Probabilistic Models

Nikita Goswami 50320909
University at Buffalo

***Abstract*— to build a classifier using Naïve Bayes method, standard normal distribution, Multivariate Normal Distribution and compare the results**

Keywords— Classification, Naïve Bayes, Normal Distribution, multivariate normal distribution

I. INTRODUCTION

We have the records of two different measurement ($F1$ and $F2$) of 1000 participants performing 5 different tasks ($C1$, $C2$, C3, C4, $C5$). Our task is to construct a classifier such that for any given values of F1 and F2, it can predict the performed task. We have simplified the learning by assuming that measurements/features are independent for each class and they can be considered to have a normal distribution. Although independence is generally a poor assumption, in practice models built using this assumption often competes well with more sophisticated classifier which do not take the independence assumption.

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

$Predicted\ Class$ = argmax$[P(Ci|X)]$, i = 1,2, ⋯ 5

II. ABOUT THE DATA

The data contains measurements F1 and F2 that are both matrices with the size of 1000x5. Each column contains the information of one of the subjects and each row corresponds to one of the tasks (1st row: 1st task, 2nd row: 2nd task, etc.).
The two measurements are independent and for each class they can be considered to have a normal distribution as follow:

$P(F1|Ci) = N(m1i, \sigma 1i2)$
$P(F2|Ci) = N(m2i, \sigma 2i2)$
for i = 1,2, ⋯ 5

III. APPROACH

There are three kind of classifiers built - First using Naïve Bayes approach for F1, Second using Standard Normal (Z-scores), First using Naïve Bayes approach for F2, Fourth using Multivariate Normal Distribution combining Naïve Bayes and Standard Normal approach.

A. Naïve Bayes Classifier

This theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Bayes theorem is shown below

P(C1|X) = P(X|C1) * P(C1) / P(X)

Similarly, for C2, C3, C4 and C5

- P(C1)=P(C2)=P(C3)=P(C4)=P(C5), as there were 100 datapoints for every class, this metric wouldn't play a role in deciding the class of any new datapoint. Since P(X) term will be same for all, this also doesn't play a role deciding the class
- The term PX|C1) will solely influence the classification, and the same metric is calculated by studying 100 datapoints for every class. As Bayes theorem works well on categorical data, the given data points are converted into discrete observations (10.49 is converted as 10, 11.02 is converted as 11, 12.88 is converted as 13), post which the frequency of every datapoint is computed for a given class. The intuition here is, if the data point is in class 1, then there is p% probability that the number is X. Using this probability distribution, the new datapoints are scored and the respective probability of being a particular class has been computed
- Once the probability of a datapoint being a particular class is computed, the datapoint is attributed to a class with maximum probability

  $Predicted\ Class$= argmax $[(Ci|X)]$,i=1,2,⋯5
- This approach was followed with F1 score for the first classifier, and with F2 for third classifier

B. Standard Normal (Z-Score):

The inherent distribution of the dataset provided is normal distribution, which is used in second classifier. 100 datapoints for every class has been used to compute mean and standard deviation, based on which every point will be standardized using the below formula

Z = (X – μ) / σ
μ - mean
σ - standard deviation

- Inbuilt function in MATLAB – normpdf has been leveraged, based on the parameters the new datapoint can be evaluated, and the probability of it being a

particular class can be computed based on the respective mean and standard deviation
- The datapoint would be attributed to the class with the maximum probability

C. *Multivariate Normal Distribution:*

F1 and F2 dimensions are independent and this assumption can be used in fourth classifier. As they are independent, the joint probability is the product of their individual probabilities

- Second classifier had used Z1 dimension to compute the probability of every datapoint being in class1, 2, 3, 4, 5. Similarly third classifier had used F2 dimension to compute the probabilities for every point being in one of the 5 classes
- When these 2 probabilities are multiplied, the resultant would be the probability where according to both Z1 and F2 the datapoint being in the respective class

