Mathematical Essay on Naive Bayes Classifier

Chaithanya Krishna Moorthy
Dept. of Physics
Indian Institute of Technology, Madras
Chennai, India
ph17b011@smail.iitm.ac.in

***Abstract*—In this assignment, I have used the Naive-Bayes Classifier on and extract of Census data in the United States (1994). The task is to use professional and personal features of a person and determine if he/she makes over $ 50k in a year. The Naive-Bayes model gives about 80% accuracy on the data.**

***Index Terms*—Naive Bayes, Classification**

I. Introduction

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

A person's income levels can be hypothesized to depend strongly on his/her gender, race, education levels, occupation and native. Building a classifier to determine if a person makes more than $50000 with high accuracy is one way to verify the hypothesis. **Naive-Bayes Classifier** is classification technique that is based on the Bayes-Theorem. It has the prefix 'Naive' because it assumes that there is not underlying correlation between the features, which is not true in most cases. Nevertheless, the simple classifier performs well on many classification tasks. Since it is simpler, it is also fast and easy to train. The classifier calculates the posterior probability of a sample to belong to all classes and assigns the the classes with highest predicted probability.

Using Naive-Bayes classifier, we can build a model to predict if an individual makes over $50k a year or not, using their education levels, occupation and native country.

In this study, I will use Naive-Bayes Classifier to build a binary classification model to explore if the socio-economic features influence the annual income of a person. Section II will be on the principles of Naive-Bayes model, section III will include Data Cleaning, Exploratory Data Analysis and applying Naive-Bayes Classifier on the cleaned data. Section IV will be conclusions and the study will end with references.

II. Principles of the Naive-Bayes Classifier

Naive-Bayes classifier operates on the assumption that all the input variables are independent of each other. While this does not always hold true, the model performs well regardless. It calculates the posterior probability of a sample to belong to each of the classes and outputs the class withe maximum probability as the predicted label.

A. Mathematical model of Naive-Bayes Classifier

The Bayes-Theorem is

$$P(C_k|x) = \frac{P(C_k) * P(x|C_k)}{P(x)} \tag{1}$$

- The index k is for each class
- The LHS $P(C_k|x)$ gives the posterior probability for a given sample x to belong to the class C_k.
- $P(C_k)$ is the prior probability of class C_k
- $P(x|C_k)$ is the likelihood of x, given class C_k
- P(x) is the prior probability of x

The class k that has maximum $P(C_k|x)$ for a given sample x, is the predicted label for x. When x is multi-dimensional, the same Bayes equation, assuming independence, takes the form

$$P(C_k|x) = \frac{P(C_k) * \Pi_{i=1}^{d} P(x_i|C_k)}{P(x)} \tag{2}$$

Here, independence among the features x_i is assumed.

To calculate the likelihood $P(x|C_k)$, there are multiple versions of Naive-Bayes classifiers, which assume

- Gaussian
- Multinomial
- Bernoulli
- Categorical

depending on the data that has to be classified. In this assignment, I have used the Categorical version, by converting some of the input variables that are continuous to categorical bins. It assumes a categorical distribution for each feature of x, conditioned on the classes.

The probability of category t in feature i, given class C_k is estimated as

$$P(x_i = t|C_k) = \frac{N_{tiC_k} + \alpha}{N_{Ck} + \alpha n_i} \tag{3}$$

where

- N_{tiC_k}No. of times category t appears in sample x_i with class C_k
- N_{Ck} is the number of samples in class C_k
- α is the smoothing parameter. In this assignment, it is set to a default 1.0
- n_i is the number of categories in feature i

B. Metrics for Evaluation

Accuracy is defined as:

Accuracy = Number of correct predictions from the model/Total number of predictions

Accuracy is a good measure to see if our model is able to make correct predictions, implying that the input features influence the output features.

Description	Key
Age	Continuous
work class	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay,Never-worked
final weight	Continuous
Level of education	Bachelors, Some-college, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, Preschool, 1st-4th, 5th-6th, 7th-8th, 10th, 9th, 11th, 12th, Masters, Doctorate,
No. of years of education	Continuous
marital status	Married-civ-spouse, Divorced, Separated, Never-married, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, , Armed-Forces Machine-op-inspct, Adm-clerical, Farming-fishing Transport-moving, Priv-house-serv, Protective-serv
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Gender	Female, Male
Capital gain	continuous
Capital loss	continuous
Working hours / week	continuous
Native Country	US, Cambodia, England ..

TABLE I
FEATURES IN THE DATA

III. USING NAIVE-BAYES CLASSIFIER FOR THE DATASET

The data has 32561 rows and contains the columns as is table 1. 'final weight' - estimate refers to population totals by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights.
The dataset was divided in 80-20 ratio as training-test set to evaluate the performance of the model on new data.

