MSc - Data Mining

Topic 06 : Classification

Part 03 : Naive Bayes

Dr Bernard Butler and Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(bbutler@tssg.org; kmurphy@wit.ie)

Spring Semester, 2021

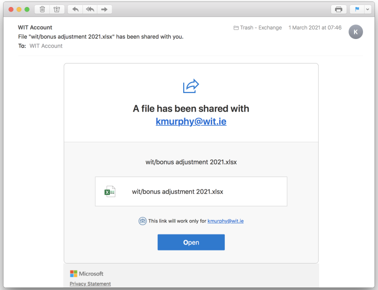
Outline

- Motivation
- Implementation

Outline

Spam Filtering

Reality



Simplified Problem

Assume that we have the following set of email previously classified as spam or ham.

Message	Class
"send us your password"	spam
"send us your review"	ham
"password review"	ham
"review us"	spam
"send your password"	spam
"send us your account"	spam

We are interested in classifying the following new email as spam or ham:

Message	Class
"review us now"	?

Count occurrences … compute probabilities

Message	Class
"send us your password"	spam
"send us your review"	ham
"password review"	ham
"review us"	spam
"send your password"	spam
"send us your account"	spam

Word	# in spam	# in ham	$\Pr(\bullet \mid \text{spam})$	$\Pr(\bullet \mid \text{ham})$
review	1	2	1/4	2/2
send	3	1	3/4	1/2
us	3	1	3/4	1/2
your	3	1	3/4	1/2
password	2	1	2/4	1/2
account	1	0	1/4	0/2

Class	Count	Probability
spam	4	$\Pr(\text{spam}) = 4/6$
ham	2	$\Pr(\text{ham}) = 2/6$

What we have

The probability that a message contains, say the word review, among our message classed as spam, i.e.,

$$\Pr(\text{review} \mid \text{spam}) = 1/4$$

What we want

The probability that a message is spam given that it contains, say the word review, i.e.,

$$\Pr(\text{spam} \mid \text{review}) = ?$$

Aside: Probability Laws

Bayes Rule

$$\overbrace{\Pr(A|B)\Pr(B)}^{\text{conditional}\times\text{marginal}} = \overbrace{\Pr(A \text{ AND } B)}^{\text{joint}} = \underbrace{\Pr(B|A)\Pr(A)}_{}^{\text{conditional}\times\text{marginal}}$$

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A) \longrightarrow \Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

Complementry Rule

$$\Pr(\text{NOT } A) = 1 - \Pr(A)$$

Independent Events

If events A and B are independent, then $\Pr(A \text{ AND } B) = \Pr(A)\Pr(B)$

Total law of probability

Let A be any event, and let $B_1, B_2, B_3, \ldots, B_n$ be a sequence of events, exactly one of which must occur, then

$$\Pr(A) = \Pr(A|B_1)\Pr(B_1) + \Pr(A|B_2)\Pr(B_2) + \cdots + \Pr(A|B_n)\Pr(B_n)$$

Classifying a Message Using a Single Word

Applying the total law of probability, with $A = \textsf{review}$, and $B_1 = \text{spam}$ and $B_2 = \text{ham}$, then we have

$$\Pr(\textsf{review}) = \Pr(\textsf{review}|\text{spam})\Pr(\text{spam}) + \Pr(\textsf{review}|\text{ham})\Pr(\text{ham}) = \frac{1}{4}\cdot\frac{4}{6} + \frac{2}{2}\cdot\frac{2}{6} = \frac{3}{6}$$

Now we can apply Bayes rule, with $A = \textsf{review}$ and $B = \text{spam}$ we have

$$\Pr(\text{spam}|\textsf{review}) = \frac{\Pr(\textsf{review}|\text{spam})\Pr(\text{spam})}{\Pr(\textsf{review})} = \frac{\frac{1}{4}\cdot\frac{4}{6}}{\frac{3}{6}} = \frac{1}{3} = 33.3\%$$

And we can apply Bayes rule, with $A = \textsf{review}$ and $B = \text{ham}$ to get

$$\Pr(\text{ham}|\textsf{review}) = \frac{\Pr(\textsf{review}|\text{ham})\Pr(\text{ham})}{\Pr(\textsf{review})} = \frac{\frac{2}{2}\cdot\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3} = 66.7\%$$

So on receiving a message containing the word review we can classify it a spam with probability 33.3% and ham with probability 66.7%.

Classifying a Message Using Multiple Words I

- We can now (hopefully) do this for each word, but what about classifying using multiple words? … here comes the naïve bit …

  Naïve Bayes assumes that presence of each word are independent events

  Recall: independent events means can multiply to get joint probabilities.
- We are interested in classifying the message

  review us now
- We don't have data on the word now so only looking at messages containing review and us and not containing send, your, password, or account.

$$\Pr(\{\text{review}, \text{us}\}|\text{spam}) = \underbrace{\left(\frac{1}{4}\right)}_{\text{review}} \underbrace{\left(1 - \frac{3}{4}\right)}_{\cancel{\text{send}}} \underbrace{\left(\frac{3}{4}\right)}_{\text{us}} \underbrace{\left(1 - \frac{3}{4}\right)}_{\cancel{\text{your}}} \underbrace{\left(1 - \frac{2}{4}\right)}_{\cancel{\text{password}}} \underbrace{\left(1 - \frac{1}{4}\right)}_{\cancel{\text{account}}} = 0.0044$$

and

$$\Pr(\{\text{review}, \text{us}\}|\text{ham}) = \underbrace{\left(\frac{2}{2}\right)}_{\text{review}} \underbrace{\left(1 - \frac{1}{2}\right)}_{\cancel{\text{send}}} \underbrace{\left(\frac{1}{2}\right)}_{\text{us}} \underbrace{\left(1 - \frac{1}{2}\right)}_{\cancel{\text{your}}} \underbrace{\left(1 - \frac{1}{2}\right)}_{\cancel{\text{password}}} \underbrace{\left(1 - \frac{0}{2}\right)}_{\cancel{\text{account}}} = 0.0625$$

Classifying a Message Using Multiple Words II

Now we can apply the total law of probability as before

$$\begin{aligned}\Pr(\{\textsf{review}, \textsf{us}\}) =& \Pr(\{\textsf{review}, \textsf{us}\}|\text{spam}) \Pr(\text{spam}) + \Pr(\{\textsf{review}, \textsf{us}\}|\text{ham}) \Pr(\text{ham}) \\ =& 0.0044 \left(\frac{4}{6}\right) + 0.0625 \left(\frac{2}{6}\right) = 0.0237\end{aligned}$$

And finally Bayes rule

$$\Pr(\text{spam}|\{\textsf{review}, \textsf{us}\}) = \frac{\Pr(\{\textsf{review}, \textsf{us}\}|\text{spam}) \Pr(\text{spam})}{\Pr(\{\textsf{review}, \textsf{us}\})} = \frac{0.0044 \left(\frac{4}{6}\right)}{0.0237} = 0.123$$

Hence, the probability that the message is spam is 12.3%, and ham is $1 = 0.123 = 0.877$ or 87.7%.

Naïve Bayes Classifier — Review

When to Consider

- Assumption of independence holds
- Categorical features.
- Spam filtering, Sentiment Analysis, and Recommendation Systems (with collaborative filtering).

Advantages

- It is easy and fast to predict class. It also perform well in multi-class prediction.
- When assumption of independence holds, performs better compare to other models like logistic regression and needs less training data.

Disadvantages

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is Laplace estimation.
- If continuous features, then assumes normality conditions — often too restrictive.
- Probability estimates via `predict_proba` are not that reliable.

Resources