IV. PROCEDURE

- Training: First 100 data points are used for training the algorithm. They are also used for normalizing the other data points by using their mean and standard deviation. Training step is performed separately for both approaches: Naïve Bayes Algorithm and Z-score approximation using 100 points
- Testing: Testing the 4500 points by calculating the probability of each class for data of the remaining subjects (columns 101-1000 of *F*1 and F2) and consequently predict the class for each data point.
- Attributing a class to a datapoint: Every datapoint will have a probability of being in one of the 5 classes, and the class with maximum probability will be attributed to the datapoint
- Calculating Accuracy: Out of 4500 datapoints, for how many datapoints the attributed class is the correct class? This metric denotes the classification accuracy for a classifier. The inverse of this metric would be the error percentage

V. RESULTS AND CONCLUSION

- For Class F1, we have used Naïve Bayes as well as Standard Normal for Classification and we see from the accuracy rates that the Standard Normal performs better than Naïve Bayes. From this observation, we can infer that normalization of Data helps in classification
- Further, we combined Standard Normal probabilities of Class F1 (Z1) along with Naïve Bayes Probabilities of class F2 into a joint probability. Since we have assumed that all classes are independent of each other, on calculating the probability, this approach resulted in better accuracy rates compared to individual Naïve Bayes and Standard normal for different measurements

Algorithm	Accuracy
Naïve Bayes for F1 (Case 1: X=F1)	52.20%
Normal Distribution (Case 2: X=Z1)	53.00%
Naïve Bayes for F2 (Case 3: X=F2)	53.42%
Multivariate Normal (Case 4: X= [Z1 F2])	77.56%

TABLE 1: ACCURACY

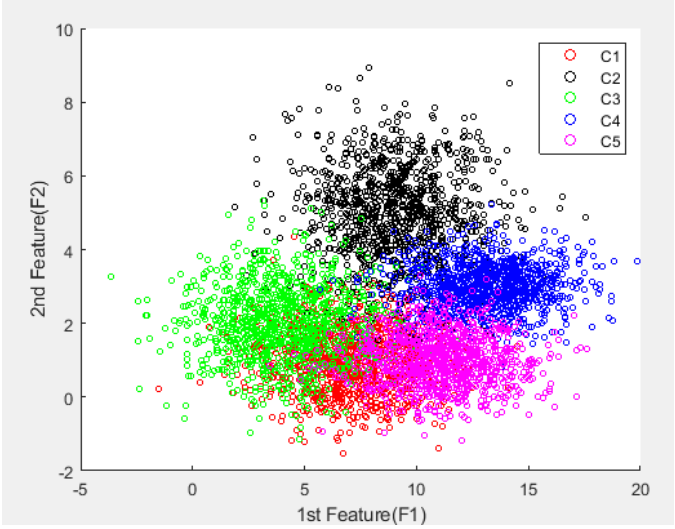

Fig 1. Plot of F1 and F2

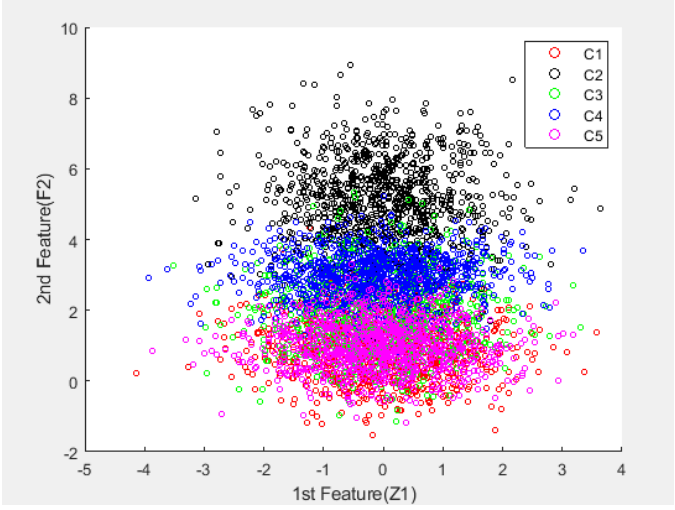

Fig 2. Plot of Z1 and F2 (Case 4)