1) Data cleaning and Exploratory Data Analysis: Histogram plots of all the features that are continuous is as in figure 1. Pie plots of all the features that are categorical is as in figure 2. Pie Chart of the output classes: '$\leq$50' and ' 50' is as in figure 3

- The columns 'Capital_Gain' and 'Capital_Loss' were removed since it was not possible to bin them appropriately
- The continuous features in figure 2 were converted to categorical by putting them in bins of quartiles.
- The categorical features (original and transformed as in the previous step) were label encoded to feed into the Categorical Naive-Bayes classifier

2) Correlations: The target features for the study is 'Income'. Rest of the columns are taken as features.
The correlation matrix heatmap for every pair of columns is shown in figure 4. As seen in the figure, the correlations between the figures is very low. Hence, we can expect the Naive-Bayes classifier to show good performance on this data set.

3) Applying Naive-Bayes Classifier: A Naive-Bayes Classifier model was built using sklearn's

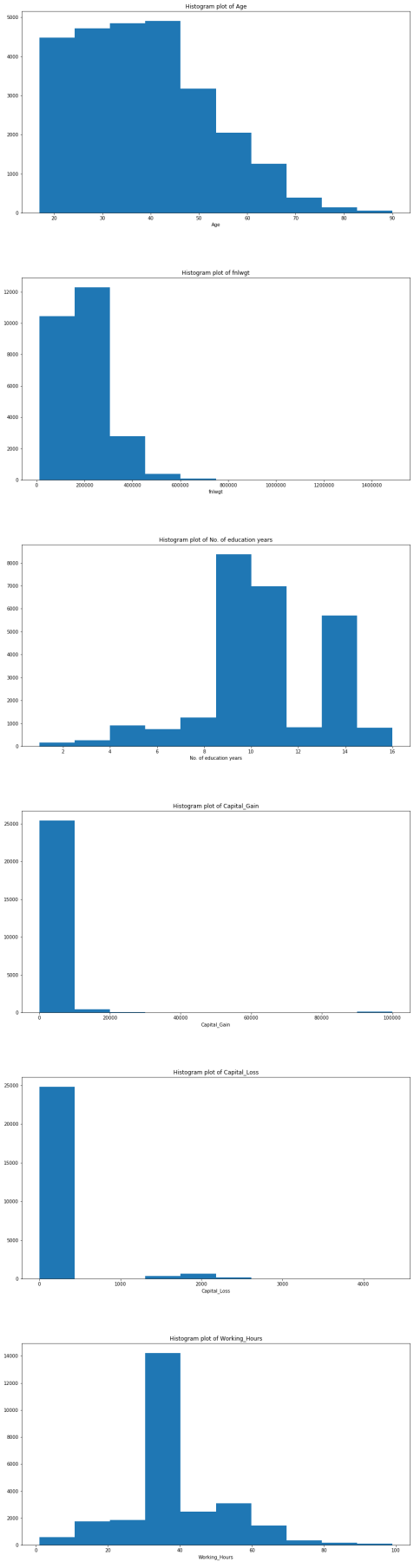

Fig. 1. Histograms of input features that are continuous

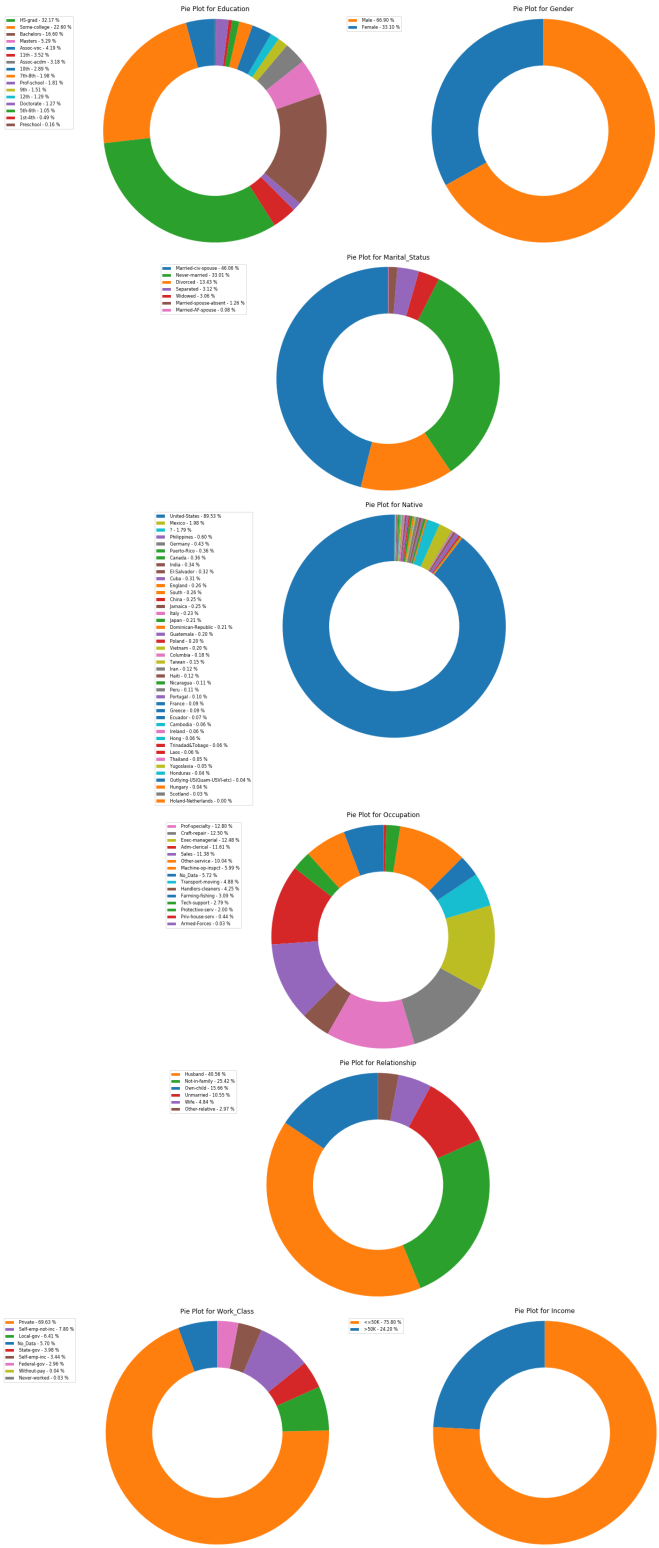

Fig. 2. Pie charts of categorical features

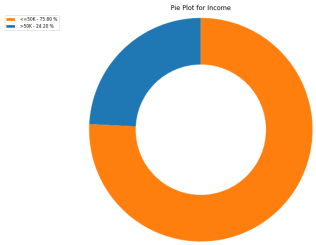

Fig. 3. Pie charts of the output classes

