A Mathematical Essay on Naive Bayes Classifier

Vignesh Kumar S
Department of Chemical Engineering
Indian Institute of Technology, Madras
Email address: ch18b118@smail.iitm.ac.in

***Abstract*—The purpose of this article is to understand the Naive Bayes classifier and the mathematics behind it. In this work, we demonstrate the application of the Bayes classifier illustrated through the 1994 Census bureau database by Ronny Kohavi and Barry Becker. Due to the complex nature and the abstractness involved in determining the income of a person, it is often difficult to obtain a good solution through manual approaches. Adopting a machine learning technique can help in obtaining insights and analysing different features in a large amount of data comprehensively. We try to establish a nexus between income range and different features involving the bio-data of a person.**

I. Introduction

In this paper, we will study Naive Bayes classifier, a technique based on Bayes theorem used to analyze and model the probability of a certain class, either a dichotomous or multiple outcomes. We will use this technique to determine whether a person makes over $50K a year using factors such as age, Work-class, education, marital-status, occupation, race, sex, native country etc. This analysis would aid in obtaining insights among a large population and understanding the trend behind income and various other factors.

The classification problem attempts to establish a relationship between a categorical target variable with one or more explanatory variable(s). Naive Bayes Classifier is a generative statistical modelling technique that can be considered as fitting a probabilistic model that optimises the joint likelihood $p(C, x)$. Conditional probability of outputs in terms of inputs can equivalently be constructed which can then be used to infer the predictions with certain assumptions on likelihood. They are one of the simplest Bayesian models but with density estimation and on certain datasets, they can work well and achieve high levels of accuracy.

The dataset used for this problem comprises age, working class, number of working hours per week, occupation, education levels, capital gain and loss. It also contains marital-status, relationship, race, sex and Native country. We use Naive Bayes classifier to learn and predict whether a given individual earns above $50K dollars a year given these input features.

This work represents the concepts behind Naive Bayes classifier and evaluation metrics involved. We then use this technique to establish the relationship to predict the class of income a person may belong to.

II. Naive Bayes Classifier

The goal in classification problem is to take an input feature $\boldsymbol{x}$ and assign it to one of the two classes (K classes in general). Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes Theorem with strong independence (naive) assumption between features. It is assumed that value of a particular feature is independent of any other feature.

By assuming Naive assumptions, we are essentially ignoring relationships among features. Because of such conditions, Naive Bayes is said to have *high bias* and this being one of the simplistic models, is said to have *low variance*.

Let $x_1, x_2, ..., x_D$ be the feature vector $\boldsymbol{x}$.

$$P(y = C_k|x_1, x_2, ..x_D) = \frac{P(y = C_k)P(x_1, x_2, ..x_D|y = C_k)}{\sum_{j=1}^{K} P(x_1, x_2, ..x_D|y = C_j)P(C_j)}$$

Assuming $x_1, x_2, ..., x_D$ are independent,

$$= \frac{P(y = C_k)\prod_{i=1}^{D} P(x_i|y = C_k)}{\sum_{j=1}^{K}\prod_{i=1}^{D} P(x_i|y = C_j)P(C_j)}$$

$$P(y = C_i|\boldsymbol{x}) \propto P(y = C_k)\prod_{i=1}^{D} P(x_i|y = C_k)$$

$P(y = C_k)$ is called the overall or the prior probability.

Now the challenge lies in identifying $P(x_i|y)$. Notice that because we assumed conditional independence, we now have to model only $P(x_i|y)$ individually for all features instead of modelling $P(x_1, ..., x_D)$ in the D-dimensional space. This greatly reduces computational complexity as we search only in 1 dimensional space d times instead of searching in a D-dimensional space one time (Volume expands cubically). This is illustrated mathematically in the below section.

$P(x_i|y)$ is generally assumed to be Gaussian for continuous data and Bernoulli distribution for binary data (Multinoulli/ Categorical distribution for categorical data). The accuracy could be improved by modelling $P(x_i|y)$ more accurately by using any density estimating methods.

