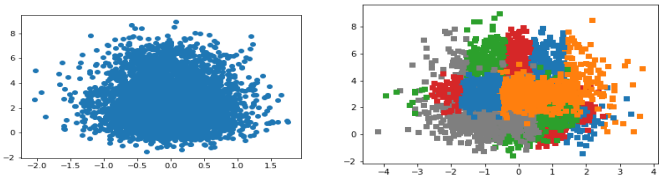
A. *Mean and variance of F1*

```
m11 = 7.093276745822095    var11 = 4.242083092666572
m12 = 9.144547521590674    var12 = 5.264687407441236
m13 = 4.287691491993289    var13 = 5.087568385990322
m14 = 13.33749006747352    var14 = 3.760788754269051
m15 = 11.24185889702683    var15 = 4.022309590554696
```

B. *Mean and variance of F2*

```
m21 = 0.9437745580004682  var21 = 0.71920000947922
m22 = 4.97942762175956    var22 = 1.3795600340845644
m23 = 1.8343812239039996  var23 = 1.0659821430276346
m24 = 3.0140986679897996  var24 = 0.27699546926623164
m25 = 1.0249099445406251  var25 = 0.4612314887016689
```

C. *Plot of Standard Normal (Z1-F2)*



Plotted image of Z1 relative to F2, the graph on the left shows the distribution of Z1 and F2, the colored graph on the right is sorted by class. In relative to the graph of two raw features, we feel like the normalized data is more compacted . Graphing it also shows a stronger line for splitting classes.

D. *Test result for case1: X = F1*

For this test case where we use F1 as X, we received and accuracy of 53%.

E. *Test result for case2: X = Z1*

For the test case of when X is equaled to Z1, the normalized value of F1, has a mere accuracy of 20%. We had originally expected Z1 to receive an greater accuracy, however after multiple tries and edits through our experiment, we could not receive a higher accuracy. The Z values were verified to be correct.

F. *Test result for case3: X = F2*

For the test case of X is equaled to F2, we received an equally high accuracy as the 1st feature F1 of 55%, we deduce that our prediction method was correct.

G. *Test result for case4: X = [Z1,F2]*

For this test case we received a mere 8% accuracy, we suspect that we had errors in the calculation of normalized data however after several tries we were unsuccessful to improve the accuracy.

H. *Compare the classification rate of the four cases*

We deduce that Normalization, although is just closing the difference in the data and to make it easier to calculate, would either require another method for calculation or we had made an mistake. Since the accuracy were relatively high on the options on raw data, however after normalizing we wold expect an more accurate answer but was not the case for us.

CONCLUSION

In conclusion, we believe with a model we could possible easily score an higher initial accuracy for the raw features(F1, F2); and also lift the accuracy of the normalized data higher.

For future references, we could do test to verify the speed of normalized data and raw data to see if normalized data significantly affects the efficiency in data processing, or even classify if other benefits exists.

Acknowledgment *(Heading 5)*

Thanks to Professor Dietrich Kuhlmann for his time and patience throughout the semester, he has given an abundance of chances and opportunity to enhance our skills and understandings on the topic.

REFERENCES

[1] https://en.wikipedia.org/wiki/Bayes%27_theoremI

[2] https://blog.csdn.net/google19890102/article/details/45672305

[3] https://blog.csdn.net/qq_31589695/article/details/80330126

[4] https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.multivariate_normal.html