Implementation of Naïve Bayes classifier on Income classification

Nithin Sai Jalukuru
School of Engineering and Applied Sciences
University at Buffalo, Buffalo, NY, USA
njalukur@buffalo.edu

Abstract : — In actual life, categorization is absolutely necessary. The Naive Bayes classifier is a mathematical classification method that uses a series of probabilistic computations to get the best classification for a given set of data inside a problem area. In the current study, Census Income data is chosen for determining whether a person is earning less than $50,000 or more than $50,000 based on input features by applying the Gaussian Nave Bayes classifier in order to guarantee the accuracy of all probabilistic calculations involved.

Keywords : Naïve Bayes, classification, Bayes Theorem, Probabilistic classification, Census Income, Adult Data

1.INTRODUCTION

The statistical method known as Bayesian decision theory is based on the quantification of trade-offs between distinct categorization judgments based on the idea of probability (Bayes Theorem) and the costs related to the decision.

In general, it is a classification method that uses the Bayes Theorem to determine the conditional probabilities. The validity of fitting a fresh piece of data into each potential categorization can then be calculated using this mathematical classifier. The classification with the highest fitness value can then be selected as the one that best fits this particular piece of data.

Although the assumption of *mutual independence* may not be true in some real-life problem domains, this assumption is commonly adopted in most of the Bayesian related computations and the results are still very trustable.

1.1 Learning Structure

Naïve Bayes works on the Bayes Theorem/Rule.

$$\mathcal{P}(\mathcal{Y}|\mathrm{x}) = \frac{\mathcal{P}(\mathcal{Y})\mathcal{P}(\mathrm{x}|\mathcal{Y})}{\mathcal{P}(\mathrm{x})} \quad \text{-- (1)}$$

Considering a strong believe that attributes are conditionally independent given the class. Mathematically,

$$\mathcal{P}(\mathrm{x}|\mathcal{Y}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathcal{P}(\mathrm{x_i}|\mathcal{Y}) \quad \text{-- (2)}$$

Where x_i is the value of i^{th} attribute in x and n in number of attributes. By substituting eq (2) in (1), equation 1 becomes,

$$\mathcal{P}(\mathcal{Y}|\mathrm{x}) = \frac{\mathcal{P}(\mathcal{Y})}{\mathcal{P}(\mathrm{x})} \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathcal{P}(\mathrm{x_i}|\mathcal{Y}) \quad \text{-- (3)}$$

Equation 3 represents the *mutual independence* between the attributes

1.2 Motivation

Motivation for selecting naïve bayes classifier because of the following properties.

I. However, classification time solely depends on the number of attributes and is independent of the number of training instances. Training time is linear with respect to both the number of training examples and the number of attributes.
II. Since the search is not used, this strategy has low variance but significant bias.
III. Naive Bayes acts on estimates of low order probabilities that are produced from the training data and adheres to incremental learning.
IV. Direct prediction of posterior probabilities.
V. Naïve Bayes considers all characteristics for all predictions, it is insensitive to noise in training data.
VI. As was previously said, this method employs all attributes for all predictions and is comparatively unaffected by missing values in the data, but the performance is negatively impacted.

2.GAUSSIAN NAÏVE BAYES APPROACH.

As the data in this study contains continuous features, the Gaussian naive bayes technique is used for classification. Therefore, we assume that the data follows a normal or gaussian distribution while working with continuous data. Hence the likelihood of the feature is assumed to be

$$\mathcal{P}(\mathrm{x_i}|\mathcal{Y}) = \frac{1}{\sqrt{2\Pi\sigma_y^2}} \exp\left(-\frac{\left(\mathrm{x_i} - \mu_y\right)^2}{2\sigma_y^2}\right)$$

Where σ is standard deviation and μ is mean and Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

Assuming that the data is characterized by a Gaussian distribution with no covariance between dimensions is one method for building a straightforward model. Finding the mean and standard deviation of the points within each

label, which is all that is required to establish such a distribution, will allow this model to be fit.

