DAL 2023 Assignment 3: A Mathematical Essay on Naive Bayes

Arun Palaniappan
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai, India
me20b036@smail.iit.ac.in

***Abstract*—This essay provides an overview of the mathematical formulations and applications of Naive Bayes classifier. First, the fundamentals of Naive Bayes are explored, as well as the specifics necessary to comprehend the research. The problem statement and data-set are introduced in a detailed way. The data-set under consideration is one of the census data extracted from the 1994 Census bureau database, to determine whether a person makes over $50K a year. To accomplish this efficiently, the data has been thoroughly evaluated, and both visual and quantitative insights have been provided in this essay. Finally, the results of the Naive Bayes model application are reported to better understand which features are essential when associating work hours per week, age, gender and several other factors with the income range. The "Data" section has been improved by including a feature engineering method called Principal Component Analysis.**

***Index Terms*—Naive Bayes, precision, recall, income, education, gender, visualization**

I. Introduction

Naive Bayes is an effective statistical technique employed to model and predict the association between various factors or variables. It falls under the category of supervised learning algorithms, primarily used for classification tasks, specifically for probabilistic classification. It is a machine learning algorithm that computes conditional probabilities based on Bayes' theorem.

The core ideation of a Naive Bayes classifier is based on probabilistic reasoning. The classifier uses Bayes' theorem to calculate the likelihood of a class given a set of observed features. The calculation of these probabilities assumes that the features are conditionally independent. We do a train-validation split of the data-set to test the robustness of the classifier. The model calculates the prior probability of each class as well as the likelihood of observing specific feature values using training data. It categorizes new data points by choosing the class with the highest posterior probability.

There are several interrelated aspects that have a role in determining an individual's annual income. The data-set provided includes details about the age, capital gain, capital loss, marital status, occupation, etc., of Americans who are in the working class. In this study, We will gain an understanding of the key factors that contribute to an individual earning more than 50,000 US Dollars annually. This allows us to better comprehend the working class and how to best represent it by using the Naive Bayes model.

In this paper, a comprehensive understanding of Naive Bayes fundamentals is provided. The paper also undertakes a detailed exploration of the targeted problem and elucidates how the resulting insights are supported by a combination of visual representations and quantitative evaluation.

II. Naive Bayes

In this section, we will describe the mathematical details of Naive Bayes. Naive Bayes is a supervised learning algorithm used for modeling the relationship between two types of variables, the target(dependent) and features(independent). Like logistic regression, target of Naive Bayes is a categorical variable. The terms and mathematics used in the Naive Bayes technique are summarized below:

- *Features/Independent variables:* The features or independent variables are used as predictor features to predict the target variable. They can be classified into three two categories i.e., continuous and categorical features.
- *Target/Dependent variables:* Dependent variables, known as target variables, exhibit either continuous or discrete traits. When implementing Naive Bayes for a specific problem, the target variable takes on a discrete distribution.

A. Assumptions in Naive Bayes

Naive Bayes is a powerful statistical technique, but its validity and reliability rests upon certain assumptions that must be met for accurate and meaningful results. Here, we will delve into the key assumptions underlying Naive Bayes:

- *Independence:* The key assumption is that the features are presumed to be conditionally independent of one another. This is a strong assumption since it suggests that all the information needed to make a classification decision is contained within each feature and that there are no interactions or relationships between features.
- *No or Little Multicollinearity:* Multicollinearity refers to a high correlation between independent variables in the regression model. This can make it challenging to isolate the individual effects of each variable, leading to unstable probability estimates.
- *Equal Weightage of Features:* Naive Bayes gives all features the same weightage and relevance. Each feature is regarded as equally constructive when the decision

is being made. This assumption may not always be true because in practice, certain features may be more informative than others.

- *Normal Distribution of Features (Gaussian-NB):* Gaussian Naive Bayes (GaussianNB) assumes that the continuous features within each class have a normal distribution. This indicates that it makes the assumption that the data is symmetrically distributed around the mean and has a bell-shaped curve.

