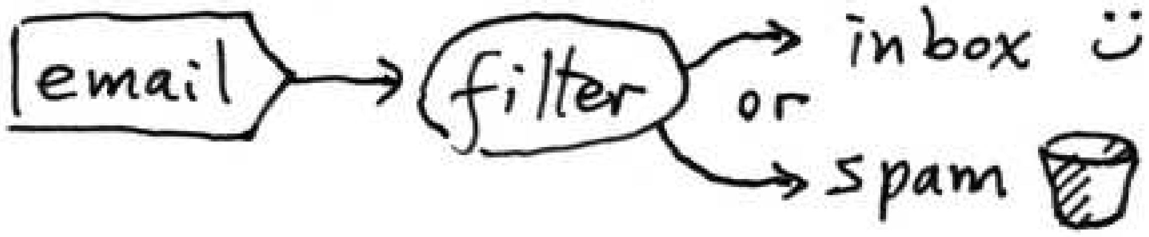
BAYES' THEOREM & BAYESIAN INFERENCE

An early example of an algorithm that tracks "authenticity" is email spam filters, motivated by the huge spam problem starting in the '90s. By 2010, close to 90% of emails were spam, with an annual societal cost of $20B. Such a filter tries to keep emails you want (eg, friends, legitimate unsolicited messages) & discard those you don't (eg, bulk ads). Clearly this is subjective – should you consider your friend's bulk marketing email as spam?

Let's learn Bayesian inference while designing a simple spam filter algorithm.

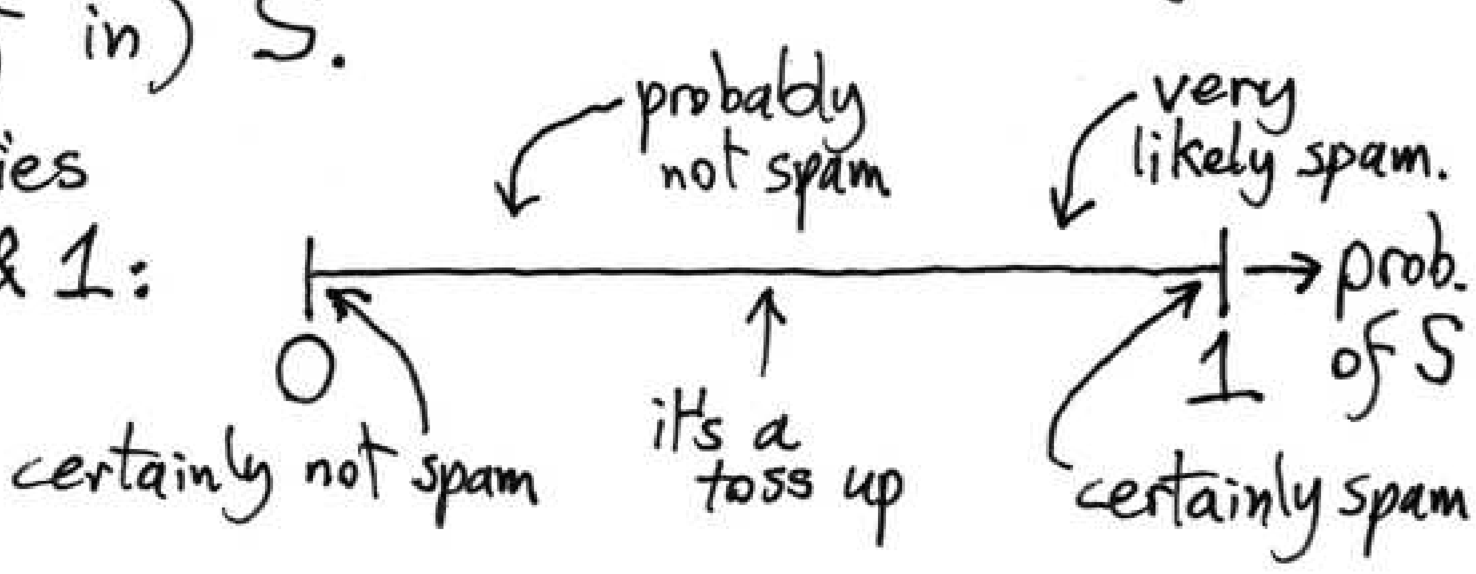


Consider one incoming email:

- the event "this email is spam" we call S
- the other possibility is simply "not S"

The Bayesian approach updates a probability of (ie, numerical belief in) S.

- All probabilities lie between 0 & 1:



- This probability will change in the light of new input, just as an opinion of an email crystallizes as you read through it.

We use $p(S)$ to denote prior probability of being spam, ie before examining the message. Given the above statistic, $p(S) = 0.9 = 90\%$ is a good prior estimate.

Now let U be the event "this email contains the word urgent".

We need to know how common 'urgent' is in spam & non-spam.

To estimate this, say we analyse 1000 random emails & find (say) these statistics:

	not S	S
all emails:	100	900
emails with 'urgent':	10	360

Armed with this "training data" we estimate,

$$p(U|S) = \frac{360}{900} = 0.4$$

← ie, 40% of spam contains 'urgent'

this means conditional probability of U occurring, given S.

