Assignment-3:A Mathematical essay on naive Bayes classifier.

Ahmed Shmels Muhe
M.Tech In Data Science
Indian Institute of Technology Madras, IIT Madras
Chennai, India
ge22m009@smail.iitm.ac.in

***Abstract*—In this revised (v. 2) work, we will give a mathematical study of the Naive Bayes classifier. We will use the Naive Bayes classifier to determine whether the person earns more than $50,000$ per year. To provide the correct data for the model, we will also try exploratory data analysis and feature selection. Finally, we will examine if our model can properly generalise to previously unknown data. The data set contained several pieces of information that could have a conditional influence on a certain number of people's earnings range at the end of the year. A model has been developed based on this correlated conditional probability to predict early savings of a certain amount given some information about a person. We go through various work stages to solve this through different steps, from cleaning, processing, exploratory analysis, and modelling the data. We found that most features have different ranges of numbers, affecting our models' results, so we have applied some feature scaling techniques to avoid various number fields that can lead the model to underfit the data.**

***Index Terms*—Probability, Exploratory analysis, Bayes' theorem**

I. Introduction

In this assignment, I attempted to solve a prediction problem and model an approach in which we had one data set containing information for slightly more than 32 thousand people about their gender, nationality, how much they worked each week, and so on, as well as important information such as whether they could earn more than $50,000 early. We wanted to understand the significant factors that had influenced early income. We wanted to build a model that, given information about one person, could predict whether they would be able to earn more than $50,000.

We took the approach. The Naive Bayes classifiers are a set of algorithms. These algorithms enabled us to split a single data set into two parts [1], the response vector, which kept track of the success in earning more than $50,000 per year. The remaining information is contained in another section, namely the feature matrix.

In the second section, we attempt to explain the mathematics underlying Bayes' theorem and the modelling process of our data. In the third section, we'll look at how the data looks and the power of visualisation of input data features.

II. NAIVE BAYES CLASSIFIER

We will assume that conditional distribution given the class belongs to the Gaussian distribution.

$$X|y=+1 = N(\mu+, I); P(y=+1) = a \tag{1}$$

$$X|y=-1 = N(\mu-, I); P(y=-1) = 1-a \tag{2}$$

in the above, we are assuming the co-variance matrix to be **I** which means that our features are conditionally independent of each other. Applying Bayes rule, we get

$$P(Y=1|X=x) = \frac{fX|y(x|+1)P(y=+1)}{fX|y(x|+1)P(y=+1) + fX|y(x|1)P(y=1)}$$
$$= \frac{aexp\frac{-1}{2}||x-\mu+||_2^2}{aexp\frac{-1}{2}||x-\mu+||^2 + (1-a)exp\frac{-1}{2}||x-\mu-||^2} \tag{3}$$

On simplifying, we get

$$\frac{1}{1+exp-(w^TX+b)} \tag{4}$$

where,

$w = \mu+ - \mu-$

$$b = \frac{-1}{-2}||\mu+||^2 + \frac{1}{2}||\mu-||^2 + \log\frac{a}{1-a}$$

Probability over complete data can be found by multiplying individual probabilities. We get the following equation:

$$p(x|y=c,\theta) = \Pi_{j=1}^{d} p(x_j|y=c,\theta_{jc}) \tag{5}$$

Now, for training this model, we can estimate the Maximum Likelihood estimate. After maximising the log of likelihood, we get the solution to the above equation as following:

$$\mu_1^{ML} = \frac{1}{N_1}\sum_{n=1}^{N} \mathbf{1}_(yn) = c_1^x n \tag{6}$$

$$\mu_2^{ML} = \frac{1}{N_1}\sum_{n=1}^{N} \mathbf{1}_{yn} = c_2^x n \tag{7}$$

$$a = \sum_{n=1}^{N} \mathbf{1}_{yn} = c_1 = \frac{N1}{N1+N2} \tag{8}$$

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32560 entries, 0 to 32559
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32560 non-null  int64
 1   workclass       32560 non-null  object
 2   fnlwgt          32560 non-null  int64
 3   education       32560 non-null  object
 4   education-num   32560 non-null  int64
 5   marital-status  32560 non-null  object
 6   occupation      32560 non-null  object
 7   relationship    32560 non-null  object
 8   race            32560 non-null  object
 9   sex             32560 non-null  object
 10  capital-gain    32560 non-null  int64
 11  capital-loss    32560 non-null  int64
 12  hours-per-week  32560 non-null  int64
 13  native-country  32560 non-null  object
 14  income          32560 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

