Jeremiah O’Connor

Artificial Intelligence: Spam Filter README

Taking into account headers improves accuracy. I parse the header, body, and any html contents for both creating training sets and parsing unseen emails. When the training set is created, I use a dictionary to keep track of terms that are collected from the email corpus. Stop words are exempted for more relevant terms to be taken into account. When an unread email comes in, parsing email terms is done similarly to previous training sets. When calculating probability of an unseen email being either ham or spam, I take into account three cases on each term. If term in the email is part of both ham and spam training sets, then the probability is calculated using both of the training sets, P(t1 | spam), P(t1 | ham). However, if the word appears in only one training set, for example the ham set, I would assign P(t1 | ham) = .99 and P(t1 | spam) = .01. If the word is in neither training sets, then I only give P(t1 | ham) = .4 because spam words are more easy to identify so it most likely to be a ham. Used A Plan for Spam by Paul Graham as a source.

In order to deal with the underflow problem I convert numbers into log domain every time I calculate the MAP and its priors.