CMSC 170: Introduction to Artificial intelligence

Week 05: Augmenting the Spam Filter with Laplace Smoothing

John O-Neil V. Geronimo
Institute of Computer Science
University of the Philippines Los Baños

Content

I. Review on Naïve Bayes Spam Filter
II. Augmenting the Spam Filter with Laplace Smoothing

Content

I. Review on Naïve Bayes Spam Filter
II. Augmenting the Spam Filter with Laplace Smoothing

Background

Naïve Bayes Theorem

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

Spam Filtering Using Naïve Bayes

$$P(Spam|message) = \frac{P(message|Spam)P(Spam)}{P(message)}$$

Formula:

$$P(Spam|message) = \frac{P(message|Spam)P(Spam)}{P(message)}$$

- $P(Spam) = \frac{count\ (Spam)}{count(Spam \cup Ham)}$
- $P(Ham) = 1 - P(Spam)$
- $P(message|Spam) = P(w_0|Spam) \dots P(w_n|Spam)$
- $P(w_n|Spam) = \frac{count(w\ in\ Spam)}{count(total\ no\ of\ words\ in\ Spam)}$
- $P(message|Ham) = P(w_0|Ham) \dots P(w_n|Ham)$
- $P(w_n|Ham) = \frac{count(w\ in\ Ham)}{count(total\ no\ of\ words\ in\ Ham)}$
- $P(message) = P(message|Spam)P(Spam) + P(message|Ham)P(Ham)$

Putting it all together,

$$P(S|m) = \frac{P(w_0|S)P(w_1|S) \dots P(w_n|S)P(S)}{P(w_0|S)P(w_1|S) \dots P(w_n|S)P(S) + P(w_0|H)P(w_1|H) \dots P(w_n|H)P(H)}$$

where m = message, S = Spam, and H = Ham

Spam Filter

Spam Dataset

- You will love this product!
- I used this cream and it removed my acne!!!
- Buy now!

Ham Dataset

- Lucky I am in love with my bestfriend
- You are still the one that I love
- Could I love you anymore?

Spam Filter

Spam Bag-of-Words

Index	Word	Frequency
0	you	1
1	will	1
2	love	1
3	this	2
4	product	1
5	i	1
6	used	1
7	cream	1

Index	Word	Frequency
8	and	1
9	it	1
10	removed	1
11	my	1
12	acne	1
13	buy	1
14	now	1

TNOW:

16

DS:

15

Spam Filter

Ham Bag-of-Words

Index	Word	Frequency
0	lucky	1
1	i	3
2	am	1
3	in	1
4	love	3
5	with	1
6	my	1
7	bestfriend	1

Index	Word	Frequency
8	you	2
9	are	1
10	still	1
11	the	1
12	one	1
13	that	1
14	could	1
15	anymore	1

TNOW:

21

DS:

16

Spam Filter

Is this message ham or spam?

I love you.

Spam Filter

Compute for P(Spam|message):

$$P(Spam|message) = \frac{P(message|Spam)P(Spam)}{P(message)}$$

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

$$= \frac{\frac{1}{4096}(0.5)}{\frac{1}{4096}(0.5) + \frac{2}{1029}(0.5)} = 0.1115931027 = 11.16\%$$

Spam Filter

Since P(Spam|message) = 11.16%

The message “I love you.”
is classified as a **ham** message!

Spam Filter

Is this message ham or spam?

I hate you.

Spam Filter

Spam Bag-of-Words

Index	Word	Frequency
0	you	1
1	will	1
2	love	1
3	this	2
4	product	1
5	i	1
6	used	1
7	cream	1

Index	Word	Frequency
8	and	1
9	it	1
10	removed	1
11	my	1
12	acne	1
13	buy	1
14	now	1

TNOW:

16

DS:

15

Spam Filter

Compute for P(message|Spam):

$$P(\text{"i"}|Spam) = \frac{count(\text{"i" } in\ Spam)}{count(total\ no\ of\ words\ in\ Spam)} = \frac{1}{16}$$

$$P(\text{"hate"}|Spam) = \frac{count(\text{"hate" } in\ Spam)}{count(total\ no\ of\ words\ in\ Spam)} = \boxed{\frac{0}{16}}$$

$$P(\text{"you"}|Spam) = \frac{count(\text{"you" } in\ Spam)}{count(total\ no\ of\ words\ in\ Spam)} = \frac{1}{16}$$

