Understanding of the Naive Bayes Classifier in Spam Filtering

Qijia Wei[a)]

International School of Kuala Lumpur, Kuala Lumpur 68000, Malaysia.

[a)]uniweiqijia@gmail.com

Abstract. Along with the development of the Internet, the information stream is experiencing an unprecedented burst. The methods of information transmission become more and more important and people receiving effective information is a hot topic in the both research and industry field. As one of the most common methods of information communication, email has its own advantages. However, spams always flood the inbox and automatic filtering is needed. This paper is going to discuss this issue from the perspective of Naive Bayes Classifier, which is one of the applications of Bayes Theorem. Concepts and process of Naive Bayes Classifier will be introduced, followed by two examples. Discussion with Machine Learning is made in the last section. Naive Bayes Classifier has been proved to be surprisingly effective, with the limitation of the interdependence among attributes which are usually email words or phrases.

INTRODUCTION

Probability is one of the major branches in mathematics. It is often used to solve many real-life situations. Mathematicians spent years building practical models to fit in the growing issues. Within this area, conditional probability represents the probability of an event happening in relation to the occurrence of another event. It assumes that no prior prediction can be made on the probability that an event is going to happen. This fundamental concept of conditional probability lies in the basis of Bayes Theorem. Bayes Theorem is a significant mathematical theorem first discovered by the British mathematician Reverend Thomas Bayes in 1763 [1]. Different from the traditional application of conditional probability which previous mathematicians used, Bayes Theorem allows mathematicians the ability to compute the unknown conditional probability of one pair of events given the known independent probability of each event and the reverse conditional probability of this pair of events.

This characteristic of the Bayes Theorem makes it the ideal choice for mathematicians and scientists to obtain the conditional probability that would otherwise remain unknown to them. Consequently, this theorem is utilized in many applications to categorize objects into different groups or to predict whether or not certain event is going to happen given what has already occurred, such as machine translation – language translation carried out by computers [2], text categorization – computers assigning texts into different categories [3], or risk management – prediction of financial risk of tasks using computer [4].

A spam is often characterized by its advertising nature or the fraud message it contains in order to deceive the users to acquire their confidential personal information. Users often identify such emails by the frequent appearance of specific phrases such as “Please enter your bank account number here” or “Congratulations, you’ve won 100,000$!”. These patterns are recognizable by computer algorithms and are used as decisive factors to judge the nature of an email. Through using such methodologies, even though the computers haven’t developed equal intelligence as humans yet at this point, the email servers can be almost as clever as human and save email users countless amount of time from being annoyed by filtering spams ourselves.

Behaving as a type of artificial intelligence, the spam filters are often computer programs designed by computer scientists, using a Bayesian analytical algorithm to generalize patterns in spams, and compute the probability that an email is actually a spam given these patterns. This algorithm performs Bayesian analysis to help email users to maintain a “healthy” inbox without being constantly annoyed by the spams or wasting their time on identifying which

6th International Conference on Computer-Aided Design, Manufacturing, Modeling and Simulation (CDMMS 2018)
AIP Conf. Proc. 1967, 020007-1–020007-7; https://doi.org/10.1063/1.5038979
Published by AIP Publishing. 978-0-7354-1672-7/$30.00

emails are spam and which are not. Although no prediction can be made on whether an email is spam or no under the assumptions of classical conditional probability, the application of Naive Bayes Classifier (NBC) here is capable of accomplishing this task. The final decision to classify this email into the "Spam" section or the "Inbox" section is made by comparing the computed probability with the threshold value. A value between 0 and 1 reflects how serious the email server believes the consequence of mistakenly classifying a non-spam email as spam is. The efficiency of this application is often measured by balancing between the percentage of "false positives", which in this example are spam emails that have been mistakenly classified as non-spam emails, and the percentage of "true positives", or spam emails that have been successfully classified into the "Spam" section [5].

This paper investigates the question of "How can Bayes Theorem be applied in Naive Bayes Classifier algorithm to classify spams?" The first part will reintroduce the concept of Bayes theorem, following by the NBC. Then, two examples will be illustrated. Discussions are made based on the machine learning algorithms to enhance the performance of NBC.

