Bayes' Classifier

Freya Genesis D Mello
School of Engineering and AppliedSciences
State University of New York
freyagen@buffalo.edu

Krina Joshi
School of Engineering and AppliedSciences
State University of New York
krinahit@buffalo.edu

Abstract**:** The Goal of project is to Construct a classifier such that for any given values of $F1$ and $F2$, it can predict the performed task ($C1$, $C2$, $\cdots$ $C5$). The classifier calculates the probability of each class given the measurement data, and output the most probable class as the predicted class.

Keywords *– Classification, Bayes Theorem*

Introduction:

Statistical inference is the process of extracting information about an unknown variable or an unknown model from available data. This is an experiment involving 1000 participants, with recorded two different measurements ($F1$ and $F2$) while participants performed 5 different tasks ($C1$, $C2$, $\cdots$ $C5$). The two measurements are independent and for each class they can be considered to have a normal distribution.

I. Training

A. Sampling the data

First we divide the data into training (100 samples) and testing data (900 samples).

B. Estimation of mean and variance

Then we calculated the mean and standard deviation of F1 for training set.
Mean: 7.0933 9.1445 4.2877 13.3375 11.2419
Standard Deviation: 2.0700 2.3060 2.2669 1.9490 2.0157

II. Testing

A. Normalizing F1

Using the Bayes' theorem, we then calculated the Z-Scores of each class for data of the testing set (101-1000 of F1) and consequently predicted the class for each data point i.e. the class having maximum probability.

Estimating the Accuracy

Then by comparing the normalized F1 data and the original data, we calculated the Accuracy and Error rates by

Classification accuracy = correct predictions / total predictions
Error rate = incorrect predictions / total predictions

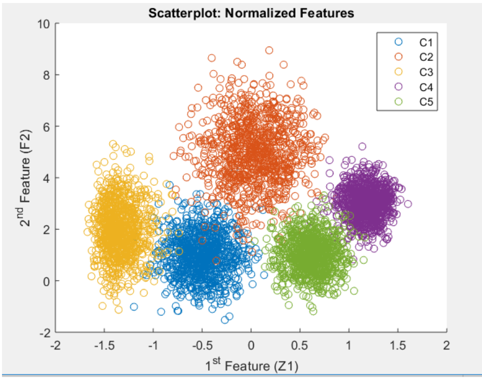
Results

Accuracy	48.2222	28.8889	76.7778	66.2222	44.8889
Error rates	51.7778	71.1111	23.2222	33.7778	55.1111

Accuracy : 53%
Error Rates: 47%

B. Plotting the Normalized Z1 and F2

We then normalized the data of each subject using the standard normal formulation i.e. by removing the mean and dividing by standard deviation. Calculated the standard normal of $F1$ (Z1) and then we plot the distribution of the data using $Z1$ and $F2$.



C. Normalizing Z1

- We then again calculated the Z-Scores of each class for data of the testing set (101-1000 of Z1) and consequently predicted the class for each data point i.e. the class having maximum probability.
- Then by comparing the normalized Z1 data and the original data, we calculated the Accuracy and Error rates :

Accuracy	84.7778	73.0000	98.5556	96.4444	88.7778
Error rates	15.2222	27.0000	1.4444	3.5556	11.2222

- Accuracy : 88.3111%
  Error Rates: 11.6889%.

D. Normalizing F2

- Then again normalized the F2 and calculated its Accuracy and Error rates.

Accuracy	23.3333	83.3333	26.5556	77.7778	64.4444
Error rates	76.6667	16.6667	73.4444	22.2222	35.5556

- Accuracy: 55.0889%
  Error Rates: 44.9111%

E. Multiplying Z1 and F2

- Multiplying Z1 and F2 (this is a multivariate normal distribution and hence we need to use the independence assumption.) We get,
- Accuracy: 97.9778%

Error Rate: 2.0222%

CONCLUSION:

Normalization helps to segregate and classify the data correctly. Because the data was initially scaled differently with respect to every data point, we got better results on normalization as it can be seen distinct clusters in the plot.

When Z1 and F2 were combined into bivariate data, it gave better accuracy and lesser error rates because bivariate data involves relationships between the two variables also it considers the correlation between two variables, while the information we would gather from univariate data would be about its distribution, such as the range and the mean.