A mathematical essay of Naive Bayes Classifier

Devansh Sanghvi
Department Of Biotechnology
Indian Institute of Technology, Madras
Chennai, India
be19b002@smail.iitm.ac.in

***Abstract*—This paper aims to develop an understanding of the mathematical properties of Naive Bayes Classifier and showcase its application in classifying the income of people from various walks of life.**

I. Introduction

Naive Bayes is a very simple, yet effective and commonly used machine learning classifier. It is a probabilistic classifier that uses Maximum A Posteriori decision rule to take a decision in the Bayesian setting. Using the features x_0 through x_n and classes c_0 through c_n, we calculate the probability of the features occurring in each class, and to return the most likely class. Hence, we want to calculate P(c_i | x_0,.....x_n). The required probability can be calculated using the provided data and applying the formula (**Bayes Rule**):

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

According to the equation, probability of the event A (class C_i in our case) given an event B (for certain values of features) is equal to probability of B given A, multiplied by P(A), divide by P(B). Since, P(B) acts as a normalization, we can ignore that term and instead state that P(c_i| x_0,.. x_n) is proportional to P(x_0,.... x_n | c_i). To simplify this computation, we assume that x_0 through x_n are **conditionally independent** given c_i and our final representation for the Naive Bayes Classifier is:

$$P(c_i|x_0, ...x_n) \propto P(x_0, ...x_n|c_i)P(c_i)$$

$$\propto P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$$

To demonstrate the effectiveness of Naive Bayes Classifiers, we analyze the "adult.csv" file that has been provided to us. The data-set contains the data of **32561 adults** across the world with personal details like: gender, level of education, their occupation, marital status amongst others. With the training data-set of the adults, we aim to **predict if the income of a particular adult is more or less than 50K**.
This paper will serve as a part of the requirements to be fulfilled for receiving credits in the course **EE4708: Data Analytics Laboratory**.

II. Naive Bayes Classifier

Naive Bayes Classifier uses the Bayes Rule to determine the probability of the occurrence of each class using the feature values given. Once we find the respective probabilities, we define regions in n-dimensions for each class. A point in a particular region will be classified as class c_i if the probability of class c_i is the most in that region. Next, we list down the assumptions, pros and cons of Naive Bayes Classifier:

A. Assumptions

- The fundamental Naive Bayes algorithm makes an assumption that each feature makes an independent contribution to the outcome.
- For numerical features, normal distribution is assumed.
- The fundamental Naive Bayes algorithm makes an assumption that each feature makes an equal contribution to the outcome.

B. Advantages of Naive Bayes

- It is easy and fast to predict the class of the test data set. It also performs well in multi class prediction.
- When the assumption of independence holds, a Naive Bayes classifier performs better when compared to other models like logistic regression.
- Less training data is required than other classification algorithms.
- It performs well in case of categorical input variables, when compared to numerical variable(s).

C. Disadvantages of Naive Bayes

- If the categorical variable has a category (in test data set), which was not observed in the training data set, then the model assigns a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency".
- Naive Bayes is also known as a bad estimator, so the probability outputs from Naive Bayes' function are not to be taken too seriously.
- Another shortcoming of Naive Bayes is the assumption that each feature is independent of each other. It is almost impossible to get features which are completely independent in the real world.

D. Evaluation Parameters in Naive Bayes Classifiers

There are multiple Bayes Classifiers that are used, and an evaluation metric is required to compare the classifiers. Some metrics that can be used to compare are:

- Confusion matrix: Four values are present in the confusion matrix. a_ij represent the i,jth value in the confusion matrix. a_11 represents the true positive (TP) and a_22 represents the true negatives (TN) value. These are the respective fractions of correct predictions when compared to the actual values i.e the prediction was true when the actual classification was true, in the case of true positives. a_12 represents false positives (FP), i.e the number of times the actual value was false, but the value predicted was true. a_21 represents true negatives (TN), i.e the number of times the actual value was true, but the value predicted was false.
- Classification matrix: this consists of three important values:
  - $Precision = TP/TP + FP$, it is the percent of predictions that were correct
  - $Recall = TP/TP+FN$, it is the percent of positive values that were determined correctly
  - $F1Score = 2 * (Recall * Precision)/(Recall + Precision)$, it is the weighted harmonic mean of precision and recall. It is used to compare two classifier models and not the global accuracy of a model. Values closer to 1 are better.
- Accuracy score: Used to compute the subset accuracy. It is equal to the Jaccard score for binary classification.