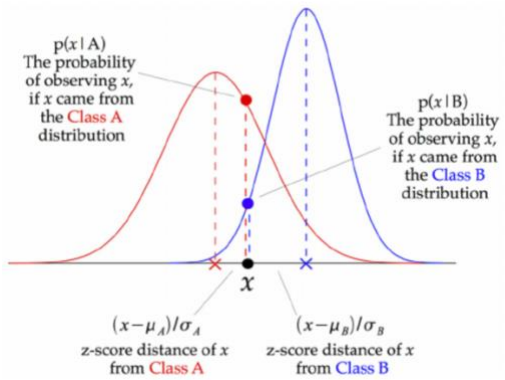

Fig1: Illustration of GNB (Gaussian naïve bayes classifier) Source (opengenus)

2. EXPERIMENTATION

2.1 Methodology

For the Naïve Bayes classifier model implementation, a data pipeline structure is constructed as illustrated in the accompanying figure 4.

Data collection and loading into the environment are both parts of data extraction After loading, the data has been labeled and cleaned so that the ML model can analyze it effectively. . In addition, feature extraction a few features/attributes have been extracted, and the most pertinent ones have been chosen to train the model.

The correct ML model has been selected, trained, and verified during the model selection, training, and validation processes by maximizing the more suitable assessment criteria.

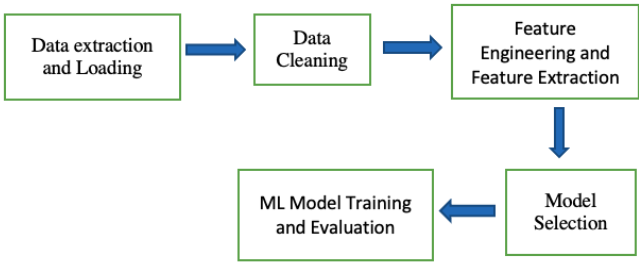

Fig2 : Data-pipeline

2.2 Data Extraction

The Adult data , which is perfect data for any classification task is collected from UC Irvine Machine Learning Repository. This data contains 32561 rows and 5 attributes. These are

a) *Age*: Which represents the age of the person(continuous feature).
b) *workclass*: This categorical feature represents the workclass of the person eg. Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
c) *education*: This categorical feature represents Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
d) *education-num* : It's a continuous variable.
e) *marital-status*: This feature represents the marital status of the person such as Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
f) *occupation*: This categorical attribute represents the occupation of the person such as Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
g) *race*: Represents the race of the person eg. White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
h) *sex*: A factor denoting the sex of a person(male or female)
i) *capital_gain, capital_loss and hours_per_week* : An continuous attributes.
j) *native-country* : This factor represents to which country does the person belongs to.

2.3 Data Cleaning and Feature Engineering

a) Raw data didn't have any columns names, so appropriate attribute names were given by following the data description.
b) Workclass, occupation and native country Factors have "?" in the data. Hence replaced them with NaN and then implemented *categorical Imputation* on those categorical attributes without removing NaN's.
c) For Feature marital_status, *categorical data binning* has performed because huge cardinality could cause the performance issues as shown in fig 3

Fig3: Data binning of marital_status factor

d) Same has done on the education feature.
e) Label Encoding has done on categorical features and the one hot encoding is not performed on those factors because ,for example native country has total 42 unique countries (high cardinality) this will eventually cause "*dimensionality curse*" for machine learning model.
f) Performed feature standardization on the numerical data attributes (agefnlwt, education_num, capital_gain, capital_loss, hours_per_week), where the feature values are re-scaled to have a mean of

$\mu = 0$ and standard deviation $\sigma = 1$.(This step is performed after EDA)

2.4 EDA and Model Selection

a) Distribution of age with respect to income

It is evident that density plot of income more than 50k follows gaussian distribution and of income less than 50k has a slight right skewness.