BAYES THEOREM

The simple version of the Bayes Theorem can be derived from basic probability concepts. To clarify, besides the aforementioned notations, $P(\mathrm{A} \cap \mathrm{B})$ represents the probability of intersection of A and B, or the probability that both event A and event A occur. The simple form of the original Bayes Theorem can be demonstrated as follow:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(B|A)P(A)}{P(B)}$$

One key condition the above proof must satisfy is that neither $P(\mathrm{A})$ nor $P(\mathrm{B})$ should be equal to zero in order for the above equation to exist and the proof to continue. This condition thus must be met for all circumstances when the Bayes Theorem is applied.

Besides, the alternative form of Bayes Theorem is generally encountered when looking at two competing statements or hypotheses:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')}$$

$P(A')$ is the corresponding probability of the initial degree of belief against A, where $P(A') = 1 - P(A)$.

For some partition $\{A_i\}$ of the sample space, the extended form of Bayes Theorem is:

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

NAIVE BAYES CLASSIFIER IN SPAM FILTERING

Naive Bayes Mode is one of the two most-used classification models. The basic concept of Naive Bayes Classifier (NBC) is applying Bayes Theorem where the objects or attributes have independence. The detailed process is show below.

a) There are two possible classes or categories, denoted by symbol A and A', to classify each email into in this application: spam and non-spam;

b) Vector denoted by $\vec{X} = < X_1,\ X_2,\ X_3,\ X_4, \ldots X_n >$ is used to represent a series of common attributes of spam emails. In this application, the attributes of emails are simply individual words or phrases;

c) Every email is represented by a vector denoted by $\vec{x} = < x_1, x_2, x_3,\ x_4, \ldots\ x_n >$, where each x_i represents the value of the attribute X_i. The email spam filter will first scan through each email searching for those specific words or phrases, and form the vector of each email using the presence or absence of the attributes it detected;

d) All attributes take the binary form in this application, meaning that each x_i is "1" if this word or phrase is present in an email, and "0" if this word or phrase is absent in an email;

e) Therefore, the vector representation of each email will look like an "ID number" consisting of a string of "1"s and "0"s, indicating which attributes are present in this particular email;

f) The string will be examined by NBC, comparing with the threshold value and resulting the final decision. The mathematic format of NBC (extended version) in the case is shown as follow:

$$P\left(C\middle|\vec{X}=\vec{x}\right)=\frac{P(\vec{X}=\vec{x}|C)\times P(C)}{P\left(\vec{X}=\vec{x}\middle|C\right)\times P(C)+P(\vec{X}=\vec{x}|C')\times P(C')}$$

where $P\left(\vec{X}=\vec{x}\middle|C\right)=\prod_{i=1}^{n}P(X_i=x_i|C)$.
Thus, the final version of the formula used to calculate the probability of an email being spam would be:

$$P\left(C\middle|\vec{X}=\vec{x}\right)=\frac{P(\vec{X}=\vec{x}|C)\times P(C)}{\prod_{i=1}^{n}P(X_i=x_i|C)\times P(C)+\prod_{i=1}^{n}P(X_i=x_i|C')\times P(C')}$$

After computing the probability, some further procedure need to be taken into consideration before making the final decision of classifying the email. Also, the criteria to evaluate the calculated probability and cautiously make NBC's decision to classify the email in order to minimize the "false positive" cases. One method to approach this issue is to compare the ratio of the probability of this email being spam to the probability of this email being non-spam, and set a threshold value. The equation of this method can be shown as follow, where α the selected threshold value is:

$$\frac{P(C|\vec{X}=\vec{x})}{P(C'|\vec{X}=\vec{x})}>\alpha$$

After formulating, the above inequality becomes:

$$P\left(C\middle|\vec{X}=\vec{x}\right)>t$$

where $t=\frac{\alpha}{1+\alpha}$.
The entire process of spam filter with NBC is show in the Figure 1.