III. Preprocessing and Exploratory Analysis on the Data

In this section, we will first preprocess the data to extract the important features in the data provided, then we will visualize the data to understand the effects of certain features and modify the data-set to optimize the results from the Naive Bayes model.

A. Preprocessing the data

The initial data-set has 32561 samples and 15 features, namely: Age, workclass, fnlwgt, education, education.num, marital status, occupation, relationship, sex, race, capital.gain, capital.loss, hours.per.week, native.country and income. Preprocessing the data-set includes (Note that the changes are made in both the training and testing data-sets):

1) Missing value assessment: Some features have a percentage of missing values and the first step is to decide what to do for these features. While none of the features have missing values, the features workclass, occupation and native.country have "?" values. We replace the workclass and native.country "?" values with the modes of these features respectively. We replace the occupation "?" values with the 'Adm-clerical' occupation.
2) Additional Variables: We do not require the 'education' feature since we know the respective education.num. Similarly 'fnlwgt' feature is unneccesary.

B. Visualization

After preprocessing the data, we visualize the data, to better understand the correlation between different features:

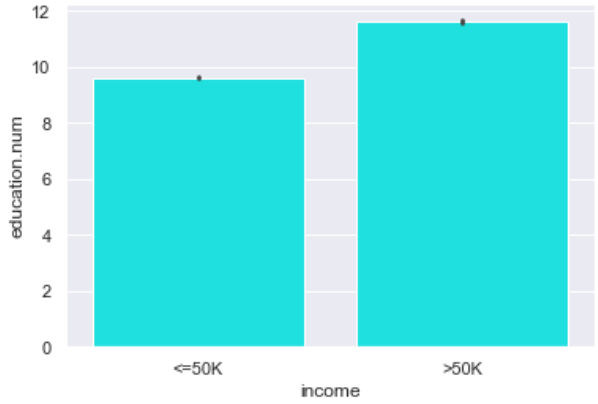

Fig. 1. Bar Plot of number of years ofcompleted education v/s the income

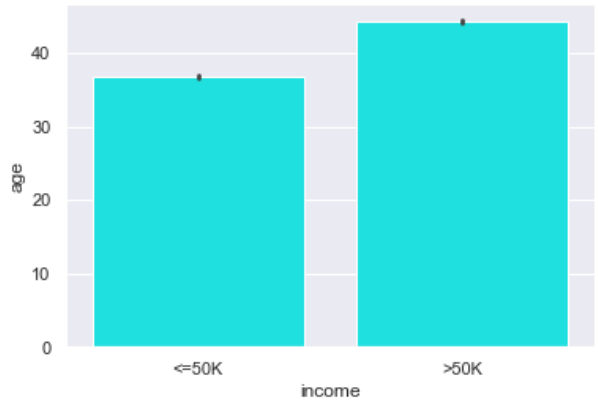

Fig. 2. Bar Plot of age v/s income

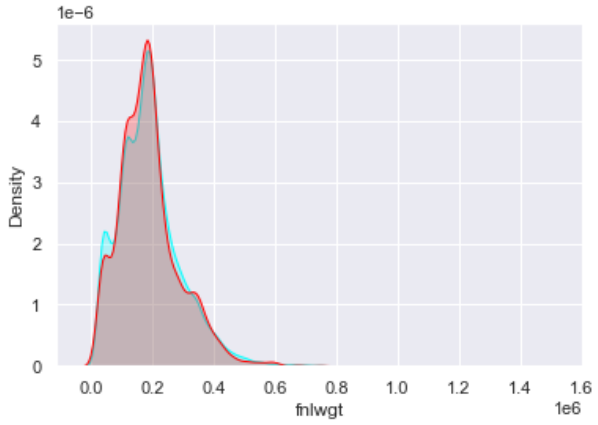

Fig. 3. Density plot of fnlwgt v/s income.Red represents >50 K income

1) Exploration of Education: The bar plot shows that the income is positively correlated with the education level

