Volume 8, Issue 12, December 2023 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Analysis And Implementation Of Naive Bayes Classifier For Continuous Data

Niraj Kumar Padme
M Tech 4th Sem
Computer Science and Engineering Department
LNCT(Bhopal) Indore Campus
Indore M.P. India

Mr Nilesh Avinash Joshi
Assistant professor
Computer Science and Engineering Department
LNCT(Bhopal) Indore Campus
Indore M.P. India

Abstract: ***Bayes theorem is given by an English statistician, philosopher, and Presbyterian minister named Mr. Thomas Bayes in 17th century. Bayes theorem provides thoughts in decision theory which is extensively used in important mathematics concepts such as Probability. Bayes theorem is widely used in Machine Learning where we need to predict classes precisely and accurately. An important concept of Bayes theorem named Bayesian method is used to calculate conditional probability in Machine Learning application that includes classification tasks. Simplified version of Bayes theorem (Naïve Bayes classification) is also used to reduce computation time and average cost of the projects In the proposed work we used Gaussian Naïve Bays Classifier. We know that Naïve Bays Classifier is a widely used classifier in machine learning, we classify data for binary class and also for multi class. Naïve Bays Classifier will efficiently work for binary as well as multiclass. Naïve Bays Classifier has certain limitations it is good when the data set has categorical or discreate value.***

Keywords: Naïve Bays, Classifier, Binary, Multiclass, Continues, Normal Distribution

I. INTRODUCTION

Naive Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks. Typical applications include filtering spam, classifying documents, sentiment prediction etc. It is based on the works of Rev. Thomas Bayes (1702) and hence the name. Naive Bayes classifiers are an assortment of simple and powerful classification algorithms based on Bayes Theorem. They are recommended as a first approach to classify complicated datasets before more refined classifiers are used.

Bayes Theorem is a collection of algorithms that share a common principle. With Bayes theorem, users find the likelihood of A happening, given that B transpired. In the equation provided below, B is the evidence and A is the hypothesis. The fundamental assumption of Bayes Theorem is that predictors are independent. In other words, the existence of one predictor will not influence the other. This is a classic example of conditional probability. So, when you say the conditional probability of A given B, it denotes the probability of A occurring given that B has already occurred. Mathematically, Conditional probability of A given B can be computed as: P(A|B) = P(A AND B) / P(B) School Example. Consider a school with a total population of 100 persons. These 100 persons can be seen either as 'Students' and 'Teachers' or as a population of 'Males' and 'Females'. With below tabulation of the 100 people, what is the conditional probability that a certain member of the school is a 'Teacher' given that he is a 'Man'?

TABLE 1
Simple data set

	Male	Female	Total
Teacher	8	12	20
Students	32	48	80
Total	40	60	100

To calculate this, you may intuitively filter the sub-population of 60 males and focus on the 12 (male) teachers. So the required conditional probability P(Teacher | Male) = 12 / 60 = 0.2.

$P(Teach\ er|Male)=P(Teach\ er\cap Male)P(Male)=1260=0.2$

This can be represented as the intersection of Teacher (A) and Male (B) divided by Male (B). Likewise, the conditional probability of B given A can be computed. The Bayes Rule that we use for Naive Bayes, can be derived from these two notations.

$P(A|B)=P(A\cap B)P(B)$

$P(B|A)=P(A\cap B)P(A)$

II. Prerequisites for Bayes Theorem

While studying the Bayes theorem, we need to understand few important concepts. These are as follows:

A. Experiment

An experiment is defined as the planned operation carried out under controlled condition such as tossing a coin, drawing a card and rolling a dice, etc.