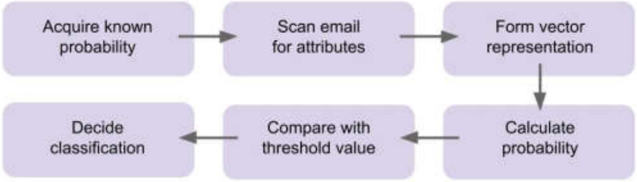
Acquire known probability
Scan email for attributes
Form vector representation
Calculate probability
Compare with threshold value
Decide classification

FIGURE 1. Flowchart of spam filtering process

EXAMPLES

The rest of this paper will be devoted to describe how do email servers employ Bayes Theorem and Bayesian analysis in their algorithms to produce the decision of whether to classify an email as spam or non-spam. All data used in the following process are constructed by the author of this essay and don't contain any aspects of realistic references to them.

Assumptions

Pertaining to the conditional independence assumption as a premise to use Bayes Theorem, all attributes are independent of each other in these examples, that the presence or absence of one word or phrase has no influence on the presence of absence of other words or phrases.

The default probability of an unknown email being a spam is 0.6, as shown in Table 1.

TABLE 1. Default probability of spams

Event	Probability
C	0.6
C'	0.4

All data used in the following process are constructed in this paper and don't contain any aspects of realistic references to them.

Three-attribute-filtering

As shown in Table 2, three phrases are used in this case. Table 3 shows the probability of whether each variable occurring in the email or not.

TABLE 2. Three attributes

Attribute	Phrase
X_1	You won $
X_2	Enter your bank account
X_3	Please donate

TABLE 3. Probability of each event (Three-attribute-filtering)

Event	Probability	
	given C	given C'
$X_1 = 1$	0.4	0.01
$X_2 = 1$	0.5	0,001
$X_3 = 1$	0.3	0.005

The above data only show the probability of an attribute being present given certain circumstances. Each can be easily substituted by 1 minus this probability in later calculations.

After evaluating the probability of each event, the anti-spam filtering algorithm performs classifying two sample emails. The vector representation of each email is shown in the Table 4.

TABLE 4. Vector representation of sample emails (Three-attribute-filtering)

Email	Vector representation
Email A	$\overrightarrow{x_A} = < 1, 0, 1 >$
Email B	$\overrightarrow{x_B} = < 1, 0, 0 >$

According to the formula above, the probability of Email A being a spam is:

$$P\left(C\middle|\vec{X} = \overrightarrow{x_A}\right) = \frac{0.6 \times (0.4 \times 0.5 \times 0.3)}{0.6 \times (0.4 \times 0.5 \times 0.3) + 0.4 \times (0.01 \times 0.999 \times 0.005)} = 0.999$$

With pre-setting the threshold value of 100, the value of t will be $t = \frac{100}{1+100} = 0.990$. Since 0.999 is greater than 0.990, the email spam filter will classify Email A as a spam.

In the same way, the probability of Email B being a spam is:

$$P\left(C\middle|\vec{X} = \overrightarrow{x_B}\right) = \frac{0.6 \times (0.4 \times 0.5 \times 0.7)}{0.6 \times (0.4 \times 0.5 \times 0.7) + 0.4 \times (0.01 \times 0.999 \times 0.995)} = 0.955$$

The computed probability of 0.955 is smaller than the t value 0.990. Thus, Email B will not be classified in the Spam section.

This example briefly illustrates the importance of the threshold value in an anti-spam filtering algorithm. Despite the high possibility of Email B being a spam, 95.5%, it is still classified as non-spam due to the extremely high t value, which is derived from a high threshold value of 100.

Ten-attribute-filtering

Compared to the three-attribute-filtering, this example will fit the practical issues more, since in real world applications, Naive Bayes Classifiers certainly identify way more than 3 attributes from spam emails. Usually, there will be more than thousands of attributes in one classifier loop.

In this example, the default probability of spams keeps same, as shown in the Table 1. Ten attributes are shown in the Table 5 and the input probabilities are shown in the Table 6.

TABLE 5. Ten attributes

Attribute	Phrase
X_1	Accept Credit Card
X_2	Apply Online
X_3	Call Now
X_4	Cash Bonus
X_5	Money Making
X_6	No Fee
X_7	Click Below
X_8	Earn Per Week
X_9	Enter Password
X_{10}	Save Up To

TABLE 6. Probability of each event (Ten-attribute-filtering)

Event	Probability	
	given C	given C'
$X_1 = 1$	0.6	0.02
$X_2 = 1$	0.3	0.1
$X_3 = 1$	0.4	0.05
$X_4 = 1$	0.1	0.01
$X_5 = 1$	0.5	0.01
$X_6 = 1$	0.7	0.3
$X_7 = 1$	0.9	0.5
$X_8 = 1$	0.2	0.05
$X_9 = 1$	0.4	0.01
$X_{10} = 1$	0.3	0.02

To exhibit the full operation of this more complex algorithm, five emails are used in this example, whose vector representations are shown in the Table 7.

