EAS595 - Probability Project

Chaitanya Pawa
School of Engineering and Applied Sciences
SUNY Buffalo
Buffalo, NY
cpawa@buffalo.edu

Vineel Patnana
School of Engineering and Applied Sciences
SUNY Buffalo
Buffalo, NY
vineelpa@buffalo.edu

***Abstract*—The goal of this project is to construct a classifier such that for any given measurements F_1 and F_2, we can predict the performed task (C_1, C_2, C_3, C_4, C_5)**

1) **Using a classifier to segregate into different classes for various tasks.**
2) **We use various classification techniques(Boosting, Decision Trees, Neural networks) to classify into different classes.**
3) **However, we used the powerful Bayes Theorem to classify into different classes taking help of log-likelihood, prior and the evidence probabilities.**
4) **We performed the classification on Multivariate distributions which shows the maximum accuracy.**

***Index Terms*—Bayes theorem, classifier, z-score, normal distribution, multivariate normal**

I. Introduction

In an experiment involving 1000 participants, we recorded two different measurement (F_1 and F_2) while participants performed 5 different tasks (C_1, C_2,.., C_5).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}, σ^2_{1i}) and P (F_2 | C_i) = N(m_{2i}, σ^2_{2i}) for i = 1,2,..,5 where m_{1i}, σ^2_{1i} are the mean and variance of F_1 for the i^{th} class and m_{2i}, σ^2_{2i} are the mean and variance of F_2 for the i^{th} class.

Using Bayes Theorem to build a Naive Bayes classifier to calculate the probability of each class given the measurement data, and output the most probable class as the predicted class

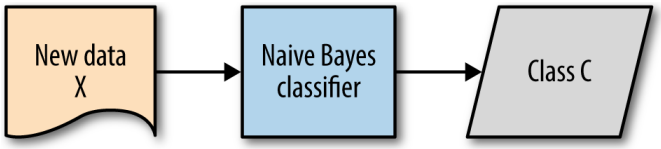

Fig. 1. Classification

II. Methods

1) We use the first 100 observations to find their mean and standard deviation to normalize the observation such that they will be comparable.
2) We use the rest 900 observations as the test set to classify the observations into different tasks.
3) We calculate the accuracy and error rates comparing the predicted class with the true class to understand the model.
4) In other to remove the effect of individual differences, we have to normalize the data of each subject using the standard normal formulation (removing the mean and dividing by standard deviation) to find Z_1 from F_1
5) We do the above for F_1, Z_1, F_2 and $\begin{pmatrix} Z_1 \\ F_2 \end{pmatrix}$

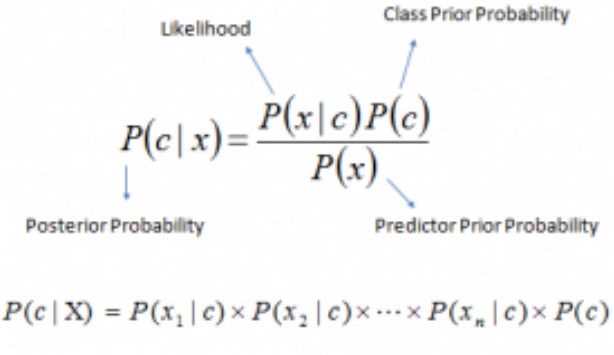

Fig. 2. Naive Bayes Classifier

6) We use likelihood, class prior probability and evidence (predictor prior probability) to find the posterior probability of the class given a particular measurement.

III. Results

TABLE I
Results

Measurements	Classification Accuracy	Error Rate
F1	52.62%	47.38%
F2	53.51%	46.49%
Z1	88.38%	11.62%
$\begin{pmatrix} Z_1 \\ F_2 \end{pmatrix}$	97.84%	2.16%

The classification accuracy was around 52% for both F1 and F2. However using Z-score to make classes comparable and then trying to predict observations have increased the accuracy to 88%.