B. Sample Space

During an experiment what we get as a result is called as possible outcomes and the set of all possible outcome of an event is known as sample space. For example, if we are rolling a dice, sample space will be: S1 = {1, 2, 3, 4, 5, 6}
Similarly, if our experiment is related to toss a coin and recording its outcomes, then sample space will be: S2 = {Head, Tail}

C. Event

Event is defined as subset of sample space in an experiment. Further, it is also called as set of outcomes. Assume in our experiment of rolling a dice, there are two event A and B such that;
A = Event when an even number is obtained = {2, 4, 6}
B = Event when a number is greater than 4 = {5, 6}

- Probability of the event A "P(A)"= Number of favourable outcomes / Total number of possible outcomes P(E) = 3/6 =1/2 =0.5
- Similarly, Probability of the event B "P(B)"= Number of favourable outcomes / Total number of possible outcomes =2/6 =1/3 =0.333
- Union of event A and B: A∪B = {2, 4, 5, 6}

Disjoint Event: If the intersection of the event A and B is an empty set or null then such events are known as disjoint event or mutually exclusive events also.

D. 4. Random Variable

It is a real value function which helps mapping between sample space and a real line of an experiment. A random variable is taken on some random values and each value having some probability. However, it is neither random nor a variable, but it behaves as a function which can either be discrete, continuous or combination of both.

E. Exhaustive Event

As per the name suggests, a set of events where at least one event occurs at a time, called exhaustive event of an experiment. Thus, two events A and B are said to be exhaustive if either A or B definitely occur at a time and both are mutually exclusive for e.g., while tossing a coin, either it will be a Head or may be a Tail.

F. Independent Event

Two events are said to be independent when occurrence of one event does not affect the occurrence of another event. In simple words we can say that the probability of outcome of both events does not depends one another.
Mathematically, two events A and B are said to be independent if $P(A \cap B)=p(AB)=P(A)*P(B)$

G. Conditional Probability:

Conditional probability is defined as the probability of an event A, given that another event B has already occurred (i.e. A conditional B). This is represented by P(A|B) and we can define it as: $P(A|B)=P(A \cap B)P(B)$

H. Marginal Probability

Marginal probability is defined as the probability of an event A occurring independent of any other event B. Further, it is considered as the probability of evidence under any consideration.

$$P(A)=p(AB)*P(B)+P(A \sim B)*P(\sim B)$$

III. Literature Survey

In 2015 Ruth Talbot et al "SWASH: A Naive Bayes Classifier for Tweet Sentiment Identification". They described a sentiment classification system designed for SemEval-2015, Task 10, Subtask B. The system employs a constrained, supervised text categorization approach. Since thorough preprocessing of tweet data was shown to be effective in previous SemEval sentiment classification tasks, various preprocessing steps were introduced to enhance the quality of lexical information. Secondly, a Naive Bayes classifier is used to detect tweet sentiment. The classifier is trained only on the training data provided by the task organizers. The system makes use of external human-generated lists of positive and negative words at several steps throughout classification[1].

In 2016 B. M. Gayathri proposed "An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer" The proposed work is to classify breast cancer with few attributes. Reducing the attributes reduces the time, so that the patient need not wait for result for a long time. For classification, the user-friendly environment is created. The user can enter the details of the patient such as Clump thickness, Uniformity in cell size etc., and the result is classified as benign or malignant. Statistical analysis: Variable selection is done by one of the variable reduction algorithms called Linear Discriminant Analysis (LDA). LDA is one of the statistical methods. The dataset is passed to LDA function repeatedly and the combination of variables which gave the good accuracy is selected [2].

In 2018 Unggul Widodo Wijayanto proposed "An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naïve Bayes". They used customer reviews from Yelp (foods), IMDb (movies) and Amazon (products). The received by reviews the company are numerous. Product management does not have much time to read customer reviews one by one. So, to speed up the reading of customer reviews were using sentiment analysis. There are many methods that used in sentiment analysis such as supervised sentiment analysis. They used TF-IDF to convert word to features implements the supervised method. Performance of the supervised method depends on the data training quality. So, to improve the accuracy of the results by improving data training quality[3].

