Random Variable And Stochastic Process

Soumyajit Chatterjee

AI22MTECH02005

Indian Institute of Technology, Hyderabad

April 30, 2022

Spam Email Filtering

Problem Statement

We get tons of messages proposing a lot of money, fantastic lottery wins, great presents and secrets of life. They may be harmful, just annoying or space-consuming, but they also can contain viruses or phishing attempts. In any case, it is not the content we want to deal with. So the demand for good spam filters is always high.

Theory

Naive Bayes classification is a simple probability algorithm based on the fact, that all features of the model are independent.

In the context of the spam filter, we suppose, that every word in the message is independent of all other words and we count them with the ignorance of the context.

The classification algorithm produces probabilities of the message to be spam or not spam by the condition of the current set of words. Calculation of the probability is based on the Bayes formula and the components of the formula are calculated based on the frequencies of the words in the whole set of messages.

Key Concept

- **Conditional Probability**
- **Independence**
- **Bayes Theorem**

Key Concept

Conditional Probability

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. It is denoted by P(A/B) and is formulated as:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Key Concept

Probability

$$\mathrm{P(X)} = \frac{\text{No. of events in favour of X}}{\text{Total number of all possible events}}$$

Independence

Two events are called independent if the probability of occurrence of one event does not depend on the probability of occurrence of another event. If A and B are two independent events with their probability of occurrence as P(A) and P(B) then:

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

Types of Emails

- **Spam:** Emails which contains malicious, spurious or questionable words in them. Usually these mails contain words like offer, sale, buy now, limited offer, limited trial, click here, click now etc. Emails like this usually contain promotion content, advertisements or are simply annoying.
- **Ham:** Ham emails are official, unofficial or informal emails which do not contains, advertisements or sales terms or offensive content. They do not make weird claims or promises or ask the user to send sensitive or private information. Anything that is not a spam can be considered as Ham email.

Bayes Theorem

Let our sample space be the set of emails, let S be the event that the mail is a spam, hence, $\overline{S}$ is the event that the mail is not a spam. Let W be the event that that a message contains a specific word W. Therefore,

$$p(S|W) = \frac{p(W|S).p(S)}{p(W|S).p(S) + p(E|\overline{S}).p(\overline{S})}$$

$p(W|S) =$ Given that it is already known that the mail is a spam, the probability of it having the spam word W.

$p(W|\overline{S}) =$ Given that it is already known that the mail is not a spam, the probability of it having the non spam word W.

Detection Method

We can detect whether an email is a spam or not by considering a single word present in the email like offer, sale etc. But a single word like that can be also present in an official mail or an informal mail due to some reasons. Therefore, classifying the entire mail based on a single word will often give incorrect results. Therefore, we consider at-least n words in our mail and modify the Bayes Theorem to include n words as-

$$p\left(S|\cap_{i=1}^{n} W_n\right) = \frac{\Pi_{i=1}^{n} p(W_i|S).p(S)}{\Pi_{i=1}^{n} p(W_i|S).p(S) + \Pi_{i=1}^{n} p(W_i|\overline{S}).p(\overline{S})}$$

Detection Method Contd.

Before trying to classify a mail as a spam or not spam by the Bayes classifier defined in the previous slide, we can use some observation to simplify our calculation even more. Whenever a new mail arrives we don't know whether it is a spam or not a spam. Therefore, the probability of it being a spam or not spam is equal. That is $p(S)$ and $p(\overline{S})$ are equal to 0.5. Therefore, the previous formula can be modified as-

$$p\left(S \middle| \cap_{i=1}^{n} W_n\right) = \frac{\Pi_{i=1}^{n} p(W_i|S)x0.5}{\Pi_{i=1}^{n} p(W_i|S)x0.5 + \Pi_{i=1}^{n} p(W_i|\overline{S})x0.5}$$

Which would be:

$$p\left(S \middle| \cap_{i=1}^{n} W_n\right) = \frac{\Pi_{i=1}^{n} p(W_i|S)}{\Pi_{i=1}^{n} p(W_i|S) + \Pi_{i=1}^{n} p(W_i|\overline{S})}$$

Training Method

The spam filter must be trained based on messages in their inbox mails to estimate probabilities. The program or user must define a threshold probability r such that if $p(S|\cap_{i=1}^{n} W_n) > r$ then the mail is considered a spam. If the Bayesian probability is less than r then it means it is not a spam. Even though the threshold value should be determined from the number of mails received as spam divided by the total number of mails. A common value would be 0.5.