B. Types of Naive Bayes

- *Gaussian Naive Bayes:* This is especially suitable for data-sets with real-valued and continuous features. In order to estimate the mean and variance of each feature for each class, it operates under the presumption that these features have a Gaussian (normal) distribution within each class.
- *Multinomial Naive Bayes:* This is used to tackle text classification tasks where the features are counts or discrete data, usually in the form of word frequencies or phrase counts. This makes it a logical choice for natural language processing applications like document categorization or sentiment analysis because it presumes that features are produced from a multinomial distribution.
- *Bernoulli Naive Bayes:* This classifier assumes that features are binary, meaning that they can only have one of two possible values$-\{0,1\}$, which are frequently used to indicate the presence or absence of particular characteristics or events. Each feature is modeled as a Bernoulli distribution.

C. Algorithm

Naive Bayes Classifier uses Bayes' theorem to estimate probabilities for each class, such as the likelihood that a given record of data belongs to a specific class. We model it as a normal distribution as assumed by Gaussian Naive Bayes. The class that has the greatest likelihood is considered the predicted class.

In our problem, let us assume there are 'n' predictor variables - $\boldsymbol{X} = [x_1, x_2, ..., x_n]$ and the target variable is $\{y\}$. Bayes theorem states that:

$$p(y|X) = \frac{p(X|y) * p(y)}{p(X)} \tag{1}$$

Where,

- $p(y|X)$ is the posterior probability which we are trying to predict for each class in the target variable.
- $p(X|y)$ is the likelihood probability which denotes what is probability of obtaining these features for a particular class.
- $p(y)$ is the prior probability of class 'y' occurring.
- $p(X)$ is the prior probability of features set 'X' occurring.

The likelihood probability can be split using the chain rule as:

$$\begin{aligned} p(X|y) &= p(x_1, x_2, ..., x_n|y) \\ &= p(x_1|x_2, ..., x_n, y) * p(x_2|x_3, ..., x_n, y) * ... \\ &\quad ... * p(x_n|y) \end{aligned}$$

The conditional probabilities, however, are independent of one another due to the Naive's conditional independence assumption. Hence,

$$p(X|y) = p(x_1|y) * p(x_2|y) * ... * p(x_n|y)$$

Hence, by conditional independence, we have:

$$p(y|X) = \frac{p(x_1|y) * p(x_2|y) * ... * p(x_n|y) * p(y)}{p(x_1) * p(x_2) * ... * p(x_n)} \tag{2}$$

The denominator can be considered as a constant of proportionality. Hence,

$$p(y|X) \propto p(y) \prod_{i=1}^{n} p(x_i|y) \tag{3}$$

Using this equation we know find the *Maximum a Posteriori* (MAP) estimate, which is the class 'y' for which the posterior probability is maximum. It can be mathematically stated as,

$$y = \arg\max_{y} p(y) \prod_{i=1}^{n} p(x_i|y) \tag{4}$$

Hence we have established the algorithm of a Naive Bayes classifier.

D. Evaluation Metrics

We can create a table of all possible cases to evaluate the model,

TABLE I
All possible outcomes

All possible cases		True class	
		(1)	(0)
Predicted class	(1)	True Pos. (TP)	False Pos. (FP)
	(0)	False Neg. (FN)	True Neg. (TN)

We now define three quantities which can be used to evaluate the model.

- *Precision:* It is defined as the ratio of True positives and the total positive predictions,

$$P = \text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- *Recall:* It is defined as the ratio of True positives and the total true positives,

$$R = \text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

- *F_1-score:* It is defined as the harmonic mean of precision and recall,

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

- *ROC-AUC score:* The ROC-AUC (Receiver Operating Characteristic Area Under the Curve) is a graphical representation of a binary classification model's performance. It plots the True Positive Rate against the False Positive Rate across various probability thresholds for classifying positive and negative instances. Higher area under the curve indicates robust performance of the model.