Fig. 1. Data Analysis and Feature Engineering

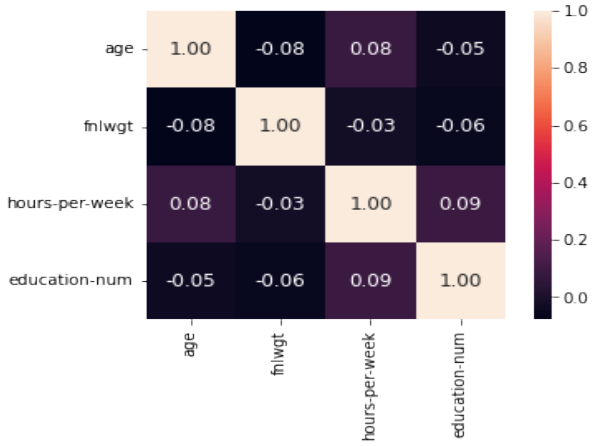

Fig. 2. Above heat map is obtained after conditioning features on class corresponding to income less than $50K$

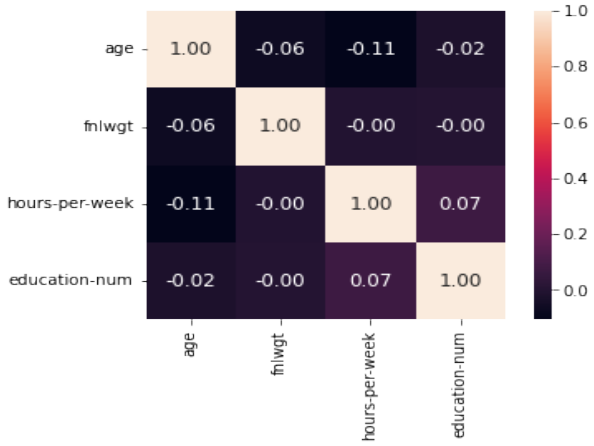

Fig. 3. Above heat map is obtained after conditioning features on class corresponding to income greater than $50K$..

III. THE PROBLEM

Given the data from the 1994 Census bureau database by Ronny Kohavi and Barry Becker, the key task is to predict whether a person is making over $50K$ annually. We will first analyse the data before attempting to fit the Naive Bayes model into it. The trained model will next be tested on an unknown dataset.

A. *Gather, Clean & Prepare data*

We worked with data that contained fifteen pieces of attributed information about 32,560 people [3]. Ronny Kohavi and Barry Becker collected this data set for the Census Bureau in 1994. Among the 32560 people in the data set, 14 variables were indicated [4]. Their age in the column "age," the type of job they have, their education level, and the years they have spent on education. The term "marital status" refers to whether a person is married, unmarried, or divorced. Some individuals may be divorced but widowed. Perhaps your spouse was not present. One variable is "race," which indicates whether a person is white, black, Asian, or from another country. The "Sex" variable represented the person's gender. The terms "capital gain" and "capital loss" denote the amount earned or lost by investing in or selling bonds, real estate, or a business. A person's native country was mentioned in the phrase "native country." At the very end of our data set, one column provided information on whether all of the probabilities mentioned in the previous variables had occurred if a person had earned more than $50,000. We aim to use this data set to develop and test a model that can predict early income to a specific amount given variable values.

We observed several columns in the dataset that has ? as values, hence we decided to replace these values with NaN values and then tried to analyse features with missing values using the following table:

Since there were very few samples with missing values, we decided to drop all those samples from the data. While deriving the Naive Bayes classifier, we assumed that all the features are conditionally independent given the class label. We tried to find the correlation among several numerical features in the data in the following figure: The above heat map is obtained after conditioning features on class corresponding to income greater than $50K$. We can see that correlation among features is very low in both figures; hence Naive Bayes could be an excellent fit for this data. Further, before fitting the model to the data, we tried to understand the categorical features in the data using bar plots. We hypothesised that this is happening because the cost of living is more in the United States than in any other country. Hence, if a person is living in the United States, then it is very likely that he/she is going to earn a large amount of money. Further, we analysed whether educated people earn more money compared to school/college dropouts.

IV. MODEL FITTING

We fitted the Naive Bayes Classifier using the sklearn library. We analysed the model after training it on unseen data. To ensure that test data is new to the model, we sampled the test data from the overall data (provided in adult.csv) before training and didn't use the test data for training.