Spam Filter

Compute for P(message|Spam):

$$P(message|Spam) = P(w_0|Spam)P(w_1|Spam) \dots P(w_n|Spam)$$

$$= P("i"|Spam)P("love"|Spam)P("you"|Spam)$$

$$= \frac{1}{16} * \frac{0}{16} * \frac{1}{16}$$

$$= \mathbf{0}$$

Spam Filter

Ham Bag-of-Words

Index	Word	Frequency
0	lucky	1
1	i	3
2	am	1
3	in	1
4	love	3
5	with	1
6	my	1
7	bestfriend	1

Index	Word	Frequency
8	you	2
9	are	1
10	still	1
11	the	1
12	one	1
13	that	1
14	could	1
15	anymore	1

TNOW:

21

DS:

16

Spam Filter

Compute for P(message|Spam):

$$P(\text{"i"}|Ham) = \frac{count(\text{"i" } in\ Ham)}{count(total\ no\ of\ words\ in\ Ham)} = \frac{3}{21}$$

$$P(\text{"hate"}|Ham) = \frac{count(\text{"hate" } in\ Ham)}{count(total\ no\ of\ words\ in\ Ham)} = \boxed{\frac{0}{21}}$$

$$P(\text{"you"}|Ham) = \frac{count(\text{"you" } in\ Ham)}{count(total\ no\ of\ words\ in\ Ham)} = \frac{2}{21}$$

Spam Filter

Compute for P(message|Ham):

$$P(message|Ham) = P(w_0|Ham)P(w_1|Ham) \dots P(w_n|Ham)$$

$$= P("i"|Ham)P("hate"|Ham)P("you"|Ham)$$

$$= \frac{3}{21} * \frac{0}{21} * \frac{2}{21}$$

$$= 0$$

Spam Filter

Compute for P(Spam|message):

$$P(Spam|message) = \frac{P(message|Spam)P(Spam)}{P(message)}$$

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

$$= \frac{0\ (0.5)}{0(0.5) + 0\ (0.5)} = \frac{0}{0} = undefined$$

overfitting

Content

I. Review on Naïve Bayes Spam Filter
II. Augmenting the Spam Filter with Laplace Smoothing

Updated Formula

Before

$$P(Spam) = \frac{count(Spam)}{count(Spam \cup Ham)}$$

$$P(Ham) = \frac{count(Ham)}{count(Spam \cup Ham)}$$

$$= 1 - P(Spam)$$

Now

$$P(Spam) = \frac{count(Spam) + k}{count(Spam \cup Ham) + 2k}$$

$$P(Ham) = \frac{count(Ham) + k}{count(Spam \cup Ham) + 2k}$$

$$= 1 - P(Spam)$$

Updated Formula

Before

$$P(w_n|Spam) = \frac{count(w\ in\ Spam)}{count(total\ no\ of\ words\ in\ Spam)}$$

Now

$$P(w_n|Spam) = \frac{count(w\ in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + (k * (dSize + count(new\ words)))}$$

Updated Formula

$$P(w_n|Spam) = \frac{count(w\ in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + (k * (dSize + count(new\ words)))}$$

dSize (or *dictionary size*) is the number of unique words found on both Spam and Ham dataset

new words refers to the words present in the message to be classified but does not exist in the Ham and Spam dictionary

Spam Filter

Spam	Ham
Secret link offers	Link for ticket offers
Click link win ticket offers	Click link for win
	Reply link for ticket
Secret link offers for secret win	Win ticket for secret place

Spam Filter

Is this message ham or spam?

Secret place haven

Consider **k = 2**

Spam Filter

Spam Bag-of-Words

Index	Word	Frequency
0	secret	3
1	link	3
2	offers	3
3	click	1
4	win	2
5	ticket	1
6	for	1

TNOW:

14

DS:

7

Spam Filter

Ham Bag-of-Words

Index	Word	Frequency
0	link	3
1	for	4
2	ticket	3
3	offers	1
4	click	1

Index	Word	Frequency
5	win	2
6	reply	1
7	secret	1
8	place	1

TNOW:

17

DS:

9

Spam Filter

Compute for P(Spam):

Since we have three emails in the spam dataset and a total of seven messages for both spam and ham dataset

$$P(Spam) = \frac{count(Spam) + k}{count(Spam \cup Ham) + 2k}$$

$$P(Spam) = \frac{3 + 2}{7 + 2(2)} = \frac{5}{11}$$

Spam Filter

Compute for P(Ham):

