The CURSE of anti-spam testing

Martijn Grooten

Virus Bulletin

25 September 2009

What are we talking about?

What are we talking about?

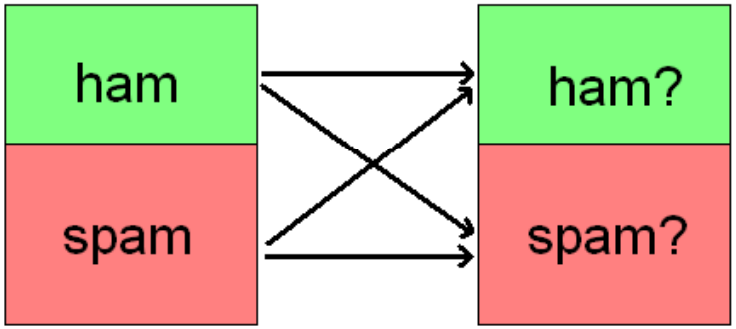
- Spam = bad; ham = good

What are we talking about?

- Spam = bad; ham = good
- A spam filter is a *binary classifier*

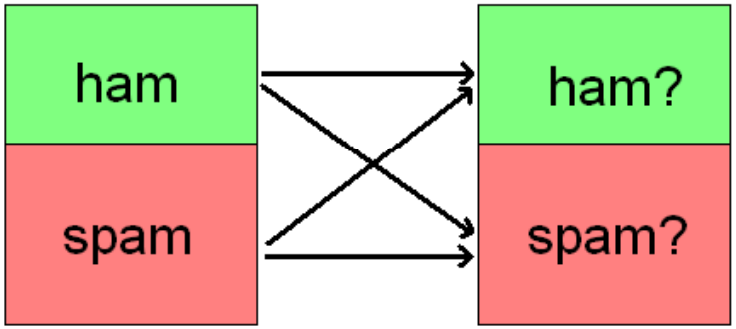
What are we talking about?

- Spam = bad; ham = good
- A spam filter is a *binary classifier*



What are we talking about?

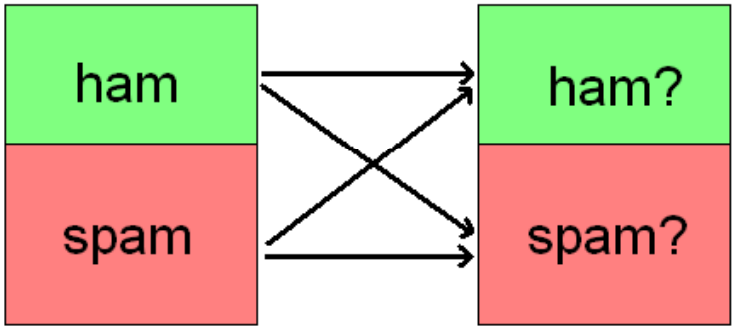
- Spam = bad; ham = good
- A spam filter is a *binary classifier*



- Spam catch rate = % spam caught

What are we talking about?

- Spam = bad; ham = good
- A spam filter is a *binary classifier*



- Spam catch rate = % spam caught
- False positive (rate) = (%) blocked ham

The CURSE of anti-spam testing

The CURSE of anti-spam testing

Comparative

The CURSE of anti-spam testing

Comparative

Unbiased

The CURSE of anti-spam testing

Comparative

Unbiased

Real email in real-time

The CURSE of anti-spam testing

Comparative

Unbiased

Real email in real-time

Statistically relevant

The CURSE of anti-spam testing

Comparative

Unbiased

Real email in real-time

Statistically relevant

Openness

The CURSE of anti-spam testing

Comparative

Unbiased

Real email in real-time

Statistically relevant

Explain what is done

99% spam catch rate?

99% spam catch rate?

This filter blocks 99 out of a 100 incoming spam emails

99% spam catch rate?

This filter blocks *an average of* 99 out of a 100 incoming spam emails

99% spam catch rate?

This filter blocks an average of 99 out of a 100 incoming spam emails *under the circumstances of the test*

99% spam catch rate?

This filter blocks an average of 99 out of a 100 incoming spam emails under the circumstances of the test

But what does it *really* mean?

99% spam catch rate

spam in mail stream	% spam in inbox
50%	

99% spam catch rate

spam in mail stream	% spam in inbox
50%	1%

99% spam catch rate

spam in mail stream	% spam in inbox
50%	1%
90%	

99% spam catch rate

spam in mail stream	% spam in inbox
50%	1%
90%	8%

99% spam catch rate

spam in mail stream	% spam in inbox
50%	1%
90%	8%
99.9%	

99% spam catch rate

spam in mail stream	% spam in inbox
50%	1%
90%	8%
99.9%	91%

99% spam catch rate

spam in mail stream	% spam in inbox	with 97% SC rate
50%	1%	
90%	8%	
99.9%	91%	

99% spam catch rate

spam in mail stream	% spam in inbox	with 97% SC rate
50%	1%	3%
90%	8%	21%
99.9%	91%	97%

But now, please stop forgetting about false positives!

Comparative testing

Comparative testing

- What is one email?

Comparative testing

- What is one email?
- How is the test set up?