TABLE 7. Vector representation of sample emails (Ten-attribute-filtering)

Email	Vector representation
Email A	$\overrightarrow{x_A} = < 1, 0, 1, 1, 1, 1, 1, 0, 0, 0 >$
Email B	$\overrightarrow{x_B} = < 0, 0, 1, 0, 1, 0, 0, 0, 0, 1 >$
Email C	$\overrightarrow{x_C} = < 0, 0, 0, 0, 1, 0, 0, 0, 0, 0 >$
Email D	$\overrightarrow{x_D} = < 1, 0, 0, 1, 0, 0, 0, 1, 1, 0 >$
Email E	$\overrightarrow{x_E} = < 0, 0, 0, 0, 0, 1, 0, 0, 1, 0 >$

The probability being a spam of each email is listed below:

$$P\left(C \middle| \vec{X} = \overrightarrow{x_A}\right) = \frac{0.6\times(0.6\times0.7\times0.4\times0.1\times0.5\times0.7\times0.9\times0.8\times0.6\times0.7)}{0.6\times(0.6\times0.7\times0.4\times0.1\times0.5\times0.7\times0.9\times0.8\times0.6\times0.7)+0.4\times(0.02\times0.9\times0.05\times0.01\times0.01\times0.3\times0.5\times0.95\times0.99\times0.98)} = 0.999$$

$$P\left(C\middle|\vec{X}=\overrightarrow{x_B}\right)=\frac{0.6\times(0.4\times0.7\times0.4\times0.9\times0.5\times0.3\times0.1\times0.8\times0.6\times0.3)}{0.6\times(0.4\times0.7\times0.4\times0.9\times0.5\times0.3\times0.1\times0.8\times0.6\times0.3)+0.4\times(0.98\times0.9\times0.05\times0.99\times0.01\times0.7\times0.5\times0.95\times0.99\times0.02)}=0.991$$

$$P\left(C\middle|\vec{X}=\overrightarrow{x_C}\right)=\frac{0.6\times(0.4\times0.7\times0.6\times0.9\times0.5\times0.3\times0.1\times0.8\times0.6\times0.7)}{0.6\times(0.4\times0.7\times0.6\times0.9\times0.5\times0.3\times0.1\times0.8\times0.6\times0.7)+0.4\times(0.98\times0.9\times0.95\times0.99\times0.01\times0.7\times0.5\times0.95\times0.99\times0.98)}=0.299$$

$$P\left(C\middle|\vec{X}=\overrightarrow{x_D}\right)=\frac{0.6\times(0.6\times0.7\times0.6\times0.1\times0.5\times0.3\times0.1\times0.2\times0.4\times0.7)}{0.6\times(0.6\times0.7\times0.6\times0.1\times0.5\times0.3\times0.1\times0.2\times0.4\times0.7)+0.4\times(0.02\times0.9\times0.95\times0.01\times0.99\times0.7\times0.5\times0.05\times0.01\times0.98)}=0.999$$

$$P\left(C\middle|\vec{X}=\overrightarrow{x_E}\right)=\frac{0.6\times(0.4\times0.7\times0.6\times0.9\times0.5\times0.7\times0.1\times0.8\times0.4\times0.7)}{0.6\times(0.4\times0.7\times0.6\times0.9\times0.5\times0.7\times0.1\times0.8\times0.4\times0.7)+0.4\times(0.98\times0.9\times0.95\times0.99\times0.99\times0.3\times0.5\times0.95\times0.01\times0.98)}=0.608$$

In this example, three different thresholds are used, 10, 100 and 1000. The corresponding t value of each α is:

$$t_1=\frac{10}{1+10}=0.909$$

$$t_2=\frac{100}{1+100}=0.990$$

$$t_3=\frac{1000}{1+1000}=0.999$$

Different threshold value results in different minimum probability required for an email to be classified as spam. The probability of Email A and Email D being spam is greater than all three t values, thus guaranteeing their classification into the "Spam" section. The probability of Email C and Email E being spam is lower than all three t values. Therefore, these two emails are definitely classified as non-spams. Nonetheless, different t values do make a difference on the classification results for Email B. Given the number of attributes present in all five emails and their final classification results, there are evidence to conclude that in this particular scenario, emails with 1 or 2 spam attributes are not classified as spam emails, whereas emails with 4 or more attributes present are certainly classified as spam.

