EAS595 Project, Fall 2018

Kishore Ravisankar
Data Science
University at Buffalo
Buffalo, USA
kravisan@buffalo.edu

Priya Karadi
Data Science
University at Buffalo
Buffalo, USA
priyakar@buffalo.edu

***Abstract*—This report 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.**

Keywords—probability, classification, prediction, Bayes' theorem, Central Limit Theorem, mean, standard deviation.

I. Introduction

A. Bayes' theorem

Bayes' theorem shows the relation between a conditional probability and its reverse form. This theorem is named after Thomas Bayes and is often called Bayes' law or Bayes' rule. The equation for Bayes' theorem for two events A and B is represented as follows:

$$P(A|B)=\frac{P(A)\,P(B|A)}{P(B)}$$

Here, P(A) is the prior probability of A. It does not take into account any information about B. P(A|B) is the conditional probability of A, given B. It is also called the posterior probability. P(B|A) is the conditional probability of B given A. It is also called likelihood. P(B) is the prior or marginal probability of B. [1]

B. Naïve Bayes Classifier

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The naive Bayes classifier combines the model formed using Bayes' theorem, with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label $\hat{y} = C_k$ for a feature x_i, for some k as given in the below equation. [2]

$$\hat{y} = argmax\ p(C_k)\prod_{i=1}^{n} p(x_i|C_k), k \in \{1,2,..K\}$$

C. Central Limit Theorem

Central limit theorem states that when an infinite number of successive random samples are taken from a population, the sampling distribution of the means of those samples will become approximately normally distributed with mean μ and standard deviation σ/√n, as the sample size (n) becomes larger, irrespective of the shape of the population distribution. [3] A normal distribution is represented in Fig.1.

D. Z-Score

The standardized value of a normally distributed random variable is called a *Z*-score and is calculated using the following formula.

$$Z = \frac{x-\mu}{\sigma}$$

Here, x is the value that is being standardized. μ and σ are the mean and standard deviation of the distribution. The probability associated with a given Z-score can be calculated by looking it up in a Z distribution table. This probability is referred to as p-value, and represents the area under the curve of a normal distribution. [4]

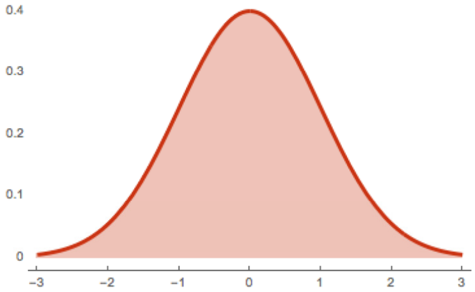

Fig. 1. Normal distribution with μ=0 and σ=1 [5]

II. Dataset

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.

III. Methodology

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 last 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. Now step 5 and 6 (as described above) were followed for this data.

IV. Observations

The classification accuracy and error rate for each case is tabulated in Table I.

TABLE I. CLASSIFICATION RESULTS

Dataset used	***Classification Accuracy***	***Error rate***
F1	0.53	0.47
F2	0.551	0.448
Z1	0.816	0.183
Z1, F2	0.939	0.06

It can be observed that the classification accuracy is poor for F1 and F2. This can be attributed to the absence of normally distributed data, that is, the scale of features are different for each class.

In the case where Z1 is used, a classification accuracy of 0.816 is obtained. This high accuracy can be attributed to the normalization of the data. Normalizing the data makes the classifier less sensitive to the scale of the features, and hence yields better consistency and accuracy. When both F1 and Z2 are used, we get an even better accuracy, which is due to the presence of highly normalized multivariate data.

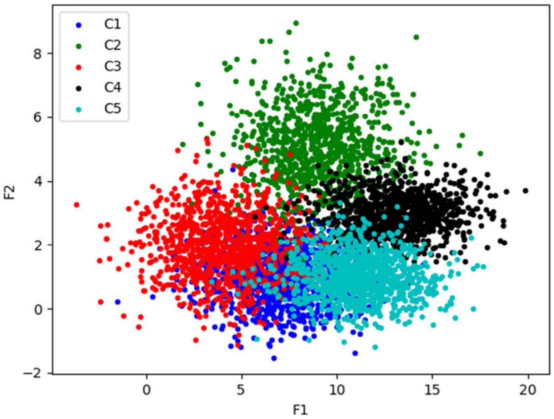

Fig. 3. Distribution of F1 against F2

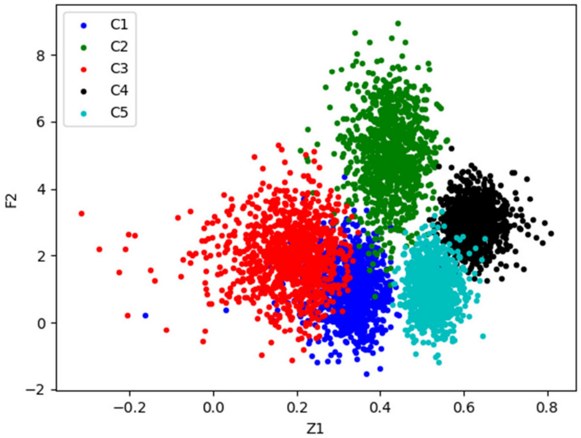

Fig. 4. Distribution of Z1 against F2

Figures 3 and 4 represent the distribution of F1 versus F2, and Z1 vs F2 respectively. It can be observed that each class is more distinguishable in Figure 4 when compared to Figure 3. To elaborate, the scale of each feature (task) is the same. This can again be attributed to the normalization of data.

V. Conclusion

In this project, the importance of normalization and performance of Bayesian classifiers were observed. If the data used in the classifier has independent features, then the classifier yields a good accuracy. It can be said that a multivariate classifier yields better accuracy compared to a univariate classifier, since a large number of data points and classes are used for prediction.

References

[1] https://simple.wikipedia.org/wiki/Bayes%27_theorem

[2] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[3] https://towardsdatascience.com/understanding-the-central-limit-theorem-642473c63ad8

[4] http://ci.columbia.edu/ci/premba_test/c0331/s6/s6_4.htm

[5] https://brilliant.org/wiki/normal-distribution