Comparative testing

- What is one email?
- How is the test set up?
- What spam is being used? What ham?

But also…

But also…

- Spam changes over time

But also…

- Spam changes over time
- Aardvarks get more spam than zebras

But also…

- Spam changes over time
- Aardvarks get more spam than zebras

There is a whole lot we don't know about spammers!

Unbiased testing

Unbiased testing

- No bias towards any product

Unbiased testing

- No bias towards any product
- No bias towards any anti-spam technology

Unbiased testing

- No bias towards any product
- No bias towards any anti-spam technology

(But what about Bayesian filters?)

Unbiased testing

- No bias towards any product
- No bias towards any anti-spam technology

(But what about Bayesian filters?)

(And what about content versus context scanning?)

Unbiased testing

- No bias towards any product
- No bias towards any anti-spam technology

(But what about Bayesian filters?)

(And what about content versus context scanning?)

(And let’s not even mention greylisting…)

Real email in real-time

Real email in real-time

- Using fixed corpora is really something of the past

Real email in real-time

- Using fixed corpora is really something of the past
- Spam changes very fast

Real email in real-time

- Using fixed corpora is really something of the past
- Spam changes very fast
- If you can automatically generate ‘ham’, then so can spammers

Real email in real-time

- Using fixed corpora is really something of the past
- Spam changes very fast
- If you can automatically generate ‘ham’, then so can spammers
- (and cheaters…)

Statistically relevant

Statistically relevant

- An anti-spam test is a statistical test

Statistically relevant

- An anti-spam test is a statistical test
- A significant difference is not necessarily relevant

Statistically relevant

- An anti-spam test is a statistical test
- A significant difference is not necessarily relevant
- What is a representative spam sample?

Statistically relevant

- An anti-spam test is a statistical test
- A significant difference is not necessarily relevant
- What is a representative spam sample? More research needed!

Statistically relevant

- An anti-spam test is a statistical test
- A significant difference is not necessarily relevant
- What is a representative spam sample? More research needed!
- About what is representative ham sample too!

Statistically relevant

- An anti-spam test is a statistical test
- A significant difference is not necessarily relevant
- What is a representative spam sample? More research needed!
- About what is representative ham sample too!
- 1 or 2 FPs never makes a significant difference

Statistically relevant

- An anti-spam test is a statistical test
- A significant difference is not necessarily relevant
- What is a representative spam sample? More research needed!
- About what is representative ham sample too!
- 1 or 2 FPs never makes a significant difference; 100 and 200 always does

Explain what you’re doing

Explain what you’re doing

- An anti-spam test is unlikely to be perfect

Explain what you’re doing

- An anti-spam test is unlikely to be perfect; don’t pretend it is!

Explain what you’re doing

- An anti-spam test is unlikely to be perfect; don’t pretend it is!
- Be open about how it might not be completely comparative

Explain what you're doing

- An anti-spam test is unlikely to be perfect; don't pretend it is!
- Be open about how it might not be completely comparative, unbiased

Explain what you're doing

- An anti-spam test is unlikely to be perfect; don't pretend it is!
- Be open about how it might not be completely comparative, unbiased, make use of real email (in real-time)

Explain what you're doing

- An anti-spam test is unlikely to be perfect; don't pretend it is!
- Be open about how it might not be completely comparative, unbiased, make use of real email (in real-time) and is statistically relevant

Explain what you're doing

- An anti-spam test is unlikely to be perfect; don't pretend it is!
- Be open about how it might not be completely comparative, unbiased, make use of real email (in real-time) and is statistically relevant
- Only then will your results be meaningful

Explain what you're doing

- An anti-spam test is unlikely to be perfect; don't pretend it is!
- Be open about how it might not be completely comparative, unbiased, make use of real email (in real-time) and is statistically relevant
- Only then will your results be meaningful
- And only then can your test get better

Can anti-spam testing be done better?

Can anti-spam testing be done better?

Yes!

Can anti-spam testing be done better?

Yes!

(But only if we all work together!)

Can spam filtering be done better?

Can spam filtering be done better?

	FP rate	SC rate
Product A	1.25%	99.60%
Product B	1.80%	99.56%
Product C	1.96%	99.51%
	…	…
Product J	0.24%	99.39%
Product K	0.39%	85.40%
Product L	0.63%	98.39%

Can spam filtering be done better?

	FP rate	**SC rate**
Product A	1.25%	99.60%
Product B	1.80%	99.56%
Product C	1.96%	99.51%
	…	…
Product J	0.24%	99.39%
Product K	0.39%	85.40%
Product L	0.63%	98.39%
Combined effort	**0.08%**	**99.59%**

Questions?

Questions?

And: discussion?

Questions?

And: discussion!

Questions?

And: discussion!

(And please take the discussion beyond the end of VB2009!)

Questions?

And: discussion!

(And please take the discussion beyond the end of VB2009!)

Questions?

And: discussion!

(And please take the discussion beyond the end of VB2009!)

Questions?

And: discussion!

(And please take the discussion beyond the end of VB2009!)

Questions?

And: discussion!

(And please take the discussion beyond the end of VB2009!)