Since we have four emails in the spam dataset and a total of seven messages for both spam and ham dataset

$$P(Ham) = \frac{count(Ham) + k}{count(Spam \cup Ham) + 2k}$$

$$P(Ham) = \frac{4 + 2}{7 + 2(2)} = \frac{6}{11}$$

Spam Filter

Compute for P(message|Spam):

$$P("secret"|Spam) = \frac{count("\text{secret}" in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + (k * (dS + nW))}$$

$$P("place"|Spam) = \frac{count("\text{place}" in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + (k * (dS + nW))}$$

$$P("haven"|Spam) = \frac{count("\text{haven}" in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + (k * (dS + nW))}$$

Updated Formula

$$P(w_n|Spam) = \frac{count(w\ in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + (k * (dSize + count(new\ words)))}$$

dSize (or *dictionary size*) is the number of unique words found on both Spam and Ham dataset

new words refers to the words present in the message to be classified but does not exist in the Ham and Spam dictionary

Spam Filter

Spam and Ham dictionary

Unique Words	
secret	for
link	win
offers	reply
click	place
ticket	

Dictionary Size: 9

Message to classify:
secret place haven

Count(new words): 1

Spam Filter

Spam Bag-of-Words

Index	Word	Frequency
0	secret	3
1	link	3
2	offers	3
3	click	1
4	win	2
5	ticket	1
6	for	1

TNOW:

14

DS:

7

Spam Filter

Compute for P(message|Spam):

$$P(\text{secret}|Spam) = \frac{count(\text{"secret"} in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + \left(k * (dS + nW)\right)} = \frac{3 + 2}{14 + (2 * (9 + 1))} = \frac{5}{34}$$

$$P(\text{place}|Spam) = \frac{count(\text{"place"} in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + \left(k * (dS + nW)\right)} = \frac{0 + 2}{14 + (2 * (9 + 1))} = \frac{2}{34}$$

$$P(\text{haven}|Spam) = \frac{count(\text{"haven"} in\ Spam) + k}{count(total\ no\ of\ words\ in\ Spam) + \left(k * (dS + nW)\right)} = \frac{0 + 2}{14 + (2 * (9 + 1))} = \frac{2}{34}$$

Spam Filter

Ham Bag-of-Words

Index	Word	Frequency
0	link	3
1	for	4
2	ticket	3
3	offers	1
4	click	1

Index	Word	Frequency
5	win	2
6	reply	1
7	secret	1
8	place	1

TNOW:

17

DS:

9

Spam Filter

Compute for P(message|Ham):

$$P(\text{secret}|Ham) = \frac{count(\text{"secret"} in\ Ham) + k}{count(total\ no\ of\ words\ in\ Ham) + \left(k * (dS + nW)\right)} = \frac{1+2}{17 + (2 * (9+1))} = \frac{3}{37}$$

$$P(\text{place}|Ham) = \frac{count(\text{"place"} in\ Ham) + k}{count(total\ no\ of\ words\ in\ Ham) + \left(k * (dS + nW)\right)} = \frac{1+2}{17 + (2 * (9+1))} = \frac{3}{37}$$

$$P(\text{haven}|Ham) = \frac{count(\text{"haven"} in\ Ham) + k}{count(total\ no\ of\ words\ in\ Ham) + \left(k * (dS + nW)\right)} = \frac{0+2}{17 + (2 * (9+1))} = \frac{2}{37}$$

Spam Filter

Compute for P(message|Ham) and P(message|Spam):

$$P(message|Spam)$$
$$= P(w_0|Spam)P(w_1|Spam) \dots P(w_n|Spam)$$
$$= \frac{5}{34} * \frac{2}{34} * \frac{2}{34}$$
$$= \frac{\mathbf{5}}{\mathbf{9826}}$$

$$P(message|Ham)$$
$$= P(w_0|Ham)P(w_1|Ham) \dots P(w_n|Ham)$$
$$= \frac{3}{37} * \frac{3}{37} * \frac{2}{37}$$
$$= \frac{\mathbf{12}}{\mathbf{50653}}$$

Spam Filter

Too small decimal values?

Use log probabilities for
Naïve Bayes

Spam Filter

Too small decimal values?