In 2019 Nafizatus Salmi et al proposed "Naïve Bayes Classifier Models for Predicting the Colon Cancer" Cancer has been known as a disease consisting of several different types. Cancer is a life-threatening disease in the world today. There are so many types of cancer in the world, one of which is colon cancer. Colon cancer is one of the number one killer in the world. However, because there aren't any obvious symptoms of colon cancer at an early stage, people do not realize that they suffer from it[4].

In 2020 Marzuki Ismail et al proposed "Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task". They used also integrated with machine learning for the purpose of opinion mining and sentiment classification as well as it utilized as a method for predicting the diseases. They explored the several different techniques that will give different results based on their respective algorithms. They used three algorithms to objectively show which algorithm is better at classifying these datasets. The parameters that were involved were feature selection and removal[5].

In 2021 Hong Chen et al proposed "Improved naive Bayes classification algorithm for traffic risk management" Naive Bayesian classification algorithm is widely used in big data analysis and other fields because of its simple and fast algorithm structure. Aiming at the shortcomings of the naive Bayes classification algorithm, they used feature weighting and Laplace calibration to improve it, and obtains the improved naive Bayes classification algorithm. Through empirical research, it is found that the improved naive Bayes classification algorithm can greatly improve the correct rate of discrimination analysis from 49.5 to 92%[6].

In 2022 M. Vijay Anand et al proposed "Gaussian Naive Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer". Cancer is a disease caused by uncontrollable cell growth. Disease is a constant subject of concern due to unavailability of treatment at a severe level. Patients who have suffered from the disease have the chance of getting saved if this fatal illness is identified in the beginning stage. Survival chance will be very low if it is detected in the final stage of cancer[7].

IV. First Approach (In case of a single feature)

Naive Bayes classifier calculates the probability of an event in the following steps:

- Calculate the prior probability for given class labels
- Find Likelihood probability with each attribute for each class Put these value in Bayes Formula and calculate posterior probability.
- See which class has a higher probability, given the input belongs to the higher probability class.

For simplifying prior and posterior probability calculation, you can use the two tables. frequency and likelihood tables. Both of these tables will help you to calculate the prior and posterior probability. The Frequency table contains the occurrence of labels for all features. There are two likelihood tables. Likelihood Table 1 is showing prior probabilities of labels and Likelihood Table 2 is showing the posterior probability.
Now suppose you want to calculate the probability of playing when the weather is overcast.

Probability of playing:

P(Yes | Overcast) = P(Overcast | Yes) P(Yes) / P (Overcast) .....(1)

Calculate Prior Probabilities:
P(Overcast) = 4/14 = 0.29
P(Yes)= 9/14 = 0.64
Calculate Posterior Probabilities:
P(Overcast |Yes) = 4/9 = 0.44
Put Prior and Posterior probabilities in equation (1)
P (Yes | Overcast) = 0.44 * 0.64 / 0.29 = 0.98(Higher)

Calculate Prior Probabilities:
P(Overcast) = 4/14 = 0.29
P(No)= 5/14 = 0.36
Calculate Posterior Probabilities:
P(Overcast |No) = 0/9 = 0
Put Prior and Posterior probabilities in equation (2)
P (No | Overcast) = 0 * 0.36 / 0.29 = 0

The probability of a 'Yes' class is higher. So you can determine here if the weather is overcast than players will play the sport.

SECOND APPROACH (IN CASE OF MULTIPLE FEATURES)

Naive Bayes classifier calculates the probability of an event in the following steps:

- Calculate prior probability for given class labels.
- Calculate conditional probability with each attribute for each class.
- Multiply same class conditional probability
- Multiply prior probability with step 3 probability