III. The Data

A census data-set with information on income range across various socioeconomic groups, age and gender are provided to us. The ultimate aim of the study is to identify key relationships between a person's income range and several factors describing the person.

A. Data description

In this section, we describe the structure of the data-set. The data-set has 32561 entries. It has been split into train(80%)-validation(20%) data-set. The columns from the data-set are:

- *age:* Age of the person.
- *workclass:* The sector in which the person is working(government, private, etc.).
- *fnlwgt:* The weight of the particular entry as per the data-set.
- *education:* The highest level of education completed by the person.
- *education-num:* The number of years spent by the person in educationg themselves.
- *marital-status:* The current marriage status of the person(married, divorced, etc.).
- *occupation:* The current job of the person(Technical, craft, armed forces, etc.).
- *relationship:* The current relationship status of the person(wife, husband, unmarried, etc.).
- *race:* The ethnic group that the person belongs to(white, black, asian, etc.).
- *sex:* States whether the person is a male or a female.
- *capital-gain:* The person's profit gained by investing.
- *capital-gain:* The person's loss by investing.
- *hours-per-week:* Number of working hours in a week for the person.
- *native-country:* The person's native country.
- *salary:* A binary variable which denotes whether the person makes more than $50,000 per year or not.

B. Data cleaning

While cleaning the data, we had to handle categorical classes. Categorical classes were encoded by integers using *LabelEncoder* from scikit-learn. For instance *sex* column was encoded using ['male':1, 'female':0].

C. Principal Component Analysis

As we have a lot of variables in our case, it makes sense to use some kind of dimensionality reduction technique. Principal Component Analysis (PCA) is a powerful mathematical technique that is widely used in multivariate data analysis for dimensionality reduction and feature extraction. The main objective of principal component analysis (PCA) is to convert a set of correlated variables into a new set of uncorrelated variables, or principal components, while preserving as much of the variability present in the original data.

Mathematically, the data matrix X can be decomposed by diagonalization as,

$$X = VDV^{-1} \tag{8}$$

$$X = \begin{bmatrix} \vert & \vert & \vert \\ v_1 & v_2 & v_3 \\ \vert & \vert & \vert \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} \vert & \vert & \vert \\ u_1 & u_2 & u_3 \\ \vert & \vert & \vert \end{bmatrix}$$

Where, the columns of V matrix denotes the eigenvectors of $X^T X$ which is the covariance matrix, and λ_i denotes the eigenvalues of $X^T X$. It is also proven that the eigenvectors are the principal components and the corresponding eigenvalues are the variance explained in that component. Hence we take the *Top-k* components for dimensionality reduction from p original features to k principal components.

D. Data Visualization

1) Correlation plot: Fig. 1 shows the heat map of the correlation values between variables in the data-set.

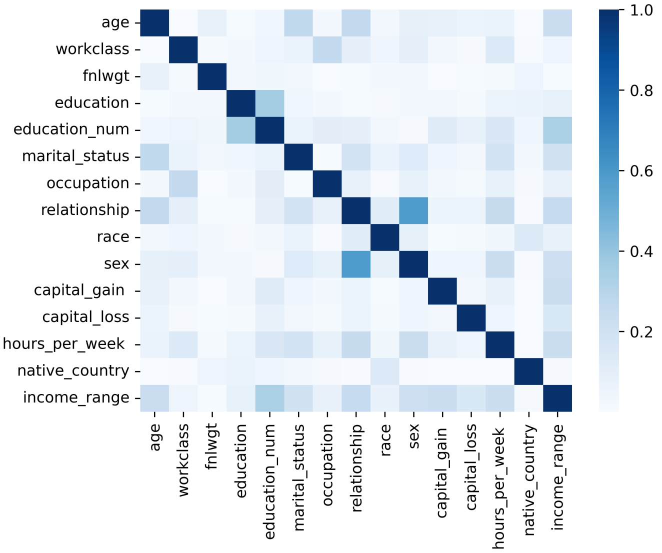

Fig. 1. Correlation heat-map within the variables

From the correlation plot we can conclude that:

- One obvious correlation is the *education* and the *education-num*. Higher the number years of education, higher is the degree completed.
- The *sex* of the person is correlated with the *relationship* of the person.
- The target variable(*income range*) is correlated with the *education-num* because, generally, people with higher degree levels are employed with a higher annual income.