Use log probabilities for
Naïve Bayes

Spam Filter

Compute for Total Spam = P(message|Spam)P(Spam):

$$P(message|Spam) = P(w_0|Spam)P(w_1|Spam) \dots P(w_n|Spam)\ P(Spam)$$

$$= \ln\left(\frac{5}{34} * \frac{2}{34} * \frac{2}{34} * \frac{5}{11}\right)$$

$$= \mathbf{-8.371806661}$$

Spam Filter

Compute for Total Ham = P(message|Ham)P(Ham):

$$P(message|Ham) = P(w_0|Ham)P(w_1|Ham) \dots P(w_n|Ham)\ P(Ham)$$

$$= \ln\left(\frac{3}{37} * \frac{3}{37} * \frac{2}{37} * \frac{6}{11}\right)$$

$$= \mathbf{-8.548517784}$$

Spam Filter

Compute for P(Spam|message):

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

$$= \frac{\exp(total\ spam)}{\exp(total\ spam)+\exp(total\ ham)}$$

$$= \frac{\exp(-8.371806661)}{\exp(-8.371806661) + \exp(-8.548517784\ )}$$

$$= 0.5440631776 = \mathbf{54.41\%}$$

Spam Filter

Since P(Spam|message) = 54.41%

The message “secret place haven”
is classified as a **spam** message!