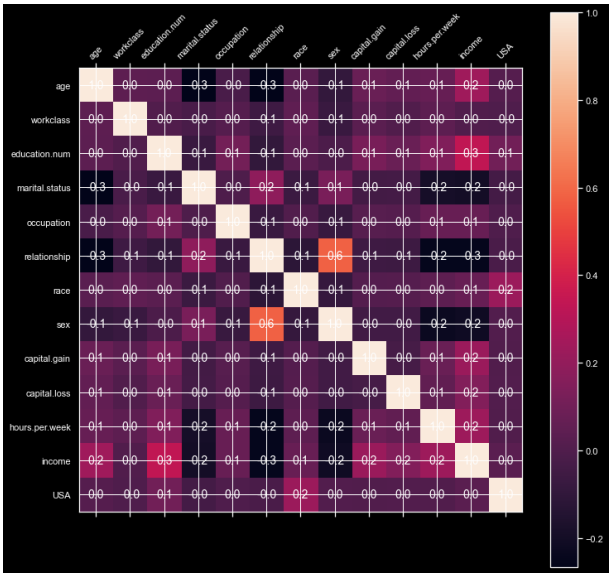

Fig. 4. Correlation matrix between all features after converting all features to numerical attributes

of an adult. Mean number of years in getting educated for people with $>$ 50K income are close to 11.75.

2) Exploration of Age: Elder men have a better income, on average. The mean ages for the two classified incomes are 36 and 43 for $<= 50K$ and $> 50K$ respectively.
3) Exploration of Gender: It can be clearly seen that females were preferred in the rescue procedure.
4) Exploration of fnlwgt: The fnlwgt values do not have any impact on the income levels of adults.

IV. The Naive Bayes Model

After preprocessing the data, we move to the heart of the problem: the prediction of whether an unseen adult (not in the training data-set) survived the crash. Here are the steps for the same:

- Feature Selection: Before applying a Bayes Classifier model to our data, we will eliminate the features that are least important using recursive feature elimination. Income levels of adults are not correlated with workclass and the native country and hence can be ignored in the final Classifier model. The correlation matrix for the same can be seen in image Fig.4.
- Naive Bayes model and results: After selecting the features, we run three different Naive Bayes models using the simple test/ train model and evaluate our results using three parameters (the values are weighted averages), which turn out to be the following:
  1) Gaussian Bayes:
     - Precision: 0.79
     - Recall: 0.8
     - F1: 0.77
     - Accuracy: 0.73
  2) Bernoulli Bayes:
     - Precision: 0.79
     - Recall: 0.73
     - F1: 0.75
     - Accuracy: 0.80
  3) Multinomial Bayes:
     - Precision: 0.81
     - Recall: 0.77
     - F1: 0.68
     - Accuracy: 0.77

V. Conclusion

The Naive Bayes model was successfully demonstrated, both visually and numerically. Bernoulli Bayes model proved to have the highest accuracy, while Gaussian Bayes had the smallest accuracy. Gaussian Bayes, however, had the highest F1 score. All three models had similar precision and recall. Bernoulli Bayes resulted in a lower recall amongst the three.

References

[1] "sklearn.metrics.accuracy_score", scikit-learn.org, 2022 [Online]. Available: **Link**.
[2] "Understanding the classification report in sklearn", personal website, 2022 [Online]. Available: **Link**.