Naive Bayes Classifier

Team: Hasnain Zeenwala (2018A7PS0307H), Dhruvikaa Ahuja(2018B3A70916H), Rupanshu Soi(2018A7PS0294H)

Naive Bayes Algorithm & Implementation

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

1. Data Preprocessing

Raw text file was imported and converted to a list of emails with their class at the end. Class(spam or ham) was separated into a different list; various punctuation marks were removed from the emails. Each e-mail was split into a list of words and from each list stop words were removed.

Emails along with their respective class were converted into a data frame of the following format. Each row denoted an email; the columns were the class label and all the distinct words appearing in all the emails. Each cell contained the count of how many times a given word has appeared in the given email.

This dataframe was split into seven nearly equal dataframes to implement *7-fold cross validation.*

2. Naive Bayes Model

By Bayes theorem we can say the following:

A-- $$\mathrm{P(Spam|w_1, w_2, ..., w_n)} \propto \mathrm{P(Spam)} \cdot \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(w_i|Spam)}$$

B-- $$\mathrm{P(Ham|w_1, w_2, ..., w_n)} \propto \mathrm{P(Ham)} \cdot \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(w_i|Ham)}$$

Where w1, w2,.....wn are the words in the email. To implement our classifier we need to calculate the RHS in A and B using our training data.

The following formulae were used to calculate RHS

$$\mathbf{P(w_i|Spam)} = \frac{\mathbf{N_{w_i|Spam}} + \boldsymbol{\alpha}}{\mathbf{N_{Spam}} + \boldsymbol{\alpha} \cdot \mathbf{N_{Vocabulary}}}$$

$$\mathbf{P(w_i|Ham)} = \frac{\mathbf{N_{w_i|Ham}} + \boldsymbol{\alpha}}{\mathbf{N_{Ham}} + \boldsymbol{\alpha} \cdot \mathbf{N_{Vocabulary}}}$$

Training data was split into spam and ham sets. Here alpha is the Laplace smoothing parameter. It ensures that if a word is absent from the span/ham sets it is assigned a non zero probability.

Accuracy of the Model:

1. Fold: 1, 0.8169014084507042
2. Fold: 2, 0.823943661971831
3. Fold: 3, 0.7464788732394366
4. Fold: 4, 0.795774647887324
5. Fold: 5, 0.8028169014084507
6. Fold: 6, 0.7676056338028169
7. Fold: 7, 0.7972972972972973

Average Accuracy: 0.7929740605796944

Limitation of Naive Bayes Classifier

1. Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. This limits the applicability of this algorithm in real-world use cases.

2. If the testing set has a variable that wasn't observed in the training set then its frequency is considered to be zero. This needs to be solved separately using smoothing techniques like Laplace smoothing.