A. Reasoning behind Naive Assumption

Let's for now do not consider conditional independence and assume the likelihood given a class k is from a Multivariate distribution.

$$X = (X_1, X_2, ..., X_D)$$

$$X|y \sim \mathcal{N}(\mu_k, \Sigma_k)$$

$$f(X|y = C_k) = \frac{1}{2\pi^{\frac{D}{2}}|\Sigma_k|^{0.5}} exp(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k))$$

where μ_k and Σ_k are class specific mean and covariance matrix. These mean and covariance matrix needs to be determined or in a machine learning linguistics, needs to be learnt from the dataset.

μ_k has D parameters and Σ_k has D^2 parameters that need to be determined. If there are K such classes, number of parameters that need to be estimated is

$$K(D + D^2)$$

This also means we need to have a huge dataset in order to estimate the parameters accurately (since MLE estimates are consistent and work well under large datasets). If we assume conditional independence, modelling $\boldsymbol{x}|y$ =C$_k$ reduces to modelling $x_i|y = C_k$ for $i\forall D$

Assuming $x_i|y \sim \mathcal{N}(\mu_k, \sigma_k)$, only 2 parameters are unknown this time. For D such features and K such classes, the number of unknown parameters are just 2KD which is a lot lesser than earlier and has huge computational advantages.

It is to be noted that the latter method, though computationally advantageous, doesn't capture model complexity as the conditionally dependent one. This trade-off needs to be well noted in understanding Naive Bayes classifier. Hence Naive Bayes is a high bias but a low variance classifier.

B. Implementation

Let's understand naive bayes classifier for two classes (K=2). Recall,

$$P(y = C_1|X = \boldsymbol{x}) = \frac{P(y = C_1)\prod_{i=1}^{D} P(x_i|y = C_1)}{\sum_{i=1}^{2} P(y = C_i)\prod_{i=1}^{D} P(x_i|y = C_i)}$$

$$\text{And } P(Y = C_2|X = \boldsymbol{x}) = 1 - P(Y = C_1|X = \boldsymbol{x}) \qquad (1)$$

Depending on the feature, we could assume various choices for $x_i|y$ distribution. Once we obtain $P(y = C_i|X = x)$, predictions could be done by

$$\hat{y} = \begin{cases} C_1 & \text{if } \hat{p} < Threshold \\ C_2 & \text{if } \hat{p} \geq Threshold \end{cases}$$

where C_1 and C_2 are the two classes. The threshold is usually set to 0.5.

C. Gaussian Naive Bayes

If the input feature is continuous and follows approximate Gaussian, we assume this distribution,

$$f_k(x|y = k) = \frac{1}{\sqrt{2\pi}\sigma_k} exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

where μ_k and σ_k^2 are the mean and variance for the k^{th} class (k=2 for our case).

$P(Y = C_1|X = x)$, $P(Y = C_2|X = x)$ is given by (1). We still do not know μ_1 and μ_2, σ_1 and σ_2 and have to be estimated.

We can use MAP estimates (Maximum A posteriori) estimates to estimate P(y) (Prior) and parameters. The estimates are given below:

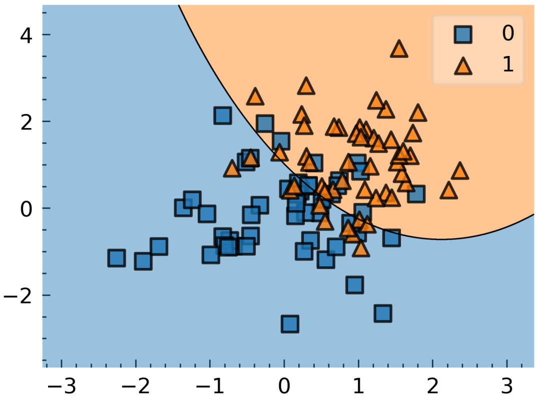
$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i\forall n_k} x_i$$

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1}\sum_{i\forall n_k}(x_i - \hat{mu}_k)^2$$

$$Prior\ P(y = C_i) = \frac{n_k}{n}$$

n ≡ total no. of observations

n_k ≡ total no. of observations in the kth class



'The above plot represents the decision boundary of a Naive Bayes classifier obtained on a synthetic dataset of two features. The x_1 and x_2 used were generated from Gaussian distribution with a small correlation. Since the correlation was small and the distribution with a small correlation. Since the correlation was small and the distribution matches, naive bayes classifier does a good job of separating the classes. Note that unlike logistic regression, naive bayes don't produce a linear decision boundary.