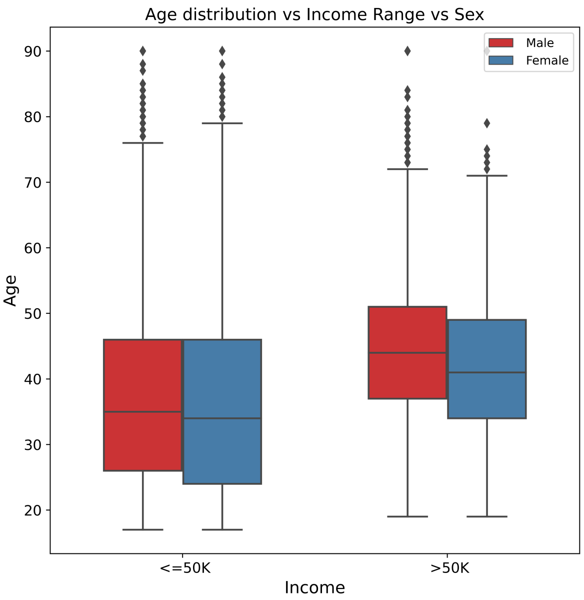

Fig. 2. Age distribution vs Income range vs Sex

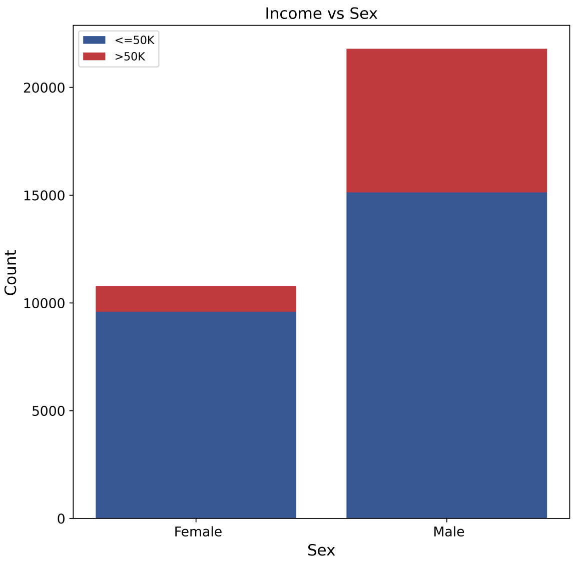

Fig. 3. Income vs Sex

2) Age distribution, Income range and Sex: From Fig. 2, it is evident that, when the income is less than $50,000, the age distribution among males and females are almost the same qualitatively, but, when the income is more than $50,000, the mean of the age distribution for both males and females are higher than the previous plots because, in a work environment a person would move to higher positions with higher annual income as they gain experience.

3) Income and Sex: We can observe from Fig. 3 that in general the male population in the working class is high, and around 30% of the males have annual income greater than $50,000 whereas it is 10% in females.