naive_bayes.CategoricalNB module. The smoothing parameter α was set to default 1.0 The accuracy on the training model obtained is 0.8123

The model was used on test data and the accuracy 0.8065

4) Observations:

- As seen in figure 3, 75% of the people earned less than $50,000.
- As in figure 2, majority (about 50%) of the people have had some college education .

IV. Conclusions

- In figure 3, about 75% people earn less than $50,000. This is resonable, since in 1994, the average annual income in US was $22,000
- The features are uncorrelated, as seen in figure 4. Hence Naive-Bayes classifier is a good model for the dataset
- An accuracy of 80% was achieved on the test dataset. It can be increased by using more complex models such as SVM, Random Forests and Neural Networks. This will be explored in the final submission

V. References

[1] Waskom, L.: seaborn: statistical data visualization, *Journal of Open Source Software, v6* [2] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion et al.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research, v12* [3] Hunter J. D: Matplotlib: A 2D graphics environment, *Computing in Science & Engineering, v9*

[4] Andreas C. Müller and Sarah Guido: Introduction to Machine Learning with Python: A Guide for Data Scientists [5] https://towardsdatascience.com/naive-bayes-classifier-how-to-successfully-use-it-in-python-ecf76a995069

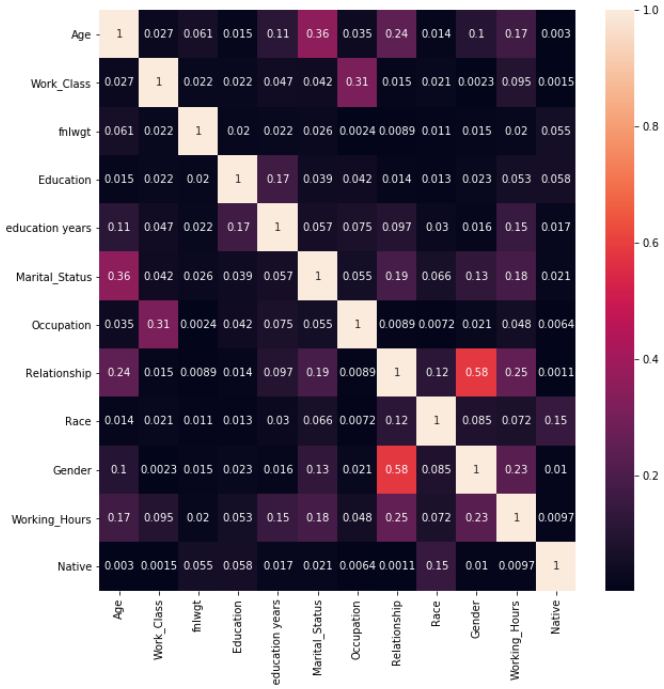

Fig. 4. Correlation between all features