BAYES CLASSIFIER

Vineet Suresh Kothari
State University of New York at Buffalo

Viswanathan Gopalakrishnan
State University of New York at Buffalo

ABSTRACT

This project describes the classification model created to predict the most probable task performed by a participant, in an experiment involving thousand participants performing five different tasks, across two different measures of evaluation. Bayes theorem was used to determine the most probable task a participant would perform, given the measure of evaluation. This probability was evaluated using central limit theorem, using the mean and standard deviation of the data. This was repeated for different versions of the measurement data. The classification rate in each case was recorded and observations were made**.**

1.1 DATA DESCRIPTION

The given dataset contains data of 1000 participants performing 5 different tasks. Two measurements F1 and F2 were recorded. Each row in table F1 and F2 corresponds to the tasks performed, whereas each column contains information of each task..

1.2 PROCEDURE

There were four cases to deal with, to identify the best Bayesian classifier. Two cases involved the use of table F1 and F2 respectively. The third case involved the use of a normalized version of F1, labelled as Z1. The fourth case involved the use of a multivariate normal distribution, which consists of Z1 and F2. The following steps were performed for the first three cases:

1. Split the dataset into train and test dataset containing 100 and 900 observations respectively.

2. Calculate the mean and standard deviation of the train data, for each task.

3. Create an array containing the indices 1,2,3,4 and 5 replicated across 900 rows. This will be referred to as true class values.

4. Calculate p-values for each observation in the test data using each mean and standard deviation obtained from the previous step, for each task.

5. Select the index (task number) of the maximum p-value out of the five p-values, for each observation in the test data. This array will be referred to as predictions.

6 .Calculate the classification accuracy and error rate by comparing the true class values and predictions.

In the latter case, that is, where the data contains both Z1 and F2, two probabilities were calculated. One set of p-values were calculated using the mean and standard deviation of Z1 (which was previously computed), whereas another set of p-values were calculated using the mean and standard deviation of F2. The final probability of each task was calculated by multiplying the above two p-values, since the distributions were independent:

1.3. RESULTS AND OUTCOMES

Github code repository: https://github.com/MicroprocessorX069/Generalized-Bayes-classifier-/blob/master/Bayes_classifier.ipynb

Github documentation:

Case 1: Training with 100 samples of feature ‘F1’

Accuracy: 0.53 Error rate : 47

Case 2 Training with 100 samples of normalized feature ‘Z1’

The feature ‘F1’ is standard normalized to the distribution Z(0,1)

Z=X - E(X) / std_dev(X)

Now since each of the classes have a different mean and standard deviation, each class’s data points have to be normalized with their respective mean and standard deviations.

Results: Accuracy: 0.20 Error rate : 0.80

Case 3: Training with 100 samples of feature ‘F2’

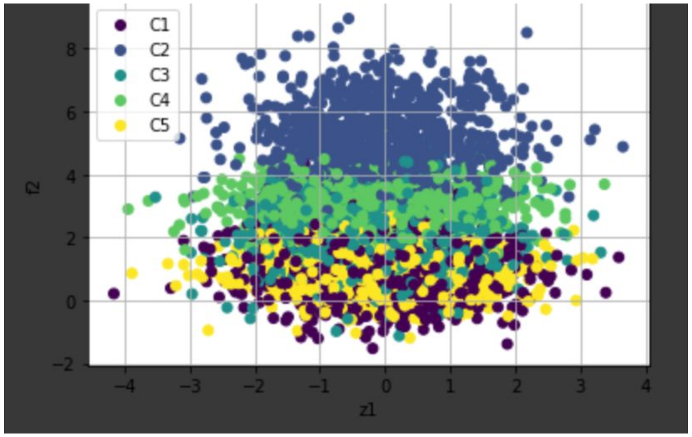
Results: Accuracy: 0.55 Error rate : 0.45

Case 4: Training with 100 samples of features [‘Z1’,’F2’]

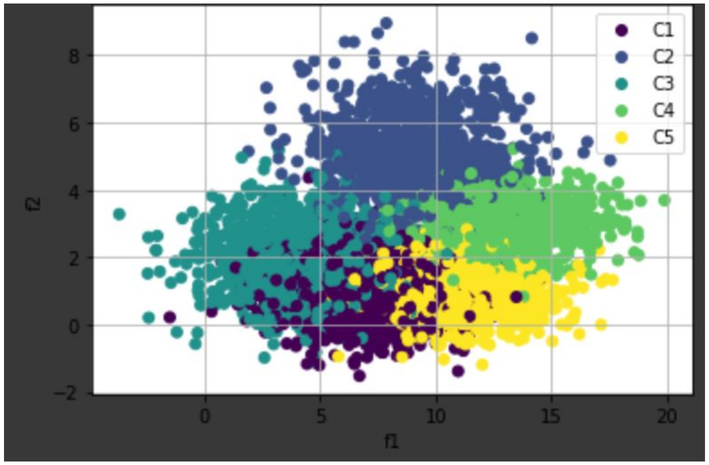
Here the formula for PDF of a multivariate normal distribution involves a list of means of the variables and covarience matrix of the variables.

Results: Accuracy: 0.55 Error rate : 0.45

Graph of features Z1 vs F2



Graph of features F1 vs F2



Vineet Kothari (50291159, vineetsu@buffalo.edu)

Viswanathan Gopalakrishnan (5081280304960148
vgopalak@buffalo.edu)