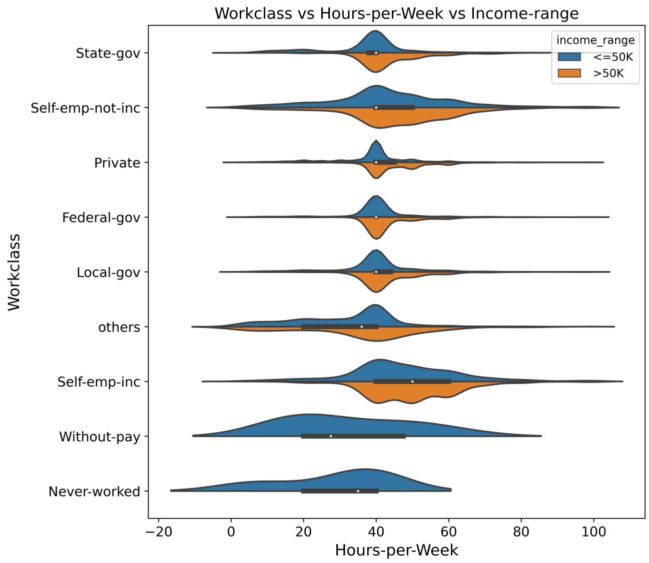

Fig. 4. Workclass vs Hours-per-week vs Income-range

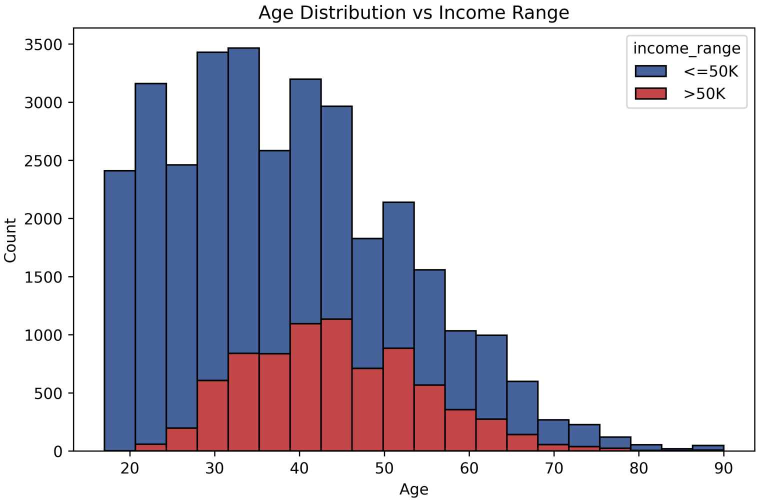

Fig. 5. Age Distribution vs Income Range

4) Workclass, Hours-per-week, Income-range and Age: The Fig. 4 shows that the mean work hours for people of all workclasses were around 40 hours irrespective of income range(as generally private sectors have the highest employment rate and a 40 hour work week). We can also notice that self employed people have slightly higher working hours which is compensated by the higher number of people in the income greater than $50,000 bracket. Fig. 6 shows that middle-aged people(35-50) have a higher average weekly work hours

5) Age Distribution and Income Range: The Fig. 5 represents a very important assumption of Gaussian Naive Bayes classifier, that the features can be modeled to be normally distributed.

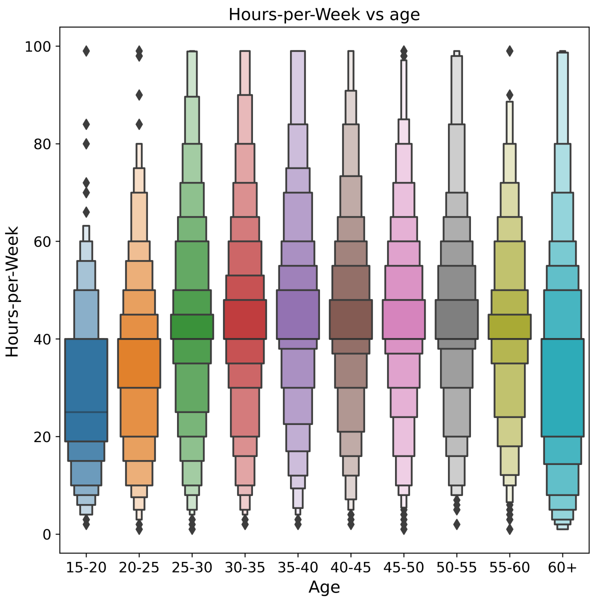

Fig. 6. Hours-per-week vs Age

IV. The Problem

A. Outline

We have done the analysing, cleaning and visualizing the data. A train-set and a validation-set are created from the data-set. The validation-set is used to make predictions after the Gaussian Naive Bayes model has been trained using the train-set. The validation-set predictions are used to calculate a number of metrics to assess model performance.