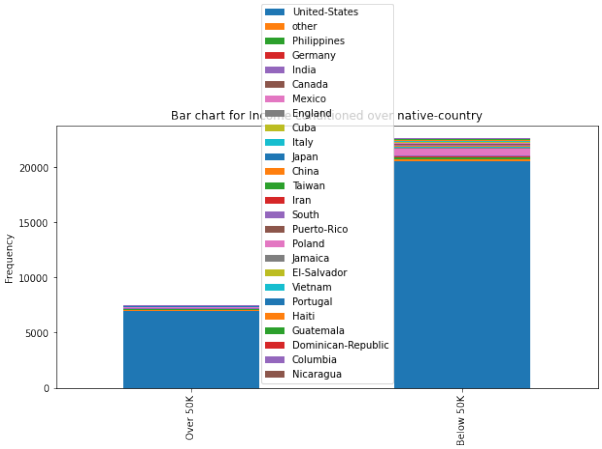

Fig. 4. In the above plot, we can visualise that income of people is more than $50K$ for more people living in the United States as compared to any other country.

	Predicted Positive	Predicted Negative
Actual Positive	4313	231
Actual Negative	1018	471

Fig. 5. The above table shows the confusion matrix of the trained model on the unseen data. In the following table, we tried to analyse the confidence of prediction belonging to each category.

We found that the model can predict the negative class with more confidence when compared to the positive category. After further inspecting the data, we found that model assumptions do not meet data with positive categories correctly. Hence, the model is not able to be utterly sure while predicting the positive category.

To further confirm our hypothesis, we plotted the probability distribution function in the following figures:

A. Model development and testing

With the training data set of 24420 entries containing both categorical and numerical variables, we used the Naive Bayes classifier algorithm to create a model that could predict a person's yearly income. We tested the accuracy, and the model

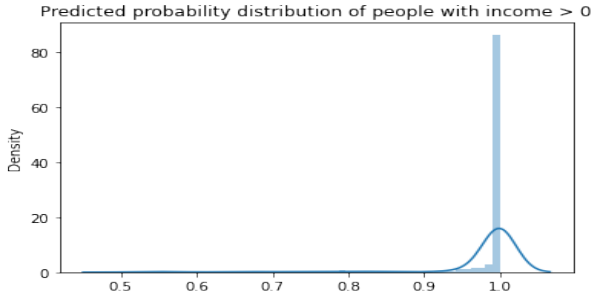

Fig. 6. Predicted probability distribution of people with income greater than zero

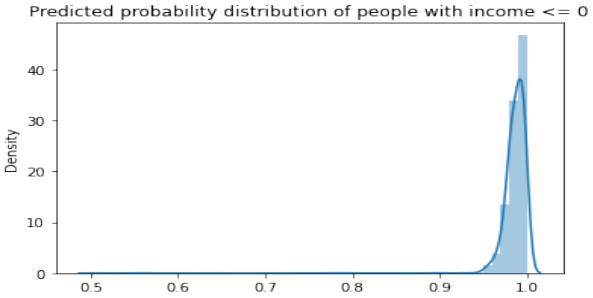

Fig. 7. Predicted probability distribution of people with income ≤ 0

accuracy score was 0.8016 for the training data, which is significant. Then we applied the same model to our test data set, which contained 8140 entries. In the model accuracy test, we received a score of 0.8045. The comparison of the scores for the training and test data sets (8016, 0.8045), which are significantly close and successful.

V. CONCLUSION

The Gaussian Nave Bayes Classifier model is used in this paper to predict whether a person earns more than $50,000 per year. The model produces very good results, as evidenced by the model accuracy, which was found to be 0.8083. The accuracy score for the training set is 0.8067, while the accuracy score for the test set is 0.8083. These two values are very similar. As a result, there is no evidence of over-fitting. I compared the model accuracy score of 0.8083 with the null accuracy score of 0.7582. As a result, we can conclude that our Gaussian Nave Bayes classifier model predicts class labels very well.

REFERENCES

[1] Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18(60), 1-8.
[2] Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007, July).Transferring naive bayes classifiers for text classification.
[3] Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007, November). Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In 2007 international conference on convergence information technology (ICCIT 2007) (pp. 1541-1546). IEEE.
[4] Kononenko, I. (1991, March). Semi-naive Bayesian classifier. In European working session on learning (pp. 206-219). Springer, Berlin, Heidelberg.
[5] Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, 44(1), 48-59.
[6] Zhang, H., & Su, J. (2004, September). Naive bayesian classifiers for ranking. In European conference on machine learning (pp. 501-512).
[7] Bouckaert, R. R. (2004, December). Naive bayes classifiers that perform well with continuous variables. In Australasian joint conference on artificial intelligence (pp. 1089-1094).