Assignment-3: A mathematical essay on Naive Bayes Classifier

Vasudev Gupta
Interdisciplinary Dual Degree - Data Science
Indian Institute of Technology Madras
Chennai, India
me18b182@smail.iitm.ac.in

***Abstract*—In this 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 $50K per year. In order to provide the correct data to 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.**

I. Introduction

Machine learning techniques are gaining popularity, helping many organisations solve challenging issues using data-driven approaches. Machine learning approaches try to estimate the function based on supplied training data by capturing the underlying distribution of the training data and then predicting unseen data using the learnt distribution. When creating the model, one wants the training data to be similar to previously unseen data and generalise well to future data.

This essay will concentrate on the Naive Bayes Classifier. The Naive Bayes Classifier is predicated on the assumption that the conditional distribution of the predicted variable concerning the input data is of the Gaussian distribution family and that all variables are conditionally independent given the class label. In the following section, we will go over a thorough mathematical analysis.

Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics) provide us with data from the 1994 Census Bureau database in this article. The main objective is to establish whether a person earns more than $50,000 per year. We are given a plethora of qualities to base our predictions on, but we must evaluate if any of them may significantly influence the model's performance.

This article will infer mathematical solutions to the Naive Bayes Classifier and use existing Python packages to fit the Bayes Classifier to the data supplied. We will also discuss the data insights we discovered and why we chose to alter or delete particular features.

II. Naive Bayes Classifier

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

$$X|y=+1 = N(\mu_+, I); P(y=+1) = a$$

$$X|y=-1 = N(\mu_-, I); P(y=-1) = 1-a$$

Note: Above, we are assuming co-variance matrix to be **I** which means that our features are conditionally independent of each other.

Applying Bayes rule, we get

$$P(Y=1|X=x) = \frac{f_{X|y}(x|+1)P(y=+1)}{f_{X|y}(x|+1)P(y=+1) + f_{X|y}(x|-1)P(y=-1)}$$

$$= \frac{a\exp\frac{-1}{2}||x-\mu_+||^2}{a\exp\frac{-1}{2}||x-\mu_+||^2 + (1-a)\exp\frac{-1}{2}||x-\mu_-||^2}$$

On simplifying, we get

$$= \frac{1}{1+\exp -(w^T X + b)}$$

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:

$$p(x|y=c,\theta) = \prod_{j=1}^{D} p(x_j|y=c,\theta_{jc})$$

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

$$\mu_1^{ML} = \frac{1}{N_1}\sum_{n=1}^{N}\mathbf{1}_{y_n=c_1}x_n$$

$$\mu_2^{ML} = \frac{1}{N_2}\sum_{n=1}^{N}\mathbf{1}_{y_n=c_2}x_n$$

$$a = \frac{1}{N_1}\sum_{n=1}^{N}\mathbf{1}_{y_n=c_1} = \frac{N_1}{N_1+N_2}$$

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 to it. The trained model will next be tested on an unknown dataset.

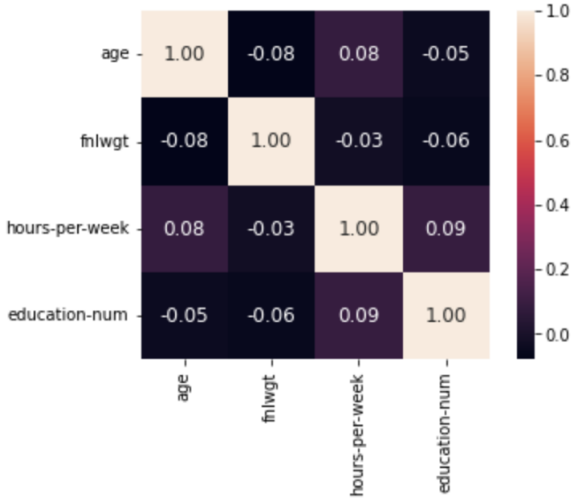
A. Data Analysis and Feature Engineering

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 following table:

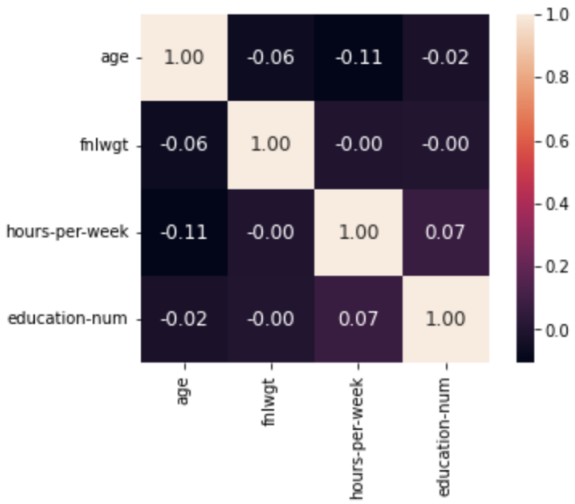
```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32560 non-null  int64
 1   workclass       30724 non-null  object
 2   fnlwgt          32560 non-null  int64
 3   education       32560 non-null  object
 4   education-num   32560 non-null  int64
 5   marital-status  32560 non-null  object
 6   occupation      30717 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  31977 non-null  object
 14  income          32560 non-null  object
```

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:



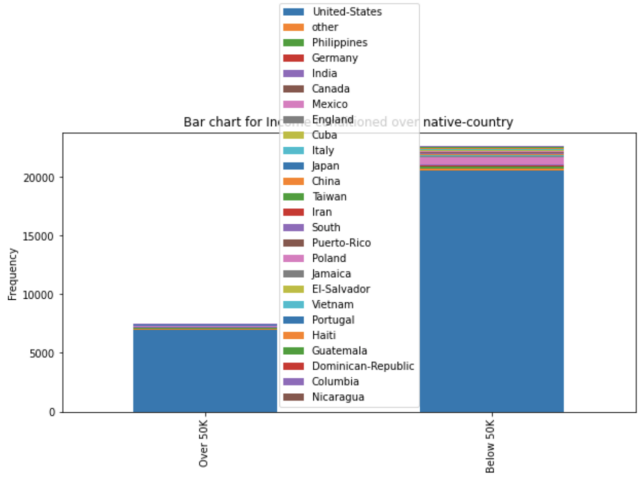
Above heat map is obtained after conditioning features on class corresponding to income less than $\$50K$.



Above heat map is obtained after conditionaling features on class corresponding to income greater than $\$50K$.

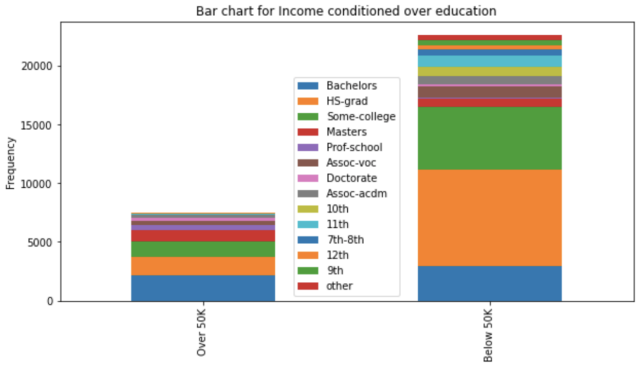
We can see that correlation among features is very low in both the figures; hence Naive Bayes could be an excellent fit for this data.

Further, before fitting the model to the data, we tried to understand the categorical features in the data using the bar plots.

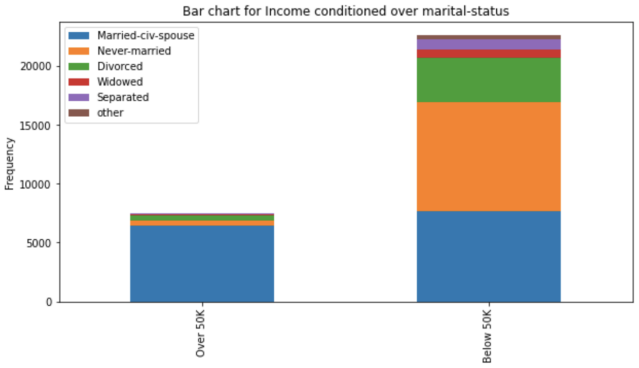


In the above plot, we can visualise that income of people is more than $\$50K$ for more people living in the United States as compared to any other country. We hypothesised that this is happening because the cost of living is more in the United States than in any other country. Hence, if a person is living in the United States, then it is very likely that he/she is going to earn a large amount of money.