The decision boundary would be linear if we had assumed a shared variance across classes instead of a different variance across each class.

D. Categorical Naive Bayes

When the input feature is categorically distributed, we use categorical Naive Bayes. CategoricalNB estimates a categorical distribution for each feature i of X conditioned on class y.

$$P(X_i|y = C_k) = \frac{N_{kt}}{N_k}$$

where N_{kt} is the number of times category t appears in the sample x_i which belongs to class k. N_c is the number of samples with class k.

Sometimes, when a new category hasn't been seen before arrives, the probability turns zero. In order to prevent this, a smoothing parameter, $\alpha > 0$ (usually 1) is added to the numerator and denominator.

$$P(X_i|y = C_k) = \frac{N_{kt} + \alpha}{N_k + \alpha}$$

E. *Time Complexities*

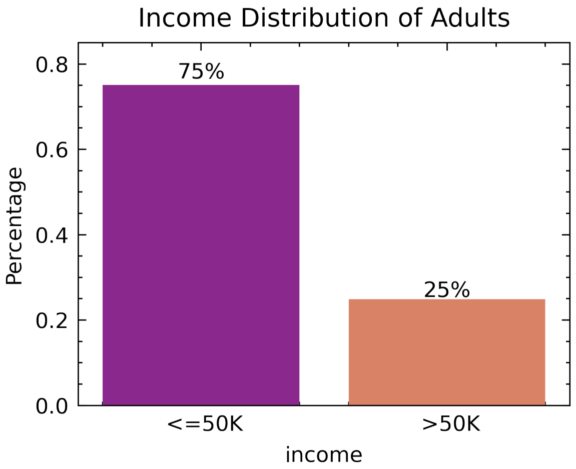
For a Categorical Naive Bayes training, the worst case is when we have n categories in n-datapoints (i.e. Each datapoint has a unique category), D dimensions and C classes. Then, inorder to compute and store all combinations, the required time complexity will be in the order of $O(n*d*c)$. The testing will require $O(d*c)$ as it involves just fitting and multiplying all d-dimensions for each class. Similar time complexity is obtained for Gaussian as well. Naive Bayes hence could learn fast and is computationally advantageous.

III. The Problem

In this section, We will analyse the Census bureau database and try to predict whether a person earns above $ 50 K per year or not.

A. *Imbalanced Dataset*

The dataset contains around 75% of people who earn below $50K and only 25% of people who earn above $50K



We know,

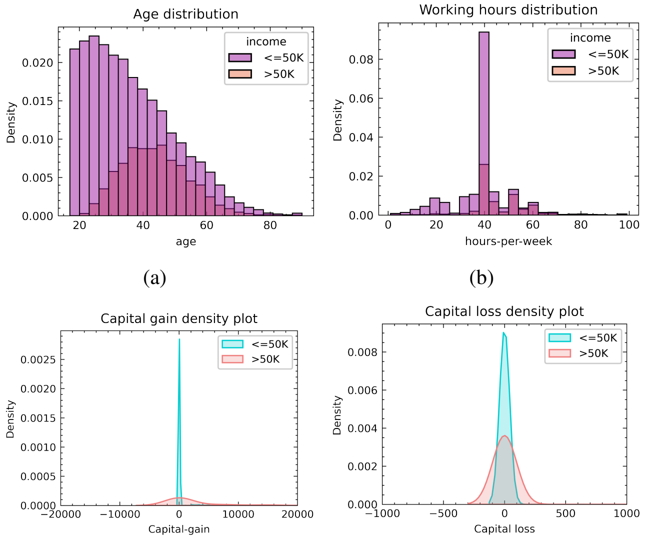
$$P(y = k|X = x) \propto P(y = k) \prod P(x_i|y = k)$$

Even though the likelihood probabilities are similar to some extent, the posterior probabilities are affected by the prior probabilities. We build two classifiers.