Example

Suppose we have the following data-
The word **sale** occurs in 250 of 2000 spam mails and the same word occurs in only 5 out of 1000 non-spam mails.
The probability that the mail has word sale in it and it is a spam-

$$p(W|S) = 250/2000 = 0.125$$

The probability that the mail has word sale in it and it is not a spam-

$$p(W|\overline{S}) = 5/1000 = 0.005$$

Example Contd.

Since, we are assuming that it is equally likely for a new mail to be a spam or not a spam, we have the probability $p(S|W)$ for a new mail as the probability that this new mail is a spam given the word **sale** is present in it.

$$p(S|W) = \frac{p(W|S).p(S)}{p(W|S).p(S) + p(W|\overline{S}).p(\overline{S})}$$

$$p(S|W) = \frac{0.125x0.5}{0.125x0.5 + 0.005x0.5} = 0.962$$

If we set our threshold to 0.5, we can see that $p(S|W) > 0.5$ then based on this one word our mail is definitely a spam.

Example Contd.

Classifying the Email based on only 1 word can often lead to erroneous classification. Therefore, we need to take into account multiple words for a more reasonable classification. Now let us repeat the above example by considering two words **for** and **sale**.

Let the word **for** appear in 200 spam mails out of 2000 spam mails.
Let the word **sale** appear in 400 spam mails out of 2000 spam mails.

Let the word **for** appear in 60 non-spam mails out of 1000 non-spam mails.
Let the word **sale** appear in 25 non-spam mails out of 1000 non-spam mails.

Example Contd.

Probability that the mail is a spam and it has the word **for** in it-

$$p(W_1|S) = 200/2000 = 0.1$$

Probability that the mail is a spam and it has the word **sale** in it-

$$p(W_2|S) = 400/2000 = 0.2$$

Probability that the mail is not a spam and it has the word **for** in it-

$$p(W_1|\overline{S}) = 25/1000 = 0.025$$

Probability that the mail is not a spam and it has the word **sale** in it-

$$p(W_2|\overline{S}) = 60/1000 = 0.06$$

Example Contd.

Now from Bayes Theorem the probability that the email is a spam given the word **for** and **sale** are present in the email is-

$$p(S|(W_1 \cap W_2) = \frac{p(W_1|S).p(S).p(W_2|S).p(S)}{p(W_1|S).p(S).p(W_2|S).p(S) + p(W_1|\overline{S}).p(\overline{S}).p(W_2|\overline{S}).p(\overline{S}}$$

Example Contd.

We again assume that the probability that the new email can either be a spam or not a spam with equal probability of 0.5, therefore, given the two words the probability that the email is a spam is-

$$p(S|W_1 \cap W_2) = \frac{0.2x0.5x0.1x0.5}{0.2x0.5x0.1x0.5 + 0.06x0.5x0.025x0.5}$$

$$p(S|W_1 \cap W_2) = 0.930$$

If we take our threshold to be 0.5, we can observe that the probability given the word **for** and **sale** present in it to be a spam is 0.93 which means that this new mail containing these two words is definitely a spam.

Conclusion

Bayesian classifier is a very fast and robust method for classification but the main disadvantage of Bayesian classifier is that it considers each word independent and does not take into account the dependence which might be present between the words. Like in our example Bayes rule considers the two words **for** and **sale** separately but does not check what will happen if **for sale** occurs together in a sentence. Due to this reason it is called a Naive Bayes classifier.

Thank You