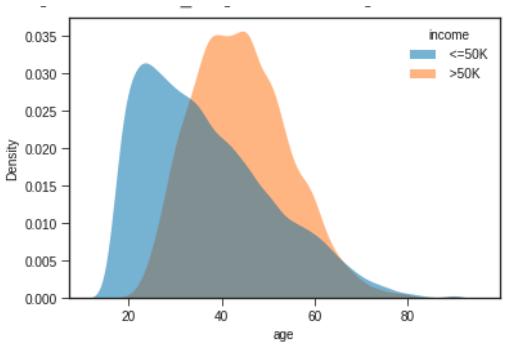

Fig4: Distribution of icome

b) Education and Income

From the plot , inference that can be drawn is that the persons who has either a masters or doctorate degree, majority of them earns more than 50k whereas a person who has a HS-Grad or a community college/Junior college degree tend to earn an income of less than 50k

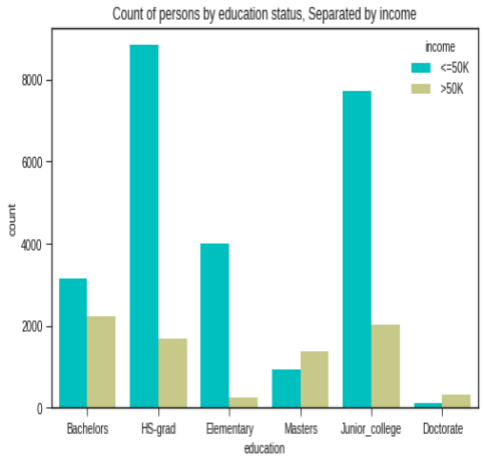

Fig5 : Count of persons by education status

c) Hours of working per week and Age

Most of the persons of age above 35 and who work 40+ hours per week , tend to earn more than 50k of income per year.

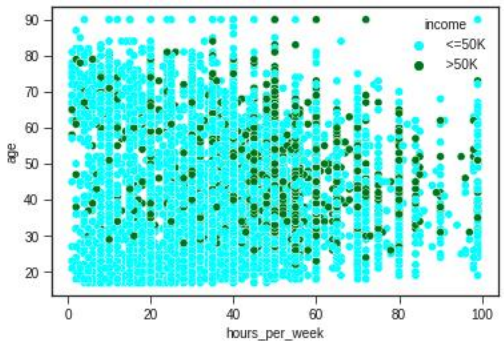

Fig6: Plot between working hours and age

d) Organization and Income

Majority of those who work in private sector/work-class earns less than 50k. In other working sectors its 50-50

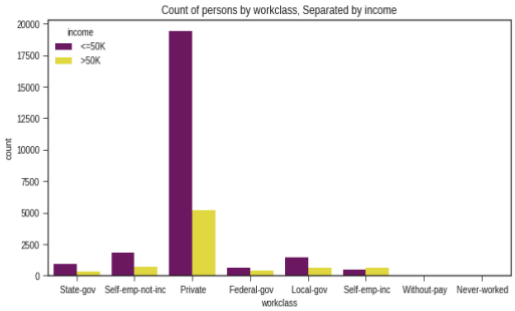

Fig7: Workclass with respect to income

e) Person's correlation plot

Features have low person co-efficient (shows almost independent) which is good in naïve bayes as this method assumes independency between the features.

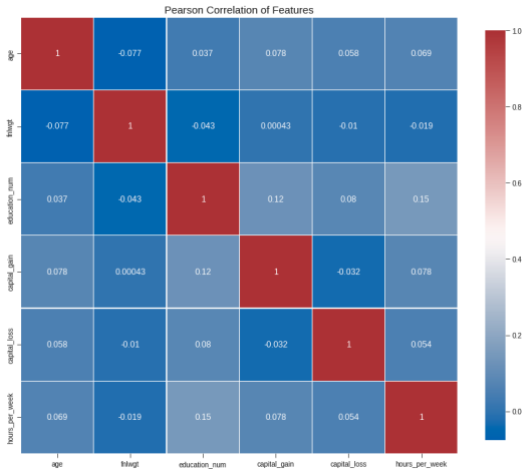

Fig8:Correlation plot

2.5 Gaussian Naïve Bayes Modelling

Finally, the data is split into 2/3rd into training the data and 1/3rd into testing the data. The data is ready GNB machine learning model.

2.6 Model Evaluation Metrics

Overall accuracy, recall, precision, f1 score, and AUC of the ROC curve are the most common evaluation measures for classification models.