DISCUSSION

The interdependent nature of these spams' attributes disagrees with the conditional independence assumption of this algorithm, which can sometimes cause mistakes or errors when classifying emails, further affecting the accuracy of Naive Bayes Classifiers. Furthermore, the value of core parameters may cause different results in NBC, such as the default probability of spams and threshold values. At the same time, corresponding issues comes up. To fix these issues and enhance the performance of spam filtering, Machine Leaning (ML) is used in resent researches and applications.

Higher accuracy

The first feature of ML is the supervision mechanism. The outputs can be used as the references to the inputs, which is called feedback. For example, the Back propagation algorithm is one of the most famous methods in image recognition [6]. In the spam filtering, ML could help with refining the prior parameters, to tune the filtering results.

Higher speed

In practical uses, thousands of attributes might be considered. Thus, a higher computing speed is necessary. Another advantage of ML is to guarantee the calculating time by using hierarchical structure. Also, it would be more useful when facing huge amount emails which need filtering. Two examples demonstrate the real-life uses of the algorithm. Furthermore, this paper has explored the critical role threshold value play in this algorithm, and how does the variance of this value affect the final classification results

CONCLUSION

To conclude, this paper has investigated the mathematical process of Naive Bayes Classifier in email servers, and how this algorithm applies the extended version of Bayes Theorem to compute the probability of each email being spam and further classify the emails into “Spam” section and “Inbox” section. Furthermore, this paper has explored the critical role threshold value play in this algorithm, and how does the variance of this value affect the final classification results.

Even though this algorithm is more sophisticated and effective in real life, it still has certain limitations, which accounts for the mistakes made by these anti-spam filters, as users can find a spam email in their “Inbox” sometimes while finding certain legitimate emails in the “Spam” section once in a while. One significant weakness of this algorithm is the conditional independence assumption, which is the premise to use Bayes Theorem. Often in reality, events, or attributes, are interrelated to and dependent on each other. Thus, the presence or absence of each word or phrase have an impact on the presence or absence of other words or phrases as words of relevant topics appear together, vice versa.

Looking towards future developments, with the ever-advancing Artificial Intelligence technology and the efforts of countless mathematicians, computer scientists, and researcher, the Naive Bayes Classifier will certain evolve and improve overtime to cater to the needs of email users. One possible direction is the better classifying criteria. Currently, the Naive Bayes Classifier simply computes the probability of an email being spam based solely on the presence and absence of words and phrases. However, as the field of AI makes more progress on the robots’ ability to understand semantic meanings of languages, better classification decisions can be made considering the more complex defining properties of spam emails as decided by the meanings of sentences. Therefore, a future with enhanced and user-friendlier spam email filters is in prospect.

REFERENCES

1. Peter, S., Werner, R., & Stephan, P. (2001). New methods and critical aspects in bayesian mathematics for 14c calibration. Radiocarbon, 43 (2A), 373-380.
2. Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., ... & Roossin, P. S. (1990). A statistical approach to machine translation. Computational linguistics, 16 (2), 79-85.
3. Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization (No. CMU-CS-96-118). Carnegie-mellon univ pittsburgh pa dept of computer science.
4. Cornalba, C., & Giudici, P. (2004). Statistical models for operational risk management. Physica A: Statistical Mechanics and its applications, 338 (1), 166-172.
5. Henderson, H. (2009). Encyclopedia of computer science and technology /. Reference Reviews, 67 (5), 1556-65.
6. Hijazi S, Kumar R, Rowen C. Using convolutional neural networks for image recognition [R]. Tech. Rep., 2015. [Online]. Available: http://ip. cadence. com/uploads/901/cnn-wp-pdf, 2015.