1) One with ignoring the prior probabilities (assuming equal weightage to both classes). We call it CLF1
2) Another with prior estimated through the dataset (0.751, 0.249). We call it CLF2.

Care is taken to ensure similar proportions are used in training and test set by doing a stratified split.

B. *Numerical Attributes*



(a)

(b)

(c)

(d)

It is important that inorder for Naive Bayes classifier to work well, the likelihood $X_i|y = C_i$ must be close to assumed Gaussian distribution.

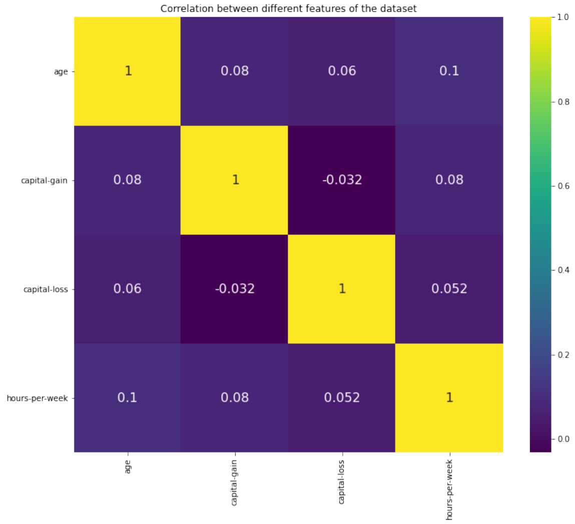
Age distribution, though shows some exponential feature, seems skewed and may not fit Gaussian perfectly. Also, more people earn above $50K at 45 years and above, when their career matures which makes sense. Box-cox transformation is a technique to make non-normal distribution into a normal-shape. Normality is an important assumption for many statistical techniques including Naive Bayes.

If $w_t$is our transformed variable and y is our target, then

$$w_t = \begin{cases} log(y_t) & \text{if } \lambda = 0 \\ \frac{y_t - 1}{\lambda} & \text{otherwise} \end{cases}$$

$\lambda$is varied from -x to x (-5 to 5) usually. The optimal value is the one which results in the best approximation of a normal distribution. Though the above formulation only works for y_t being positive, this could be easily fixed by considering $y_t + c$ instead of y_t Capital-gain and Capital-loss seem to follow a normal distribution. Hours-per-week seems a little bumpy but the normal distribution would be a decent approximation due to the spike in middle.

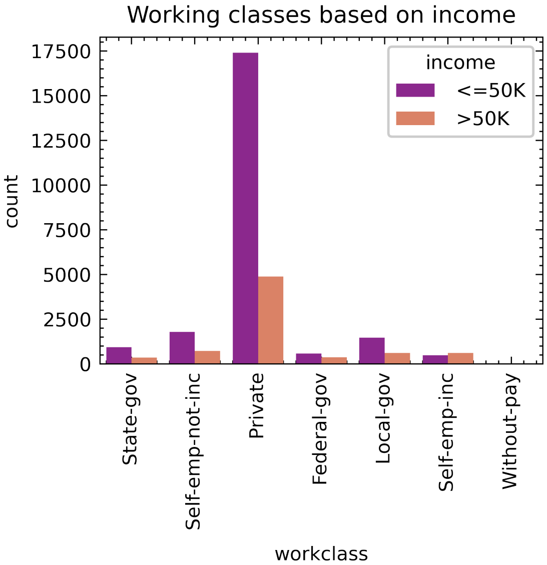
Education-num (Number of years of education) feature is dropped since we are including categorical education level feature and correlation could exist between them.