When performing a multivariate normal the accuracy increased to 98%

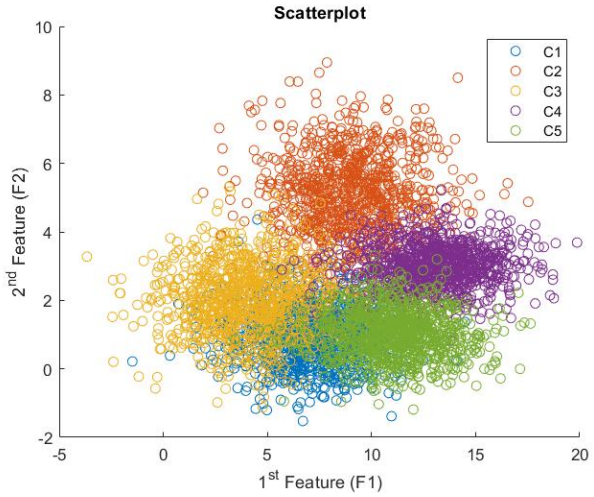

Fig. 3. Distribution of data using F_1 and F_2

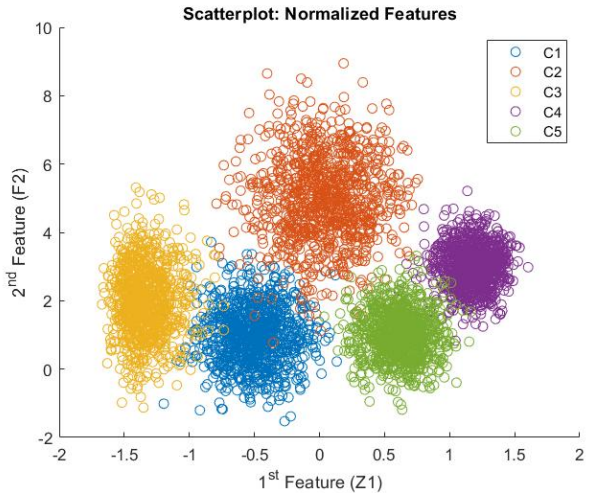

Fig. 4. Distribution of data using Z_1 and F_2

From the figures 3, 4 and 5 it can be clearly seen that normalizing F_1 to Z_1 and F_2 to Z_2 the data points move such that their mean is centered at 0 which makes values from different classes easily comparable and distinguished from each other since all the classes are having data under same scale thereby giving better accuracy rates.

Using the normal distribution pdf we can calculate the probabilities.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

For multivariate since it is given that Z_1 and F_2 are independent because of which ρ (correlation is 0) which changes the below formula for Multivariate to product of two normal distributions.

$$f(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

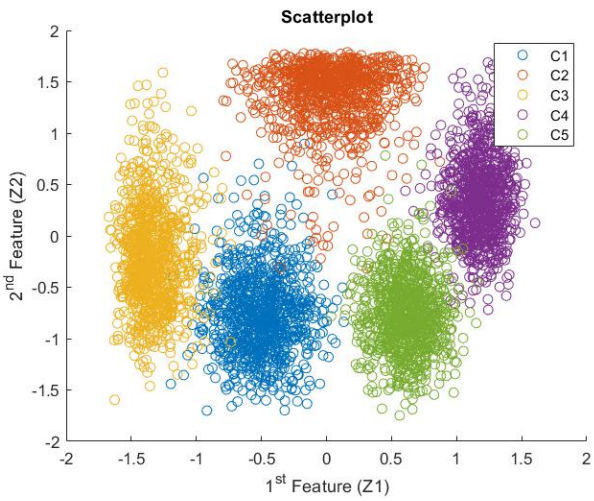

Fig. 5. Distribution of data using Z_1 and Z_2

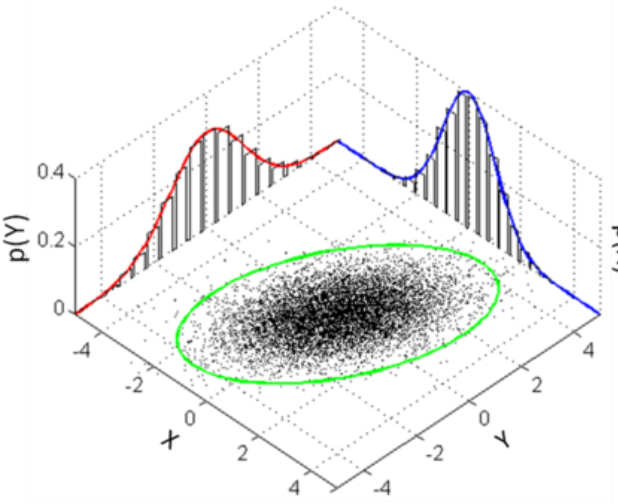

Fig. 6. Multivariate normal distribution

Hence we have multiplied the probabilities of Z_1 and F_2 and calculated argmax to find the exact class an observation belongs to. Due to multivariate normal distribution the inter-cluster distance increases and intra-cluster distance decreases making them predict more accurately than univaraiate (Refer. Fig.6).

IV. Conclusion

F_1, F_2 give 52% accuracy which can be increased using the normalization of F_1 into Z_1 there by by making the values in different classes comparable (Refer Fig. 4) the accuracy of the prediction class can be increased.

Multivariate (Bi-variate) Normal performs as the best classifier among all because it considers the relationship between different features in multiple data-sets and this relationship makes it predict better.

REFERENCES

- https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54
- http://www.statsoft.com/textbook/naive-bayes-classifier
- https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
- https://en.wikipedia.org/wiki/Multivariate_normal_distribution