EAS595: Final Project - Bayesian Classifier

Dithya Sridharan
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, United States
dithyasr@buffalo.edu

Hitesh Kailash Santwani
School of Engineering and Applied Sciences
State University of New York at Buffalo
Buffalo, United States
hiteshka@buffalo.edu

***Abstract*—Classification algorithms such as Bayes is an efficient model that is easy to learn and has a high accuracy in many domains. We propose a project to construct a classifier based on Bayes and Gaussian probability distribution to predict the performed task(Ci) for two independent measurements. We output the most probable class as the predicted class.**

***Index Terms*—Naïve Bayes, classes, independent attributes, Gaussian**

INTRODUCTION

A classifier, in general, produces a mapping from data (attributes) to two or more predefined classes. The Bayes and Gaussian classifier is a classification model that has several advantages [2]: it is easy to learn and understand, it is very efficient, it is in general robust and has a high accuracy.

Baye's theory is widely implemented as a classification method. Based on class conditional density estimation and class prior probability, the posterior class probability of a test data point can be derived and the test data will be assigned to the class with the maximum posterior class probability.

ABOUT BAYE'S CLASSIFIER

In probability theory, Bayes theorem relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations. Let x=(x1,x2,...,xd) be a d-dimensional instance which has no class label, and our goal is to build a classifier to predict its unknown class label based on Bayes theorem. Let C=C1,C2,...,CK be the set of the class labels. P(Ck) is the prior probability of Ck(k=1,2,...,K) that are inferred before new evidence; P(x|Ck) be the conditional probability of seeing the evidence x if the hypothesis Ck is true. A technique for constructing such classifiers to employ Bayes' theorem to obtain: P(Ck|x)=P(x|Ck)P(Ck)kP(x|Ck)P(Ck) Predicted Class = argmax[P(Ck/x)], k=1,2,3,4,5.

CLASSIFICATION EXPERIMENT

This project deals with the classification of two measurements (F1 and F2) involving 1000 participants whenthey performed five different tasks(C1,C2,....C5).It is assumed that P(F1/Ci) and P(F2/Ci) follows normal distribution. For this problem we use Bayesian classifier and Gaussian probability distribution to predict the tasks of each value of measurement(that is, which class they fall into). We have also considered multivariate classification and built a classification model where we will predict the task performed given the multinormal data (Z1, F2) where Z1 is the normalized value of F1 observation. Hence we will look into and analyse four cases :

1) Case 1 : X = F1
2) Case 2 : X = Z1(Z1 is normalized F1)
3) Case 3 : X = F2
4) Case 4 : X = [Z1, F2]

I. CASE 1

In this case, we build a Bayesian classifier where X = F1. We calculate the mean and the variance of the training set(We have taken the first 100 observations) and build our model. We get a classification accuracy of 30.26% and an error rate of 69.73% on the test data.

II. CASE 2

In this case, we build a Bayesian classifier where X = Z1. Here Z1 is the normalized value of F1. We normalize F1 by calculating the z-score. We subtract each value in each observation with the mean of all observations and then divide by the standard deviation of the observations. Fig 1 gives the plot between F1 and F2 and we can observe that most of the classes overlap and does not seem to give a good classification based on classes. Fig 2 is the plot between Z1 and F2 and we can see a more clear distinction between the five classes. When we test the model for X = Z1 on the test data, we get a classification accuracy of 86.2% and an error rate of 13.8%.

III. CASE 3

In this case, we build a Bayesian classifier where X = F2. We calculate the mean and the variance of the training set(We have taken the first 100 observations) and build our model. We get a classification accuracy of 34.42% and an error rate of 65.57% on the test data.

IV. CASE 4

In this case, we build a Bayesian classifier where X = [Z1 F2]. We calculate the mean and the variance of the training set(We have taken the first 100 observations) and fit a model on the training data. When we test the model for X = [Z1, F2] on the gtest data, we get a classification accuracy of 97.26% and an error rate of 2.73%.

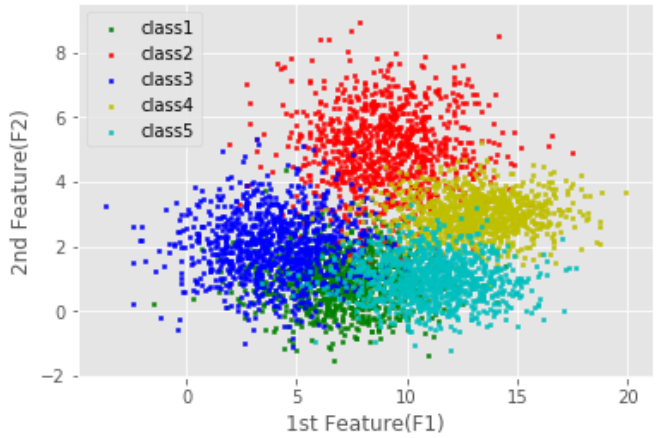

Fig. 1. F1 vs F2

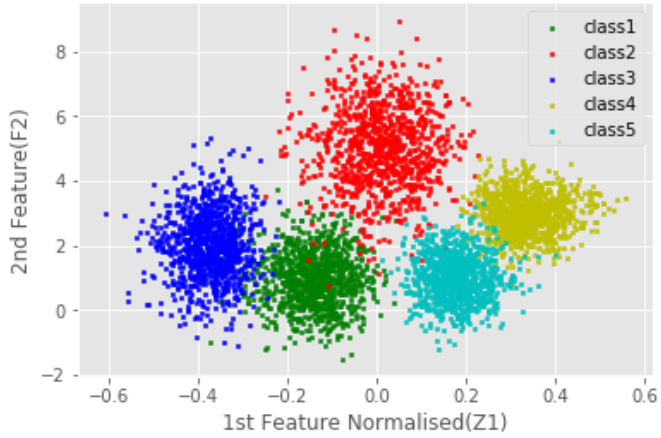

Fig. 2. Z1 vs F2

V. Conclusion

This experiment deals with two models : uni-variate and multi-variate classifiers and how significant data normalization is for modelling. We can see that in case 2, we get an accuracy improvement just by normalizing the data. We got an impressive result in case 4 too, because our classes are well separated and satisfies assumption of Bayesian classifier. This shows how important it is to bring down the variables to the same scale before processing it.

If we compare the four cases, we see that multivariate classification gives the best result. This is expected because we have more information to predict classes. The next best performance is by Z1(normalized F1).

Acknowledgment

This project has been completed as part of EAS595 : Introduction to Probability Theory for Data Science taught at State University of New York at Buffalo under the guidance of Professor Ehsan Esfahani and Professor Abani Patra.

References

[1] http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learnnote09-2up.pdf