Now suppose you want to calculate the probability of playing when the weather is overcast, and the temperature is mild.
Probability of playing:
*P(Play= Yes | Weather=Overcast, Temp=Mild) = P(Weather=Overcast, Temp=Mild | Play= Yes)*P(Play=Yes) .....(1)
P(Weather=Overcast, Temp=Mild | Play= Yes)= P(Overcast |Yes) P(Mild |Yes) ……..(2)
1. Calculate Prior Probabilities: P(Yes)= 9/14 = 0.64
2. Calculate Posterior Probabilities: P(Overcast |Yes) = 4/9 = 0.44 P(Mild |Yes) = 4/9 = 0.44
Put Posterior probabilities in equation (2) P(Weather=Overcast, Temp=Mild | Play= Yes) = 0.44 * 0.44 = 0.1936(Higher)
4. Put Prior and Posterior probabilities in equation (1) P(Play= Yes | Weather=Overcast, Temp=Mild) = 0.1936*0.64 = 0.124
Similarly, can calculate the probability of not playing:

Probability of not playing:
*P(Play= No | Weather=Overcast, Temp=Mild) = P(Weather=Overcast, Temp=Mild | Play= No)*P(Play=No) .....(3)
P(Weather=Overcast, Temp=Mild | Play= No)= P(Weather=Overcast |Play=No) P(Temp=Mild | Play=No) ……..(4)
1. Calculate Prior Probabilities: P(No)= 5/14 = 0.36
2. Calculate Posterior Probabilities: P(Weather=Overcast |Play=No) = 0/9 = 0 P(Temp=Mild | Play=No)=2/5=0.4
3. Put posterior probabilities in equation (4) P(Weather=Overcast, Temp=Mild | Play= No) = 0 * 0.4= 0
4. Put prior and posterior probabilities in equation (3) P(Play= No | Weather=Overcast, Temp=Mild) = 0*0.36=0

The probability of a 'Yes' class is higher. So you can say here that if the weather is overcast than players will play the sport.

V. RESULT ANALYSIS

The dataset contains three attributes-Gender, Age, and Estimated Salary of people surveyed. The target class is Purchased where based on the other three attributes the buying decision of a person is observed. The output has two classes-0 and 1. 0 means the customer did not buy the product, 1 means they buy one. So, this is clearly a classification problem we trained the model by 75% of training data set.

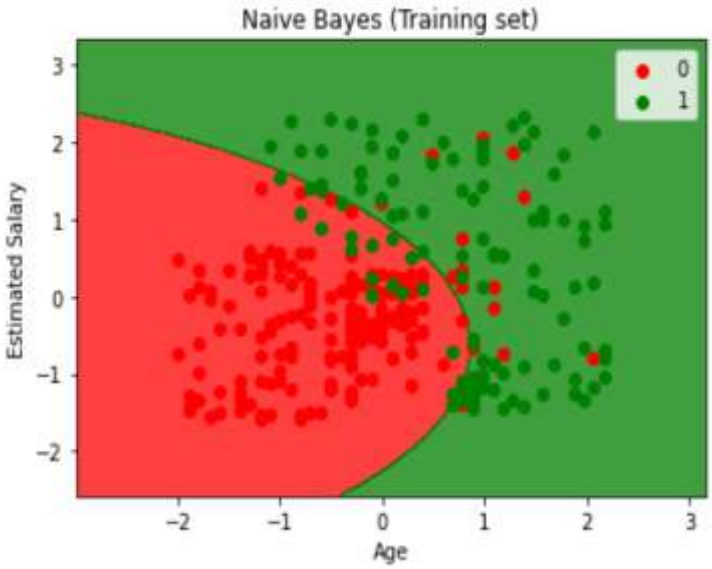

Fig 1 Plotting Age vs Estimated salary for training dataset

CONCLUSION

An important concept of Bayes theorem named Bayesian method is used to calculate conditional probability in Machine Learning application that includes classification tasks. Simplified version of Bayes theorem (Naïve Bayes classification) is also used to reduce computation time and average cost of the projects. Naïve Bayes are robust to isolated noise points because such points are averaged out when estimating conditional probabilities from data. It can also manage missing values by deleting the instances during model constructing and classification. Naïve Bayes is robust to irrelevant attributes. If Xi is an inappropriate attribute, therefore P (Xi|Y) becomes consistently distributed. The class conditional probability for Xi has no impact on the complete calculation of the posterior probability. In the proposed work we used