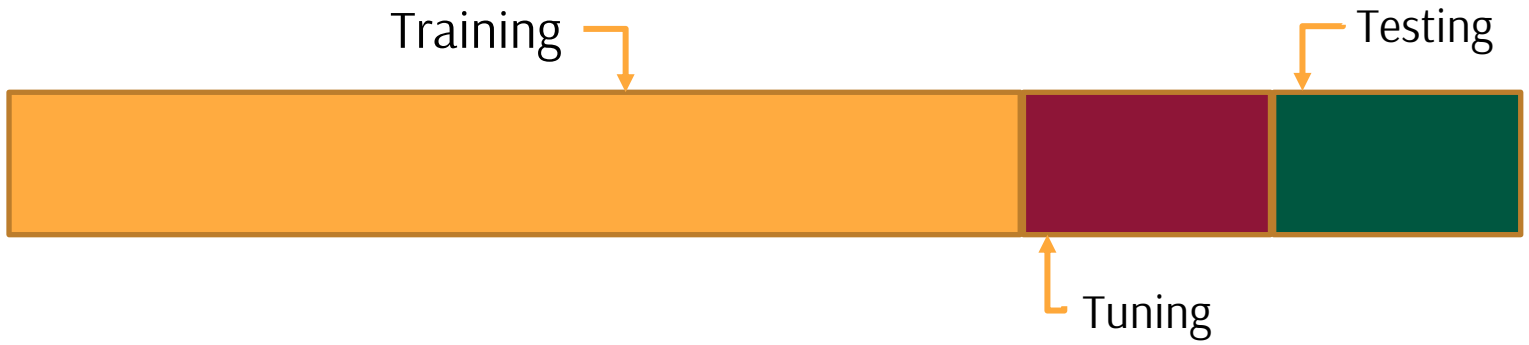
Content

I. Review on Naïve Bayes Spam Filter
II. Augmenting the Spam Filter with Laplace Smoothing

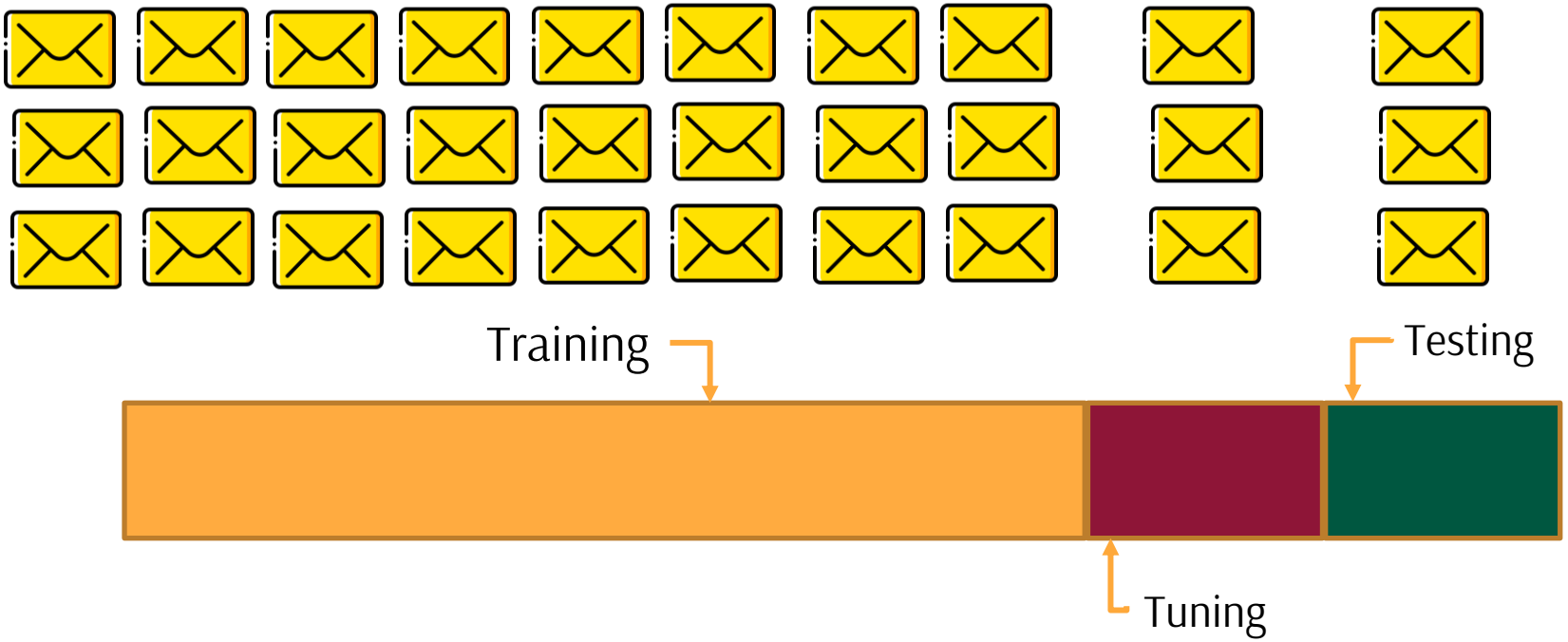
Exercise

Cross validation

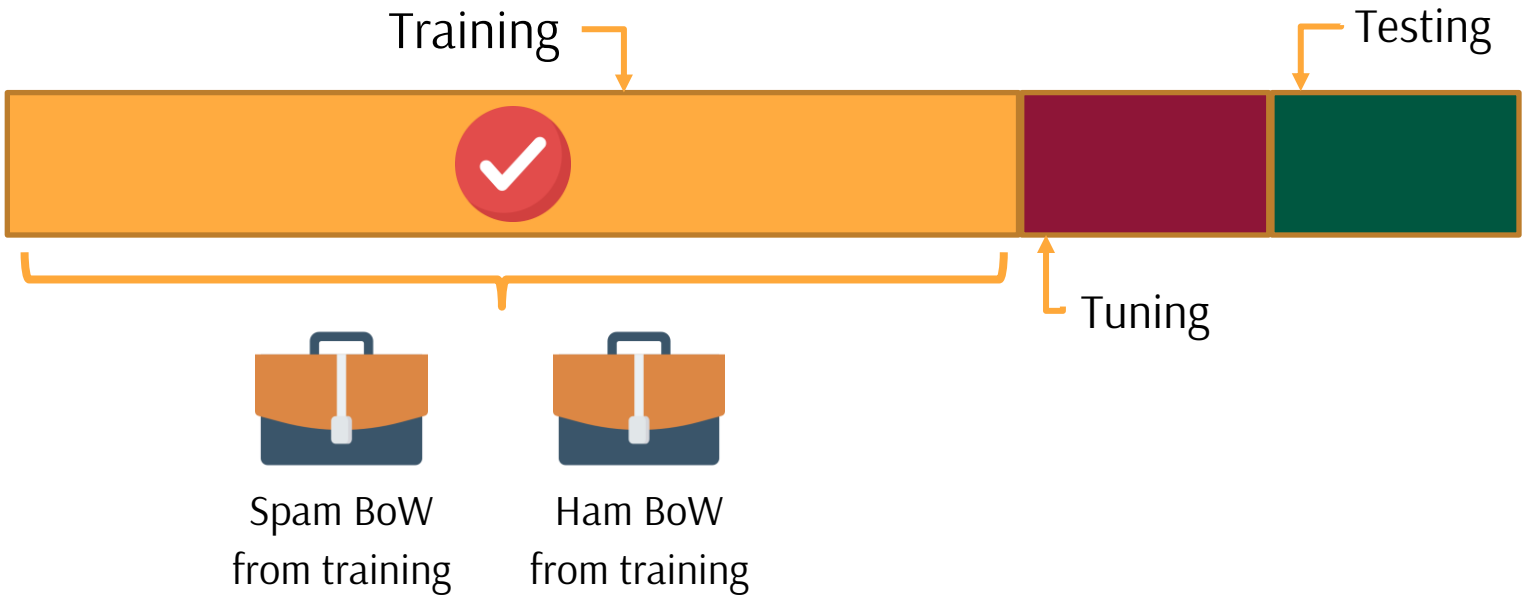
Spam Filter



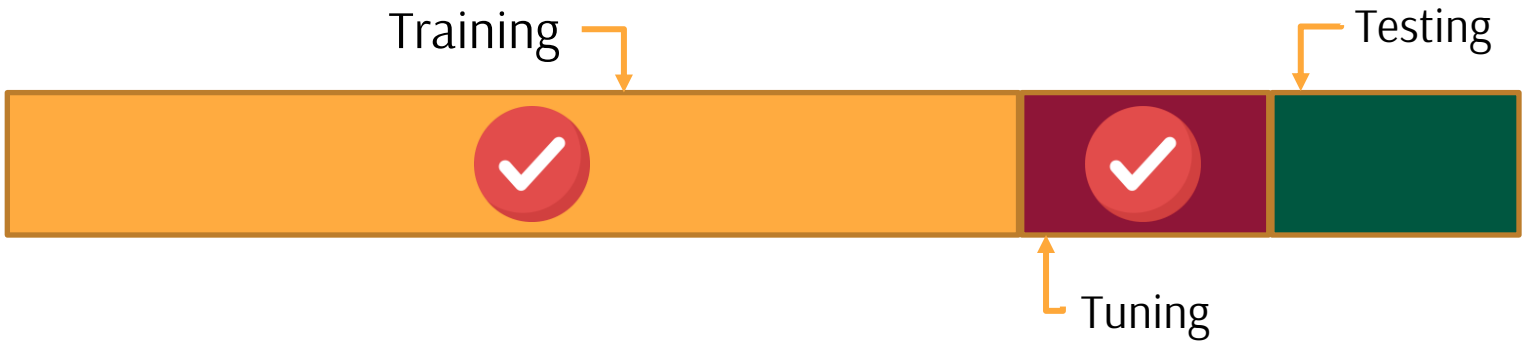
Spam Filter



Spam Filter


Training
Testing
Tuning
Spam BoW
from training