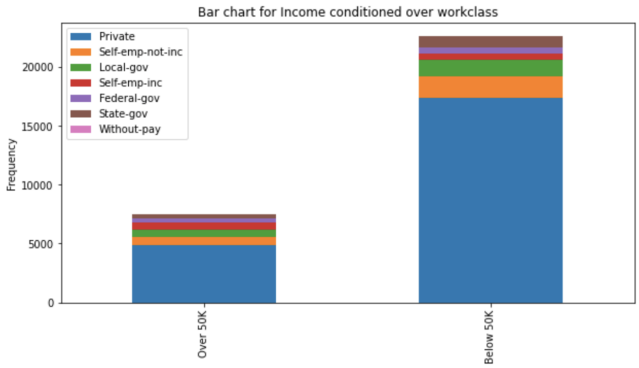
Further, we analysed whether educated people earn more money compared to school/college dropouts.



We observed that people with a higher education degree earn more compared to people who leave education early. This is expected as people who have completed higher education are more skilled and value businesses more.



We further checked if married people earn more compared to unmarried people/students. We found that married people make more money compared to other categories. We hypothesised that married people are usually of more age and hence they are more likely to earn more due to the experience gained over years of job.



Further, we tried to understand which occupations are more rewarding in terms of income. We found that people in the private sector earn more money than people from other sectors or working for govt.

B. Model Fitting

We fitted the Naive Bayes Classifier using the `sklearn` library. We analysed the model after training it on the 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.

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

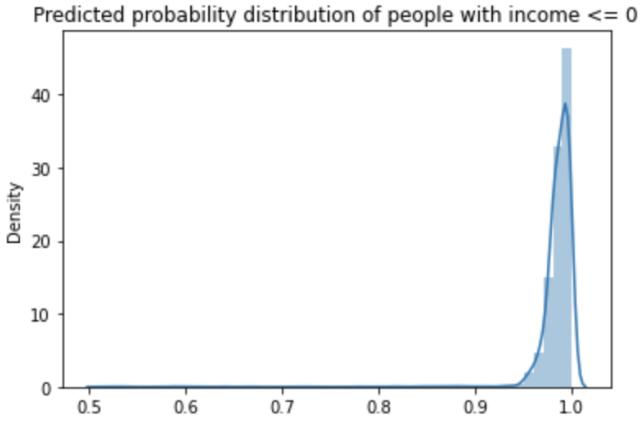
The above table shows the confusion matrix of the trained model on the unseen data. In the following table, we tried to analyse the confidence of prediction belonging to each category.

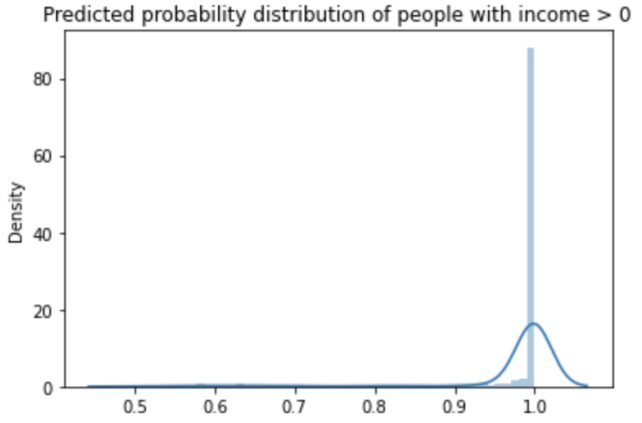
Percentage of people with predicted probability > 0.9

Income $< \$50K$	99%
Income $> \$50K$	92%

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 category 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:





The above figures clearly show that conditional distribution over positive class violates Gaussian assumption, and hence the model is less sure of predicting this category.

IV. Conclusions

In this essay, we derived the solution of Naive Bayes mathematically based on some assumptions. We also applied Naive Bayes techniques to real-world data and inspected whether the fitted model can predict well. We conclude that a simple model like Naive Bayes may be used on complicated real-world datasets if the underlying assumptions are met.

References

[1] Christopher M. Bishop, "Pattern Recognition and Machine Learning"
[2] Kevin P. Murphy, "Machine Learning, A Probabilistic Perspective"