Notes on Bayesian e-mail filter ("A plan for Spam," http://www.paulgraham.com/spam.html)

Let

\[ P(w_i|S) = \text{probability of word } w_i \text{ occurrence in spam e-mail.} \]
\[ P(w_i|\bar{S}) = \text{probability of word } w_i \text{ occurrence in non-spam e-mail.} \]

I. By Bayes Thm, the probability \( P(S|w_i) \) that an e-mail is spam if it contains \( w_i \) satisfies

\[
P(S|w_i) = \frac{P(w_i|S)P(S)}{P(w_i)} = \frac{P(w_i|S)P(S)}{P(w_i|S)P(S) + P(w_i|\bar{S})P(\bar{S})} = \frac{P(w_i|S)}{P(w_i|S) + \frac{1-P(S)}{P(\bar{S})}P(w_i|\bar{S})}
\]

(For \( P(S) = P(\bar{S}) = \frac{1}{2} \), this reduces to \( P(S|w_i) = P(w_i|S)/(P(w_i|S) + P(w_i|\bar{S})) \).

II. To combine the individual probabilities \( p_i \equiv P(S|w_i) \) for multiple pieces of evidence, \( \vec{w} = \{w_1, \ldots, w_n\} \), again first use Bayes’ theorem in the form

\[
P(S|\vec{w}) = \frac{P(\vec{w}|S)}{P(\vec{w}|S) + \frac{1-P(S)}{P(\bar{S})}P(\vec{w}|\bar{S})}
\]

This problem is underdetermined so assume statistical independence,

\[
P(\vec{w}|S) = \prod_{i=1}^{n} P(w_i|S), \quad P(\vec{w}|\bar{S}) = \prod_{i=1}^{n} P(w_i|\bar{S}),
\]

and again, using Bayes’ Thm, substitute \( P(w_i|S) = p_i P(w_i)/P(S) \) and \( P(w_i|\bar{S}) = (1 - p_i) P(w_i)/P(\bar{S}) \) (where \( P(S|w_i) = 1 - P(S|w_i) = 1 - p_i \)) to find

\[
P(S|\vec{w}) = \frac{\prod_i p_i \prod_i \frac{P(w_i)}{P(S)}}{\prod_i p_i \prod_i \frac{P(w_i)}{P(S)} + \frac{1-P(S)}{P(\bar{S})} \prod_i (1-p_i) \prod_i \frac{P(w_i)}{1-P(S)}}
\]

\[
= \frac{\prod_i p_i}{\prod_i p_i + \left(\frac{P(S)}{1-P(S)}\right)^{n-1} \prod_i (1-p_i)}.
\]

In the case \( P(S) = P(\bar{S}) = \frac{1}{2} \) ("flat prior"), this reduces to ("Bayes’ Rule"):

\[
P(S|\vec{w}) = \frac{\prod_i p_i}{\prod_i p_i + \prod_i (1-p_i)}
\]