EAS 595 Final Project

Abhishek Dhyani

Nupur Sunil Agrawal

***Abstract*—For this project we have compared accuracy of Naive Bayes classifier on different data i.e F1, F2, normalized F1 and [Z1 , F2] considering independence condition.**

I. Introduction

In an experiment involving 1000 participants, data was recorded two different measurements while participants performed 5 different tasks. For this project, we were supposed to classify the score in any of the five classes namely C1, C2, C3, C4 and C5.

II. Idea Behind Classification

In machine learning, a Bayes classifier is a simple probabilistic classifier, which is based on applying Bayes' theorem. The feature model used by a naive Bayes classifier makes strong independence assumptions. This means that the existence of a particular feature of a class is independent or unrelated to the existence of every other feature.
Naive Bayes classifiers are linear classifiers that are known for being simple yet very efficient. The probabilistic model of naive Bayes classifiers is based on Bayes theorem, and the adjective naive comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well under this unrealistic assumption. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful alternatives .

III. Flow Of The Code

- We have used python to implement the project.
- For dealing with tables we have used pandas library to use data as a pandas dataframe.
- For performing mathematical function we have used numpy.
- In the beginning we have made functions to calculate basic statistical values and predict classes based on probabilities given by Baye's theorem.
- We have also made a function to determine accuracy of the classifier by comparing the true labels with the predicted labels.
- We have predicted labels for all of the 4 cases given and calculated the respective accuracy.
- To visualize the data, we have used seaborn library.

IV. Case 1 : X = F1

For this the input vector chosen was F1. To calculate mean and standard deviation of each class, we used the first 100 elements.
After training the data on 100 elements we used the other 900 entries to make predictions using Baye's theorem. Probability of an entry belonging to every class was calculated and the class which gave maximum probability was given as prediction.
We then calculated accuracy by checking the original class labels with the predicted labels. The accuracy came out to be 50 percent.

V. Case 2 : X = Z1

Here the input dataset is a normalized version of F1 called Z1. To normalize the data, we converted the data into standard normal form by subtracting the mean of each row of the dataset and dividing it by the standard deviation of the row. We then used the same approach to make preditions and calculate accuracy.
The accuracy came out to be 85.977 percent.

VI. Case 3 : X = F2

For this the input vector chosen was F2. To calculate mean and standard deviation of each class, we used the first 100 elements. We used the other 900 entries to make predictions using Baye's theorem. Probability of an entry belonging to every class was calculated and the class which gave maximum probability was given as prediction.
We then calculated accuracy by checking the original class labels with the predicted labels. The accuracy came out to be 55.08 percent.

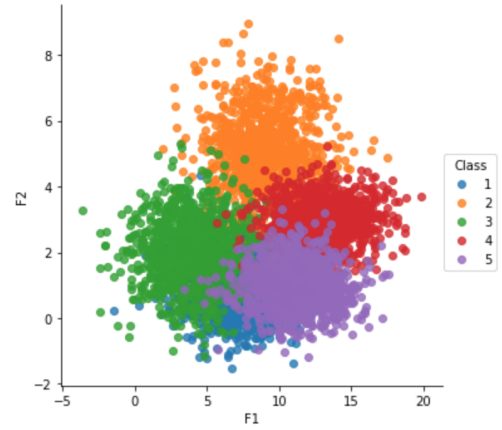
VII. Case 4 : X = Z1,F2

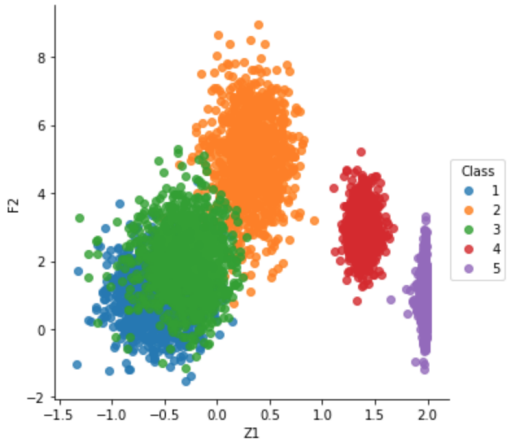
For this case we had to use multivariate data to perform classification. To create the multivariate data, we first made Z1 and calculated probability of an entity in Z1 of belonging to all the 5 classes. This process was repeated for each entity of Z1.
After this, the same procedure was adopted for F2. Now these probabilities were multiplied and the maximum was selected. Following this, first 100 values were used to get mean and standard deviations and rest 900 were used for prediction.
The accuracy came out to be 89.75.

VIII. Visualization

To get an idea of how the data is distributed we made 2 plots.

- F2 vs F1.
- F2 vs Z1.





IX. Results And Discussion

- Accuracy of F1 : 53.
- Accuracy of Z1 : 85.97.
- Accuracy of F2 : 55.08.
- Accuracy of multivariate : 89.75.
- From the plots we can clearly see that in F2 vs Z1 the classes are better separated than F2 vs F1. So the accuracy of Z1 is more than F1.
- While normalizing the data we make sure that data is not skewed, therefore it give better predictions.
- For the fourth case, since data is multivariate, it gives the best accuracy.
- Since we can make complex decision boundaries for multivariate data, hence it performs best.