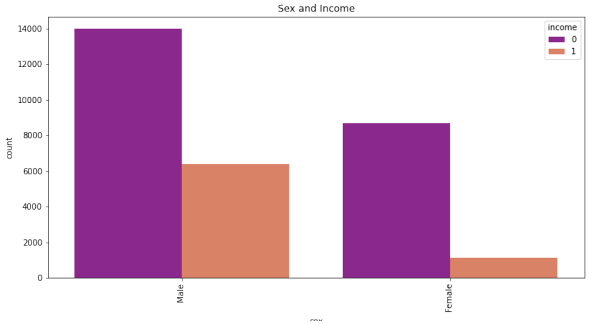
The above heat map reveals that the correlation between numerical features is very small. Since naive bayes assumes independence, in order for it to work well, this needs to hold good.

C. Categorical Features

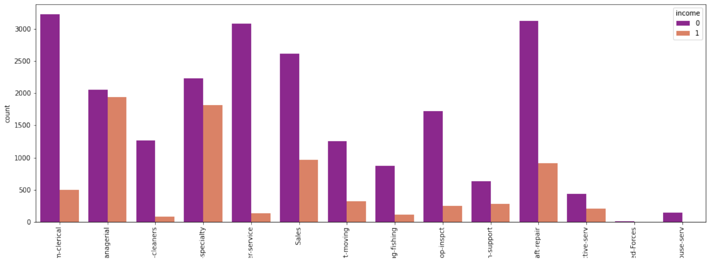
Insights regarding few categorical features are discussed below.



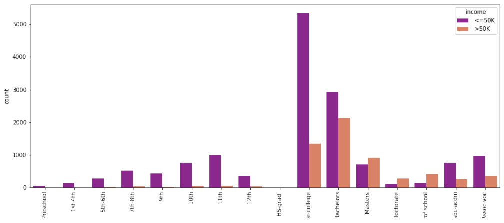
People in the private sector face significant differences in the pay scale. A lot of people earn below 50k and around one third earn above. There aren't as many differences that exist in the government sector compared to private ones. Self employed inc is the only category where there is more number of people earning above 50k than below. We could say when you handle your own business, start-up, you are more likely to earn better if it gets successful.



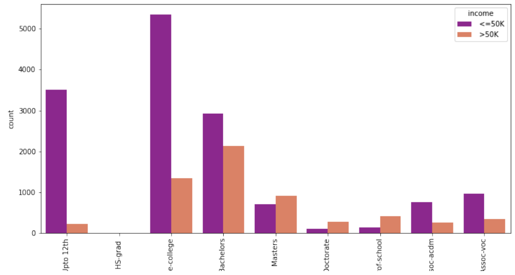
Less than 10 % of females earn more than 50K dollars a year while 33% of males earn more than 50K dollars a year.



Executive managerial role and Prof-specialty roles are likely to earn more than 50K dollars a year while clerical, Handlers-cleaners, Farming-fishing, Machine inspection, Transport-moving are less likely to earn a lot.



(e)

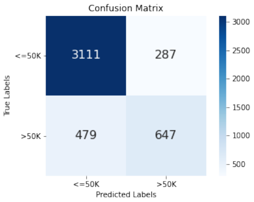


(f)

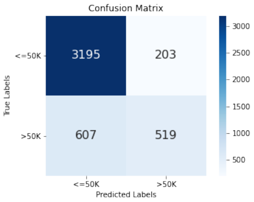
People who have completed less than 12th standard rarely earn above 50K. The distribution for people below 12th is almost similar with most earning considerably less. We could group education less than 12th into a single group. This would make probability estimates accurate, especially useful in a naive bayes setting. People who have completed Bachelors, Masters, Doctorate, Prof-School significantly are on the higher pay scale implying education does matter to a certain extent.

D. Model Evaluation

After Data cleaning, Preprocessing and Feature Engineering, two Naive Bayes classifiers (one with uniform prior, another with prior learnt from data) were trained. The following confusion matrix was obtained on the test set.