REFERENCE

1. Ruth Talbot , Chloe Acheampong and Richard Wicentowski "SWASH: A Naive Bayes Classifier for Tweet Sentiment Identification" Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 626–630, Denver, Colorado, June 4-5, 2015. c 2015 Association for Computational Linguistics.
2. B. M. Gayathr C. P. Sumathi, PhD An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer International Journal of Computer Applications (0975 – 8887) Volume 148 – No.6, August 2016
3. Unggul Widodo Wijayanto Riyanarto Sarno An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naïve Bayes 2018 International Seminar on Application for Technology of Information and Communication (iSemantic)
4. Nafizatus Salmi and Zuherman Rustam Naïve Bayes Classifier Models for Predicting the Colon Cancer 9th Annual Basic Science International Conference 2019 (BaSIC 2019) IOP Conf. Series: Materials Science and Engineering 546 (2019) 052068 IOP Publishing doi:10.1088/1757-899X/546/5/052068
5. Marzuki Ismail , Norlida Hassan , Salem Saleh Bafjaish Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task JOURNAL OF SOFT COMPUTING AND DATA MINING VOL.1 NO. 2 (2020) 1-10 © Universiti Tun Hussein Onn Malaysia Publisher's Office JSCDM Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm Journal of Soft Computing and Data Mining e-ISSN : 2716-621X
6. Hong Chen , Songhua Hu , Rui Hua and Xiuju Zhao Hong Chen , Songhua Hu , Rui Hua and Xiuju Zhao Chen et al. EURASIP Journal on Advances in Signal Processing (2021) 2021:30 https://doi.org/10.1186/s13634-021-00742-6
7. M. Vijay Anand, B. KiranBala, S. R. Srividhya , Kavitha C. ,Mohammed Younus, and Md Habibur Rahman Gaussian Nave Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer Hindawi Mobile Information Systems Volume 2022, Article ID 2436946, 7 pages https://doi.org/10.1155/2022/2436946.
8. Allsela Meiriza , Endang Lestari , Pacu Putra , Ayu Monaputri , and Dini Ayu Lestari Prediction Graduate Student Use Naive Bayes Classifier Advances in Intelligent Systems Research, volume 172 Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)
9. M. Vijay Anand, B. KiranBala, S. R. Srividhya , Kavitha C. ,Mohammed Younus, and Md Habibur Rahman Gaussian Nave Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer Hindawi Mobile Information Systems Volume 2022, Article ID 2436946, 7 pages https://doi.org/10.1155/2022/2436946.
10. Naulak, Chingmuankim A comparative study of Naive Bayes Classifiers with improved technique on Text Classification https://www.techrxiv.org/articles/preprint/ 10.36227/techrxiv.19918360.v1.
11. Daniel Jurafsky & James H. Martin. Naive Bayes and Sentiment Classification Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright © 2023. All rights reserved. Draft of January 7, 2023.
12. Kalakonda Shashank1, Anumandla Sahithya2, Shaik Shakeel3, Dr. R. Lakshmipriya4 Reviews analysis Using gaussian naïve bayesin machine Learning International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 10 Issue: 01 | Jan 2023 www.irjet.net p-ISSN: 2395-0072
13. M. Vedaraj, 2C.S. Anita, 3A. Muralidhar, 4V. Lavanya, 5K. Balasaranya, 6P. Jagadeesan Early Prediction of Lung Cancer Using Gaussian Naive Bayes Classification Algorithm International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING ISSN:2147-67992147-6799 www.ijisae.org Original Research Paper International Journal of Intelligent Systems and Applications in Engineering.
14. Mokhairi Makhtar, Hasnah Nawang, Syadiah Nor Analysis On Students Performance Using Naïve Bayes Classifier Journal of Theoretical and Applied Information Technology 31st August 2017. Vol.95. No.16 © 2005 - Ongoing JATIT & LLS ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195