Ham BoW
from training

Spam Filter

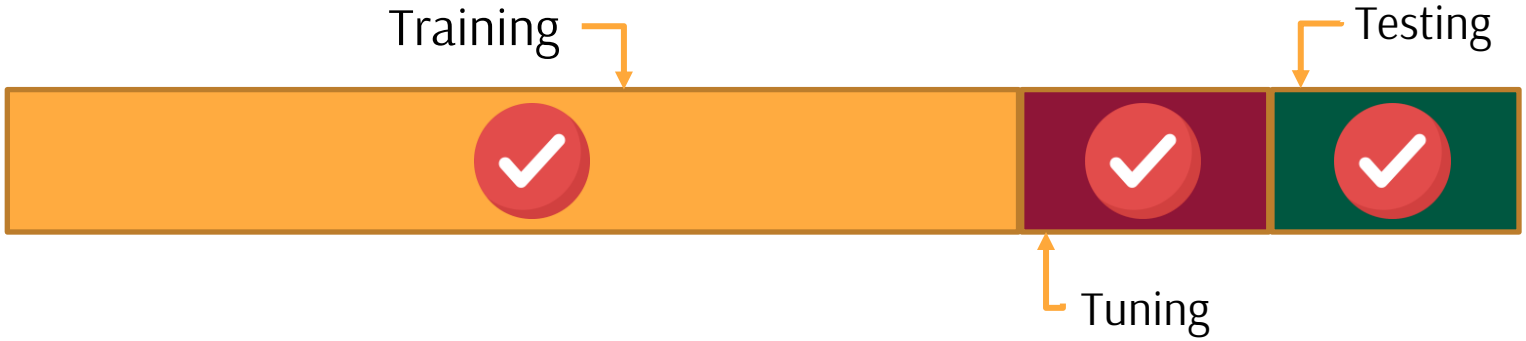


k =

Spam BoW from training

Ham BoW from training

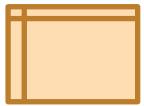
Spam Filter



best **k value** from tuning

Spam BoW
from training

Ham BoW
from training



Confusion matrix


UNIVERSITY OF THE PHILIPPINES
1908

INSTITUTE OF COMPUTER SCIENCE UPLB
1995

Keep safe!