Naïve Bayes Classifier for Spam Detection

Made By –

Anusha Agrawal (2018A3PS0032H)

Kriti Jethlia (2018A7PS0223H)

Tanay Gupta (2018AAPS0343H)

Model Description –

- Naïve Bayes classifier assumes that words in a sentence are independent of each other. Probability of a sentence to be spam or not spam is denoted by -

  P(spam|sentence) = P(spam|w1,w2,w3,….wn) = $\prod_1^n$ P(spam|wi)]

  P(spam|wi) = P(wi|spam)*P(spam)/P(wi)

But if a word hasn't occurred in a spam sentence than the above probability will be zero, so we use **Laplacian Smoothing.**

P(spam|wi) = (Frequency of word in spam sentence + 1)/(Number of words in vocabulary + ∑Frequency of each word in spam sentence)

- Similar computation is done for non-spam sentences.
- Also, we use log(P(spam|wi)) because multiplying probabilities directly may cause numerical underflow.
- A sentence is considered spam if P(spam|sentence)>P(not spam|sentence), this implies log(P(spam|sentence))-log(P(not spam|sentence))>0.
- To mitigate the issue of uneven distribution of spam and not spam sentences in the training set, we set a prior log(P(spam)/P(not spam)).
- So the final formula becomes –

  log(P(spam|sentence)/P(not spam|sentence)) = log(P(spam)/P(not spam)) + ∑log(P(spam|wi)/ P(not_spam|wi))

  If the above equation results in a number >0 then the given sentence is spam else not spam.

Implementation –

Sentences are first preprocessed to remove punctuations, stopwords and digits as these characters don't provide any useful information about whether a sentence is spam or not.

The sentences are then shuffled and divided into 7 cross folds and cross fold validation is used to train and test the model.

Probabilities required by the model and all other required information is generated using the train set. While testing a sentence if a word in it is not present in the vocabulary it is ignored.

Finally the model is tested on the test set.

The probabilities for each word is saved in a dictionary for fast retrieval.

Accuracy of Model –

Mean Accuracy = 82.6%

Minimum Accuracy = 78.87%

Maximum Accuracy = 86.61%

Accuracy over each crossfold = 82.39%, 81.69%, 78.87%, 79.57%, 86.61%, 84.50%, 80.28%

Limitations –

1. Naive Bayes assumes that all predictors (or features) are independent which rarely happens in real life.
2. Naïve Bayes implementation without any kind of smoothing suffers from zero frequency problem, meaning that if a word was not encountered in the spam class of train dataset than the probability of that word occurring in the spam class is zero, P(word|spam) = 0.
3. Naïve Bayes cannot deal with continuous features and binning is used to convert them into discrete features, but if this is not done properly then a lot of information can be lost in the process.