EAS 595 Project Report

Introduction to Probability Theory for Data Science

Shubham Sharma
Graduate Student - Data Sciences
University at Buffalo
New York, United States
ss628@buffalo.edu

***Abstract*—The given problem statement requires us to process the data in various steps and obtain the conclusion for the best accuracy and error rate:**

- **Step 1: Construct a training set of 100 rows from F_1, F_2**
- **Step 2.1: Test the remaining 900 rows of the data set i.e. F_1, F_2. Calculate the probability of each class of the data and choose the class with maximum probability, i.e. the most probable class.**
- **Step 2.2: Determine the classification accuracy and error rate of the test data i.e. rows 101 - 1000**
- **Step 3: Normalize the F_1 matrix row-wise for each subject which is called as Z_1. Compare the distribution of the Z_1 vs F_2 to F_1 vs F_2.**
- **Step 4: Repeat the steps 2.1 and 2.2 for**

$$X = Z_1, X = F_2, X = \begin{bmatrix} Z_1 \\ F_2 \end{bmatrix} \tag{1}$$

- **Step 5: - Compare the classification rates of all the four cases**

I. Step by Step by procedure

- Step 1: F1 data set has been used to create the model, the data set has been divided into training (100 rows) and test (900 rows). Since equal number of samples are there in each class, it is reasonable to assume that all the classes as equiprobable. As given in the question, continuous values associated with each class are assumed to be distributed according to a Guassian distribution. Mean and variance for each class is computed.

- Step 2.1: The corresponding probability distribution for all values is computed using the Normal distribution parameterized by the mean and variance computed in the training step. Thereafter, out of the five values of the conditional probability obtained, the one with the maximum value is treated as the most likely class. By Bayes theorem, the conditional probability P($F_1 \| C_i$) is proportional to P($C_i \| F_1$) since P(C_i) and P(F_1) are constants.

- Step 2.2: Accuracy is found as the number of correct predictions made/total predictions. The error rate is incorrect predictions/total predictions.

- Step 3: The data in F_1 was normalized to make each row centre around the same mean (0) and have the same standard deviation (1) which segregates the different classes better as is evident in the figure 1. This helps in putting all the subjects on a comparable scale. Without normalization, the range of values for each subject was large and inconsistent with other subjects due to which each class had overlap with each other and the resulting plot was diffused as shown in figure 2. After normalization, each class gets a different range of values for itself over the different subjects as the effect of individual differences is removed. This is because F_1 is a subjective measure and the range of values is different for different individuals.

- Step 4: The same practice was repeated for F_2 as well as Z_1. In case of the multivariate normal distribution of (Z_1;F_2) we assume that the features are independent of each other thus the resulting conditional probabilities for the multivariate distribution would simply be proportional to the product of the conditional probabilities for Z_1 and F_2 under the conditional independence assumption.

Dataset	Accuracy (in %)	Error Rate (in %)
F_1	53.00	47.00
Z_1	88.33	11.67
F_2	55.16	44.84
X = (Z_1;F_2)	97.98	2.02

TABLE I
Accuracy and Error Rates for each case

II. Results

The results are summarized in table I which tells the accuracy as well as the error rate for each case. Following can be concluded for each case:

- The accuracy is lowest for the case of dataset F_1 since it is a subjective measure and have a range of values for each subject which overlap within classes. Thus, the model obtained is not able to predict the correct classes with high accuracy. The case for F_2 have a similar line of reasoning and for the same reason the errors for both are close.

- The accuracy is improved for the normalized version Z_1 since each class has a better representation now and

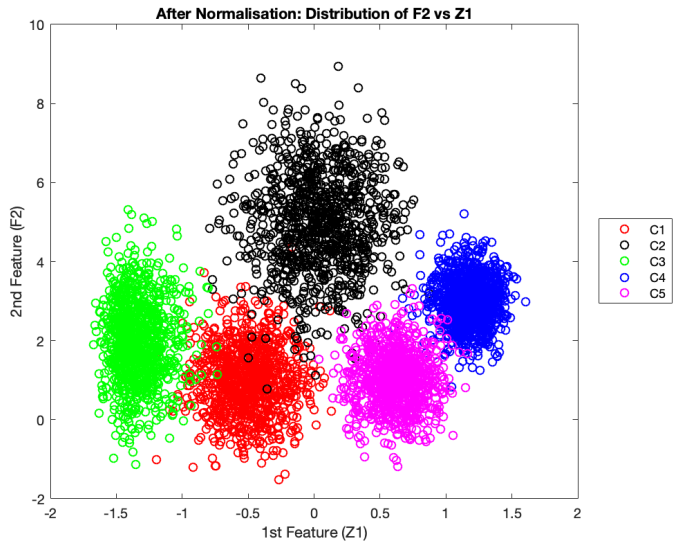

Fig. 1. Distribution of F_2 vs Z_1 after normalization

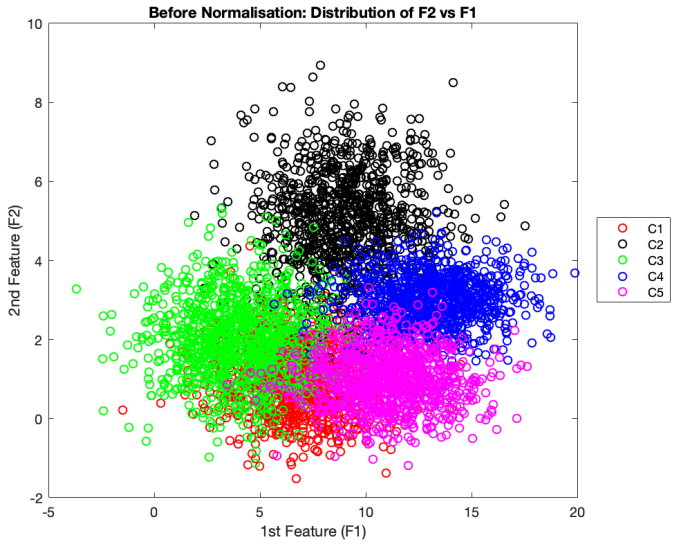

Fig. 2. Distribution of F_2 vs F_1 before normalisation

the overlap between different classes is removed. Each row is now centered with a common mean of 0 and standard deviation of 1 due to which each class gets a well defined range of values.

- The accuracy for the multivariate distribution, $X = (Z_1;F_2)$ is highest because it has more variables which it is considering in the model because of which the predictions come out to be better. Also, because of the independence assumption there is no correlation between the two features.

III. Discussion

In this case we consider each distribution as a one-dimensional distribution independent of the other distribution. Also, the only requirement for correct classification in this case is that the correct class be more probable than any other class. This is true regardless of whether the probability estimate is slightly, or even grossly inaccurate. The classifier, here tries to find posterior probabilities based on prior probabilities using Bayes Theorem. The posterior probabilities are characterized in terms of only two numbers, the mean and the variance.

References

Following book was used for reference. [1].

References

[1] Introduction to Probability, Bertsekas, D.P. and Tsitsiklis, J.N.