(g) Uniform Prior- CLF1

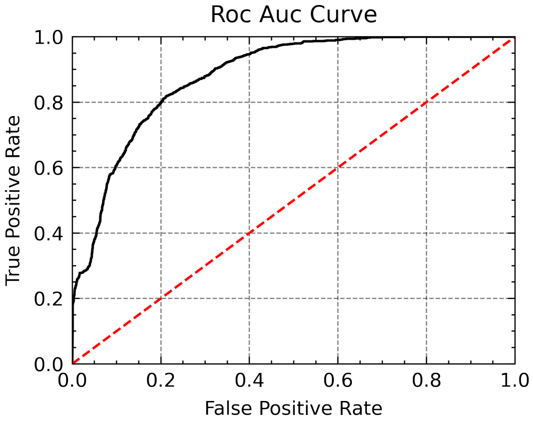


(h) Learnt Prior - CLF2

Naive Bayes Classifier		
Metric	CLF1 Score	CLF2 Score
Accuracy	0.8278	0.8177
Precision	0.6824	0.7107
Recall	0.5772	0.4516
F1 score	0.6254	0.5523

In a classification setting, accuracy may not be the best score to consider especially in a skewed class situation. The training data in our case is imbalanced. Precision is the ratio of correct positive predictions to the total positive predictions of our model. The recall is the ratio of positive instances that are correctly detected by our classifier. The F1 score is simply a harmonic average of these two. Classifier 1 (that assumes uniform prior for both classes) seems to perform better both in terms of Accuracy and F1 score than the classifier 2 (that learnt its prior)

ROC curve is another common tool used with binary classifiers. The ROC curve plots the true positive rate (another name for recall) against the false positive rate. The FPR is the ratio of negative instances that are incorrectly classified as positive. It is equal to one minus the true negative rate, which is the ratio of negative instances that are correctly classified as negative. The TNR is also called specificity. Hence the ROC curve plots sensitivity (recall) versus 1 – specificity.



The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner). One way to compare classifiers is to measure the area under the curve (AUC). A perfect classifier will have a ROC AUC equal to 1, whereas a purely random classifier will have a ROC AUC equal to 0.5. The area under the ROC curve of our Naive Bayes classifier is *0.8821*.

Both the classifiers have improved significantly after correcting for normality. The best classifier has improved approximately 0.7 percent in F1 score and also in Area under ROC. This reinstates how important distribution assumptions are in Naive Bayes setting.

E. Comparison with SVM

Let's try doing the same dataset with SVC's. This algorithm was chosen because it doesn't emphasis on Naive assumptions anymore and could capture non-linearities that we would get in Naive Bayes. A Support Vector Classifier with RBF kernel was gridsearched over γ and C.

Support Vector Classifier Comparison	
Metric	CLF Score
Accuracy	0.8566
Precision	0.6484
Recall	0.92
F1 score	0.7628

This clearly outperforms the Naive Bayes classifier at most levels. It is to be noted that the SVC classifier for single run in this dataset on the same machine took 57.6s on average while Naive Bayes was almost instantaneous. This is to conclude that even though SVM's performed better, Naive Bayes does possess discriminatory characteristics to separate out the classes and would be useful to set a minimum benchmark in large datasets.

IV. Conclusions

Based on our analysis, the key takeaways we had from this exercise are the following:

1) Naive Bayes assumes conditional independence and it may work well only when the assumptions are well met.

2) Imbalanced data could affect the predictions of Naive Bayes classifier. As in our case, we saw that adjusting to a uniform prior resulted in better predictions compared to the settings where no interventions were taken.
3) Factors such as age, education, working hours, sex and race seem to have impacted the income levels. While education and working hours are positives that could motivate an individual, issues related to disparities based on sex and race need to be addressed.

Possible avenues of research could include transforming numerical features to fit Gaussian well (such as box-cox), applying density estimation techniques for better fit and comparing results of other machine learning techniques to Naive Bayes classifier.

References

[1] An Introduction to Statistical Learning, Gareth James et al. pp.138-151

[2] Aurelian Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems, pp. 112–180

[3] Christopher M. Bishop, Pattern Recognition and Machine Learning