CS 5100: Foundations of Artificial Intelligence

Nisarg Shah, Sugandha Kher

05 March 2018

Project Proposal

Abstract

Spam emails make up to 73% of global emails. According to a study by the Radicati Research Group Inc., spam costs businesses $20.5 billion annually in decreased productivity as well as in technical expenses. We plan to develop a spam identification engine that would employ Naive Bayes' classifier to identify spam.

1 Methodology

We deem this as a classification problem and plan to use the Naive Bayes' classifier to solve this. To begin, we need to teach our filter what a spam email looks like and what a non-spam email looks like. So, the basic idea would be to train the filter on 80% of dataset of emails pre-labelled as spam or ham (not spam) and then, test the filter on rest of data.
Bayes' Theorem: This can be written as follows:

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

This is essentially a formula to calculate the conditional probabilities.
Above formula could be written in another form as:

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

Naive Bayes' Classifier: Here, if

$$P(Spam|EmailContent) > P(\neg Spam|EmailContent),$$

then we can classify the email as spam!
We would use the TF-IDF model to calculate the

$$P(EmailContent|Spam) \text{ and } P(EmailContent|\neg Spam)$$

TF-IDF model basically reflects on how important a word is w.r.t the entire collection.

2 Evaluation

The following evaluation factors could be used while evaluating the filter.

1. Precision: Fraction of relevant instances among the retrieved instances
2. Recall: Fraction of relevant instances that have been retrieved over the total amount of relevant instances

Following definitions are explained w.r.t the problem domain

1. False Positive: Emails marked as Ham even if they are not Spam
2. False Negative: Spam Emails are marked as Ham
3. True Positive: Spam Emails are marked as Spam
4. True Negative: Ham Emails are marked as Ham

Concrete ideas on evaluation

1. Measure the Precision over Training set and Test set. Here, the precision should not be off by a large factor.
2. Measure the percentage of false negatives when the false positive is made significantly less if not zero.