B. Naive Bayes modelling

We accomplish this by using *scikit-learn* package in python. We now instantiate a Naive Bayes model (GaussianNB), for prediciting the *income-range*. We use F1-score and ROC-AUC score as our evaluation criteria on the validation data-set. After testing on the Validation Split, the Naive Bayes model for *income-range* obtains an F1-score of 0.88 and ROC-AUC score of 0.84. While improving the paper, we performed dimensionality reduction using Principal Component Analysis to just 2 features from the original 14 features and still we maintained an F1-score of 0.88. We have also significantly reduced the computation time and also increased the robustness of the model.

C. Confusion matrix

TABLE II
Confusion matrix

All possible cases		True class	
		(0)	(1)
Predicted class	(0)	4811	131
	(1)	1219	352

This confusion matrix generated gives count to all the possible outcomes in the validation set(20% of the data).

D. ROC-AUC score

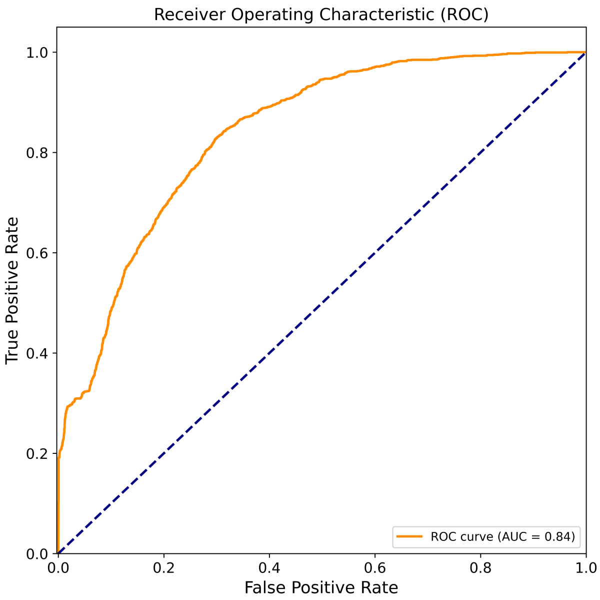

Fig. 7. ROC-AUC

We can notice that the area under the curve (Fig. 7) is close to 1, and the curve tends towards the LHS top corner, which implies it has a high true positive rate and low false positive rate, hence more clearer differentiation between classes.

V. Conclusion

The purpose of this report was to determine whether a person's yearly income was below or above $50,000. The conclusions that can be drawn from all the above analysis are:

- Over 75% of people earn less than $50,000 annually. As a result, there is some class imbalance.
- Due to class imbalance, the model's performance in the majority class was better than for the minority class.
- While the age distribution of those making more than $50,000 is roughly Gaussian and mean is around 40 years, the distribution of those making less than $50,000 tends towards the left and mean is around 25 years.
- The private working class is of the highest density, and married people tend to earn more to support the family.
- The plots shown, support all the above quantitative evidence in a visual way.
- The model has been run on the validation-set and has been evaluated.

Future research could use *One-Hot Encoding* for categorical variables, that is creating dummy columns for denoting the presence of a class to avoid any unnecessary bias. We have observed that, there is class imbalance in the income-range column. This can create bias inherent bias in the model, hence we can adopt class-balancing techniques to overcome this bias. We could assign weights to some variables if they seem more important that other features as currently we assign equal weight to all features.

References

[1] Census bureau (1994) database by Ronny Kohavi and Barry Becker. (Data Mining and Visualization, Silicon Graphics)

[2] ”Get Started With Naive Bayes Algorithm: Theory & Implementation” *https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/* (accessed 24 Sept. 2023)

[3] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning : with Applications in R*. New York :Springer, 2013.