Machine Learning & Data Mining

Spam Filtering using Naïve Bayes classification

Group members: Nguyễn Vũ Minh - 20194801
Lê Huy Hoàng - 20194766

Table of contents

01 Problem Domain

Problem Domain

???????

02 Algorithm

Algorithm

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

Bayes formula

$$P(Spam|Content) = \frac{P(Content|Spam) * P(Spam)}{P(Content)}$$

Algorithm (Cont.)

Compute & Compare

$$P(Spam|Content) = \frac{P(Content|Spam) * P(Spam)}{P(Content)}$$

&

$$P(Normal|Content) = \frac{P(Content|Normal) * P(Normal)}{P(Content)}$$

$$P(Spam|Content) = P(Content|Spam) * P(Spam)$$

&

$$P(Normal|Content) = P(Content|Normal) * P(Normal)$$

Algorithm Breakdown

For P(Spam):

$$P(Spam) = \frac{N_{Spam}}{N_{Spam} + N_{Ham}}$$

For P(Content | Spam):

Content = $w_1 w_2 w_3 \ldots w_n$

$$\Rightarrow \quad P(Content|Spam) = \prod_{i=1}^{n} P(w_i \mid Spam)$$

With

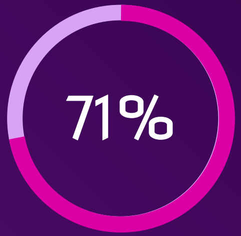
$$P(w_i \mid Spam) = \frac{N_{w_i|Spam}}{N_{Spam}}$$

Dataset

https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv

Ignore all stopwords from the dataset

Spam



Ham

Email No.	the	to	ect	and	for	of	a	you	hou	in	on	is	this	enron	i	be	that	will	have	with	your	at	we	s	are	it	by	com	as
Email 1	0	0	1	0	0	0	2	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	
Email 2	8	13	24	6	6	2	102	1	27	18	21	13	0	1	61	4	2	0	0	2	0	12	9	95	4	3	3	3	
Email 3	0	0	1	0	0	0	8	0	0	4	2	0	0	0	8	0	0	0	0	0	0	2	0	2	0	0	0	0	
Email 4	0	5	22	0	5	1	51	2	10	1	5	9	2	0	16	2	0	0	1	1	0	2	1	36	3	1	2	0	
Email 5	7	6	17	1	5	2	57	0	9	3	12	2	2	0	30	8	0	0	2	0	0	7	0	19	2	4	2	0	
Email 6	4	5	1	4	2	3	45	1	0	16	12	8	1	0	52	2	0	0	0	1	0	5	5	56	2	7	1	1	
Email 7	5	3	1	3	2	1	37	0	0	9	4	6	2	0	27	1	0	0	0	0	0	7	1	40	0	0	0	0	
Email 8	0	2	2	3	1	2	21	6	0	2	6	2	0	0	28	1	0	1	0	0	5	1	0	23	0	1	0	0	
Email 9	2	2	3	0	0	1	18	0	0	3	3	2	1	0	15	0	1	0	0	0	0	3	2	6	0	0	0	0	
Email 10	4	4	35	0	1	0	49	1	16	9	4	1	0	0	35	10	0	2	1	1	0	3	1	37	0	1	1	0	
Email 11	22	14	2	9	2	2	104	0	2	35	13	21	9	0	96	6	8	2	2	3	0	27	4	76	2	13	0	5	
Email 12	33	28	27	11	10	12	173	6	12	28	47	27	7	4	160	11	1	6	1	3	3	18	4	145	3	21	1	3	
Email 13	27	17	3	7	5	8	106	3	0	22	33	16	5	0	102	7	0	6	1	3	2	11	1	91	1	10	1	2	
Email 14	4	5	7	1	5	1	37	1	3	8	8	6	1	0	43	1	0	1	0	4	0	2	4	46	0	5	1	0	

Difficulties

Since we have $$P(w_i \mid Spam) = \frac{N_{w_i|Spam}}{N_{Spam}}$$

If w_i never appear in a spam email in the dataset:

$P(w_i \mid Spam) = 0$ → $P(Content|Spam) = 0$ → Faulty result

Solution:

$$P(w_i \mid Spam) = \frac{N_{w_i|Spam} + \alpha}{N_{Spam} + \alpha * N_{Vocabulary}}$$

Basically we count the world from α instead of 0

Difficulties (Cont.)

$P(Content|Spam) = \prod_{i=1}^{n} P(w_i \mid Spam)$ → Number very small → Floating-point underflow

Solution: since if a > b then log(a) > log(b)

we calculate and compare log(P(Spam|Content)) and log(P(Ham|Content)) instead

$$log(P(Spam|Content)) = log(P(Spam)) + \sum_{i=1}^{n} \log(P(w_i \mid Spam))$$

&

$$log(P(Ham|Content)) = log(P(Ham)) + \sum_{i=1}^{n} \log(P(w_i \mid Ham))$$

03 Results

Results

Methods: k-fold cross-validation with stratified sampling

Spam		Classified by the system	
		Spam	Normal
True class	Spam	1417	83
	Normal	214	3458

Normal		Classified by the system	
		Normal	Spam
True class	Normal	3458	214
	Spam	83	1417

$$Precision\ (Spam) = \frac{1417}{1417 + 214} = 86.88\ \%$$

$$Recall\ (Spam) = \frac{1417}{1417 + 83} = 94.47\ \%$$

$$Precision\ (Normal) = \frac{3458}{3458 + 83} = 97.66\ \%$$

$$Recall\ (Normal) = \frac{3458}{3458 + 214} = 94.17\ \%$$

$$=> \ Precision\ (Macro) = \frac{86.88 + 97.66}{2} = 92.27\ \%$$

$$=> \ Recall\ (Macro) = \frac{94.47 + 94.17}{2} = 93.32\ \%$$

$$=> F1 = \frac{2 * 92.27 * 93.32}{92.27 + 94.32} = 92.79\ \%$$

04 Conclusion

Conclusion

- The AI model achieve a pretty good result although false positive is a problem
- Naïve Bayes classification might have a problem of ignore the order of words in an email, which might be crucial to detect if a mail is spam or not

THANK YOU FOR LISTENING!