False positives (FP) and false negatives (FN) are outcomes that were mistakenly classified by the model, while true positives (TP) and true negatives (TN) are outcomes of the positive class and negative class, respectively.

a) Overall Accuracy (OA): This is defined by the following the equation

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

The model could attain almost perfect overall accuracy if it consistently predicts the majority of classes.

b) The issue is more severe the more unbalanced the data. So, we require additional measurements. include Recall. It measures the proportion of accurately predicted positive classes to all positively categorized items.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall is important when we believe False Negatives are more important than False Positives

c) Precision: It is the ratio of correctly predicted positive classes to all items predicted to be positive

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

It tells us how correct or precise that our model's positive predictions are. When we think False Positives are more significant than False Negatives, precision is crucial.

d) F1- Score : The F1-score is a single performance statistic that considers both recall and precision. It is also frequently referred to as the F-Measure. It is calculated by averaging the two metrics harmonically.

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

with values closer to one indicating better performance, and values closer to zero indicating poorer performance

e) ROC Curve: The ROC Curve is helpful since it not only provides a summary of our model's performance but also makes it simple to compare the effectiveness of other classifiers.

2.7 Test Results

Let's look at the classification report

Classification Report:	precision	recall	f1-score	support
0	0.84	0.94	0.89	8277
1	0.68	0.41	0.51	2577
accuracy			0.81	10854
macro avg	0.76	0.68	0.70	10854
weighted avg	0.80	0.81	0.80	10854

Fig9 : Classification report

Naïve Bayes classifier had a overall accuracy of 81% with a good tradeoff between recall and precision (recall = 68% and precision = 76%). It is always good for a model to have a tradeoff between these two.

Also look at Confusion Matrix

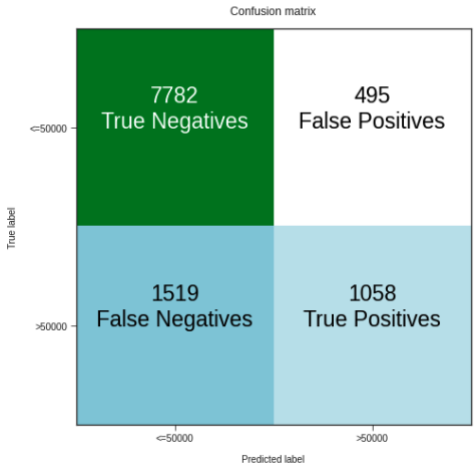

Fig10 : Confusion Matrix

From the confusion matrix , it is evident that model is successful is predicting True negatives and True Positives. Model has predicted 1519 False negatives (which means for persons with income greater than 50k, model has predicted it wrong by classifying into less than 50k category).

From ROC and AUC , AUC is nearly to 85% (>50%) which is excellent for any classification model.

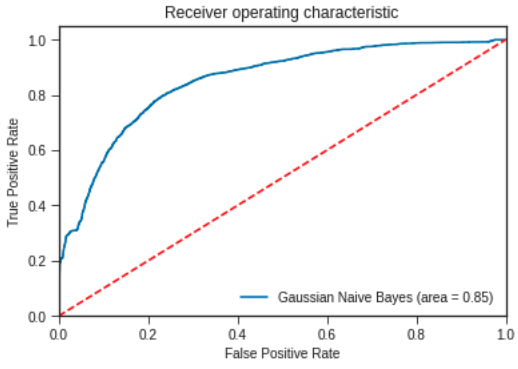

Fig11: Area under the ROC

3. CONCLUSION

Naïve Bayes classifier has done a excellent work in binary classification of income with accuracy of 81% and AUC of 85%. This score can be improved by other classification methods (Logistic, SVM, Random Forest etc...)

4. REFERENCES

1.CSE 474/574 lecture slides (Prof. Chen)
2. Jason Brownlee, Master Machine Learning Algorithms
3. “Multiclass Approaches for Support Vector Machine Based Land Cover Classification”
4. https://www.baeldung.com/cs/svm-multiclass-classification
5. Pier Francesco at.el, “Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry”.