We'll also need the probability of U without knowledge of S, which we can also estimate from our table:

$$p(U) = \frac{10+360}{1000} = 0.37$$

Time for some inference! There's two cases:

- An incoming email contains 'urgent'. Bayes' theorem (derived at the end) tells us then,

"likelihood"

$$p(S|U) = \frac{p(U|S)}{p(U)}\,p(S) = \frac{0.4}{0.37}\,0.9 \approx 0.973$$

prior

called the "posterior", this is the updated prob. of S, in light of U.

97.3% chance of being spam.

- Alternatively, the email doesn't contain 'urgent', and we apply the same theorem but with different data,

$$\underset{\text{"posterior"}}{p(S \mid \text{not } U)} = \frac{p(\text{not } U \mid S)}{p(\text{not } U)} p(S) = \frac{0.6}{0.63} 0.9 \approx 0.857$$

here we use $p(\text{not } U) = 1 - p(U)$.

high, but less than our prior.

Finally the algorithm must pick a threshold: if posterior > 0.95, say, it goes to spam. So far, this is not a good filter: you lose every email containing 'urgent', which includes 10% of legitimate emails! (an unacceptable "false positive rate").

But the updating step (inference) can be repeated with many other search words, and the filter becomes much better.

- Here's a sketch of that idea:

Let V be the event "the email contains 'viagra'"

note, $p(V \mid \text{not } S)$ is tiny!

A useful (but wrong) model is to assume independence:

$$p(U \text{ and } V \mid \cdots) = p(U \mid \cdots)\, p(V \mid \cdots)$$

Then Bayes gives, for an email with both 'urgent' & 'viagra',

$$p(S \mid V \text{ and } U) = \frac{p(V \text{ and } U / S)}{p(V \text{ and } U)} p(S) = \underbrace{\frac{p(V/S)}{p(V)}}_{\text{viagra update factor.}} \cdot \underbrace{\frac{p(U/S)}{p(U)} p(S)}_{\substack{\text{old posterior} \\ p(S \mid U)}}$$

new posterior, will be very close to 1.

Combining these factors for many "spammy" words gives a decent Bayesian spam filter.

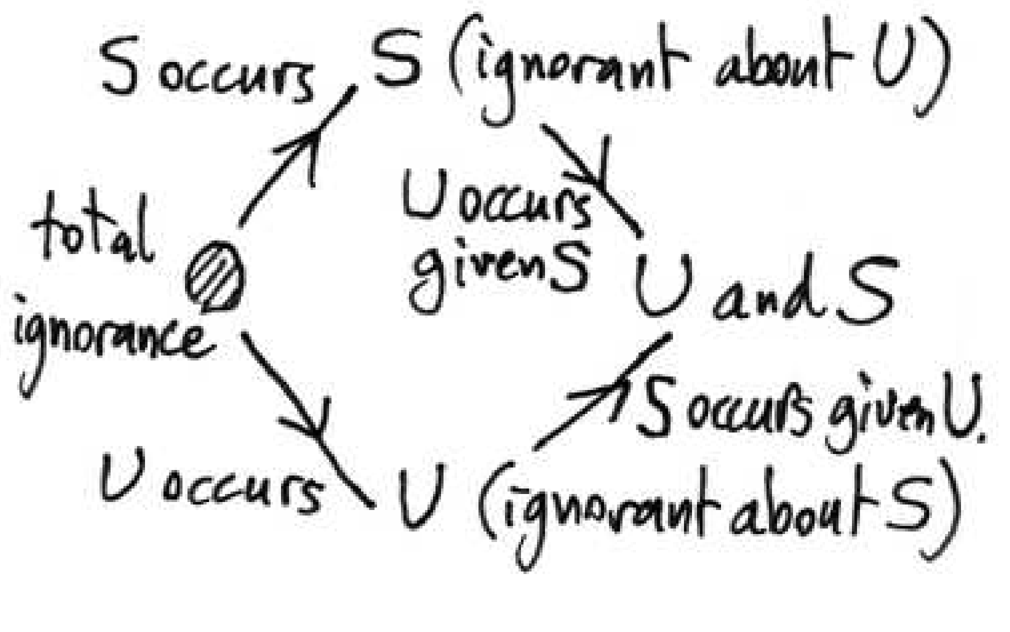
Notes on spam filters:

- We have just seen Bayesian inference in action: it shows how to update probabilities in the light of new data.
- Training data (statistics on spam and non-spam) is needed for the "likelihoods" $p(U|\neg S)$, etc, needed in the updates.
- There can be many unintended consequences! Legitimate emails end up lost, but also spammers change tactics by misspelling words (V1agra), including random text ("Bayesian poisoning")... an arms race of evolving viral warfare.
- Real filters are fancier than above, using phrases, URLs, blacklisted senders, presence of CAPS, etc...

Derivation of Bayes' Theorem

Let's use the symbols U and S for any two events.

One can reveal knowledge by two different routes to get to the event "U and S occur":



This gives 2 ways to factor the joint probability:

$$p(U \text{ and } S) = p(U|S)\,p(S)$$
$$p(U \text{ and } S) = p(S|U)\,p(U)$$

} equating these and rearranging gives the formula 2 pages back.